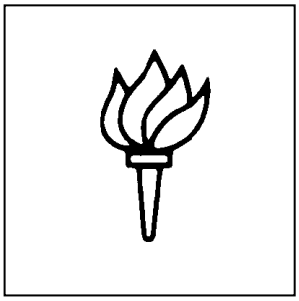
NEW YORK UNIVERSITY SCHOOL OF LAW

PUBLIC LAW & LEGAL THEORY RESEARCH PAPER SERIES
WORKING PAPER NO. 15-36



Anonymization and Risk

Ira S. Rubinstein and Woodrow Hartzog

September 2015

ANONYMIZATION AND RISK

Ira S. Rubinstein* & Woodrow Hartzog**

Abstract: Perfect anonymization of data sets that contain personal information has failed. But the process of protecting data subjects in shared information remains integral to privacy practice and policy. While the deidentification debate has been vigorous and productive, there is no clear direction for policy. As a result, the law has been slow to adapt a holistic approach to protecting data subjects when data sets are released to others. Currently, the law is focused on whether an individual can be identified within a given set. We argue that the best way to move data release policy past the alleged failures of anonymization is to focus on the process of minimizing risk of reidentification and sensitive attribute disclosure, not preventing harm. Process-based data release policy, which resembles the law of data security, will help us move past the limitations of focusing on whether data sets have been "anonymized." It draws upon different tactics to protect the privacy of data subjects, including accurate deidentification rhetoric, contracts prohibiting reidentification and sensitive attribute disclosure, data enclaves, and query-based strategies to match required protections with the level of risk. By focusing on process, data release policy can better balance privacy and utility where nearly all data exchanges carry some risk.

* Adjunct Professor of Law and Senior Fellow, Information Law Institute, New York University School of Law.

** Associate Professor, Samford University's Cumberland School of Law, Affiliate Scholar, The Center for Internet and Society at Stanford Law School. The authors wish to thank Derek Bambauer, Jane Bambauer, Daniel Barth-Jones, Steve Bellovin, Gregg Brown, Edward Felten, Simson Garfinkel, Robert Gellman, Sue Glueck, Seda Gürses, Michael Hintze, Cameron Kerry, Susan Landau, Orit Levin, Yves-Alexandre de Montjoye, Krish Muralidhar, Paul Ohm, Jules Polonetsky, Stuart Shapiro, Peter Swire, Omer Tene, Salil Vadhan and the participants of the Eighth Annual Privacy Law Scholars Conference and the New York University Privacy Research Group for their many valuable comments. The authors would also like to thank Lydia Wimberly, Megan Fitzpatrick, Aaron Alva, and Philip Cernera for their research assistance and Kristin Earle for helping us to navigate the genomics literature. We are grateful to Microsoft Corporation and the Future of Privacy Forum for supporting the research for this paper. However, the views expressed herein are solely our own.

INTRODUCTION

For years, it was widely believed that as long as data sets were "anonymized," they posed no risk to anyone's privacy. If data sets were anonymized, then they did not reveal the identity of individuals connected to the data. Unfortunately, the notion of perfect anonymization has been exposed as a myth. Over the past twenty years, researchers have shown that individuals can be identified in many different data sets once thought to have been "anonymized."[1] For example, in 2006, America Online (AOL) famously published a sample of its search queries. Although AOL replaced screen names with random numbers in the published search logs, this minimal step did not suffice to protect its users, and within days the *New York Times* discovered and revealed the identity of a 62-year-old AOL customer in the data set, Thelma Arnold.[2] Similar high-profile anonymization failures were

1. *See infra* Part I.

2. *See* Michael Barbaro & Tom Zeller, Jr., *A Face Is Exposed for AOL Searcher No. 4417749*, N.Y. TIMES (Aug. 9, 2006), http://www.nytimes.com/2006/08/09/technology/09aol.html?_r=0 [https://perma.cc/DHF9-8YEV]. For a full account of the AOL reidentification, see Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1717–18 (2010) (noting that AOL released twenty million search queries for 650,000

attributed to data sets released by Netflix[3] and by the New York Taxi and Limousine Commission.[4]

The possibility of correctly identifying people and attributes from anonymized data sets has sparked one of the most lively and important debates in privacy law. The credibility of anonymization, which anchors much of privacy law, is now open to attack. How should the law respond?

The failure of anonymization has identified a weakness in the focus of the law surrounding data releases. Some critics argue that it is impossible to eliminate privacy harms from publicly released data using anonymization techniques. They point out that other data sets containing related data will inevitably be released, allowing someone to link data in both sets and reidentify individuals in the first data set.[5] Defenders of anonymization counter that despite the theoretical and demonstrated ability to mount such attacks, the likelihood of reidentification for most data sets remains minimal and, as a practical matter, most data sets will remain anonymized using established techniques.[6]

These divergent views might lead us to different regulatory approaches. Those that focus on the remote possibility of reidentification might prefer an approach that reserves punishment only in the rare instance of harm, such as a negligence or strict liability regime revolving around harm triggers. Critics of anonymization might suggest we abandon deidentification-based approaches altogether, in favor of different privacy protections focused on collection, use, and disclosure that draw from the Fair Information Practice Principles, often called the FIPPs.[7]

---

users).

3. For the details of the Netflix incident, see *infra* text accompanying notes 36–39.

4. *See* Anthony Tockar, *Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset*, NEUSTAR (Sept. 15, 2014), http://research.neustar.biz/author/atockar [https://perma.cc/EJP5-5A3W] (describing the reidentification of a dataset consisting of "details about every taxi ride (yellow cabs) in New York in 2013, including the pickup and drop off times, locations, fare and tip amounts, as well as anonymized (hashed) versions of the taxi's license and medallion numbers").

5. *See infra* Section I.A.1.

6. *See, e.g.*, Jane Yakowitz, *Tragedy of the Data Commons*, 25 HARV. J.L. & TECH. 1, 2–3 (2011), http://jolt.law.harvard.edu/articles/pdf/v25/25HarvJLTech1.pdf [https://perma.cc/76ZM-LSYW]; ANN CAVOUKIAN & KHALED EL EMAM, INFO. & PRIVACY COMM'R OF ONT., DISPELLING THE MYTHS SURROUNDING DEIDENTIFICATION: ANONYMIZATION REMAINS A STRONG TOOL FOR PROTECTING PRIVACY (2011), http://www.ipc.on.ca/images/Resources/anonymization.pdf [https://perma.cc/43XQ-CGEH].

7. *See generally* ROBERT GELLMAN, FAIR INFORMATION PRACTICES: A BASIC HISTORY (2005), http://bobgellman.com/rg-docs/rg-FIPShistory.pdf [https://perma.cc/4VPE-FKAB].

There is a better focus for the data release law and policy: the process of minimizing risk. The main thesis of this Article is that the best way to move data release policy past the alleged failures of anonymization is to focus on the process of minimizing risk, not preventing harm. We argue that focusing on process and risk can bridge the concerns of formalists (for whom mathematical proof is the touchstone of any meaningful policy) and pragmatists (for whom workable solutions should prevail over theoretical concerns).[8] This change in focus reframes the debate away from the endpoint of perfect anonymity and toward the process of risk management.

In order to develop a clear, flexible, and workable legal framework for data releases, we propose drawing from the related, more established area of data security. Data security law is process-based, contextual, and tolerant of harm, so long as procedures to minimize risk are implemented ex ante. The law of data security focuses on requiring reasonable processes that decrease the likelihood of harm, even if threats are remote. Because there is no such thing as perfect data protection, data security policy is focused on regular risk assessment, the implementation of technical, physical, and procedural safeguards, and the appropriate response once a system or data set has been compromised.

Data security policy also largely refrains from overly specific rules, deferring instead to a reasonable adherence to industry standards. As the motivation for a consistent approach to releasing personal data increases, industry standards will inevitably develop in coordination with public policy and consumer protection goals. In short, the law of data release should look more like the law of data security.

The path for a process-based data release policy can be seen in nascent efforts by regulators. For example, according to the Federal Trade Commission (FTC):

> [D]ata is not "reasonably linkable" [and thus excluded from additional data protection frameworks] to the extent that a company: (1) takes reasonable measures to ensure that the data is de-identified; (2) publicly commits not to try to re-identify the data; and (3) contractually prohibits downstream recipients from trying to re-identify the data.[9]

---

8. *See infra* text accompanying notes 45–54.

9. FED. TRADE COMM'N, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE: RECOMMENDATIONS FOR BUSINESSES AND POLICYMAKERS iv, 20–21 (2012), http://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf

This multi-pronged approach is promising, but sound data release policy requires more nuance as well as attention to techniques other than deidentification (a term we use in this paper to refer to alteration or removal of identifying information to protect the identity of a data subject).[10] The full spectrum of possible data release protections should be utilized to tailor a company's obligations to the likely level of risk.

We advocate a system where perfect anonymization is not the enemy of sound data release policy.[11] However, we do not fully embrace the pragmatism advocated by defenders of anonymization. We first take issue with the current framing of the anonymization debate. The terms "anonymous" and "anonymization" should be largely abandoned in our data release policy and discourse. Almost all uses of the terms to describe the safety of data sets are misleading, and often they are deceptive.[12] Focusing on the language of process and risk will better set expectations.

Additionally, anonymization critics have rightfully pointed out that it is a mistake to rely too much upon risk assessments that are not scalable and are not able to account for either new data inputs or increasingly sophisticated analytical techniques.[13] An effective risk-based approach to releasing data—combined with a transition away from existing privacy laws that treat personally identifiable data (PII) as their subject matter while leaving non-PII unregulated—should accommodate risk models and support important baseline protections for consumers.

In this Article, we aim to use the lessons learned from the criticism and defense of anonymization to propose a policy-driven and comprehensive process-based framework for minimizing the risk of reidentification and sensitive attribute disclosure. We identify the

---

[https://perma.cc/R32U-M64B].

10. *See* Khaled El Emam & Bradley Malin, *Appendix B: Concepts and Methods for De-identifying Clinical Trial Data*, *in* SHARING CLINICAL TRIAL DATA: MAXIMIZING BENEFITS, MINIMIZING RISK 203, 214 (Inst. of Med. ed., 2015) [hereinafter IOM STUDY] (distinguishing identity versus attribute disclosure); Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 U. COLO. L. REV. 1117, 1152 (2013) (same); SIMSON L. GARFINKEL, NAT'L INST. OF STANDARDS & TECH., DE-IDENTIFICATION OF PERSONAL INFORMATION iii, 1–2 (2015), http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf [https://perma.cc/4RFL-X6AS] [hereinafter NIST REPORT]. Like Wu and El Emam & Malin, we focus on sensitive attribute disclosure.

