B. How the Failure of Anonymization Disrupts Privacy Law

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

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

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

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

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

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

[204] 18 U.S.C. § 2725(3) (2006).

[205] *Id.*

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

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

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

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

[206] 20 U.S.C. § 1232g(a)(5)(A) (2006).

[207] 21 C.F.R. § 21.70(a)(3)(i) (2009).

[208] EU Data Protection Directive, *supra* note 3, art. 2(a).

[209] 18 U.S.C. § 2702(c) (2006).

C. The End of PII

1. Quitting the PII Whack-a-Mole Game

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[217] *Id.* at 184.

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

[219] *Id.*

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

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

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

2. Abandoning "Anonymize" and "Deidentify"

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

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

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

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

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

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

[225] *Id.*

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

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

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

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

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

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