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

PAUL OHM*

Draft: Please do not cite or quote without permission.

ABSTRACT

Computer scientists have recently undermined our faith in the privacy-protecting power of anonymization, the name for techniques for protecting the privacy of individuals in large databases by deleting information like names and social security numbers. These scientists have demonstrated they can often "reidentify" or "deanonymize" individuals hidden in anonymized data with astonishing ease. By understanding this research, we will 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 Harvard University's Center for Research and Computer Science, Harvard University's Berkman Center, Princeton University's Center for Information Technology Policy, U.C. Berkeley's Privacy Law Scholars Conference, the University of Washington School of Law, University of Washington Computer Science & Engineering Department, and University of Colorado Law School. I thank all participants for their comments.

Thanks in particular to Danielle Citron, Alissa Cooper, Nestor Davidson, Cynthia Dwork, Ed Felten, Victor Fleischer, Susan Freiwald, Brett Frischmann, Michael Froomkin, Simson Garfinkel, Eric Goldman, Marcia Hofmann, Chris Hoofnagle, Clare Huntington, David Johnson, Jerry Kang, Jon Kleinberg, Aleecia McDonald, Scott Moss, Arvind Narayanan, Scott Peppett, Jules Polonetsky, Joel Reidenberg, David Robinson, Andrew Schwartz, Ari Schwartz, Vitaly Shmatikov, Chris Soghoian, Dan Solove, Harry Surden, Peter Swire, and Phil Weiser for their comments. 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. Imagine further that the records 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 find the data still useful, but unscrupulous marketers and malevolent identity thieves will find it impossible to identify the people tracked. As it always has before, 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 subspeciality of com-

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

Reidentification science disrupts the privacy policy landscape by undermining the faith that we have placed in anonymization. This is no small faith, for technologists rely on it to justify sharing data indiscriminately and storing data perpetually, all while promising their 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. In the European Union, the famously privacy-protective Data Protection Directive extends a similar safe harbor through the way it defines "personal data."[2] 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 like 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.

How many other people in the United States share your specific combination of ZIP code, birth date, and sex? According to a landmark study, for 87% of the American population, the answer is zero; these three pieces of information uniquely identify each of them.[3] How many

---

[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 & 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 (2005). In addition, a few law student have discussed some of the reidentification studies discussed in this Article but without connecting these studies to larger questions about information privacy. Note, Christine Porter, 5 SHIDLER J.L. COM. & TECH. 3 (2008) (discussing the AOL and Netflix stories); Note, Benjamin Charkow, *The Control over the De-Identification of Data*, 21 CARDOZO ARTS & ENT. L.J. 195 (2003).

[2] 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].

[3] Latanya Sweeney, Uniqueness of Simple Demographics in the U.S. Population, Laboratory for International Data Privacy Working Paper, LIDAP-WP4 (2000). For more on this study, see *infra* Part I.B.1.b. More recently, Phillipe Golle revisited Dr. Sweeney's study, and recalculated the statistics based on year 2000

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% of Netflix users.[4] 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 served as the central focus of information privacy law for almost forty years, is a fallacy that has run its course and 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, breeds future successes. 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."[5]

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 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 which 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 they 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

---

census data. Dr. Golle could not replicate the earlier 87% statistic, but he did calculate that 61% of the population in 1990 and 63% in 2000 were uniquely identified by ZIP, birth date, and sex. Phillipe Golle, *Revisiting the Uniqueness of Simple Demographics in the US Population*, 2006 WORKSHOP ON PRIVACY IN THE ELEC. SOC'Y PROC.

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

[5] *See infra* Part III.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 raises three appealing but ultimately incomplete responses to correct these imbalances. Finally, Part IV offers a way forward, proposing a new test for deciding when to impose new privacy restrictions on information flow, demonstrating the test with examples from health and internet privacy.

I. ANONYMIZATION AND REIDENTIFICATION

A. The Past: Robust Anonymization

Something important has changed. To understand what has changed, consider first the importance of the thing changed. We are about to lose something—our faith in robust anonymization—which is important and widespread. For decades, technologists have believed that they could robustly protect the privacy of people by making small changes to their data, using techniques surveyed below. I call this the *robust anonymization assumption.* Embracing this assumption, for years they have promised privacy to their users, and in turn, privacy is what their users have come to expect. Today, anonymization is a ubiquitous practice.

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 obliterating the foundational robust anonymization assumption, this result will topple the edifices of promise and expectation we have built upon it. For all of these reasons, the power of easy reidentification will disrupt an important tool we use to order vital societal relationships.

The easy reidentification result will also wreak havoc on our legal systems, because the faith in robust anonymization has thoroughly infiltrated our privacy laws and regulations, as Part II will explore. Whenever lawmakers have enacted new privacy protections, here and abroad, they have built their own edifices of rules, standards, and exceptions based on the robust anonymization assumption, reassured by the technologists of the soundness of this approach.

This section looks backwards; it tells the story of how we built these grand structures; it does this before we deploy the wrecking balls, to explain what we are about to lose.

1. Ubiquitous Anonymization

Anonymization plays a central and ubiquitous role in modern data handling, featuring prominently in 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* or *database*, but it may also maintain information in any format using any technology or no sophisticated technology at all.

Data administrators try to protect the privacy of data subjects by *anonymizing* the data. Although I will later argue against the future use of this term,[6] I am not quite ready to let it go, so for now, anonymize means manipulate the information in a database to make it difficult to identify data subjects.

Database experts have developed scores of different anonymization techniques, which vary along many dimensions such as cost, complexity, ease-of-use, and robustness. For starters, consider a very common technique: *suppression*.[7] 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 the data in order to anonymize it.

The opposite of anonymization is *reidentification* or *deanonymization*.[8] A person, known in the scientific literature as an *adversary*,[9] reidentifies anonymized data by linking anonymized records to outside information, and in the extreme case, discovering the true identity of the data subjects.

b) The Reasons to Anonymize

Data administrators anonymize to protect the privacy of data subjects when they store or disclose data. They disclose data to three groups: First, they release data to third parties. Health researchers share patient data with other health researchers,[10] websites sell user

---

[6] *See infra* Part II.C.2.

[7] Latanya Sweeney, *Achieving* k-*Anonymity Privacy Protection Using Generalization and Suppression*, 10 INT'L J. ON UNCERTAINTY, FUZZINESS AND KNOWLEDGE-BASED SYSTEMS 571, 572 (2002)

[8] *E.g., Netflix Prize Study, supra* note 4, at 2.

[9] *Id.*

[10] Nat'l Inst. Health, HIPAA Privacy Rules for Researchers, http://privacyruleandresearch.nih.gov/faq.asp (last visited July 22, 2009).

transactional data to advertisers,[11] and phone companies can be compelled to disclose telephone call data to law enforcement officials.[12]

Second, they sometimes release anonymized data to the public.[13] 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.[14]

Third, they disclose anonymized data to others within their organization.[15] Particularly within large organizations, the data-collection part of the organization may want to protect the privacy of the data subjects even from others in the organization.[16] For example, large banks may want to share some data with their marketing departments, but only after anonymizing it to protect customer privacy.

At least four pressures compel administrators to anonymize: norms and ethics, the market, architecture, and law.[17] Anonymization norms and ethics operate often through "best practices," documents that recommend anonymization as a technique for protecting privacy. For example, biomedical guidelines often recommend "coding" genetic data—associating stored genes with non-identifying numbers—to protect privacy.[18] Other guidelines recommend anonymization in various other contexts including electronic commerce,[19] internet service provision,[20] data mining,[21] and national intelligence data sharing.[22] Academic researchers rely heavily on anonymization to protect human research subjects, and their research guidelines recommend anonymization gen-

---

[11] *E.g.* Susan Wojcicki, Vice President, Product Management, *Making Ads More Interesting*, THE OFFICIAL GOOGLE BLOG, March 11, 2009, http://googleblog.blogspot.com/2009/03/making-ads-more-interesting.html (announcing new Google initiative to tailor ads to "they types of sites you visit and the pages you view").

[12] *E.g.* In re Application of United States for an Order for Disclosure of Telecommunications Records, 405 F. Supp. 2d 435 (S.D.N.Y. 2005) (granting government authority to compel provider to provide information suggesting location of customer's cell phone).

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

[14] *See passim* JAMES SUROWIECKI, THE WISDOM OF CROWDS (2005).

[15] *See* Philip Lenssen, *Google-internal Data Restrictions,* GOOGLE BLOGOSCOPED BLOG, June 27, 2007, http://blogoscoped.com/archive/2007-06-27-n27.html (detailing how Google and Microsoft limit internal access to sensitive data).

[16] *Id.*

[17] *See* LAWRENCE LESSIG, CODE, AND OTHER LAWS OF CYBERSPACE V 2.0 at 123 (2007) (listing four regulators of online behavior: markets, norms, laws, and code).

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

[19] ALEX BERSON & LARRY DUBOV, MASTER DATA MANAGEMENT AND CUSTOMER DATA INTEGRATION FOR A GLOBAL ENTERPRISE 338-39 (2007)

[20] *See infra* Part II.A.3.b.

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

[22] Markle Task Force, National Security in the Information Age, 134, 144 (2003).

erally[23] and specifically in education,[24] computer network monitoring,[25] and health studies[26]. Professional statisticians are duty bound to anonymize data as a matter of professional ethics.[27]

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.[28] One way these companies add value is by aggregating and republishing data about their customers to help individual users compare how much they spend with other similarly situated people.[29] To make customers comfortable with this type of data sharing, both mint.com and wesabe.com promise to anonymize the data before sharing it.[30]

Architecture, defined in Larry Lessig's sense as technological constraints,[31] 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 about your visit into what is called a logfile.[32] As it turns out, 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.[33]

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 confiden-

---

[23] LISA GIVEN, THE SAGE ENCYCLOPEDIA OF QUALITATIVE RESEARCH METHODS 196 (entry for "Data Security") (2008).

[24] LOUIS COHEN ET AL., RESEARCH METHODS IN EDUCATION 189 (2003).

[25] Ruoming Pang et al., *The Devil and Packet Trace Anonymization*, COMP. COMM. REV. (2006)

[26] INST. OF MEDICINE, PROTECTING DATA PRIVACY 178 (2000).

[27] 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).

[28] Eric Benderoff, *Social Finance Sites Help Readers Save Money,* THE SEATTLE TIMES, Nov. 8, 2008.

[29] Carolyn Y. Johnson, *Online Social Networking Meets Personal Finance*, INT'L HERALD TRIBUNE, Aug. 7, 2007.

[30] Wesabe, Security and Privacy, http://www.wesabe.com/page/security (visited July 23, 2009); Mint.com, How Mint Personal Finance Management Protects Your Financial Safety, http://www.mint.com/privacy/ (visited July 23, 2009).

[31] LESSIG, *supra* note 17.

[32] STEPHEN SPAINHOUR & ROBERT ECKSTEIN, WEBMASTER IN A NUTSHELL 458-59 (2003).