11. "Data release policy" typically refers to the release of data and related resources to the scientific community for research purposes. We use the term more broadly to refer to the voluntary or mandatory release of data to scientists, business partners, or the general public for any legitimate reason.

12. *See infra* Section III.B.

13. *See infra* Section I.A.3.

relevant risk factors and techniques that can mitigate that risk. By shifting from output to process, we can move past the anonymization stalemate between the formalists and the pragmatists driving this debate.

This approach recognizes that there is no perfect anonymity. It focuses on process rather than output. Yet effective risk-based data release policy also avoids a ruthless pragmatism by acknowledging the limits of current risk projection models and building in important protections for individual privacy. This policy-driven, integrated, and comprehensive approach will help us to better protect data while preserving its utility.

Our argument proceeds in three parts. In Part I, we review the anonymization debate and its stagnant policy. We argue that anonymization policy should be re-conceptualized as a data release policy. In Part II, we propose that data release policy should be focused on the process of minimizing risk. Drawing from data security law, we develop process-based data release policy as a holistic, contextual, and risk-tolerant approach. Finally, in Part III, we build upon the FTC's process-based approach to protecting data subjects to identify how the full spectrum of techniques from the field of statistical disclosure limitations can be used to tailor data release obligations to risk. We identify specific risk vectors such as data volume, data sensitivity, type of data recipient, data use, data treatment technique, data access controls, and consent and consumer expectations.[14] We propose several legal reforms to implement process-based data release policy, including a general requirement for "reasonable" data release protections and a prohibition on deceptive deidentification.

The revelation that purportedly anonymized data sets were vulnerable to attack was a wake-up call for companies, regulators, and the public. Yet despite years of scholarly attention, policy has yet to respond fully. By focusing on the steps required to mitigate risk rather than the actual elimination of harm, data sets can be better shared while still protecting data subjects.

I. THE ANONYMIZATION DEBATE IS STAGNANT AND IS NOT ADVANCING POLICY

Anonymization was not always a contentious concept. For years, scholars, professionals, and policymakers were content with the notion

14. *See infra* text accompanying notes 162–174.

that anonymized data sets were safe.[15] But around fifteen years ago, anonymization began to seem fallible. High-profile cases of reidentification attracted media attention and became lightning rods for critics and defenders of deidentification as a technique to protect personal information.[16] The alleged failure of anonymization seemingly threw deidentification policy discussions into chaos. Fifteen years in, the debate has led to polarization, and policy discussions are now splintered. While policymakers like the FTC and the Article 29 Working Group have taken note of deidentification's limits,[17] they have largely ignored developments in adjacent fields such as differential privacy. They also lack an integrated vision of the full spectrum of techniques for safely releasing data sets. Meanwhile, privacy law remains largely unchanged.

Why has the anonymization debate had such little impact on privacy law? Part of the reason might be that the debate too often fixates on high-profile cases in which a researcher develops and applies a method for reidentifying individuals in a deidentified data set or demonstrates the feasibility of an attack by publishing a proof-of-concept. The news media turns these research results into anecdotes proving the failure (if not the death) of anonymity.[18] A major problem with this narrative is that it overemphasizes one method (deidentification) at the expense of other methods in the full spectrum of data release techniques.

Because of their outsized role in policy discussions, the high-profile cases are key to understanding the shortcomings of the current policy debate. Thus, this Part revisits a few of the original attacks and proof-of-concept papers with a critical eye to understanding how and why

---

15. Ohm, *supra* note 2, at 1710–11.

16. *See infra* text accompanying notes 18–25.

17. For the FTC, see *supra* note 9; for the Article 29 Working Group, see *infra* note 189.

18. For objections to the "death of anonymization" narrative, see, for example, Jane Yakowitz Bambauer, *Is De-Identification Dead Again?*, INFO/L. BLOG (Apr. 28, 2015), https://blogs.law.harvard.edu/infolaw/2015/04/28/is-de-identification-dead-again/ [https://perma.cc/CQ47-B53U]; Jane Yakowitz Bambauer, *Reporting Fail: The Reidentification of Personal Genome Project Participants*, INFO/L. BLOG (May 1, 2013), https://blogs.law.harvard.edu/infolaw/2013/05/01/reporting-fail-the-reidentification-of-personal-genome-project-participants/ [https://perma.cc/JJ9N-UZZS]; Daniel Barth-Jones, *The Antidote for "Anecdata": A Little Science Can Separate Data Privacy Facts from Folklore*, INFO/L. BLOG (Nov. 21, 2014), https://blogs.law.harvard.edu/infolaw/2014/11/21/the-antidote-for-anecdata-a-little-science-can-separate-data-privacy-facts-from-folklore/ [https://perma.cc/D5EC-5LGV]; Daniel C. Barth-Jones, *Press and Reporting Considerations for Recent Re-Identification Demonstration Attacks: Part 2 (Re-Identification Symposium)*, BILL HEALTH HARV. L. BLOG (Oct. 1, 2013), http://blogs.law.harvard.edu/billofhealth/2013/10/01/press-and-reporting-considerations-for-recent-re-identification-demonstration-attacks-part-2-re-identification-symposium/ [https://perma.cc/QZN9-P9SF].

deidentification failed, what this implies for data release policy, and the strengths and weaknesses of alternative methods.

A. Survey of Data Release Problems and Solutions

This Section begins by explaining in layman's terms how deidentification works and why deidentified data sets are vulnerable to reidentification attacks as exemplified by two well-known cases. We also examine the impasse between the two leading camps in the deidentification debate—we dub them "pragmatists" and "formalists"—and their sharp disagreement over the risks of reidentification. Next, we situate the deidentification debate within the spectrum of data release techniques, which includes not only deidentification but also access controls and query-based methods such as differential privacy. Finally, we consider whether "open data" is a precondition of scientific progress, developing a case study around recent developments in genomic data sharing policy.

1. Deidentification and Reidentification

The term *deidentification*[19] has been defined several different ways. In this paper, we adopt the usage in a recent National Institute of Standards and Technology (NIST) Report, which defines deidentification as "a tool that organizations can use to remove personal information from data that they collect, use, archive, and share with other organizations."[20] As we describe below, we consider the term deidentification distinct from the concept of "anonymity" or "anonymization," which we argue implicitly guarantees protection of identity. Others use deidentification and anonymization interchangeably; we do not.

The most basic step in deidentification is to remove *direct identifiers* (i.e., those data that directly identify a unique individual, such as name or social security number) or replace them with pseudonyms or random values. This step is often unwisely passed off as anonymizing data.[21] Unfortunately, it often proves inadequate against *reidentification*, which

---

19. Terms in italics are defined in Appendix: A Glossary of Terms.

20. NIST REPORT, *supra* note 10, at 1; Wu, *supra* note 10, at 1152 (distinguishing identity versus attribute disclosure); *see also* IOM STUDY, *supra* note 10, at 214 (same).

21. *See* Daniel C. Barth-Jones, *Public Policy Considerations for Recent Re-Identification Demonstration Attacks on Genomic Data Sets: Part 1 (Re-Identification Symposium)*, BILL HEALTH HARV. L. BLOG (May 29, 2013), http://blogs.law.harvard.edu/billofhealth/2013/05/29/public-policy-considerations-for-recent-re-identification-demonstration-attacks-on-genomic-data-sets-part-1-re-identification-symposium/ [https://perma.cc/Y85F-DVPD].

is the process of attempting to determine the identities of the data subjects whose identifiers have been removed from the data set.[22] For example, in a *linkage attack*, an *adversary* (any individual or entity trying to reidentify a data subject) takes advantage of *auxiliary* (or *background* or *outside*) *information* to link an individual to a record in the deidentified data set.[23]

A well-known example in the literature concerns the hospitalization records of Governor Weld of Massachusetts.[24] A state insurance agency was obligated to release certain hospitalization records to the public for research purposes after first removing direct identifiers while leaving demographic data (birthday, ZIP code, gender) and sensitive health data. Latanya Sweeney obtained the deidentified hospital records, matched them with publicly available voter registration records (which contained similar demographic data), and reidentified Governor Weld by isolating his record in the voter rolls and matching it with his deidentified hospital record.[25]

Linkage attacks, however, are much more complicated than they sound. The scenario above assumes that the targeted data subject is represented in both data sets (the hospital records and the voter rolls), that the matching variables are recorded identically in both, and that the linked data elements uniquely distinguish an individual. Sweeney's successful linkage attack met all of these conditions, but the rate of success in reidentification attacks is very low, for reasons discussed in the next Section.

2. *Quasi-Identifiers and the Auxiliary Information Problem*

The usual way to hinder linkage attacks is to alter common attributes (like birthday, ZIP code, and gender) and other *quasi-identifiers*. A quasi-identifier does not itself "identify a specific individual but can be aggregated and 'linked' with other information to identify data subjects."[26] Indeed, one of the most complicated parts of protecting

22. NIST REPORT, *supra* note 10, at 9.

23. *Id*. at 17–18. Voter registration records are a good example of auxiliary information. Other sources include any public record (whether maintained by a government agency or a commercial data broker), newspapers, social media, or data deliberately shared on social networking sites.

24. *See* Latanya Sweeney, k-*Anonymity: A Model for Protecting Privacy*, 10 INT'L J. ON UNCERTAINTY, FUZZINESS & KNOWLEDGE-BASED SYSTEMS 557, 558–59 (2002). For a full account of the Weld reidentification, see Ohm, *supra* note 2, at 1719–20.

25. Sweeney, *supra* note 24, at 558–59.

26. NIST REPORT, *supra* note 10, at 19.

against linkage attacks is distinguishing identifiers from potentially identifying links to a person.