[33] Apache, Apache HTTP Server Version 1.3 Log Files, http://httpd.apache.org/docs/1.3/logs.html (describing the default "common log format" which logs less information than the alternative "combined log format") (last visited July 23, 2009).

tiality."[34] 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."[35]

Government officials and policymakers embrace anonymization as well. Two influential data mining task forces made up of luminaries have endorsed anonymization. In 2004, the Technology and Privacy Advisory Committee (TAPAC), a Defense Department-led group established in the wake over controversy surrounding the government's Total Information Awareness program, produced an influential report about government data mining.[36] The report recommends data anonymization "whenever practicable" and restricts its recommendations only to databases "known or reasonably likely to include personally identifiable information."[37]

Likewise, the Markle Foundation task force, which included now-Attorney General Eric Holder, produced a similar report.[38] 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] 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."[39]

Law Professors and other legal scholars share this faith in anonymization.[40] Ira Rubenstein, 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."[41] 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 re-identification would be impossible."[42] Many other legal scholars have made similar claims premised on deep faith in robust

---

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

[35] Chris Soghoian, *Debunking Google's Log Anonymization Propaganda*, SURVEILLANCE STATE BLOG, Sept. 11, 2008, http://news.cnet.com/8301-13739_3-10038963-46.html.

[36] Technology and Privacy Advisory Committee, Report, Safeguarding Privacy in the Fight against Terrorism ("TAPAC Report") 35-36 (Mar 1, 2004), *available at* http://www.cdt.org/security/usapatriot/20040300tapac.pdf (visited Aug. 5, 2009).

[37] *Id.* at 50 (Recommendation 2.2).

[38] Task Force on National Security in the Information Age, Creating a Trusted Network for Homeland Security: Second Report of the Markle Foundation Task Force 34 (2003).

[39] *Id.* at 31 ("Guidelines for Database Access and Use" recommendation 8).

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

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

[42] Barbara J. Evans, *Congress' New Infrastructural Model of Medical Privacy*, 84 NOTRE DAME L. REV. 585, 619-20 (2009).

anonymization.[43] 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 of 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 which I call *release-and-forget* anonymization. As the name suggests, when a data administrator practices these techniques, she *releases* records—either publicly, 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.[44] Because these techniques promise privacy while allowing the broad dissemination and downstream use of data, data administrators embrace their easy tradeoff.[45] Second, these techniques are often flawed. Many of the recent advances in the science of reidentification target release-and-forget anonymization in particular.[46]

Consider some common release-and-forget techniques.[47] First, we need a sample database to anonymize, a simplified and hypothetical model of a hospital's database for tracking visits and diagnoses:[48]

---

[43] *See, e.g.,* Matthew P. Gordon, *A Legal Duty to Disclose Individual Research Findings to Research Subjects?*, 64 FOOD & DRUG L.J. 225, 258-59 (2009); Fred H. Cate, *Government Data Mining: The Need for a Legal Framework*, 43 HARV. C.R.-C.L.L. REV. 435, 487 (2008); Susan M. Wolf et al., *Incidental Findings in Human Subjects Research: From Imaging to Genomics*, 36 J.L. MED. & ETHICS 219, 226-27 (2008); Comment, Irfan Tukdi, *Transatlantic Turbulence: The PassengerName Record Conflict*, 45 HOUS. L. REV. 587, 618-19; 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).

[44] Laks V.S. Lakshmanan et al., *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.").

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

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

[47] This is only a survey. This discussion will make an expert of no one.

[48] All of the hypothetical data in this table aside from the "Name" column comes from a paper by Latanya Sweeney. Latanya Sweeney, *k-Anonymity: A Model for Protecting Privacy*, 10 INT'L J. ON UNCERTAINTY, FUZZINESS, AND KNOWLEDGE-BASED SYSTEMS 557, 567 fig.4 (2002). Where the first names come from is left as an exercise for the reader.

Name	**Race**	**Birth Date**	**Sex**	**ZIP Code**	**Diagnosis**
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	Obesity
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	Obesity
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

Table 1: Original (Nonanonymized) Data

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*, which are identified by labels (in bold) called *field names* or *attribute names*; each record has a particular *value* for a given attribute.[49]

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 single out a record in the table.[50] 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.[51] 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.

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 from one of these sources to treat the following four fields as potential identifiers: name, birth date, sex, and ZIP code.[52]

Suppression: Next, the administrator will modify the identifying fields. She might suppress them, removing the fields from the table

---

[49] GAVIN POWELL, BEGINNING DATABASE DESIGN 38-41 (2005).

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

[51] *Id.* Because these sorts of identifiers do not link directly to identity, researchers sometimes refer to these as quasi-identifiers.

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

altogether.[53] In our example, the administrator might delete all four potential identifiers, producing this table:

Race	**Diagnosis**
Black	Short of breath
Black	Chest pain
Black	Painful eye
Black	Wheezing
Black	Obesity
Black	Chest pain
White	Short of breath
White	Hypertension
White	Obesity
White	Fever
White	Vomiting
White	Back pain

Table 2: Suppressing Four Identifier Fields

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 diagnosis. But on the other hand, aggressive suppression has rendered this data almost useless for research.[54] Although a researcher can use the data remaining in this table to track the incidence of a particular disease 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.[55] This means she will alter rather than delete identifier values to increase privacy but preserve utility. For example, the anonymizer may choose to suppress the name field, generalize the birth date into the year of birth, and generalize ZIP codes by retaining only the first three digits.[56] The resulting data would look like this:

[53] *Sweeney*, *supra* note 7, at 3.
[54] *See, infra*, Part III.B.1 (discussing relationship between utility and privacy).
[55] *Sweeney*, *supra* note 7, at 3.
[56] Under the HIPAA Privacy Rule, these three changes would qualify the resulting table as de-identified health information. U.S. Health and Human Services; Standards for Privacy of Individually Identifiable Health Information; Final Rule, 45 C.F.R. pts. 160 & 164 (2009). For more on HIPAA and the Privacy Rule, *see infra* Part II.A.3.a.

Race	Birth Year	Sex	ZIP Code*	Diagnosis
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*	Obesity
Black	1964	Female	021*	Chest pain
White	1964	Male	021*	Short of breath
White	1965	Female	021*	Hypertension
White	1964	Male	021*	Obesity
White	1964	Male	021*	Fever
White	1967	Male	021*	Vomiting
White	1967	Male	021*	Back pain

Table 3: Generalized

Now, even one who knows Forest's birth date, ZIP code, sex, and race, will still have trouble plucking out Forest's specific diagnosis. The people 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.[57] Thus, if researchers only need to know how many men complained of shortness of breath, data administrators could release this:

Men Short of Breath	2

Table 4: Aggregate Statistic

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

Privacy lawyers tend to refer to release-and-forget anonymization techniques using two other names: deidentification[58] and the removal of personally-identifiable information (PII).[59] Deidentification has

[57] *E.g.,* Nabil R. Adam & John C. Wortmann, *Security-Control Methods for Statistical Databases: A Comparative Study*, 21 ACM COMPUTING SURVEYS 515 (1989); Ivan P. Fellegi, *On the Question of Statistical Confidentiality*, 67 J. AM. STAT. ASS'N 7 (1972); Tore Dalenius, *Towards a Methodology for Statistical Disclosure Control*, 15 STATISTIK TIDSKRIFT 222 (1977).

[58] Nat'l Inst. Health, De-identifying Protected Health Information Under the Privacy Rule, http://privacyruleandresearch.nih.gov/pr_08.asp (last visited July 23, 2009).

[59] Nat'l Inst. Standards & Tech., Guide to Protecting the Confidentiality of Personally Identifiable Information (PII), Special Pub. 800-122 (January 2009) (draft), *available at* http://csrc.nist.gov/publications/drafts/800-122/Draft-SP800-122.pdf.

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.[60]

B. The Present and Future: Easy Reidentification

Until a decade ago, the robust anonymization assumption worked well for everybody involved. Data administrators could responsibly protect privacy while disclosing information to third parties; data subjects could rest assured that their secrets were secured beneath impenetrable layers of protective anonymization; legislators struck the balance between privacy and other interests such as security and the advancement of knowledge by deregulating the trade in anonymized records;[61] and regulators could pressure data handlers to anonymize more, dividing the world into the responsible (those who anonymize) 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 entire state of affairs has been built. Very recently, these researchers have done more than chip away, they have essentially blown up the robust anonymization assumption, proving some important theoretical limits of the power of anonymization, 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 enough to suggest we should reject anonymization as a privacy-providing panacea.

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 reappear throughout the Article, provide two important lessons: they suggest the spread 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.

These stories demonstrate well the power of reidentification, but they do not demonstrate how reidentification can be used to harm people. The researchers described below are all professional journalists or academic researchers, and ethical rules and good moral judgment limited the harm they caused. But do not be misled by how benign these studies may seem. Later, we will connect this back to harm, demonstrating how the same techniques (and indeed, these very same studies) can be used as links in chains of inferences connecting individuals to harmful facts.[62]

---

[60] 45 C.F.R. §§ 164.502(d)(2), 164.514(a), (b) (2009). *See infra* Part II.A.3.a.

[61] *See infra* Part II.A.

[62] *See infra* Part III.A (describing "the database of ruin").

a) The AOL Data Release

On August 3, 2006, America Online (AOL) announced a new initiative called "AOL Research."[63] To "embrac[e] the vision of an open research community" AOL Research publicly posted to a website 20 million search queries for 650,000 users of AOL's search engine summarizing three months of activity.[64] Researchers of internet behavior rejoiced to receive this much of this type of information, the kind of information usually treated by search engines as a closely-guarded secret.[65] 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[66] in the released data.[67] In order to preserve the usefulness of the data for research, however, it replaced these identifiers with unique identification numbers which allowed researchers to correlate different searches to individual users.[68]

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 "hunt[ing] for particularly entertaining or shocking search histories."[69] Due to all of this blogging and subsequent news reporting, these 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.'"[70] User 17556639 searched for "how to

---

[63] E-mail from Abdur Chowdhury, cabdur@aol.com, to SIGIR-IRList, irlist-editor@acm.org (Aug. 3, 2006) (available at http://sifaka.cs.uiuc.edu/xshen/aol/20060803_SIG-IRListEmail.txt).

[64] *Id.* The data released actually contained 36 million entries. Paul Boutin, *You Are What You Search*, SLATE, Aug. 11, 2006, http://www.slate.com/id/2147590.

[65] Katie Hafner, *Researchers Yearn to Use AOL Logs, but They Hesitate*, N.Y. TIMES, Aug 23, 2008 (describing difficulty academic researchers experience accessing raw search data).

[66] IP addresses are numbers which identify computers on the internet and can be used to track internet activity. Part II.A.3 will discuss IP addresses in greater depth.

[67] Michael Barbaro and Tom Zeller, Jr., *A Face is Exposed for AOL Searcher No. 4417749*, N.Y. TIMES, Aug. 9, 2006. IP Addresses are discussed *infra* in Part II.A.3.b.

[68] *Id.*

[69] *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 differences between these two classes of information. *See infra* Part II.A.

[70] *See* Barbaro and Zeller, *supra* note 67.

kill your wife" followed by a string of searches for things like "pictures of dead people" and "car crash photo."[71]

While most of the blogosphere quickly and roundly condemned AOL,[72] a few bloggers argued that the released data, while titillating, did not violate privacy because nobody had linked actual individuals with their anonymized queries.[73] This argument was quickly silenced by New York Times reporters Michael Barbaro and Tom Zeller who recognized many clues to identity in User 4417749's queries, such as "'landscapers in Lilburn, Ga,' several people with the last name Arnold and 'homes sold in shadow lake subdivision gwinnett county georgia.'"[74] They quickly tracked down Thelma Arnold, a 62-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."[75]

The fallout was swift and crushing. AOL fired the researcher who released the data and also his supervisor.[76] Chief Technology Officer Maureen Govern resigned.[77] The fledgling AOL Research division has been silenced, and a year after the incident, the group still had no working website.[78]

b) ZIP, Sex, 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% of people in the United States were uniquely identified by their combined five-digit ZIP code, birth date (including year), and sex.[79] According to her study, even less specific information can often

---

[71] Markus Frind, *AOL Search Data Shows Users Planning to Commit Murder*, THE PARADIGM SHIFT BLOG, Aug. 7, 2006, http://plentyoffish.wordpress.com/2006/08/07/aol-search-data-shows-users-planning-to-commit-murder/.

[72] *E.g.*, Michael Arrington, *AOL Proudly Releases Massive Amounts of Private Data*, TECHCRUNCH, 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.").

[73] 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, Aug. 4, 2006, http://glinden.blogspot.com/2006/08/chance-to-play-with-big-data.html.

[74] Barbaro and Zeller, *supra* note 67.

[75] *Id.*

[76] Tom Zeller, Jr., *AOL Executive Quits After Posting of Search Data*, INTERNATIONAL HERALD TRIBUNE, Aug. 22, 2006.

[77] *Id.*

[78] Chris Soghoian, *AOL, Netflix and the End of Open Access to Research Data*, CNET NEWS SURVEILLANCE STATE blog, Nov. 30, 2007, http://news.cnet.com/8301-13739_3-9826608-46.html.

[79] Latanya Sweeney, Uniqueness of Simple Demographics in the U.S. Population, Laboratory for International Data Privacy Working Paper, LIDAP-WP4 (2000). A subsequent study placed the number at 61% (for 1990 census data) and 63% (for 2000 census data). Golle, *supra* note 3.

reveal identity, as 53% of American citizens are uniquely identified by their *city*, birth date, and sex, and 18% by their *county*, birth date, and sex.[80]

Like the reporters who discovered Thelma Arnold, Dr. Sweeney offered a hyper-salient example to drive home the power (and the threat) of her reidentification techniques. In Massachusetts, a government agency called the Group Insurance Commission (GIC) purchased health insurance for state employees.[81] At some point in the mid-1990s, GIC decided to release records summarizing every state employee's hospital visits to any researcher who requested it for free.[82] By removing fields containing name, address, social security number, and other "explicit identifiers," GIC had 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, ZIP code, birth date, and sex.[83]

At the time GIC released the data, William Weld, then Governor of Massachusetts, assured the public that GIC had protected patient privacy by deleting identifiers.[84] In response, then-graduate student Sweeney started hunting for the Governor's hospital records in the GIC data.[85] She knew that Governor Weld resided in Cambridge, Massachusetts, a city of 54,000 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 Weld with ease. Only six people in Cambridge shared his birth date, only three of them men, and of them, only he lived in his ZIP code.[86] In a theatrical flourish, Dr. Sweeney sent the Governor's health records (which included diagnoses and prescriptions) to his office.[87]

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 re-

---

[80] Sweeney, *supra* note 79.

[81] Massachusetts Executive Office for Administration and Finance, Who is the GIC?, http://www.mass.gov/?pageID=afmodulechunk&L=4&L0=Home&L1=Insurance+%26+Retirement&L2=Oversight+Agencies&L3=Group+Insurance+Commission&sid=Eoaf&b=terminalcontent&f=gic_whoisgic_who_is_gic&csid=Eoaf (last visited Feb. 19, 2009).

[82] *Recommendations to Identify and Combat Privacy Problems in the Commonwealth Before the H. Select Comm. on Information Security* (Penn. 2005) (statement of Latanya Sweeney, Associate Professor, Carnegie Mellon University).

[83] *Id.*

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

[85] *Id.*

[86] Sweeney, *supra* note 79.

[87] Greely, *supra* note 84.

leased 100 million records revealing how nearly 500,000 of its users had rated movies from December 1999 to December 2005.[88] In each record, Netflix disclosed the movie rated, the rating assigned from one to five, and the date of the rating.[89] 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.[90] 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.[91] 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 100 million records to launch what it called the "Netflix Prize," a prize that took almost three years to claim.[92] The first team which used the data to create a movie recommendation algorithm that beat Netflix's algorithm by ten percent won one million dollars. 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.[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."[94] 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 re-identify people in the database, and these results seem surprising and novel even to computer

---

[88] The Netflix Prize Rules, http://www.netflixprize.com/rules (last visited Aug. 3, 2009).

[89] *Id.*

[90] Netflix Prize: FAQ, http://www.netflixprize.com/faq (last visited Aug. 3, 2009) (answering question, "Is there any customer information in the dataset that should be kept private?").

[91] Clive Thompson, *If You Liked This, You're Sure to Love That*, N.Y. TIMES MAG., Nov. 23, 2008, at MM74.

[92] Steve Lohr, *Netflix Challenge Ends, but Winner is in Doubt*, N.Y. TIMES BITS BLOG, July 27, 2009, http://bits.blogs.nytimes.com/2009/07/27/netflix-challenge-ends-but-winner-is-in-doubt.

[93] Thompson, *supra* note 91.

[94] Arvind Narayanan and Vitaly Shmatikov, *How to Break the Anonymity of the Netflix Prize Dataset*, October 16, 2006, ARVIX, 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 4.

scientists, judging from the way the paper has been celebrated[95] and cited[96] by other researchers.

If an adversary knows the precise ratings a person in the database has assigned to six obscure movies[97], and nothing else, he will be able to identify that person 84% of the time.[98] 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 99% of the people in the database.[99] 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% of the users.[100]

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 Movie Database (IMDb),[101] a movie-related website, which also gives users the chance to rate movies. Unlike Netflix, IMDb posts these ratings publicly on its website, like Amazon does with user-submitted book ratings.

Narayanan and Shmatikov obtained ratings for fifty IMDb users.[102] From this tiny sample,[103] they identified two users who were

---

[95] 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. 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

[96] *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).

[97] By obscure movie, I mean a movie outside the top 500 movies rated in the database, ranked by number of ratings given. Narayanan and Shmatikov, *Netflix Prize Study*, *supra* note 4.

[98] *Id.* at 11, 12 fig.8. The fact that some people do not rate so many obscure movies is no hindrance, because the authors note that 90% of the rating users rated five or more obscure movies and 80% rated ten or more obscure movies. *Id.* at 11 (table).

[99] *Id.* at 11, 10 fig.4.

[100] *Id.*

[101] The Internet Movie Database, http://www.imdb.com/ (last visited Aug. 3, 2009).

[102] 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. Narayanan and Shmatikov, *Netflix Prize Study*, *supra* note 94 at 12. On Feb. 11, 2009, the IMDb terms of service prohibited, among other things, "data mining, robots, screen scraping, or similar data gathering and extraction tools". IMDb Copyright and Conditions of Use, http://www.imdb.com/help/show_article?conditions.

[103] IMDb reports 57 million users visit its site each month. IMDb, IMDb History, http://www.imdb.com/help/show_leaf?history (visited Feb. 11, 2009).

identifiable, to a statistical near-certainty, in the Netflix database.[104] Because neither database comprised a perfect subset of the other, one could learn things from Netflix unknowable only from IMDb, and vice versa,[105] including some things these users probably did not want revealed.[106] 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").[107]

2. Reidentification Techniques

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

Even though the administrator removes any data fields she thinks might uniquely identify individuals, the researchers unlocked identity in each of the three cases above by discovering pockets of surprising uniqueness remaining in the data. Just as human fingerprints can uniquely identify a single person and link that person with "anonymous" information—a print left at a crime scene—so too do data subjects generate "data fingerprints"—combinations of values of data shared by nobody else in their table.[108]

Of course even before this research, many understood the basic intuition behind a data fingerprint; this intuition lay at the heart of the endless debates about personally identifiable information (PII). What has startled observers about the new results, however, is that researchers have found data fingerprints in pools of *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; Narayanan and Shmatikov unearthed the surprising uniqueness of the set of movies a person had seen and rated; and Barbaro and Zeller relied upon the uniqueness of a person's search queries. 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.

---

[104] Narayanan and Shmatikov, *Netflix Prize v1*, *supra* note 94, at 13.
[105] *Id.*
[106] *Id.*
[107] *Id.*
[108] *See* BBN Tech., Anonymization & Deidentification, http://www.bbn.com/technology/hci/security/anon (last visited August 2, 2009) (referring to services to remove "'fingerprints' in the data").

a) The Adversary

Computer scientists model anonymization and reidentification as an adversarial game, with anonymization simply an opening move.[109] They call the person trying to reidentify the data the "adversary."[110] They tend 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 is 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."[111] 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.[112] 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 which regulators should adopt: we cannot predict the type and amount of outside information the adversary can access.[113] It is a naïve assumption to assume that the adversary will find it difficult to find the particular piece of data needed to unlock anonymized data.[114] In computer security, this discredited attitude is called "security through obscurity".[115] Not only do reidentification scientists spurn security through obscurity, but in fact they often assume that the adversary possesses the *exact* piece of data—if it exists—needed to unlock anonymized identities, and they try to design responses that protect identity even in this worst case.[116]

It seems wise to adopt this aggressively pessimistic assumption of perfect outside information given the avalanche of information now

---

[109] *See* Irit Dinur & Kobbi Nissim, *Revealing Information while Preserving Privacy*, 2003 ACM SYMP. ON PRINCIPLES DATABASE SYS. 202, 203.

[110] *Id.*

[111] Arvind Narayanan & Vitaly Shmatikov, *De-Anonymizing Social Networks,* 2009 IEEE SECURITY & PRIVACY at 6 [hereinafter *De-Anonymizing Social Networks*].

[112] *See Netflix Prize Study*, *supra* note 4, at 2.

[113] *Id.*

[114] *Id.*

[115] SIMSON GARFINKEL, PRACTICAL UNIX & INTERNET SECURITY 61 (2003) (describing "[t]he problem with security through obscurity").

[116] *Cf.* Cynthia Dwork, *Differential Privacy*, 33RD INT'L COLLOQUIUM ON AUTOMATA, LANGUAGES AND PROGRAMMING PROC. 2 (2006).

available on the Internet[117] 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.[118] Alessandro Acquisti and Ralph Gross—researchers who developed an efficient algorithm for guessing some people's social security numbers[119]—call this the "age of self-revelation."[120]

As only one example from among many, in early 2009, many Facebook users began posting lists called "25 random things about me."[121] The implicit point of the exercise was to bare ones soul—at least a little—by revealing secrets about oneself that ones friends would not already know.[122] "25 random things about me" acts like a reidentification virus,[123] 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.[124]

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.

Database administrators call the hand-folding operation an inner join.[125] An inner join is an operation combining two database tables, connecting rows from one to rows from the other by matching shared information.[126] When the rows in the tables represent people, an inner join assumes that two rows which match along critical fields refer to the

---

[117] *See* Lakshmanan et al., *supra* note 44, at 13:3 ("The assumption that there is no partial [outside] information out there is simply unrealistic in this Internet era.").

[118] *Cf. De-Anonymizing Social Networks*, *supra* note 111, at Part 2 (describing sharing of information obtained from social networks).

[119] Alesandro Acquisti & Ralph Gross, *Predicting Social Security Numbers from Public Data*, 106 NAT'L ACAD. SCI. 10975 (July 7, 2009).

[120] Alessandro Acquisti & Ralph Gross, SSN Study – FAQ, http://www.heinz.cmu.edu/~acquisti/ssnstudy/.