The challenge in altering quasi-identifiers is that they convey useful information that might be important for later analysis. Thus, rather than remove the quasi-identifiers (which may severely limit the utility of the released data set), data custodians rely on generalization (e.g., changing date of birth to month or year of birth), suppression (e.g., removing a value in a record that makes it an outlier, such as a diagnosis of a very rare disease), and more sophisticated techniques including rounding, randomization (adding noise to the data), sampling, and swapping.[27]

A popular method for altering quasi-identifiers using generalization and suppression is Sweeney's concept of *k-anonymity*[28] which "requires the data administrator to ensure that, given what the adversary already knows, the adversary can never narrow down the set of potential target records to fewer than *k* records in the released data."[29] A weakness in this approach is that *k*-anonymity assumes that only a small number of attributes may be used as quasi-identifiers for purposes of a linkage attack. Several researchers have taken issue with this claim.

For example, Cynthia Dwork has demonstrated that some formal definitions of privacy are impossible, in part because there is simply too much auxiliary information attackers can draw from.[30] It is virtually always possible to learn *something* about individuals from deidentified data sets. In a later paper, Dwork describes the auxiliary information problem as follows: "in any 'reasonable' setting there is a piece of information that is in itself innocent, yet in conjunction with even a modified (noisy) version of the data yields a privacy breach."[31]

Similarly, Charu Aggarwal has argued that it is a mistake to assume there are a limited number of quasi-identifiers in high dimensional or "sparse" data sets.[32] In such contexts almost any variable may function

---

27. *Id*. at 20. For an eleven-step, risk-based process for deidentifying data using these techniques, see IOM STUDY, *supra* note 10, at 240–43.

28. *See* Sweeney, *supra* note 24, at 572.

29. Wu, *supra* note 10, at 1142.

30. Cynthia Dwork, *Differential Privacy*, *in* 33RD INTERNATIONAL COLLOQUIUM ON AUTOMATA, LANGUAGES AND PROGRAMMING PART II, at 1, 2 (2006), research.microsoft.com/pubs/64346/dwork.pdf [https://perma.cc/TCB7-PKAX].

31. Cynthia Dwork & Moni Naor, *On the Difficulties of Disclosure Prevention in Statistical Databases or the Case for Differential Privacy*, 2 J. PRIVACY & CONFIDENTIALITY 93, 93 (2010).

32. *See* Charu C. Aggarwal, *On* k-*Anonymity and the Curse of Dimensionality*, *in* PROCEEDINGS OF THE 31ST INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES 901, 909 (2005), http://www.vldb2005.org/program/paper/fri/p901-aggarwal.pdf [https://perma.cc/QZ9E-HQDV]. A "sparse" data set is one in which each individual record contains values only for a small fraction of

as a quasi-identifier.[33] Arvind Narayanan and Vitaly Shmatikov have made a similar point.[34] In a later paper they concluded "*any attribute can be identifying in combination with others*."[35] This potentially devastating objection to deidentification is known as the auxiliary information problem.

In this age of big data, the privacy risks of large data sets are especially relevant. Narayanan and Shmatikov demonstrated this by showing how a small amount of auxiliary information could be used to reidentify individuals in the Netflix Prize data set. Netflix offered a prize for improvements to its recommendation algorithm and provided contestants with access to a data set consisting of "more than 100 million ratings from over 480 thousand randomly-chosen, anonymous customers on nearly 18 thousand movie titles."[36] It "anonymized" the data set by removing all PII from customer records and replacing all identifiers with randomly assigned IDs, leaving only movie ratings and the date of rating for each customer.

Did Narayanan and Shmatikov succeed in re-identifying all 480,000 Netflix customer names in the released data set? No, but this was never their goal.[37] Rather, they obtained the records of about fifty users of the publicly available Internet Movie Database (IMDb) and linked this data to two users in the Netflix database.[38] Still, their results may be viewed as a proof-of-concept for how to reidentify *all* records in the Netflix Prize data set by linking them with IMDb or other publicly available data.[39]

Yves-Alexandre de Montjoye and his colleagues have extended this work by publishing important studies of deidentified mobile phone and

---

attributes. For example, Amazon's shopping database is sparse because while Amazon sells millions of items, the shopping history of any single customer contains only a tiny fraction of them. Sparse data sets include not only recommendation systems but also any real-world data sets of individual transactions or preferences. *See* Arvind Narayanan & Vitaly Shmatikov, *Robust De-Anonymization of Large Sparse Datasets*, 2008 29TH IEEE SYMP. ON SECURITY & PRIVACY 111.

33. Aggarwal, *supra* note 32, at 909.

34. *See* Narayanan & Shmatikov, *supra* note 32.

35. Arvind Narayanan & Vitaly Shmatikov, *Myths and Fallacies of "Personally Identifiable Information*,*"* 53 COMM. ACM 24, 26 (2010) (emphasis in original).

36. *The Netflix Prize Rules*, NETFLIX (2006), http://www.netflixprize.com/assets/rules.pdf [https://perma.cc/8XUU-G4GK].

37. Narayanan & Shmatikov, *supra* note 32, at 122.

38. *Id*.

39. Their paper describes a robust "de-anonymization" algorithm that succeeded in identifying ninety-nine percent of the records in the Netflix data set from "8 movie ratings (of which 2 may be completely wrong) and dates that may have a 14-day error." *Id*. at 121.

credit card metadata. De Montjoye introduced the concept of "unicity" to quantify "how much outside information one would need, on average, to reidentify a specific and known user in a simply anonymized data set."[40] Not surprisingly, the higher a data set's unicity, the easier it is to reidentify data subjects in the anonymized data. Mobile phone metadata is highly unique and therefore can be reidentified using little outside information.[41] The same is roughly true of credit card data.[42] Although de Montjoye recognizes that further work is needed, he and his colleagues consider it likely "that most large-scale metadata sets—for example, browsing history, financial records, and transportation and mobility data—will have a high unicity."[43] Social network data should also be added to this list.[44]

3. The Debate Between Formalists and Pragmatists

Does the auxiliary information problem sound the death knell of deidentification, or does it remain a viable strategy for protecting the privacy of data subjects? More than a dozen interchanges among the experts show that they are deeply divided, not only in how they view the implications of the auxiliary information problem, but in their goals, methods, interests, and measures of success.[45]

---

40. Yves-Alexandre de Montjoye et al., *Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata*, 347 SCI. 536, 537 (2015). A "simply anonymized data set" is one from which obvious identifiers have been removed—names, home, address, phone numbers, and other forms of PII. *Id*.

41. *See* Yves-Alexandre de Montjoye et al., *Unique in the Crowd: The Privacy Bounds of Human Mobility*, 3 SCI. REP. 1, 2 (2013) (showing that only four spatiotemporal points are enough to uniquely reidentify ninety-five percent of mobile phone users). This is intuitively obvious: A's mobile phone data consists of the set of A's locations at specific times as recorded by a mobile operator whenever A initiates or receives a call or a text message, or otherwise connect to a cell tower. And there are very few people besides A who are in the same place at the same time on multiple occasions.

42. *See* de Montjoye et al., *supra* note 40, at 537 (showing that only four spatiotemporal points are enough to uniquely reidentify ninety percent of shoppers using credit cards).

43. *Id*. at 539.

44. *See, e.g*., Arvind Narayanan & Vitaly Shmatikov, *De-Anonymizing Social Networks*, *in* 2009 30TH IEEE SYMPOSIUM ON SECURITY & PRIVACY 173 (demonstrating effectiveness of new reidentification algorithm targeting anonymized social network graphs by showing that a third of verified users with accounts on both Twitter and Flickr can be reidentified in the anonymous Twitter graph with only a twelve percent error rate).

45. *See, e.g*., Daniel Barth-Jones et al., Letter to the Editor, *Assessing Data Intrusion Threats*, 348 SCI. 194 (2015); Yves-Alexandre de Montjoye & Alex "Sandy" Pentland, Letter to the Editor, *Response*, 348 SCI. 195 (2015); ANN CAVOUKIAN & DAN CASTRO, INFO. & PRIVACY COMM'N OF ONT., BIG DATA AND INNOVATION, SETTING THE RECORD STRAIGHT: DEIDENTIFICATION DOES WORK (2014), http://www2.itif.org/2014-big-data-deidentification.pdf [https://perma.cc/UK2F-

The computer scientists, epidemiologists, and statisticians whom we refer to as pragmatists—including El Emam and Barth-Jones—share an expertise in deidentification methods and value practical solutions for sharing useful data to advance the public good. Accordingly, they devote a great deal of effort to devising methods for measuring and managing the risk of reidentification for clinical trials and other specific disclosure scenarios.[46] Unlike those who invent linkage attacks, pragmatists consider it difficult to gain access to auxiliary information and give little weight to attacks demonstrating that data subjects are distinguishable and unique but that fail to reidentify anyone.[47] Rather, they argue that empirical studies and meta-analyses show that the risk of reidentification in properly deidentified data sets is, in fact, very low.[48]

PK5V]; CAVOUKIAN & EL EMAM, *supra* note 6; ARVIND NARAYANAN & EDWARD W. FELTEN, NO SILVER BULLET: DE-IDENTIFICATION STILL DOESN'T WORK (2014), http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf [https://perma.cc/N365-448N]; Khaled El Emam & Luke Arbuckle, *De-Identification: A Critical Debate*, FUTURE PRIVACY F. (July 24, 2014), https://fpf.org/2014/07/24/de-identification-a-critical-debate/ [https://perma.cc/L873-KCVQ]; Barth-Jones, *supra* note 21; Daniel Barth-Jones, *Re-Identification Risks and Myths, Superusers and Super Stories (Part I: Risks and Myths)*, CONCURRING OPINIONS (Sept. 6, 2012), http://concurringopinions.com/archives/2012/09/re-identification-risks-and-myths-superusers-and-super-stories-part-i-risks-and-myths.html [https://perma.cc/3ZK5-5PX7]; Ed Felten, Reader Comment, *Re-Identification Risks and Myths, Superusers and Super Stories (Part I: Risks and Myths)*, CONCURRING OPINIONS (Sept. 6, 2012, 8:20 PM and Sept. 7, 2012, 8:57 PM), http://concurringopinions.com/archives/2012/09/re-identification-risks-and-myths-superusers-and-super-stories-part-i-risks-and-myths.html [https://perma.cc/3ZK5-5PX7]; Arvind Narayanan, *Reidentification as Basic Science (Re-Identification Symposium)*, BILL HEALTH HARV. L. BLOG (May 26, 2013), http://blogs.law.harvard.edu/billofhealth/2013/05/26/reidentification-as-basic-science/ [https://perma.cc/T6JJ-3BCC].