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

[122] *Id.*

[123] *E.g.,* Michael Kruse, *'25 Random Things About Me' Fad Takes Off Like a Virus on Facebook.com*, ST. PETERSBURG TIMES, Feb. 23, 2009.

[124] *De-Anonymizing Social Networks*, *supra* note 111; Lars Backstrom, Cynthia Dwork & Jon Kleinberg, *Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography*, 2007 INT'L WORLD WIDE WEB CONF. 181.

[125] ALAN BELIEU, LEARNING SQL 77 (2005).

[126] *Id.*

same person and can be combined into one row in the output table.[127] For example, if an adversary has one table that looks like this:

Race	Birth Date	Sex	ZIP Code	Diagnosis
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	Obesity
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	Obesity
White	5/5/1964	Male	02139	Fever
White	2/13/1967	Male	02138	Vomiting
White	3/21/1967	Male	02138	Back pain

Table 5: Anonymized Database

and a separate table that looks like this:

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

Table 6: Anonymized Database

and she performed an inner join on the three columns, birth date, sex, and ZIP code, she would have produced this:

Name	Race	Birth Date	Sex	ZIP	Diagnosis	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	Obesity	No
Forest	White	10/23/1964	Male	02138	Short of breath	Yes
Hilary	White	3/15/1965	Female	02139	Hypertension	No

Table 7: Inner Join of Tables 7 and 8 on birth date/ZIP/sex

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.

---

[127] *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 4, at 10.

d) The Myth of the Superuser

Some might object that just because reidentification is possible, it does not mean that it is 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 I have called the "Myth of the Superuser."[128]

The power of reidentification is not a Myth of the Superuser story for three reasons: (1) reidentification techniques are not Superuser techniques. The Netflix study reveals that it is startlingly easy to reidentify people in anonymized data.[129] 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 Excel or Access;[130] (2) there seem to be great financial motivations pushing people to reidentify; and (3) the AOL release reminds us about the power of a small group of bored bloggers.

Moreover, some have misunderstood the Myth of the Superuser argument. 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.[131] I concede that my claim of reidentification power should be held to the high standard of proof, but I argue that I have met that standard.

II. HOW THE FAILURE OF ANONYMIZATION DISRUPTS PRIVACY LAW

Regulators cannot simply ignore easy reidentification, because for decades they enacted many laws 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, and has thrown carefully balanced statutes out of equilibrium.

---

[128] Paul Ohm, *The Myth of the Superuser*, 41 U.C. DAVIS L. REV. 1327 (2008).

[129] *Netflix Prize Study*, *supra* note 4, at 2.

[130] The INNER JOIN command is taught in beginner database texts. *E.g.,* ANDY OPPEL & ROBERT SHELDON, SQL: A BEGINNER'S GUIDE 264 (2008); PAUL WILSON & JOHN W. COLBY, BEGINNING SQL 501 (2005); ALLEN G. TAYLOR, SQL ALL-IN-ONE DESK REFERENCE FOR DUMMIES 309 (2007).

[131] Ohm, *supra* note 128, at ____.

At the very least, legislators must abandon the idea that we protect privacy when we 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 information privacy in the United States and Europe has evolved from discussions in the pages of law reviews, to new, limited common law torts, to broad statutory schemes. Before we can decide how to respond to the rise of easy reidentification, we must first understand two themes from this history of privacy law. First, while privacy torts focus solely on compensating the injured victims of completed privacy harms, the more recent privacy statutes focus more on "problem prevention." The shift from redress to prevention has taken the form of a hunt for PII, a kind of quasi-scientific exercise in information categorization. Second, legislatures have tried to inject balance into privacy statutes, often by relying on robust anonymization.

1. The Privacy Torts: Compensation for Harm

Most privacy law scholars point to a celebrated 19th century law review article by Samuel Warren and Louis Brandeis, *The Right to Privacy*,[132] as the wellspring of information privacy law. In the article, Warren and Brandeis, alarmed by the rise of tabloid journalism, concocted a new right of privacy, urging courts to allow plaintiffs to bring new privacy torts.[133] 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 the injuries suffered by those whose privacy rights are violated. They experience "mental suffering,"[134] "mental pain and distress, far greater than could be inflicted by mere bodily injury,"[135] and "injury to the feelings."[136] 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."[137]

Seventy years later, William Prosser synthesized the case law that followed Warren and Brandeis into four privacy torts,[138] the same 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 which places the plaintiff in a false light in the

---

[132] Samuel D. Warren & Louis D. Brandeis, *The Right to Privacy*, 4 HARV. L. REV. 193, 195-96 (1890).

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

[134] Warren & Brandeis, *supra* note 132, at 213.

[135] *Id.* at 196.

[136] *Id.* at 197.

[137] *Id.* at 219.

[138] 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).

public eye, and (4) appropriation, for the defendant's advantage, of the plaintiff's name or likeness."[139] All four require actual injury, as do all torts.[140]

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

The courts took the lead during the evolution of the privacy torts,[141] while the legislatures stayed mostly in the background, doing little more than occasionally codifying privacy torts.[142] 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 1960's the U.S. government began computerizing records about its citizens, combining this data into massive databases; these actions sparked great privacy concerns.[143] Throughout the decade, many commentators described the threats to privacy from computerization, and helped defeat several government proposals.[144] 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 it called "Fair Information Principles" (FIPS).[145] The FIPS have been enormously influential, having launched statutes,[146] law review articles,[147] and multiple refinements.[148]

FIPS requires a data protection scheme that provides, among other things, notice and consent, access, data integrity, enforcement, and remedies,[149] 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 which 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

---

[139] Prosser, *supra* note 138, 389.

[140] 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 . . . .").

[141] Prosser, *supra* note 138, 389.

[142] *E.g.*, N.Y. CIV. RIGHTS LAW §§ 50-51 (McKinney 1976 & Supp. 1990).

[143] PATRICIA REGAN, LEGISLATING PRIVACY 82 (1995).

[144] Daniel J. Solove, *A Taxonomy of Privacy*, 154 U. PA. L. REV. 477, 505-06 & nn.138-45 (2006).

[145] U.S. Dep't of Health, Educ., & Welfare, Records, Computers, and the Rights of Citizens (1973).

[146] *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)).

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

[148] FTC, Fair Information Practice Principles (1998), available at http://www.ftc.gov/reports/privacy3/fairinfo.shtm; Organisation for Economic Cooperation and Development, OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data (2001).

[149] FTC, *supra* note 148.

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."[150]

Thus, beginning in the 1970's, Congress began to enact statutes designed to reduce the *risk* of *potential* 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,"[151]—Congress tries to enumerate and categorize all of the types of information that raise the risk. They categorize on a macro level—treating health information different from education information which is different from financial information—and on a micro level—making finely wrought distinctions between names, account numbers, and other classes of information. Through this process, they have filled many pages of the U.S. Code with lists and taxonomies, singling out the categories of information that deserve special treatment because of their special ability to lead to harm.[152]

Congress has thus embraced a wholly data-centric approach, the PII-approach, to protecting privacy. This approach assumes lawmakers can evaluate the inherent riskiness of a category of data in a vacuum, assessing with mathematical precision whether or not a particular field of data contributes to the problem enough to be regulated. It tends to ignore messier, human factors that should also factor into a risk assessment, such as the likelihood that anybody will be motivated enough to care about a particular dataset.[153]

It is necessary, however, to disentangle two very different legislative motivations for singling out specific types of information. The easy reidentification result calls into question only one of these motivations. Sometimes, statutes restrict *sensitive* information, the kind of information that causes harm when disclosed.[154] For example, the Driver's Privacy Protection Act (DPPA) provides special protection for "highly restricted personal information," a category that includes sensitive data categories like "photograph" and "medical or disability information."[155] Easy reidentification has not disrupted the logic of provisions like this one. Even though robust anonymization has failed, it still makes sense to categorize and 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* or identifiability of anonymized data. The very same act, the DPPA, defines a second category of "personal

---

[150] Solove, *supra* note 144 at 487-88.

[151] Solove, *supra* note 144 at 487-88.

[152] *See infra* notes 188-192 (giving examples of statutes that list categories of information).

[153] *See infra* Part IV.B.4 (discussing motive).

[154] *De-Anonymizing Social Networks*, *supra* note 111, at app. B (noting that some laws single out information that "itself is sensitive" which is distinguishable from laws that target "deductive disclosure").

[155] 18 U.S.C. § 2725(3)-(4) (2009).

information," which includes all of the information in the "highly restricted" list but also linkable data fields like social security number and driver identification number, for special, but less restrictive, treatment.[156] The law implicitly assumes that this list of linkable personal information is the complete list of information that can link data 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."[157] So too is every privacy statute an "exercise [in] the power of censorship."[158] These laws restrict the free flow of information. This should give lawmakers 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.[159] 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 taken a short cut that has absolved them of the need to engage in rigorous balancing: reliance on the supposed power of anonymization. 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 almost not 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 health care

---

[156] *Id.*

[157] Prosser, *supra* note 138, at 413.

[158] *Id.*

[159] 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 pt. II (enumerating the "benefits of shared information").

and health insurance in this country.[160] Among the other things it accomplishes, HIPAA is a significant privacy law, because Title II of the Act mandates compliance with health privacy regulations, which have been promulgated by the Department of Health and Human Services ("HHS") and are now known as the HIPAA Privacy Rule.[161]

In many ways, the HIPAA Privacy Rule represents the high water mark for how regulators use PII to balance risks to privacy against valuable uses of information.[162] In fact, HIPAA even demonstrates Congress's early sensitivity to the power of reidentification, through its treatment of what it calls "de-identified health information." (DHI)[163]

HIPAA itself exempts DHI from any regulation whatsoever.[164] The statute defines DHI as information which "does not identify an individual" nor provides "a reasonable basis to believe that the information can be used to identify an individual."[165] The Privacy Rule elaborates this vague reasonability standard 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 the data is not individually identifiable.[166]

Second, data is DHI under the so-called "safe harbor standard" if the covered entity suppresses or generalizes eighteen enumerated identifiers.[167] 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

---

[160] Pub. L. 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.

[161] *Id.* § 264 (directing Secretary of Health and Human Services to submit standards for protecting privacy); 45 C.F.R. pts. 160 & 164 (2009) (HIPAA Privacy Rule).

[162] Jay Cline, *Privacy Matters: When is Personal Data Truly De-Identified?*, COMPUTERWORLD, July 24, 2009 ("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, 15 U.S.C. § 6801 et seq.; Children's Online Privacy Protection Act of 1998, Pub. L. No. 106-170, 15 U.S.C. § 6501 et seq.

[163] 45 C.F.R. §§ 164.502(d)(2), 164.514(a), (b) (2009).

[164] *Id.*

[165] 45 U.S.C. § 1320d(6) (2009).

[166] 45 C.F.R. § 164.514(b)(1) (2009).

[167] *Id.* at § 164.514(b)(2).

requires the researcher to generalize birth dates to years[168] and zip codes to their initial three digits.[169]

In the Privacy Rule, regulators relied on their faith in the power of anonymization as a stand-in for a meaningful costs-benefits balancing. This is an opportunity lost, because it is hard to imagine another privacy problem with such starkly presented benefits and costs. On the one hand, when medical researchers can freely trade information, they can develop treatments to ease human suffering and save lives. On the other hand, our medical secrets are among the most sensitive we hold. It would have been quite instructive to see how regulators might have weighed such stark choices.

But thanks to the power of anonymization, regulators bypassed the nuances of the balancing. Congress and HHS concluded simply that by making data unidentifiable, they could allow health professionals to 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.[170]

Alas, the failure of anonymization unravels these expectations. 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.[171]

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

EU lawmakers have also relied upon the power of anonymization to skip past many difficult balancing questions. Unlike the American approach with HIPAA, however, the EU enacted a broad, industry-spanning law,[172] the Data Protection Directive, which purports to cover any "personal data," held by any data administrator.[173] Data is personal data if it can be used to identify, "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."[174]

The EU never intended the Directive to apply to all data. Instead it meant for "personal data" to exclude at least some data—data which

---

[168] *Id.* at § 164.514(b)(2)(C).

[169] *Id.* at § 164.514(b)(2)(B) (requiring additional perturbation for ZIP codes with 20,000 or fewer residents).

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

[171] *See infra* Part IV.D.1.

[172] 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 Aug. 13, 2009).

[173] EU Data Protection Directive, *supra* note 2, art. 2(a).

[174] *Id.*

was not "directly or indirectly" identifiable, anonymized data—from regulation. Like their U.S. counterparts, the 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 some space in society for the storage and transfer of data—anonymized data—that had nothing to do with people, thereby providing room for unencumbered innovation and free expression.

Whether and to what extent the Directive retains such a preserve has been debated in the internet privacy context.[175] 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.[176] 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.[177] Like 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.[178] Every computer reveals its IP address to every other computer it contacts,[179] so, every time I visit Google, my computer reveals its IP address to a Google computer.[180] Following longstanding industry practice, Google records my IP address along with details about what I am doing when using Google's services.[181]

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.[182] Google's competitors, Microsoft and Yahoo, are much more thorough, throwing away entire IP addresses.[183] Specifically, an IP address is composed of four equal pieces called *octets*,[184] and Google

---

[175] Note, Frederick Lah, *Are IP Addresses "Personally Identifiable Information?"*, 4 I/S: J.L. & POL'Y FOR INFO. SOC'Y 681 (2008-09).

[176] *E.g.*, Saul Hansell, *Europe: Your IP Address is Personal*, N.Y. TIMES BITS BLOG, Jan. 22, 2008, http://bits.blogs.nytimes.com/2008/01/22/europe-your-ip-address-is-personal/.

[177] DOUGLAS COMER, INTERNETWORKING WITH TCP/IP, ch. 3 (2006)

[178] *Id.*

[179] *Id.*

[180] *Id.*

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

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

[183] Peter Cullen, *Microsoft Supports Strong Industry Search Data Anonymization Standards*, MICROSOFT THE DATA PRIVACY IMPERATIVE BLOG, Dec. 8, 2008, http://blogs.technet.com/privacyimperative/archive/2008/12/08/microsoft-supports-strong-industry-search-data-anonymization-standards.aspx; *Behavioral Advertising: Industry Practice and Consumers' Expectations 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.).

[184] DOUGLAS COMER, INTERNETWORKING WITH TCP/IP 53 (2006)

stores the first three octets and deletes the last, claiming that this practice protects user privacy sufficiently.[185]

Again, at its core this is a debate about balance—between the wonderful new innovations Google promises it can deliver by studying our behavior[186] versus the possible harm to users whose IP addresses are known or revealed—and again, claims that we should trust robust anonymization stand in for nuanced, careful costs-benefits balancing arguments. Google promises we can have our cake while it eats it too thanks to 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[187] ever written in the U.S. and 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 upsets the balance of 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,[188] 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 can be used by an adversary with rich outside information 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,"[189] while protecting much less "the 5-digit zip code" and "information on vehicular accidents, driving violations, and

---

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

[186] In 2008, to try to placate those worried about privacy, Google posted a series of blog posts "about how [they] harness the data we collect to improve our products and services for [their] users." *E.g.,* Google, *Using Data to Fight Webspam,* THE OFFICIAL GOOGLE BLOG, June 27, 2008, http://googleblog.blogspot.com/2008/06/using-data-to-fight-webspam.html (linking to earlier posts in series).

[187] 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.

[188] 45 C.F.R. §§ 164.502(d)(2), 164.514(a), (b) (2009).

[189] 18 U.S.C. § 2725(3) (2009).

driver's status."[190] Similarly, the Federal Education Rights and Privacy Act (FERPA) singles out for special treatment "directory information" including, among other things, "name, address, telephone listing, date and place of birth, [and] major field of study."[191] 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."[192] 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, like HIPAA, these provisions seem arbitrary and underprotective.

In contrast, anonymization 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,[193] 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. A careful balance struck by legislators between privacy and information flow shifts wildly to impose 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 line-drawing protections promised by HIPAA illusory and underinclusive, because it deregulates the dissemination of certain 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. Nothing can escape government regulation when laws are written so expansively. We should tolerate neither result, because they both evade the careful balance supposedly at the heart of both types of laws.

Most other laws match one of these two forms. Even the few that do not fit so neatly into one or the other often contain terms that seem indeterminate and unpredictable in light of easy reidentification. As only one example, the Stored Communications Act in the U.S. applies to "record[s] or other information pertaining to a subscriber to or customer" without specifying how directly to identity the records must relate to the subscriber or consumer.[194] Courts will struggle to decide whether anonymized records fall within this definition. The vagueness of provisions like these will invite costly litigation and may result in irrational distinctions between jurisdictions and between laws.

---

[190] *Id.*
[191] 20 U.S.C. § 1232g(a)(5)(A) (2009).
[192] 21 C.F.R. § 21.70 (2009).
[193] EU Data Protection Directive, *supra* note 2, art. 2(a).
[194] 18 U.S.C. § 2702(c) (2009).

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 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.[195] Today, two years after computer scientists exposed the power of these categories of data to identify, no law or regulation yet treats them as PII.

Maybe two years has not been enough time to give regulators the chance to react. For example, HIPAA's Privacy Rule does incorporate Dr. Sweeney's research.[196] 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.[197] Should this be the model of future privacy law reform—whenever reidentification science identifies 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, and just like the carnival game, even if you manage to whack one mole, another will pop right up. No matter how effectively you follow the latest reidentification research, folding newly identified data fields into your laws and regulations, reidentification researchers will always find another data field type you do not yet cover.[198] The list of potential PII will never stop growing until it includes everything.[199]

Consider Narayanan and Shmatikov's latest reidentification study.[200] The researchers have reidentified anonymized users of an online social network based almost solely on the stripped-down graph of connections between people.[201] By comparing the structure of this graph to the non-anonymized graph of a different social network, they could reidentify many people even ignoring almost all usernames, activity

---

[195] The Video Privacy Protection Act, enacted in 1988, protects lists of movies watched not because they are PII, but because they are sensitive. For more on the distinction, see *infra* Part II.A.2.

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

[197] 45 C.F.R. §§ 164.502(d)(2), 164.514(a), (b) (2009).

[198] *See De-Anonymizing Social Networks*, *supra* note 111, at app.B ("While some data elements may be uniquely identifying on their own, *any* element can be identifying in combination with others." (emphasis in original)).

[199] *Id.*; Dinur & Nissim, *supra* note 109, at 202 ("[T]here usually exist other means of identifying patients, via indirectly identifying attributes stored in the database.").

[200] *De-Anonymizing Social Networks*, *supra* note 111, at app.B.

[201] *Id.*

information, photos, and every other single piece of identifying information.[202]

To prove the power of the method, the researchers obtained and anonymized the entire Twitter social graph, reducing it essentially to nameless, identity-less nodes representing people connected to other nodes representing Twitter's "follow" relationships. Next, they compared this mostly de-identified husk of a graph[203] 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.[204] 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[205]—to the HIPAA Privacy Rule list and to the lists of every other PII-focused law? Of course not.

Instead, lawmakers and regulators should reevaluate any law or regulation that draws distinctions based on whether particular data types can be linked to identity, and they should never draft a new law or rule that draws such a distinction. This is an admittedly disruptive prescription. PII has long served as the center of mass around which the entire data privacy debate has orbited.[206] From this point forward, we need a new organizing principle; we should abandon deidentification, the PII whack-a-mole game, and release-and-forget anonymization.

Although this proposal is disruptive, it is also necessary. Too often, the only thing that gives us comfort about the privacy provided by a data administrator's practices is that the administrator has gone through the motions of identifying and deleting PII, and in those cases, we deserve no comfort at all. Anonymization has become privacy theater;[207] it should no longer be confused as a way to provide meaningful guarantees of privacy.

---

[202] *Id.*

[203] *Id.* at § 6.2.2. In order to make this 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 §§ 5.1, 6.2.2.

[204] *Id.*

[205] *Id.* at § 2 (surveying examples of how social-networks data is shared).

[206] Leslie Ann Reis, *Personally Identifiable Information*, *in* ENCYCLOPEDIA OF PRIVACY 383-85 (William G. Staples ed., 2006).

[207] Noted security expert Bruce Schneier refers often to "security theater," meaning security measures that create the illusion of security without actually improving security. *E.g.*, Bruce Schenier, *In Praise of Security Theater*, SCHNEIER ON SECURITY BLOG, Jan. 25, 2007, http://www.schneier.com/blog/archives/2007/01/in_praise_of_se.html.

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; [208] let us simply strike it from our debates. A word which should mean, "try to achieve anonymity" too often has been used to mean, "achieve anonymity," by technologists and non-technologists alike. We need a new word that conjures effort, not achievement.

Latanya Sweeney has similarly argued against using forms of the word anonymous when they are not literally true.[209] 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."[210] Owing to her influence, the HIPAA Privacy Rule explicitly talks about the "de-identification of protected health information."[211]

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 seem to treat claims of deidentification as promises of robustness, when in reality, people can deidentify robustly or weakly.[212] Whenever a person uses the unmodified word, deidentified, we should demand details and elaboration.

Better yet, we need to start using a new word for privacy-motivated data manipulation that connotes only effort but not success. I propose "scrub." Unlike anonymize or deidentify, it conjures only effort. One can scrub a little, a lot, not enough, 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.[213] In fact, Dr. Sweeney herself has created a system she calls

---

[208] 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 medical." *Id.*

[209] 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." (emphasis in original)).

[210] *Id.*

[211] 45 C.F.R. § 164.514 (2009) (defining term).

[212] 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, 299 (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.

[213] *See, e.g.,* Jeremy Kirk, *Yahoo to Scrub Personal Data After Three Months*, IDG NEWS SERVICE, Dec. 17, 2008,

Scrub for "locating and replacing personally-identifying information in medical records."[214]

III. HALF MEASURES AND FALSE STARTS

Abandoning 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 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, not 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 compensates harm alone. 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 the technologists *will* develop better privacy-protection techniques, they will run 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 ban reidentification. Such a ban will almost certainly fail.

A. Punish Those who Harm Strictly

If reidentification will make it easier for malevolent actors like identity thieves, blackmailers, and unscrupulous advertisers to cause harm, perhaps we need to step up our enforcement of pre-existing laws prohibiting identity theft,[215] blackmail,[216] and unfair marketing practices.[217] Anything we can do to deter those who harm and to provide reme-

---

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, *available at* http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=78230.

[214] Latanya Sweeney, *Replacing Personally-Identifying Information in Medical Records, the Scrub System*, 1996 J. Am. Med. Informatics Ass'n Proc. 333..