46. *See* IOM STUDY, *supra* note 10, at 233–34 (describing the level of acceptable risks in terms of factors such as the available deidentification techniques; the extent to which a disclosure would invade the privacy to data subjects—which in turn depends on the sensitivity of the data, the potential injury from an inappropriate disclosure, and the nature and scope of any consent that participants may have provided—and the motives and capacity of likely adversaries).

47. *See, e.g.*, Barth-Jones, *supra* note 21.

48. *See, e.g.*, Kathleen Benitez & Bradley Malin, *Evaluating Re-Identification Risks with Respect to the HIPAA Privacy Rule*, 17 J. AM. MED. INFORMATICS ASS'N 169, 169 (2010) (estimating that the percentage of a state's population vulnerable to unique reidentification using a voter registration list to perform a linkage attack ranged from 0.01% to 0.25%); Deborah Lafkey, *The Safe Harbor Method of De-Identification: An Empirical Test* (Oct. 8, 2009), www.ehcca.com/presentations/HIPAAWest4/lafky_2.pdf [https://perma.cc/5Y25-GPZE] (statistical experts analyzing approximately 15,000 deidentified patient records found only 216 unique profiles in the deidentified data set, and only 28 potential matches—using age, gender, and ZIP as quasi-identifiers—and were able to accurately reidentify only two data subjects, giving a verified match rate of 0.013%); Khaled El Emam et al., *A Systematic Review of Re-Identification Attacks on Health Data*, 6 PLOS ONE 1, 8–9 (2011) (meta-analysis of fourteen reidentification attempts found relatively high rate of reidentification (twenty-six percent across all studies and thirty-four percent for attacks on health data) but successful reidentification events typically involved small data sets that had not been

Formalists, on the other hand, are all computer scientists like Dwork, Narayanan (and his colleague Edward Felten), Shmatikov, and de Montjoye.[49] They insist on mathematical rigor in defining privacy, modeling adversaries, and quantifying the probability of reidentification. Dwork, in particular, seeks provable privacy guarantees using methods first developed in cryptography.[50] Formalists argue that efforts to quantify the efficacy of deidentification "are unscientific and promote a false sense of security by assuming unrealistic, artificially constrained models of what an adversary might do."[51] Similarly, they take seriously proof-of-concept demonstrations while minimizing the importance of empirical studies showing low rates of reidentification.

Their disagreements arise because pragmatists and formalists represent distinctive disciplines with very different histories, questions, methods, and objectives. Their disagreements play out in what Seda Gürses calls "a terrain of contestations."[52] Even though there are important differences between them, both approaches offer valuable insights in formulating data release policies. From a policy standpoint, it is misguided to fixate on which approach is correct, and far more productive to figure out where they come together.

Granted, the pragmatists see value in their practical approach, although the problem of auxiliary information cautions against over-confidence in how they think about risk assessment. At the same time, some leading pragmatists concede that a highly formal approach like differential privacy "has a number of important strengths, but also faces a number of empirical and practical barriers to its deployment in healthcare settings."[53] On the other hand, formalists see value in their

---

deidentified according to existing standards).

49. We omit Latanya Sweeney because she has a foot in both camps.

50. Differential privacy is the paradigmatic example of formalism. It seeks to place privacy on a *mathematically rigorous* foundation, thereby enabling computer scientists "to argue formally about the degree of risk in a sequence of queries." Cynthia Dwork & Rebecca Pottenger, *Towards Practicing Privacy*, 20 J. AM. MED. INFORMATICS ASS'N 102, 102 (2013), http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3555331/pdf/amiajnl-2012-001047.pdf [https://perma.cc/Z2TL-5CRY]. In this paper, Dwork and Pottenger dismiss deidentification as a "sanitization pipe dream." *Id*. On the other hand, they concede that setting a "privacy budget" based on the "different types of data, or even different types of queries against data, may make sense, but these are policy questions that the math does not attempt to address." *Id*. at 106.

51. *See* NARAYANAN & FELTEN, *supra* note 45; de Montjoye & Pentland, *supra* note 45.

52. *See* Seda Gürses, "Privacy Is Don't Ask, Confidentiality Is Don't Tell": An Empirical Study of Privacy Definitions, Assumptions and Methods in Computer Science Research (2013) (unpublished manuscript) (on file with authors).

53. Bradley A. Malin et al., *Biomedical Data Privacy*: *Problems, Perspectives, and Recent Advances*, 20 J. AM. MED. INFORMATICS ASS'N 1, 5 (2013); *see also* Fida K. Dankar & Khaled El

more rigorous approach, notwithstanding practical implementation challenges.[54] At the same time, even Dwork concedes that the literature on statistics "contains a wealth of privacy supportive techniques and investigations of their impact on the statistics of the data set" while insisting that "[r]igorous definitions of privacy and modeling of the adversary are not prominent features of this portion of the literature."[55]

Is there a way forward that recognizes the limits of deidentification without abandoning it while embracing the full spectrum of techniques to protect the identity and attributes of data subjects? We believe the first step is recognizing that deidentification techniques are only part of a larger approach to protecting the privacy and confidentiality of data subjects known as *statistical disclosure limitation* (SDL).[56] We argue below that SDL provides the broader context in which to understand and evaluate a range of protective techniques. Our brief examination of SDL sets the stage for overcoming the divide between pragmatists and formalists and reformulating the policy debate along more productive lines.

4. Statistical Disclosure Limitation

SDL comprises the principles and techniques that researchers have developed for disseminating official statistics and other data for research purposes while protecting the privacy and confidentiality of data subjects. Satkartar Kinney describes SDL in terms of three major forms of interaction between researchers (whom she refers to as users) and personal data: direct access, dissemination-based access, and query-based access.[57]

Direct access encompasses both licensed data, which allows users who click-through the applicable licensing terms to perform any data query and receive full results, and authorized access to research data

---

Emam, *Practicing Differential Privacy in Health Care: A Review*, 5 TRANSACTIONS ON DATA PRIVACY 35 (2013).

54. Making differential privacy more practical is an ongoing area of research. *See, e.g.*, *Putting Differential Privacy to Work*, U. PA. DEP'T COMPUTER & INFO. SCI., http://privacy.cis.upenn.edu/ [https://perma.cc/F5TK-KC79] (last visited Apr. 11, 2016).

55. Dwork & Naor, *supra* note 31, at 94.

56. *See generally* Satkartar K. Kinney et al., *Data Confidentiality: The Next Five Years Summary and Guide to Papers*, 1 J. PRIVACY & CONFIDENTIALITY 125 (2009) (describing SDL methods). This field of research is also more intuitively known as statistical disclosure control. *See, e.g.*, ANCO HUNDEPOOL ET AL., STATISTICAL DISCLOSURE CONTROL (1st ed. 2012).

57. Kinney et al., *supra* note 56, at 127 fig.1.

centers, which also allows any query but only returns vetted results.[58] Direct access imposes the fewest restrictions on data but limits data access to qualified investigators who must agree to licensing terms or execute a Data Use Agreement (DUA), which may also stipulate security measures and prohibit redistribution of the data sets or attempts to reidentify or contact data subjects.[59] Alternatively, an agency (such as the Census Bureau) may host the data at a research center and provide access to data sets under agreement at secure enclaves,[60] or license users to access data remotely via secure internet portals.[61] In any case, direct access methods avoid many of the reidentification issues discussed above by never releasing data sets to the general public, thereby thwarting linkage attacks.

Dissemination-based access refers to the practice of publicly releasing reduced, altered, or synthetic data (i.e., hypothetical data that have similar characteristics to the real data). Like direct access, it allows full results to any query.[62] The data custodian applies various techniques to construct the transformed data set before publicly releasing it (although users may have to register or consent to terms of use that contain few if any of the restrictions in DUAs). In short, this form of access combines public release of data with masking of data sets by methods including generalization and suppression. Deidentification falls into the SDL sub-category of dissemination-based access.

Query-based access allows users to interact with the data by posing queries, typically over a secure internet connection.[63] Kinney identifies several sub-categories of query-based access, including remote analysis servers and differential privacy. Remote analysis servers allow researchers to analyze confidential data without ever seeing the underlying data, although both the queries they can pose and the results they can obtain may be subject to limitations. Another sub-category of query-based access, differential privacy, is a set of techniques developed by Dwork.[64] In this framework, the query results (analyses) are altered, often by adding noise, so that released information does not reveal any

58. Vetted results typically involve "forbidding users access to confidentiality-threatening items." *Id*.

59. *Id*. at 128.

60. *Id*.

61. *Id*.

62. *Id*. at 128–29.

63. *Id*. at 129.

64. *See* Dwork, *supra* note 30, at 3.

person's data with certainty. According to Dwork, differential privacy "addresses all concerns that any participant might have about the leakage of his or her personal information, regardless of any auxiliary information known to an adversary: [e]ven if the participant removed her data from the dataset, no outputs . . . would become significantly more or less likely."[65] The key point about query-based access is that users rely on techniques that allow useful statistical queries without the need for having any direct access to the underlying data sets. This too avoids most of the reidentification issues discussed above.[66]

Kinney's analysis helps clarify several contested issues in the current debate over deidentification. First, as Kinney points out, the most urgent need is for research that "provides agencies methods and tools for making sound decisions about SDL."[67] Second, her taxonomy calls attention to the fact that researchers in statistics and computer science pursue very different approaches to confidentiality and privacy and often in isolation from one another. They might achieve better results by collaborating across methodological divides.[68] Third, the legal scholars who have written most forcefully on this topic tend to evaluate the pros and cons of deidentification in isolation from other SDL methods.[69] Debates that focus exclusively on the merits of deidentification are only part of the story.[70]

5. Open Data

Much of the deidentification debate overlaps with discussions about *open data*, which refers to "information that is accessible to everyone,

---

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

66. Not all query-based methods are immune from attack. *See, e.g.*, Amatai Ziv, *Israel's 'Anonymous' Statistics Surveys Aren't So Anonymous*, HAARETZ (Jan. 7, 2013), http://www.haaretz.com/news/israel/israel-s-anonymous-statistics-surveys-aren-t-so-anonymous-1.492256 [https://perma.cc/TR4G-E6SA] (describing an attack based on sophisticated queries from which the attacker can infer census responses and match them with auxiliary information).

67. Kinney et al., *supra* note 56, at 131.

68. *Id*. at 132.