[215] *E.g.,* 18 U.S.C. § 1028 (2009); CAL. PENAL CODE § 530.5 (2009); MASS. GEN. LAWS ch. 266, § 37E (2009); N.Y. PENAL LAW §§ 190.77-190.84 (2009).

[216] *E.g.,* 18 U.S.C. § 872 (2009) (extortion by federal government officials).

[217] *E.g.,* 15 U.S.C. § 45 (2009) (FTC unfair methods of competition provision); CAL. BUS. & PROF. CODE §§ 17200-210 (2009).

dies for those harmed is, in light of the increased power of reidentification, imperative. But this is merely a necessary response; it is not sufficient alone.

The failure of anonymization might tempt legislators to abandon the forty-year preventative turn in privacy law, returning to the way we once protected privacy solely like a tort, compensating those harmed and punishing and deterring those who harm. Regulators must know that this would be a grave mistake. They must understand how 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."[218] This is why we should worry even about reidentification events that seem to expose only non-sensitive 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 non-threatening and trivial,[219] because they did not care if others knew the movies they had rented. 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,

---

[218] *Netflix Prize Study*, *supra* note 4, at 9 (emphasis in original).

[219] *E.g.,* Comment of user chef-ele to Netflix Prize Community message board, Nov. 28, 2007, 09:04:54, http://www.netflixprize.com/community/viewtopic.php?id=809 ("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 user jimmyjot to personal website of study author Arvind Narayanan, Feb. 17, 2008, http://arxivblog.com/?p=142 ("Choice of movies also does not tell a whole lot."). See also various comments to post Anonymity of Netflix Prize Dataset Broken, Nov. 27, 2007, 9:23 AM, http://it.slashdot.org/article.pl?sid=07/11/27/1334244&from=rss.

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.[220]

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.[221] User john_doe20 on IMDb is likely to be john_doe20 on many other websites as well.[222] Relying on this assumption, the researchers could have linked each living, breathing person revealed through Facebook, through the Netflix Prize data, through the IMDb username, to a pseudonymous user at another website. Perhaps they could have unearthed the identity of the person who had savagely harassed people on a message board.[223] Maybe they could have determined who had helped plan an attack on a computer system on a 4chan message board.[224] 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.

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.[225] 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."[226] 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.

---

[220] Of course, they could have skipped this step if they could have connected records in the IMDb database directly to Facebook records, without any need to route through the Netflix data. 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 Yankowitz for this point.

[221] Arvind Narayanan, *Lendingclub.com: A De-Anonymization Walkthrough*, 33 BITS OF ENTROPY BLOG, Nov. 12, 2008, http://33bits.org/2008/11/12/57/ ("Many people use a unique username everywhere . . . ."); *De-Anonymizing Social Networks*, *supra* note 111, at 6-7 (relying on fact that users tend to reuse usernames on different social networks).

[222] *See* Narayanan, *supra* note 221.

[223] Danielle Citron, *Cyber-Civil Rights*, 89 B.U.L. REV. 61, 71-75 (discussing harassing comments on AutoAdmit internet discussion board).

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

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

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

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.

For almost every person on earth, there is at least one fact about them stored in a computer database that an adversary could use to blackmail, discriminate against, harass, or steal the identity of him or her. 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 access to it to our worst enemies.

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 non-sensitive 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. They need to understand the flaw in this argument. Any reidentifier brings his target closer to harm through accretion even when he fails to complete the path.

Computer scientists formalize this idea using a complicated but useful construct called entropy.[227] In thermodynamics, entropy measures disorder in a system; in information theory, it tracks the amount of information needed to describe possible outcomes.[228] In reidentification science, entropy measures how close an adversary is to connecting a given fact to a given individual;[229] it measures the length of the inference chains heading in opposite directions; it quantifies the remaining uncertainty.

Consider entropy in the children's game, Twenty Questions.[230] At the start of a game, the Answerer thinks of some subject the Questioner must discover through yes/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 Answerer is like the anonymizer and the Questioner is like the reidentifier, connecting outside information to the anonymized database, reducing entropic uncertainty about the identity of his target.

---

[227] Arvind Narayanan, *About 33 Bits*, 33 BITS OF ENTROPY BLOG http://33bits.org/about/ (explaining the concept of entropy).

[228] The concept originated with a seminal paper by Claude Shannon. *A Mathematical Theory of Communication*, 27 BELL SYS. TECH. J. 379 (1948).

[229] Narayanan, *supra* note 227.

[230] I am indebted to Anna Karion for the analogy.

Entropy formalizes the accretion problem. We should worry about reidentification attacks which do less than connect 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.

4. The Need to Regulate Before Completed Harm

If we fail to regulate reidentification that has not yet ripened into harm, then adversaries will nudge each of us closer to the brink of connection to our private database of ruin with impunity. It will take some time before most people become precariously compromised, and whether it will take months, years, or decades is difficult to measure or predict. Because some people have more to hide than others, the burden of decreasing entropy will not be distributed equally across society.[231]

But when we finally arrive at that day when most of us are on the brink of being connected to our own personal 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 do anything to break that link.[232] Anybody who wants can replicate their result by downloading the Netflix Prize data[233] and mining the IMDb 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 large amount of entropic uncertainty, we can rest easy. But as data is connected to data, and as adversaries whittle down entropy, we will soon, every one of us, 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 the parallel advances in reidentification and anonymization as an arms race, and even though the reidentifiers have raced ahead

---

[231] 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. I tend to believe that the numbers of people in these groups are so small in our modern society that they are like myths—the unicorns and mermaids of information privacy. Still, those who disagree—who believe that society is full of people free of secrets or not described in data—will probably disagree with my assertion of an urgent need to regulate. Ultimately, the size of these groups is a difficult empirical question we might try to answer.

[232] 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 4, at 9.

[233] Netflix Prize, http://www.netflixprize.com/index (last visited August 4, 2009).

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 of 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.[234] 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."[235] 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.

Some researchers go further: At least for useful databases, perfect anonymization is impossible.[236] Theorists call this the impossibility result.[237] There is always some piece of outside information which could

[234] Shuchi Chawla et al., *Toward Privacy in Public Databases*, 2 THEORY CRYPTOGRAPHY CONF. 363 (2005).
[235] *Id.* at 363.
[236] Cynthia Dwork, *Differential Privacy*, 33RD INT'L COLLOQUIUM ON AUTOMATA, LANGUAGES AND PROGRAMMING PROC. (2006).
[237] *Id.*

be combined with anonymized data that will reveal private information about an individual.[238] Cynthia Dwork offers one proof.[239]

Although useful data can never be perfectly private, it is important to understand what Dwork means by impossibility.[240] For some situations, she is talking about a kind of theoretical, pathological impossibility involving a type of privacy breach that may concern policymakers very little. To use her 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.[241] Although we would properly say that the utility of the anonymized data revealed a private fact when combined with outside information, we would be foolhardy to regulate or forbid the release of databases containing aggregated height data to avoid this possibility. In this case, the richness of the outside information creates almost all of the privacy breach, and the statistic itself contributes very little.

Thus, 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 it instead, as Dwork puts it, "leads naturally to a new approach to formulating privacy's goals."[242] She calls her preferred goal, "differential privacy," and she ties it to so-called interactive techniques. Differential privacy and interactive techniques are discussed shortly.

b) The Inverse and Imbalanced Relationship

Other theoretical work suggests that release-and-forget anonymization techniques are particularly ill-suited for protecting the privacy while preserving the utility of data. Professor Shmatikov, one of the Netflix Prize researchers, co-authored a study with Justin Brickell which 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."[243]

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, for a health study perhaps the diagnoses, for an education study the grade point averages, and for a labor study the salaries.[244] We would hope that real-

---

[238] Dinur & Nissim, *supra* note 109, at 203 (showing, for a particular model, "tight impossibility results," meaning that privacy would require "totally ruining the database usability.").
[239] Dwork, *supra* note 236.
[240] *Id.*
[241] *Id.*
[242] *Id.*
[243] Brickell & Shmatikov, *supra* note 46, at 70.
[244] *Id.*

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.[245] Although the full details are beyond the scope of this Article, consider the intuition revealed in the following graph:

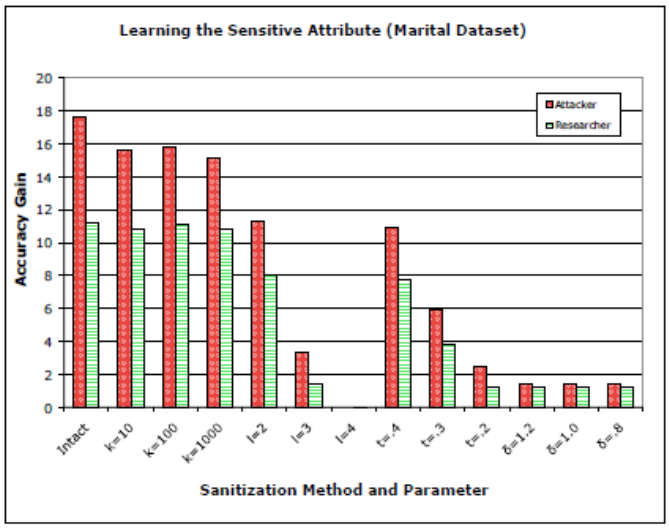

Figure 1: Effects on Privacy and Utility of Anonymization[246]

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, solid bar represents the privacy of the data, with smaller bars signifying more privacy. The right, striped 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, solid bar next to a long, striped 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, represent 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.[247] Although the second trio of bars, those labeled with the letter l,[248] show much greater improvements in privacy than the first trio, such improvements come only at great losses to utility.

---

[245] *Id.*

[246] *Id.*

[247] These bars represent techniques that achieve *k-anonymity*, a widely embraced metric for strong anonymity. Brickell & Shmatikov, *supra* note 46, at 70; Sweeney, *supra* note 7 (defining *k-anonymity*).

[248] These bars represent *l-diversity*, another widely adopted metric. The final six bars represent *t-closeness*. Brickell & Shmatikov, *supra* note 46, at 70.

These results prove that for traditional, widespread, release-and-forget anonymization, not only are privacy and utility related, but also 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, suppression-and-generalization, release-and-forget, PII-focused anonymization techniques, any data which 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 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, the 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*.[249] 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%. In most cases, reidentifiers will find it much more difficult to link answers like these to identity than if they had access to the raw underlying data.

Researchers can do even better. Using one class of interactive techniques, those that satisfy a requirement called differential privacy,[250] 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 use the data to unearth information about the people in the database.[251]

Finally, just as these techniques relax the requirement of release, other techniques work by monitoring what happens to data after release—they refuse to forget. These techniques involve the use of *access controls* and *audit trails* and borrow from computer security research.[252] Using these techniques, data administrators release their data but only after first protecting it with software that limits access and tracks usage.

---

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

[250] *See* Dwork, *supra* note 116, at 3.

[251] *See* Adam & Wortmann, *supra* note 57, at 540 (describing "output-perturbation approach").

[252] *See* RICK LEHTINEN ET AL., COMPUTER SECURITY BASICS 66 (2006).

The data analyst who receives the protected data will be able to interact with it only in limited, circumscribed ways, and the analyst's every move will be watched by the audit trail and reported back to the data administrator or a third-party watchdog.

3. The Limitations of the Improved Techniques

Unfortunately, none of these alternatives replace all of the broken promises of release-and-forget anonymization. For starters, these techniques tend to be less flexible than traditional anonymization. Interactive techniques require the constant participation of 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.

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 War.[253]

Similarly, interactive techniques can be gamed—for example through repeated questions that build on one another—to reveal more identity than it seems. These techniques take advantage of so-called "signaling information" to reveal parts of the database that are not expressly revealed.

Interactive techniques that introduce noise may not be useful in some situations.[254] For example, 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. Techniques that satisfy differential privacy also require complex calculations that can be costly to perform.[255]

Finally, computer security researchers have thoroughly documented the problem with creating robust access controls.[256] Simply put, even the best computer security solutions are prone to have bugs, and the best computer security solutions are expensive to create and dep-

---

[253] William Seltzer & Margo Anderson, *After Pearl Harbor: The Proper Role of Population Data Systems in Time of War*, 2000 POPULATION ASSOC. AM. 1.

[254] For example, researchers give the example of a city analyzing census data to determine where precisely to run a bus line to serve elderly residents. Noise introduced to provide privacy may inadvertently produce the wrong answer to this question. Chawla et al., *supra* note 234, at 4.

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

[256] BRUCE SCHENEIER, SECRETS AND LIES: THINKING SENSIBLY ABOUT SECURITY IN AN UNCERTAIN WORLD (2003).

loy.[257] 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, but it will do nothing to protect all of the data which has been stored or disclosed in the past: a database which has already been 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 a silver bullet alternative. Technology cannot save the day, and regulation must play a role.

C. Ban Reidentification

Finally, perhaps we simply need to ban reidentification, as some have urged.[258] Lawmakers can offer a straightforward argument for a ban: by anonymizing data, a data administrator announces her intent to protect the privacy of her data subjects, and this announcement often convinces data subjects to consent to provide their data to the database in the first place. A reidentifying adversary therefore 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 when somebody reidentifies?[259] Reidentification can happen completely in the shadows. Imagine Amazon.com anonymizes its customer purchase database before transmitting it to a marketing firm. Imagine further that although the marketing firm promises not to reidentify the people in Amazon's database, it could increase profits significantly if it could access reidentified rather than anonymized data. If the marketing firm breaks its promise and reidentifies, how will Amazon or anybody else ever know? The marketing firm can conduct the reidentification completely in secret, and the gains in revenue may not be detectable to the vendor.

This profits-in-private, practical problem appears insurmountable. Three forces might ameliorate the problem a little. First, lawmakers might pair a ban with better enforcement, for example by declaring reidentification a crime and a felony and providing extra money to the

---

[257] *Id.*; *Cf.* FREDERICK P. BROOKS, THE MYTHICAL MAN-MONTH (1975) (discussing how software engineering principles lead to bugs).

[258] Earl Lane, *A Question of Identity: Computer-Based Pinpointing of 'Anonymous' Health Records Prompts Calls for Tighter Security*, NEWSDAY, Nov. 10, 2000 at C8 ("Our goal has been to get a national policy making it illegal to re-identify an anonymized database." (quoting Janlori Goldman, head of the Health Privacy Project at Georgetown University)).

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

FBI and FTC for enforcement.[260] Moreover, lawmakers can give private citizens a cause of action against those who reidentify.[261] Second, lawmakers can mandate software audit trails for those who use anonymized data. Finally, a smaller scale ban, one imposed only on trusted recipients of specific databases, may be much easier to enforce. For example, a ban prohibiting government data-miners from reidentifying may be enforceable.[262]

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 PII

Once regulators take PII off the table, and after they conclude that the three solutions provided above are not enough to restore balance to privacy law, they must do more. They should weigh the benefits of unfettered information flow against the costs from 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 determine when 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. Principles of Post-PII Privacy Regulation

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—

---

[260] The DMCA provides criminal penalties along with civil remedies. 17 U.S.C. § 1204 (2009).

[261] 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 (2009).

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

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.[263]

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?[264] Usually, these competing choices are labeled, respectively, the European and United States approaches. In a post-anonymization 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.[265] Solove says that we should simultaneously "resolve privacy issues by looking to the specific context,"[266] while at the same time use "a general framework to identify privacy harms or problems and to understand why they are problematic."[267]

Thus, the U.S.'s exclusively sectoral approach is flawed, because it allows entire industries to escape privacy regulation completely. It rests upon the illusion that some data, harmless data, is so bland and non-threatening that it is not likely to lead to harm if it falls into the wrong hands. Entire industries remain subject to no federal privacy regulation based on this rationale. 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.

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 obligations of the Directive.[268] It might have made good

---

[263] *See* Chawla et al., *supra* note 234, at 5 (noting a relative advantage in one interactive technique because with it "the real data can be deleted or locked in a vault, and so may be less vulnerable to bribery of the database administrator").

[264] *See, e.g.*, Schwartz, *supra* note 147.

[265] DANIEL J. SOLOVE, UNDERSTANDING PRIVACY 46-49 (2008).

[266] *Id.* at 48.

[267] *Id.* at 49.

[268] *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").

sense to impose such strict requirements (notice, consent, disclosure, accountability, etc.) 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 *any* database containing facts, no matter how well scrubbed, relating to people, no matter how directly or indirectly, very likely now fall 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, for 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 in fact mean "almost all data," the obligations of the Directive might seem a burden too many.

For these reasons, the European Union might want to reconsider whether it should *lower* the floor of its comprehensive data handling obligations. Even if it does not do this (but especially if it does) the EU should also begin to tackle privacy regulations sectorally much more often than they do today. 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.[269] 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" accounts.[270] This approach also stands in stark distinction to the great weight of advice given by other information privacy scholars and activists who tend to valorize sweeping, society-wide approaches to protecting privacy and have nothing complimentary to say 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

---

[269] *Id.* at Part III.D (applying approach to privacy to three case studies).

[270] *Cf.* Helen Nissenbaum, *Privacy as Contextual Integrity*, 79 WASH. L. REV. 119, 155 (2004).

reader elsewhere for that topic.[271] The rest of the article explores how to enact 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-PII, privacy law look like?

3. The Test

The problem with PII-based solutions is that they focus on the nature and quality of data in a vacuum, ignoring human factors like motive and trust. With this new test the lawmaker or regulator should consider instead several factors that together answer a new question, what is the risk of reidentification in this particular context? Of course, such risks can never be calculated with mathematical precision, particularly when messy human factors are considered, but they can at least produce a rough high/medium/low risk answer.

Then, regulators should multiply this risk—impressionistically, not mathematically—by the sensitivity of the information put at risk. They should regulate medical diagnoses more stringently than television watching habits, for example, because the path to harm is more direct for the former than the latter. This multiplied factor is the *cost* to privacy of not regulating. Finally, against the cost, regulators should weigh the *benefits* of unfettered information flow in the particular context, to decide whether and how much to regulate.

B. Factors for Assessing the Risk of Privacy Harm

Regulators should weigh the following factors to determine the risk of reidentification in a given context. The list is not exhaustive; other factors might be relevant.[272] 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 (factor three), the risk of reidentification increases. Regulators should step through 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 (factor two) while declining to regulate private uses of data.

---

[271] *E.g.*, Solove & Hoofnagle, *supra* note 146.

[272] The European privacy watchdog group, 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.

European Union Article 29 Data Protection Working Party, *Opinion 4/2007 on the Concept of Personal Data*, 01248/07/EN WP 136 at 15 (June 20, 2007).

1. Data Handling Techniques

How do different data handling techniques affect the risks of reidentification? The 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% risk while interactive techniques satisfying differential privacy raise a 1% risk. Still, one hopes that computer scientists can refine their research to generate a rough relative ordering of different techniques. Perhaps, at the very least, researchers can grade data handling practices on a high risk/medium risk/low risk of reidentification scale.

There are encouraging signs that computer scientists will be able to devise such a rubric.[273] As discussed earlier, computer scientists have made strides in aggregation and interactive techniques, and they have developed useful benchmarks like differential privacy.

2. 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 can limit the flow of information through regulation 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."[274] Companies like Netflix release great stores of information they once held closely to try to harness these masses.[275]

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[276] and the appointment of federal officials like CTO Aneesh Chopra[277] and CIO Vivek Kundra,[278] the ad-

---

[273] Some computer scientists have already tentatively offered studies that attempt to categorize the risk of reidentification of different techniques. Lakshmanan et al., *supra* note 44; Adam & Wortmann, *supra* note 57. These studies do not take into account the latest advances in reidentification, but they are models for future work.

[274] SUROWIECKI, *supra* note 14.

[275] *See* Thompson, *supra* note 91.

[276] Data.gov, About, http://www.data.gov/about (last visited Aug. 10, 2009) ("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.").

[277] Nate Anderson, *Obama Appoints Virginia's Aneesh Chopra US CTO*, ARS-TECHNICA, April 20, 2009, http://arstechnica.com/tech-policy/news/2009/04/obama-appoints-virginias-aneesh-chopra-us-cto.ars.

ministration has promised to release massive databases heralding a twenty-first century mode of government openness.[279] Amidst the accolades that have been showered upon the government for these efforts,[280] one should pause to consider the costs. We must remember that utility and privacy are two sides of the same coin,[281] 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,[282] the increase in sunlight and utility.

3. Quantity

In every reidentification study cited, the researchers were aided by the size of the database. Would-be reidentifiers will find it easier to match data to outside information when they can access many records indicating the personal preferences and behaviors of many people. As a rough rule, more outside information is better than less.

Most privacy laws regulate data quality but not quantity.[283] Laws dictate what data administrators can do with data depending on 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. Lawmakers should consider enacting new quantitative limits on data collection and retention.[284] They might consider laws, for example, mandating data destruction after a set period of time. They should also support imposing quantitative caps on data collection.

---

[278] Brian Knowlton, *White House Names First Chief Information Officer*, N.Y. TIMES THE CAUCUS BLOG, March 5, 2009, http://thecaucus.blogs.nytimes.com/2009/03/05/white-house-names-first-chief-information-officer/.

[279] *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.").

[280] *E.g.,* SunlightLabs.com, Redesigning the Government: Data.gov, http://www.sunlightlabs.com/blog/2009/04/16/redesigning-government-datagov/ (April 16, 2009); Information Aesthetics, Data.gov: How to Open Up Government Data, http://infosthetics.com/archives/2009/03/open_up_government_data.html (March 13, 2009). The Center for Democracy and Technology has posted a supportive but more cautious memo, flagging concerns about Data.gov involving deidentification and reidentification. Government Information, Data.gov and Privacy Implications, Policy Post 15.13 (July 13, 2009), http://cdt.org/publications/policyposts/2009/13 ("While Data.gov has great potential, there are important privacy implications associated with data disclosure.").

[281] *See supra* Part III.B.1.a.

[282] *See supra* Part III.B.1.b.

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

[284] *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) (arguing search engines should store queries for only six months).

4. Motive

In many contexts, sensitive data is held only by a small number of actors who lack the motive to reidentify.[285] 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 which 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.[286]

5. 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 under the shared hallucination of anonymization, we trusted the technology so we did not have to trust the recipients of data, but 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.

C. Applying the Test

Once again, regulators should consider the cost to privacy of not regulating to be the risk of reidentification multiplied by the sensitivity of the data. This (rough) product represents the cost to privacy, the likelihood of significant privacy harm. They should compare this to the 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.