69. *See infra* Section I.B.

70. As Salil Vadhan and his colleagues in Harvard University's Privacy Tools for Sharing Research Data project point out, techniques such as "privacy-aware methods for contingency tables, synthetic data, data visualizations, interactive mechanisms, and multiparty computations[] have been successfully used to share data while protecting privacy, with no major compromises as far as we know." Salil Vadhan et al., Comment Letter on Advance Notice of Proposed Rulemaking on Human Subjects Research Protections (Oct. 26, 2011), http://privacytools.seas.harvard.edu/files/privacytools/files/commonruleanprm.pdf [https://perma.cc/2AJT-NAC4].

machine readable, offered online at zero cost, and has no limits on reuse and redistribution."[71] Adherents of an open data philosophy typically promote greater access to government data in order to advance the public good.[72] A key argument in favor of open data within the scientific community is that it "promote[s] transparency, reproducibility, and more rapid advancement of new knowledge and discovery."[73] Scientific journals and funding agencies may also require that experimental data be made publicly available; however, additional requirements apply to data sets that derive from clinical studies to ensure that researchers have taken all steps necessary to protect the privacy of data subjects.[74] Nor is it clear that the only way to make data available and shareable for the purposes of advancing scientific research is by adopting open data principles.

Genetic research provides a powerful example of the advantages of controlled access. More generally, the following brief case study of genomic data sharing illustrates the role of legal and institutional arrangements in limiting the flow and use of personal data consistent with the privacy expectations of data subjects.

The proliferation of genomic information for research, clinical care, and personal interest has opened up new reidentification attack vectors on DNA and other genetic data sets,[75] forcing the scientific community to reconsider the privacy assurances they can offer participants in DNA studies.[76] Two of the many recent genetic privacy breaches are highly relevant. In the first case, a group of geneticists discovered a statistical

---

71. Emmie Tran & Ginny Scholtes, *Open Data Literature Review*, *in* 19TH ANNUAL BCLT/BTLJ SYMPOSIUM: OPEN DATA: ADDRESSING PRIVACY, SECURITY, AND CIVIL RIGHTS CHALLENGES 1 (2015), https://www.law.berkeley.edu/wp-content/uploads/2015/04/Final_OpenDataLitReview_2015-04-14_1.1.pdf [https://perma.cc/UL5X-P5SS]; *see also* BUDAPEST OPEN ACCESS INITIATIVE, http://www.budapestopenaccessinitiative.org/ (last visited May 10, 2015).

72. *See* Robert M. Goerge, *Data for the Public Good: Challenges and Barriers in the Context of Cities*, *in* PRIVACY, BIG DATA, AND THE PUBLIC GOOD: FRAMEWORKS FOR ENGAGEMENT 153 (Julia Lane et al. eds., 2014) (discussing various efforts to use data analysis to improve public safety, education, urban transportation, public housing, and so on).

73. *See* IOM STUDY, *supra* note 10, at 141.

74. *See, e.g.*, Theo Bloom, *Data Access for the Open Access Literature: PLOS's Data Policy*, PLOS (Dec. 12, 2013), https://www.plos.org/data-access-for-the-open-access-literature-ploss-data-policy/ [https://perma.cc/DD89-4U7E]; IOM STUDY, *supra* note 10, at 141 (recommending a restricted access model for holders of clinical data as opposed to open access).

75. For an excellent survey, see generally Yaniv Erlich & Arvind Narayanan, *Routes for Breaching and Protecting Genetic Privacy*, 15 NATURE REVIEWS GENETICS 409 (2014).

76. Gina Kolata, *Poking Holes in Genetic Privacy*, N.Y. TIMES (June 16, 2013), http://www.nytimes.com/2013/06/18/science/poking-holes-in-the-privacy-of-dna.html [https://perma.cc/PQ8U-9JXH].

method for analyzing a complex mixture of DNA samples from the HapMap database[77] and confirming whether or not an individual's DNA is present in the mixture.[78] This study led the National Institute of Health (NIH) to remove aggregate genomic data from its public databases and place it in a controlled access database, where there are "protections and policies in place for appropriate data access."[79]

The second case occurred five years later, when a group of genetics researchers described a new statistical method for identifying individual data subjects from donated DNA samples. They began with Y-chromosome data hosted in a HapMap database and searched for matching records in recreational genetic genealogy databases (which allow the public to enter their own DNA information and find relatives with the same surname). When the researchers found a match, they combined the surnames with additional demographic data to reidentify the sample originator.[80]

These two cases prompted geneticists and associated research institutions to reconsider existing paradigms for sharing genomic data, culminating in a new genomic data sharing policy, announced by the NIH in 2014.[81] NIH's final rule on genomic data sharing cites the Gymrek attack in the context of explaining a change in policy requiring investigators to obtain informed consent from prospective subjects, even

---

77. HapMap catalogues common genetic variants that occur in human beings and provides information that researchers can use to link genetic variants to the risk for specific illnesses, with the aim of developing new methods of preventing, diagnosing, and treating disease. *See generally What Is the HapMap?*, INT'L HAPMAP PROJECT, http://hapmap.ncbi.nlm.nih.gov/whatishapmap.html [https://perma.cc/MV7G-NZ93] (last visited Apr. 11, 2016).

78. *See* Kolata, *supra* note 76. For the technical paper describing the relevant techniques, see Nils Homer et al., *Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays*, 4 PLOS GENETICS 1 (2008).

79. Elias A. Zerhouni & Elizabeth G. Nabel, Letter to the Editor, *Protecting Aggregate Genomic Data*, 322 SCI. 43, 44 (2008). A year earlier, NIH had created a database of genotypes and phenotypes (dbGaP), which relied on a "tiered access" system to provide unprecedented access to the large-scale genetic and phenotypic data sets required for so-called genome wide associations studies, in which researchers examined many common genetic variants in different individuals to see if any variant is associated with a genetic trait. *See* Matthew D. Mailman et al., *The NCBI dbGaP Database of Genotypes and Phenotypes*, 39 NATURE GENETICS 1181 (2007). Tiered access allows anyone to access less sensitive study protocols and summary data without restriction, but requires preauthorization from sponsoring NIH programs for access to more sensitive, individual-level data. *Id*. NIH also protected the confidentiality of study subjects by accepting only deidentified individual data into the dbGaP and releasing such data as encrypted files to authorized users who also had to comply with additional data security requirements. *Id*. at 1183.

80. *See* Kolata, *supra* note 76. For the technical paper describing the relevant techniques, see Melissa Gymrek et al., *Identifying Personal Genomes by Surname Inference*, 339 SCI. 321 (2013).

81. Final NIH Genomic Data Sharing Policy, 79 Fed. Reg. 51,345 (Aug. 28, 2014).

if the data in question would be deidentified.[82] While the new NIH policy promotes the use of consent for broad sharing, it also requires researchers to explain to prospective participants the risks of reidentification and whether or not their deidentified data will be shared through unrestricted or controlled-access repositories.[83] Thus, deidentification, consent, and tiered access work together to provide multiple layers of protection.

This brief case study of genetic reidentification illustrates two points. The first is that it is possible to achieve most of the benefits of open access without releasing data to the public with no restrictions. As the former director of the National Institute of Statistical Sciences observed, data availability in the purist sense of "openness" is not what matters most. Rather, the most important goal is sound "decisions by governments, businesses, and individuals that are based on the data."[84] The second is that even in the face of reidentification attacks, it remains possible to balance participant privacy and broad accessibility of genomic data for research purposes by combining technical *and* policy safeguards. Rather than give up deidentification entirely, the new NIH policy supplements it with other mechanisms such as informed consent protocols and tiered access, along with new security requirements,[85] code of conduct for approved users,[86] and DUAs.[87] The scientific community generally favors this balanced approach,[88] although some

---

82. *Id*. at 51,347.

83. *Id*. at 51,351; *see also* NAT'L INST. OF HEALTH, NIH SECURITY BEST PRACTICES FOR CONTROLLED-ACCESS DATA SUBJECT TO THE NIH GENOMIC DATA SHARING (GDS) POLICY (2015), http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?document_name=dbgap_2b_security_procedures.pdf [https://perma.cc/BLC7-LLVC].

84. ALAN F. KARR, NAT'L INST. OF STATISTICAL SCIS., WHY DATA AVAILABILITY IS SUCH A HARD PROBLEM, TECHNICAL REPORT 186 (2014), http://www.niss.org/sites/default/files/tr186.pdf [https://perma.cc/93CY-Z68F]; *see also* NAT'L INST. OF HEALTH, *supra* note 83.

85. *See* NAT'L INST. OF HEALTH, *supra* note 83.

86. *See* NAT'L INST. OF HEALTH, GENOMIC DATA USER CODE OF CONDUCT (2010) [hereinafter NIH CODE OF CONDUCT], http://gds.nih.gov/pdf/Genomic_Data_User_Code_of_Conduct.pdf [https://perma.cc/4CFP-GR6J].

87. *See* NAT'L INST. OF HEALTH, MODEL DATA USE CERTIFICATION AGREEMENT (2013), https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?view_pdf&stacc=phs000016.v1.p1 [https://perma.cc/2MHL-R6LG]. Both the NIH CODE OF CONDUCT, *supra* note 86, and the DUA explicitly prohibit the use of genomic data sets to identify or contact data subjects.

88. *See, e.g.*, George Church et al., *Public Access to Genome-Wide Data*: *Five Views on Balancing Research with Privacy and Protection*, 5 PLOS GENETICS 1 (2009), http://journals.plos.org/plosgenetics/article/asset?id=10.1371%2Fjournal.pgen.1000665.PDF [https://perma.cc/2K8L-7WVX]; Catherine Heeney et al., *Assessing the Privacy Risks of Data Sharing in Genomics*, 14 PUB. HEALTH GENOMICS 17 (2010); WILLIAM W. LOWRANCE, MED. RESEARCH COUNCIL, ACCESS TO COLLECTIONS OF DATA AND MATERIALS FOR HEALTH RESEARCH:

geneticists would prefer greater use of open access[89] and/or a more dynamic form of consent.[90]

B. Moving Past the Deidentification Debate

The deidentification debate—which pits those who reject deidentification as irretrievably flawed against those who defend both its ongoing validity and practical value—has greatly overshadowed successful policy outcomes like NIH's new genomic data sharing policy. Experts in the field of genomics achieved the latter by careful deliberation and compromise. In contrast, the privacy scholarship seems fixated on the deidentification debates, with opposing sides taking extreme positions and making overly general claims about data release policy across all disciplines.