---

[285] *See* EU Data Protection Directive, *supra* note 2, recital 26 ("[T]o determine whether a person is identifiable, account should be taken on all the means likely reasonably to be used . . . to identify the said person.").

[286] 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.

Jay Cline, *Privacy Matters: When is Personal Data Truly De-Identified?*, COMPUTERWORLD, July 24, 2009.

1. Option One: Surrender

If regulators conclude that the benefits of unfettered information significantly outweigh the costs to privacy in a particular context, they might decide to surrender.[287] Perhaps lawmakers will see reidentification as yet 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 to life in a society with very little privacy.

A famous version of this argument comes from David Brin.[288] A decade ago, Brin foresaw how technology would chip away at our expectations of privacy. Brin argued that in the face of decreasing privacy, we should redirect the energy we would waste on preventing privacy invasions, a futile endeavor, into instead turning the telescope around, to using technology to watch the watchers.[289] The very same technology that can and will be used to spy on us, Brin argued, can also shine light on what governments and corporations are doing.[290] Do not ban reidentification or impose anonymization, Brin might have argued, but instead force companies to reveal exactly what they are doing, and allow us to peer as deeply into their secrets as they can peer into ours.

The problem with surrender is how it neglects or downplays the possibility of harm. If we allow unfettered reidentification, adversaries will discover and reveal harmful, damaging, sometimes even reputation-destroying secrets. Of course, these harms will not be distributed equally across society, and some of us will not be harmed significantly, but I believe that enough people will suffer enough harm to justify even more government regulation than we have today, especially because the accretion problem means today's petty indignity provides the key for unlocking tomorrow's harmful secret.

I recognize, however, that in some narrow circumstances, surrender may be appropriate. When countervailing values are so important that they outweigh even the risk of grave harm from reidentification, surrender may be justifiable. In particular, when national security, public safety, or public health are at risk, regulators may try to tailor laws to allow reidentification to advance security and safety. I would hope that these exceptions would be very narrowly drawn, exquisitely justified, and well-targeted to advancing the countervailing value.[291]

2. Option Two: Carefully Restrict the Flow of Information

Much more often, regulators will conclude that the costs to privacy outweigh the benefits of unfettered information flow, particularly given the thumb on the scale regulators should place to prevent easy

---

[287] *Cf.* Peter Dizikes, *Your DNA is a Snitch*, SALON.COM, Feb. 17, 2009, http://www.salon.com/env/feature/2009/02/17/genetic_testing/ (describing Harvard's Personal Genome Project's advice to contributors to its genetic database to "forget about privacy guarantees" because of possibility of reidentification).

[288] DAVID BRIN, THE TRANSPARENT SOCIETY (1999).

[289] *Id.*

[290] *Id.*

[291] For an example of such an exception, see the solution proposed for academic research and health privacy *infra* Part IV.D.1.

access to the database of ruin. When they come to such a conclusion, regulators should consider rules and laws that reduce the risk by restricting the amount of information flowing through society.

Of course, 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.

When the costs significantly outweigh the benefits of information flow, regulators might ban completely the dissemination or storage of the particularly 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.[292] 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 am re-tilling this well-planted ground to show how to regulate data privacy after the fall of PII and 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 instead protect databases full of sensitive symptoms, diagnoses, and treatments? Focus in particular on one class of users of such information—medical researchers seeking new treatments and cures for disease. In this particular 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 will justifiably be reluctant to throttle information flow too much in this context when the toll of such choices might be measureable in human lives lost.

---

[292] *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) (arguing search engines should store queries for only six months).

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 try to build upon these human networks of trust in a revised HIPAA, allowing data transfer where trust is high and forbidding it when 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?[293] 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* controls against unauthorized access such as encryption and password protection, as well as *ex post* means for review such as audit trail mechanisms.

Regulators can vary these additional protections based on 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 all of these new, burdensome requirements might *stifle* research, they permit another change from the status quo that might greatly *expand* research instead: with the new HIPAA regulators should 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.[294] With more data and more specific data,

---

[293] According to Federal Rule, federally-funded research involving human subjects must be approved by an IRB. 45 C.F.R. § 46.101 et seq. (2009).

[294] It makes sense to continue to prohibit the transfer of some data such as names that might reveal identity without any outside information at all. Other data fields that might qualify for these reasons for continued suppression are home address and photograph.

researchers will be able to draw more accurate results, and thereby hopefully come to quicker and better conclusions.[295]

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, Dr. B cannot give the data to Dr. C, who must instead ask Dr. A for the data. Furthermore, releases to non-researchers 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 which 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[296] and online companies, notably Google, who argue that they should not.[297] European officials have split on the question,[298] with courts and regulators in Sweden[299] and Spain[300] deciding that IP addresses fall within the Directive and those in France,[301] Germany,[302] and the UK[303] finding they do not.

---

[295] 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.

[296] European Union Article 29 Data Protection Working Party, *Opinion 1/2008 on Data Protection Issues Relating to Search Engines*, 00737/EN WP 148 at 21 (April 4, 2008); Elec. Privacy Info. Center, Search Engine Privacy, http://epic.org/privacy/search_engine/ (last visited Aug. 3, 2009).

[297] *See infra* note 304.

[298] For a good summary, see Joseph Cutler, *Was That Your Computer Talking to Me? The EU and IP Addresses as "Personal Data",* PERKINS COIE DIGESTIBLE LAW BLOG, JUNE 24, 2008, http://www.perkinscoie.com/ediscovery/blogQ.aspx?entry=5147.

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

[300] Agencia de Protection de Datos, Statement on Search Engines, available at https://www.agpd.es/upload/Canal_Documentacion/Recomendaciones/declaracion_aepd_buscadores_en.pdf (Dec. 1, 2007) (opinion of Spanish Data Protection Agency that IP addresses held by search engines are personal data).

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

[302] Jeremy Mittman, *German Court Rules that IP Addresses are not Personal Data*, PROSKAUER PRIVACY LAW BLOG, Oct. 17, 2008,

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.[304] 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 it should not be regulated. By embracing this idea, Google has ignored 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.[305] Google's Global Privacy Officer Peter Fleischer offers hypothetical situations where 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."[306] 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 ignores further the lessons of easy reidentification by assuming it has no access to information that it can use to tie IP addresses to identity. On

---

http://privacylaw.proskauer.com/2008/10/articles/european-union/german-court-rules-that-ip-addresses-are-not-personal-data/.

[303] Information Commissioner's Office, Data Protection Good Practice: Collecting Personal Information Using Websites 3 (June 5, 2007) *available at* http://www.ico.gov.uk/upload/documents/library/data_protection/practical_application/collecting_personal_information_from_websites_v1.0.pdf.

[304] Peter Fleischer, *Can a Website Identify a User Based on IP Address?*, PETER FLEISCHER: PRIVACY . . . ? BLOG, Feb. 15, 2008, http://peterfleischer.blogspot.com/2008/02/can-website-identify-user-based-on-ip.html ("Privacy laws should be about protecting identifiable individuals and their information, not about undermining individualization."); Alma Whitten, *Are IP Addresses Personal?*, GOOGLE PUBLIC POLICY BLOG, Feb. 22, 2008, http://googlepublicpolicy.blogspot.com/2008/02/are-ip-addresses-personal.html (tying the discussion to the broad question, "as the world's information moves online, how should we protect our privacy?").

[305] *See* Fleischer, *supra* note 304 (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.").

[306] *Id.*

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."[307] 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, for example, from its log files 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.[308]

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.[309] Just because Google does not store these data associations on its own servers should hardly be the point. Otherwise, national ID numbers in the hands of private parties would not be "personal data" because only the government can authoritatively map these numbers to identities.[310]

Google can find entropy-reducing information which narrows IP addresses to identity in many other places: public databases reveal which ISP owns an IP address[311] and sometimes even narrow down an address to a geographic region,[312] 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.[313] In light of the richness of outside information relating to IP

---

[307] Whitten, *supra* note 304.

[308] European Union Article 29 Data Protection Working Party, *Opinion 1/2008 on Data Protection Issues Relating to Search Engines*, 00737/EN WP 148 at 21 (April 4, 2008) ("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.").

[309] *Id* at 8.

[310] Fleischer correctly points out that ISPs are often forbidden from disclosing the user associated with an IP address. Fleischer, *supra* note 304 ("[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 than all account numbers which are authoritatively tied to identity only by the entity that issued the number. All other entities must make educated guesses.

[311] ARIN WHOIS Database Search, http://ws.arin.net/whois/ (last visited Aug. 3, 2009) ("ARIN's WHOIS service provides a mechanism for finding contact and registrationinformation for resources registered with ARIN.").

[312] ERIC COLE & RONALD KRUTZ, NETWORK SECURITY BIBLE 316-18 (2005) (discussing reverse DNS queries).

[313] *E.g.* IP2Location.com home page, http://www.ip2location.com/ (last visited Aug. 3, 2009); Quova home page, http://www.quova.com/ (last visited Aug. 3, 2009).

addresses, and given the power of reidentification, Google's arguments amount to overstatements and legalistic evasions.

Google's arguments that it protects privacy further by deleting a single octet of information from IP addresses are 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.[314] 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.[315]

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, mental, economic, cultural or social identity."[316] 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 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 benefits of allowing *search engines* in particular to store and disclose information, including IP addresses, associated with *search queries* outweigh the potential harms to privacy.[317]

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 PII, 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

---

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

[315] *See supra* note 183

[316] EU Data Protection Directive, *supra* note 2, *a*rt. I(a).

[317] In the EU, the Article 29 Working Group privacy watchdog has proposed similarly special treatment for search engines. European Union Article 29 Data Protection Working Party, *supra* note 292, at 8.

behavior and preferences.[318] 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 if not greater for search query information than for health information. As the AOL breach revealed, stored search queries often contain user-reported health symptoms.[319] In fact, Google takes advantage of this to track and map influenza outbreaks in the U.S.[320] 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 even than 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 torrents of the mundane and unthreatening.[321] 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.[322] Regulators have under-appreciated the sensitive nature of this data, but events like the AOL data release have reawakened them to the special quality of stored search queries.[323]

Because the costs to privacy of allowing unfettered access to search queries are as high as in the health information context, EU and U.S. regulators should consider enacting search engine-specific laws to govern the storage and transfer of this information. Because the benefits of access to this data are less than for access to health information, regulators should be willing to restrict the storage and flow of search query information even more than HIPAA restricts the flow of 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 more than a prescribed amount of time.[324] They should set the specific amount of time this data can be stored only after considering search engine claims that they must keep data at least for a few months

---

[318] *Supra* note 186.

[319] Barbaro & Zeller, *supra* note 67 ("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.").

[320] Google.org, Flu Trends, http://www.google.org/flutrends/ (last visited Aug. 3, 2009).

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

[322] Paul Ohm, *The Rise and Fall of Invasive Internet Surveillance*, 2009 U. ILL. L. REV. ____ (forthcoming 2009).

[323] European Union Article 29 Data Protection Working Party, *supra* note 292, at 8 ("Search engines play a crucial role as a first point of contact to access information freely on the internet.").

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

to serve their vital business needs. They should significantly limit the third parties who can receive search query data.

CONCLUSION

Easy reidentification represents a sea change not only in technology but in our understandings of privacy. It undermines decades of assumptions about robust anonymization, assumptions that have charted the course for legions of business relationships, individual choices, and government regulations. Regulators must respond rapidly and forcefully to this disruptive technological shift, to restore an upended balance to the law and to save us all 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.

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.