For example, Paul Ohm insists that deidentification is a failure and should be abandoned.[91] In the opposing corner, Jane (Yakowitz) Bambauer and Daniel Barth-Jones have argued that the famous trio of reidentification attacks (Weld, AOL, and Netflix) distorts the policy debate because they are not representative or have been misrepresented in popular media.[92] Like Ohm, we credit these attacks for demonstrating shortcomings with deidentification techniques. But we argue they should be used differently. Instead of focusing on what they illustrate about the failure of anonymization, the attacks show what data custodians can learn from past mistakes, while encouraging them to experiment with new techniques and institutional arrangements.

In this Part, we review the deidentification literature to see if it is really as divided as it seems. There are distinct arguments and ideologies, but they are often isolated or concern more specific aspects of deidentification. We suggest that a careful reading of the privacy scholarship against the backdrop of our earlier analysis of SDL and related topics reveals a rough consensus that can be used to develop data release policy around the concept of minimizing risk.

A REPORT TO THE MEDICAL RESEARCH COUNCIL AND THE WELLCOME TRUST (2006), http://www.wellcome.ac.uk/stellent/groups/corporatesite/@msh_grants/documents/web_document/wtx030842.pdf [https://perma.cc/FWV6-58H4] (acknowledging the importance of controlling access to sensitive health information).

89. *See*, *e.g.*, Laura L. Rodriguez et al., *The Complexities of Genomic Identifiability*, 339 SCI. 275 (2013).

90. *See*, *e.g.*, Amy L. McGuire & Richard A. Gibbs, *No Longer De-Identified*, 312 SCI. 370 (2006); Stacey Pereira et al., *Open Access Data Sharing in Genomic Research*, 5 GENES 739 (2014).

91. Ohm, *supra* note 2.

92. *See supra* notes 15–21.

1. Ohm v. Yakowitz

Ohm's highly influential article treats deidentification—or what he calls "release-and-forget anonymization"—as fool's gold.[93] He reads the computer science literature as proving the "theoretical limits" of the power of deidentification techniques,[94] and argues that we should not expect any technical breakthroughs to "save us" or to replace the need for a regulatory response premised on a more realistic assessment of the risks of reidentification and the appropriate response.[95] Ohm's analysis accomplishes a great deal by alerting privacy regulators to the dangers of treating anonymization as a silver bullet. The scope of many privacy laws depends on whether information is identifiable or not, and Ohm's critique raises legitimate questions about defining the scope of privacy laws by reference to this distinction. He also wisely suggests that privacy regulators reject this longstanding binary distinction between PII and non-PII in favor of a more risk-based approach.[96]

Yakowitz sought to rebut Ohm's arguments by offering two main points in defense of anonymization. First, she claimed that Ohm (and other critics) neglected the value of the data commons, which she described as the "diffuse collections of data made broadly available to researchers with only minimal barriers to entry."[97] According to Yakowitz, the benefits flowing from the data commons are immense and range across diverse fields. Thus, if policymakers were to end or even restrict public data releases of deidentified data sets, society would suffer a new tragedy of the data commons.[98] Second, she argues that the risks of reidentification are mainly theoretical and in any case highly exaggerated. She thus advances a proposal that would make it easier, not harder, to disseminate anonymized data sets.[99] Like Ohm, Yakowitz makes a valuable contribution to the public policy debate by alerting

---

93. Ohm, *supra* note 2, at 1711–12 (noting that "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").

94. *Id*. at 1751.

95. *Id*. at 1759–69.

96. *Id*. at 1764–68.

97. Yakowitz, *supra* note 6, at 2–3.

98. *Id*. at 4.

99. Yakowitz's proposal imposes two conditions on a data producer: "(1) strip all direct identifiers, and (2) either check for minimum subgroup sizes on a preset list of common indirect identifiers—such as race, sex, geographic indicators, and other indirect identifiers commonly found in public records—or use an effective random sampling frame." *Id*. at 44.

policy makers to the opportunity costs of reduced data sharing.

2. *A Different Path*

Ohm sought to kill deidentification and used strong rhetoric as a weapon.[100] Yakowitz also made a forceful argument, but hers was at the opposite pole.[101] However, these extreme positions undermine the policy debate. By limiting their respective analyses almost exclusively to the release-and-forget model, both Ohm and Yakowitz largely neglect the full gamut of SDL techniques. Rather, they favor the dissemination-based model in which deidentification techniques must bear the entire weight of balancing privacy and utility, with no help from direct access (which employs administrative, technical, and physical controls in support of controlled access) or query-based methods like differential privacy (which refrain from releasing data at all).

Ohm rejected these other forms of SDL out of hand, not because they fail on technical grounds, but on the grounds they are "slower, more complex, and more expensive than simple anonymization," "useless for many types of data analysis problems," and "cannot match the sweeping privacy promises that once were made regarding release-and-forget anonymization."[102] Of course, it is ironic for Ohm to raise these objections given his utter lack of faith in release-and-forget anonymization.

Similarly, Yakowitz does not endorse other SDL methods. This might be because some perceive them as inconsistent with open data. According to Yakowitz: "[n]early every recent public policy debate has benefited from mass dissemination of anonymized data."[103] But the necessity of open data in its purest sense is debatable. At least some of the examples cited by Yakowitz as evidence of this claim do not depend

---

100. According to Ohm, deidentification methods are not merely flawed but a "shared hallucination." Ohm, *supra* note 2, at 1768. The distinction between PII and non-PII is not just in need of adjustment, but must be completely abandoned because the list of potential PII (or quasi-identifiers) "will never stop growing until it includes everything." *Id*. at 1742. And not only the HIPAA Privacy Rule, but "every privacy law and regulation" needs reassessment and revision. *Id*. at 1731.

101. She not only criticized the computer science literature, but set out to debunk five "myths" about reidentification risk. Yakowitz, *supra* note 6, at 23–35. True risks posed by anonymization are not merely lower than reported but "nonexistent." *Id*. at 4. And concerns over anonymization are not only disproportionate to the true risks, but "have all the characteristics of a moral panic." *Id*. at 5.

102. Ohm, *supra* note 2, at 1751.

103. Yakowitz, *supra* note 6, at 9.

on any public release of anonymized data.[104] More generally, as noted above, the values supporting openness do not rest on the public availability of anonymized data. Finally, the database of genotypes and phenotypes (dbGaP)[105] and the favorable treatment of controlled access in the NIH genomic data sharing policy,[106] and the even more recent IOM Study,[107] show the value that can be had from relatively controlled releases of information.

We agree with later commentators such as Felix Wu that both Ohm and Yakowitz have "misinterpreted, or at least overread" the relevant computer science literature, although in different ways.[108] In particular, Ohm and Yakowitz deploy the problem of auxiliary information in different and problematic ways. Ohm's article neglects the empirical research around deidentified health data, which shows that the risk of reidentification is in fact very small (although Ohm's article preceded some, but not all, of this research).[109] Yakowitz, on the other hand, treats the Netflix study as a "theoretical contribution,"[110] while embracing the empirical studies of health data over the more "hypothetical risks" identified by popular reidentifications.[111] But these risks are not merely hypothetical in light of the impressive theorems and proofs of computer scientists working in this field, and hence not so easily dismissed.[112]

We highlight the opposing positions of Ohm and Yakowitz to show why the policy debate has stagnated. Is there an alternative path forward? The answer is "yes," and the relevant headline is "Reidentification Is Not the End of the Story." There is no denying that deidentification techniques have significant limits, especially with regard to internet scale data sets.[113] But the trio of high-profile cases point in a

---

104. In at least two of the sentencing studies cited by Yakowitz, researchers were granted special permission to access non-public data sets. *Id*. at 9.

105. *See supra* note 79.

106. *See supra* text accompanying notes 79–83.

107. *See supra* text accompanying note 73.

108. Wu, *supra* note 10, at 1124. Wu advanced the discussion by carefully delineating the meaning of privacy and utility in different contexts, thereby enabling policymakers "to choose among these competing definitions." *Id*. at 1125.

109. *See supra* note 48.

110. Yakowitz, *supra* note 6, at 26.

111. *Id*. at 35.

112. See, for example, Dwork's proof of the auxiliary information problem, *supra* text accompanying notes 30–31, Narayanan and Shmatikov's deanonymization algorithm and proof-of-concept deidentification of the Netflix dataset, *supra* text accompanying notes 36–39, and de Montjoye's study of unicity in large data sets, *supra* text accompanying notes 40–43.

113. *See supra* Section I.A.1.

different direction from the usual death of anonymization narrative.

For example, the exposure of Weld's medical records directly influenced the Health Insurance Portability and Accountability Act of 1996[114] (HIPAA) Privacy Rule by ensuring that it included deidentification requirements designed to limit the risk of linkage attacks, and thereby improving the privacy of health records.[115] Both the AOL debacle and the Netflix attack inspired research on, respectively, the safe release of search logs,[116] and privacy-preserving recommendations systems.[117] Furthermore, Overstock.com learned a lesson from the Netflix experience by organizing a one million dollar contest for an improved product recommendation system in which it minimized risk by refraining from releasing the anonymized prize data set to contestants.[118] Rather, it relied on synthetic data and a secure cloud

---

114. Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191, 110 Stat. 1936.

115. HIPAA exempts deidentified health data from the Privacy Rule if it meets either the Safe Harbor standard, *see* 45 C.F.R. § 164.514(b)(2) (2015) (requiring the removal of eighteen enumerated data elements including name, geographic subdivisions smaller than a state, all date elements directly related to an individual other than year, contact information, and various identifiers), or the expert determination standard, *see id*. § 164.514(b)(1) (requiring an expert determination using "generally accepted statistical and scientific principles and methods" of deidentification that there is a "very small" risk that the deidentified information "could be used, alone or in combination with other reasonably available information, . . . to identify an individual who is a subject of the information"). Sweeney's work on the Weld reidentification heavily influenced the formation of the HIPAA Safe Harbor standard. *See* Daniel Barth-Jones, The "Re-Identification" of Governor William Weld's Medical Information (2012) (unpublished manuscript), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2076397 [https://perma.cc/MN9A-7CTF] (arguing that if the Weld reidentification attack had taken place *after* the HIPAA Privacy Rule took effect, it would have been extremely difficult to undertake a successful linkage attack).

116. *See, e.g.*, Michaela Götz et al., *Publishing Search Logs—A Comparative Study of Privacy Guarantees*, 24 IEEE TRANSACTIONS ON KNOWLEDGE & DATA ENGINEERING 520 (2012); Aleksandra Korolova et al., *Releasing Search Queries and Clicks Privately*, *in* PROCEEDINGS OF THE 18TH INTERNATIONAL WORLD WIDE WEB CONFERENCE 171 (2009), http://theory.stanford.edu/~korolova/Releasing_search_queries_and_clicks_privately_WWW2009.pdf [https://perma.cc/22CB-FMG9].

117. *See* Frank McSherry & Ilya Mironov, *Differentially Private Recommender Systems: Building Privacy into the Netflix Prize Contenders*, *in* PROCEEDINGS OF THE 15TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING (KDD) 627 (2009), http://research.microsoft.com/pubs/80511/NetflixPrivacy.pdf [https://perma.cc/F7P4-P3FJ] (describing new techniques based on differential privacy that allow researchers to work on improvements to the accuracy of recommendation systems without compromising privacy).

118. *See* Steve Lohr, *The Privacy Challenge in Online Prize Contests*, N.Y. TIMES (May 21, 2011), http://bits.blogs.nytimes.com/2011/05/21/the-privacy-challenge-in-online-prize-contests/ [https://perma.cc/RHS9-ZX29]; Rich Relevance, *Overstock.com and RichRelevance Offer $1 Million Prize to Speed Innovation in Retail Personalization*, RICHRELEVANCE.COM (May 12, 2011), http://www.richrelevance.com/blog/2011/05/overstock-com-and-richrelevance-offer-1-million-prize-to-speed-innovation-in-retail-personalization/ [https://perma.cc/2PC5-TZ8M].

environment to run a contest without endangering customer privacy.[119] Finally, the Data for Development (D4D) Challenge encouraged researchers to explore international development applications using mobile data across a wide range of subject matters (including health, agriculture, transportation and urban planning, energy, and national statistics), while protecting the privacy of data subjects.[120] With help from a team of experts at MIT, D4D released a modified set of mobile phone data[121] to qualified researchers subject to a DUA imposing confidentiality obligations and restricting their use of the data to approved projects.[122] The result was a widely praised competition with over sixty entries from leading academics and practitioners around the world and valuable research conducted with reasonable privacy guarantees.[123] In short, the deidentification debate as currently conceived overlooks and obfuscates success stories involving improved regulations, new research, and improved contests and challenges that (in the case of D4D) both avoided past errors and achieved significant results.

II. A PROCESS-BASED APPROACH TO MINIMIZE RISK

There is another way for data release policy to advance. Instead of

---

119. *See* Darren Vengroff, *The Inspiration Behind RecLab: Don't Bring the Data to the Code, Bring the Code to the Data*, RICHRELEVANCE.COM (Jan. 31, 2011), http://www.richrelevance.com/blog/2011/01/the-inspiration-behind-reclab-dont-bring-the-data-to-the-code-bring-the-code-to-the-data/ [https://perma.cc/W4CV-2VDT]. On the use of synthetic data for anonymization purposes, see Ashwin Machanavajjhala et al., *Privacy: Theory Meets Practice on the Map*, in PROCEEDINGS OF THE IEEE 24TH INTERNATIONAL CONFERENCE ON DATA ENGINERING 277 (2008), http://www.cse.psu.edu/~duk17/papers/PrivacyOnTheMap.pdf [https://perma.cc/XP5N-C7LA].

120. *See, e.g.*, *The D4D Challenge Is a Great Success!*, ORANGE, http://www.d4d.orange.com/en/Accueil [https://perma.cc/SVV9-W2QQ] (last visited Apr. 24, 2016) (describing a "Data for Development" challenge organized by Orange (a multinational mobile operator) and Sonatel (Senegal's mobile operator), with a grant from the Gates Foundation).

121. For a description of "promising computational privacy approaches to make the re-identification of mobile phone metadata harder," see YVES-ALEXANDRE DE MONTJOYE ET AL., CTR. FOR TECH. INNOVATION AT BROOKINGS, ENABLING HUMANITARIAN USE OF MOBILE PHONE DATA 1, 5–6 (2014).

122. *See Conditions for the Availability of Data–Data for Development (D4D)*, ORANGE, http://www.d4d.orange.com/en/content/download/29438/273168/version/12/file/D4DSonatel_06062014Engl.pdf [https://perma.cc/FXF4-5TEL] (last visited Apr. 24, 2016).

123. *See* ORANGE, ORANGE DATA FOR DEVELOPMENT CHALLENGE IN SENEGAL, http://d4d.orange.com/content/download/43330/405662/version/3/file/D4Dchallenge_leaflet_A4_V2Eweblite.pdf [https://perma.cc/4VS2-Y3ZB]. For other examples of similar projects, see Global Pulse, *Mobile Phone Network Data for Development*, LINKEDIN: SLIDESHARE (Oct. 2013), http://www.slideshare.net/unglobalpulse/mobile-data-for-development-primer-october-2013 [https://perma.cc/Q86G-WS42].

focusing on the ultimate goal of anonymization, the law could be designed around the processes necessary to lower the risk of reidentification and sensitive attribute disclosure. One of the reasons the debate about anonymization is so lively is that the concept inherently over-promises. To say something is anonymized is to imply a certain threshold of protection has been obtained.

Think of this as a regulatory choice between output and process.[124] When data release policy focuses on endpoints like minimizing harm and avoiding actual reidentification, there are no rules about the specific ways in which data is protected. Output regimes sanction data security efforts so long as the information is made anonymous or, in more reasonable regimes, the resulting protection achieves a pre-specified threshold such as a "very small" risk that "information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information."[125]

While outcome-based approaches to releasing data might be good enough for many purposes, they are not ideal as the centerpiece for data release policy. As we discussed above, perfect anonymization is a myth. Even when more reasonable thresholds are set, scholars have shown that such estimates of protection are notoriously slippery given systemic obstacles (like the auxiliary information problem) and the number of variables that can effect just how well information is actually protected.

A more sustainable approach would focus on the preconditions and processes necessary for protection. It is hard to ensure protection. It is easier, however, to ensure that data custodians follow appropriate processes for minimizing risk, which may include both deidentification in combination with legal and administrative tools, or reliance on query-based methods like differential privacy when it is suitable for the task. We argue that data release policy should focus on processes, not outputs. Of course, there is another familiar area of information policy that focuses on process: data security.

In this Part we argue that data release policy should look more like data security policy. We explore the additional elements data release policy must incorporate beyond data treatment techniques, and we list the components of process-based deidentification.

---

124. *See, e.g.*, Lauren E. Willis, *Performance-Based Consumer Law*, 82 U. CHI. L. REV. 1309 (2015).

125. 45 C.F.R. § 164.514(b)(1)(i) (2015).

A. The Poor Fit of Traditional Privacy Law for Anonymization

The law should evolve to focus on risk and process because traditional goals and strategies of privacy law do not really fit the specific concerns related to the release of data sets. Most existing privacy laws focus on specific data subjects and discrete types of information, rather than data sets as a whole.[126] Nor would it be a good idea to focus on the harms that follow poorly deidentified data. To begin with, harm is a contentious concept in privacy law.[127] Many privacy harms are incremental or difficult to quantify and articulate. For example, if hackers steal your information and then sell that information in the black market, have you been harmed? What if you do not lose any money? Is your privacy violated if your personal information is used to create an incorrect profile of your likes and dislikes, which is used and sold by data brokers? These sorts of injuries often fall through the cracks of harm-based privacy regimes with high injury thresholds.

Additionally, harms related to insufficient anonymization can also be very difficult to detect, because reidentification usually remains hidden unless a researcher or adversary claims credit for a successful attack. Attackers can thwart anonymization attempts in secret, on their own computers in unknown places. They can also exploit the reidentification of people and attributes in largely undetectable ways. Thus, harms from failed anonymization attempts might not come to light until many years after the fact, if ever. By that time, it might be impossible to tell who "caused" the harm in a traditional legal sense, even assuming the relevant harm is articulable and legally cognizable.

Focusing solely on transparency and disclosures is also unwise. The failures of notice and choice regimes are well noted.[128] Consumers only have a limited ability to make meaningful decisions regarding their own privacy due to the incredible volume, impenetrability, and interconnectedness of data collection and transfers.[129] And the number of

---

126. Fair Credit Reporting Act, 15 U.S.C. § 1681 (2012); Gramm-Leach-Bliley Act, 15 U.S.C. §§ 6801–6809 (2012).

127. M. Ryan Calo, *The Boundaries of Privacy Harm*, 86 IND. L.J. 1131, 1135 (2011).

128. In a notice and choice regime, companies can engage in nearly any activity so long as a person has notice of the company's actions and the choice to avoid it, such as by not using a particular service. *See, e.g.*, Julie Brill, Comm'r, Fed. Trade Comm'n, Keynote Address at Proskauer on Privacy (Oct. 19, 2010), https://www.ftc.gov/sites/default/files/documents/public_statements/remarks-commissioner-julie-brill/101019proskauerspeech.pdf [https://perma.cc/3WRF-4RG2].

129. *See, e.g.*, Daniel J. Solove, *Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1880, 1885 (2013).

potential additional disclosures that would be needed to address questionable efforts to deidentify their data would quickly overwhelm them. Control over information soon becomes a burden on consumers given the responsibility of exerting that control in seemingly unlimited contexts.

The permission-based model that governs medical research under the heading of informed consent also presents numerous problems. In order to conduct medical research, companies and researchers must seek permission either from a regulatory body or the data subject, unless an exception applies. In the private sphere, companies easily extract consent from people, even though it is regularly meaningless.[130] While consent might have an important role to play in data release policy, it should not be over-leveraged.

Yet blanket and robust prohibitions on information collection and disclosure would be incredibly costly to organizations and society as a whole. Shutting down research and the information economy would be devastating. Even if such restrictions were wise and politically palatable, they would likely be ineffective given the existing data ecosystem.

In short, approaches that focus on transparency, disclosures, harm, and permission all seem inadequate, at least by themselves, to respond to the failure of anonymization. Traditional privacy law focuses too much on individuals and the nature of the information collected, used, or disclosed. Nor are ex post, individualized remedies very effective when specific harms can be hard to articulate or even locate. Instead, process and risk can guide the best path forward.

B. Data Release Policy Should Look Like Data Security

Data security law involves the protection of privacy, yet it is analytically distinct from traditional privacy concerns in several different ways. As Derek Bambauer has argued, "[w]hile legal scholars tend to conflate privacy and security, they are distinct concerns. Privacy establishes a normative framework for deciding who should legitimately have the capability to access and alter information. Security implements those choices."[131] According to Bambauer, security comprises "the set of technological mechanisms (including, at times, physical ones) that mediates requests for access or control."[132] Data security policy

130. *Id*. at 1894.

131. Derek E. Bambauer, *Privacy Versus Security*, 103 J. CRIM. L. & CRIMINOLOGY 667, 668–69 (2013).

132. *Id*. at 669.

addresses the selection and implementation of those mechanisms by determining who is able to "access, use, and alter data. When security settings permit an actor without a legitimate claim to data to engage in one of these activities, we do not view that fact as altering the normative calculus. The actor's moral claim does not change. The access or use is simply error."[133]

Applying a process-based data security approach to deidentification would be appropriate, even though deidentification is more often discussed as a privacy problem. The concept of an attacker is deployed in both data security and deidentification fields and many technologists already consider deidentification a data security issue.[134]

A process-based data security approach has a number of advantages over traditional privacy-related output-based approaches. For one, those who attempt to violate security have fewer ethical claims than many who are accused of violating more nebulous notions of privacy. Data security breaches and reidentifications lack the justifications often supplied for activities like surveillance and ad targeting. As Bambauer observed, "security failures generally leave everyone involved (except for the attacker) worse off."[135] Of course, security concerns also involve competing considerations like cost and usability. But this calculus is typically incorporated into applicable "reasonableness" standards common in data security policy and practice.

Data releases straddle both privacy and data security worlds. In many ways it can be difficult to distinguish the privacy and security issues at play. Consider two scenarios. First, Alpha Research Institute plans to release data, worries about confidentiality of sensitive records, relies solely on deidentification methods, which fail, resulting in individuals being harmed because their reidentified data sets have been accessed by those without authorization. Second, Beta Research Institute holds similar data, which is hacked via an elevation of privilege attack. Beta failed to encrypt its data, resulting in disclosure. Setting aside questions

---

133. *Id*. at 676.

134. NIST REPORT, *supra* note 10, at 9 ("The term 'attack' is borrowed from the literature of computer security . . . ."); *cf*. Stuart S. Shapiro, *Separating the Baby from the Bathwater: Toward a Generic and Practical Framework for Anonymization*, *in* PROCEEDINGS OF THE 2011 IEEE INTERNATIONAL CONFERENCE ON TECHNOLOGIES FOR HOMELAND SECURITY (2011) [hereinafter Shapiro, *Separating the Baby from the Bathwater*]; Stuart S. Shapiro, *Situating Anonymization Within a Privacy Risk Model*, *in* PROCEEDINGS OF THE 2012 IEEE INTERNATIONAL SYSTEMS CONFERENCE (SYSCON) [hereinafter Shapiro, *Situating Anonymization*], https://www.mitre.org/sites/default/files/pdf/12_0353.pdf [https://perma.cc/H7B6-RACN].

135. Bambauer, *supra* note 131, at 681. Deidentification and data security are still costly, of course.

of difficulty or harm, is one a privacy incident and the other a security incident?

Data release and deidentification are usually conceptualized as privacy issues. In a sense, of course, they are. Embarrassing and private information can be harmfully linked to real people through reidentification attacks. But, at least to the extent that data custodians avoid release-and-forget anonymization, we argue that data release is largely a data security issue insofar as it is concerned with who can actually access, use, and alter data. Similar issues of data integrity, identification of assets and risk, and the need for safeguards and probabilistic protections apply. Below we discuss several important aspects of data security and why they should be incorporated into data-release policy. In particular, data security is process based, contextual, and risk tolerant.

Process Based. At the level of policy, data security is conceived of as a process of continually identifying risk; minimizing data collection and retention; developing and implementing administrative, technical, and physical safeguards to protect against data breaches; and developing a response plan if a breach does occur.[136] When a company fails to provide legally obligated reasonable data security, its culpable conduct is not in its failure to reach a predetermined level of protection, but rather in the failure to take the steps generally recognized in the industry to sufficiently reduce risk.

In other words, in process-based regimes like data security, companies can be liable even in the absence of an actual breach because the law mandates procedures, not outputs.[137] The actual harm is relevant only insofar as it gives clues as to which procedures might not have been properly implemented.

Compare this to output-based regimes focused on safety and harm. Under tort law, people are generally free to act as recklessly as they want, so long as they do not harm anyone. The failure of tort law in cases of data breaches demonstrates this point. Claims against companies for negligent data security practices usually fail unless the

---

136. *See* 15 U.S.C. §§ 6801–6809 (2012); 45 C.F.R. § 164.306(a) (2015); Press Release, Fed. Trade Comm'n, Commission Statement Marking the FTC's 50th Data Security Settlement (Jan. 31, 2014), http://www.ftc.gov/system/files/documents/cases/140131gmrstatement.pdf [https://perma.cc/FGB8-JB4K].

137. *See* Woodrow Hartzog & Daniel J. Solove, *The Scope and Potential of FTC Data Protection*, 83 GEO. WASH. L. REV. 2230 (2015); Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583 (2014) [hereinafter Solove & Hartzog, *Common Law of Privacy*].

plaintiff can demonstrate actual individualized harm, such as financial loss.[138] Things like uncertainty, anxiety, or increased risk of identity theft shared across large numbers of people that are significant in the aggregate but small for each affected individual are usually not recognized as sufficient to clear the harm threshold.

Process-based regimes are also more suitable than output-based regimes when parties have custodian-like responsibilities to protect people from others rather than responsibilities to keep from directly harming others. Tort law is largely based upon the concept that a party should not directly harm another party. Data security is based upon the idea that parties should take steps to protect those who have entrusted them with data. In other words, data security regimes do not have to wrestle with the same kinds of causation issues demanded in output-based regimes like tort law. Process failures or violation of reasonableness standards are treated as culpable behavior regardless of the actions of others.

Data releases fit better into a data security model than a tort law model. The party releasing a data set should be responsible for protecting people through adequate deidentification procedures, in combination with restrictions on access or use, or reliance on query-based methods where appropriate. Of course, those who engage in reidentification are also culpable. However, they are often much more difficult to locate and direct causation is not always clear. When many data sets are combined through linkage attacks to reidentify a person, it is difficult to apportion comparative fault. Focusing on process helps avoid such intractable analyses.

Contextual. Data security and related policy is contextually sensitive. FTC Chairwoman Edith Ramirez has stated that, "[t]he level of security required [by a company] depends on the sensitivity of the data, the size and nature of a company's business operations, and the types of risks a company faces."[139]

---

138. *See, e.g.*, Katz v. Pershing, LLC, 672 F.3d 64, 75 (1st Cir. 2012) (finding no harm from increased risk of identity theft); Holmes v. Countrywide Fin. Corp., No. 5:08-CV-00205-R, 2012 WL 2873892, at *13 (W.D. Ky. July 12, 2012) (rejecting theory of harm for time and efforts expended to deal with breach); Amburgy v. Express Scripts, Inc., 671 F. Supp. 2d 1046, 1053 (E.D. Mo. 2009) (rejecting standing for increased risk of identity theft); McLoughlin v. People's United Bank, Inc., No. 3:08-cv-00944(VLB), 2009 WL 2843269, at *3–4 (D. Conn. Aug. 31, 2009) (rejecting theory of harm of loss of benefit of the bargain); Bell v. Acxiom Corp., No. 4:06CCV004850WRW, 2006 WL 2850042, at *5 (E.D. Ark. Oct. 3, 2006) (rejecting theory of harm for increased risk of junk mail).

139. *Discussion Draft of H.R. __, A Bill to Require Greater Protection for Sensitive Consumer Data and Timely Notification in Case of Breach: Hearing Before the H. Comm. on Energy and*

Data release policy should be similarly dependent upon context, because sound deidentification is similarly contingent upon a large number of factors. These include different motivations for attacks,[140] different approaches for computing reidentification risk,[141] the different standards that have been used to describe the abilities of the "attacker,"[142] the variety of harms that can result from the use or distribution of deidentified data,[143] the effort that the organization can spend performing and testing the deidentification process, the utility desired for the deidentified data, the ability to use other controls that can minimize risk, the likelihood that an attacker will attempt to reidentify the data, and amount of effort the attacker might be willing to expend.[144]

Wu noted that another contextually dependent deidentification variable is the extent to which probabilistic knowledge should be treated as a privacy violation and reidentification.[145] In other words, if an attacker is fifty-one percent sure that a record is pointing to a particular person, has that person been reidentified? What if an attacker can determine there is a ninety percent chance of reidentification?[146] The answer surely depends upon the variables mentioned above, including the number of people subject to reidentification, possible harms of reidentification, and motivation of the attacker.

Commerce, Subcomm. on Commerce, Manufacturing, and Trade, 112th Cong. 42, 50 (June 15, 2011) (statement of Edith Ramirez, Fed. Trade Comm'n).

140. NIST REPORT, *supra* note 10, at 10; *see also* INFO. COMM'RS OFFICE, ANONYMISATION: MANAGING DATA PROTECTION RISK CODE OF PRACTICE (2012) [hereinafter ICO CODE], https://ico.org.uk/media/1061/anonymisation-code.pdf [https://perma.cc/6RDM-RWQ2]. A novel contribution of the ICO Code is its "motivated intruder" test, which is proposed as a default position for assessing risk of reidentification subject to modification according to context. ICO CODE, *supra*, at 22. The ICO noted:

> The "motivated intruder" test is useful because it sets the bar for the risk of identification higher than considering whether a "relatively inexpert" member of the public can achieve re-identification, but lower than considering whether someone with access to a great deal of specialist expertise, analytical power or prior knowledge could do so.

Id. at 23.

141. ICO CODE, *supra* note 140, at 23.

142. *Id*.

143. NIST REPORT, *supra* note 10, at 9–14 (the variety of harms might include incomplete deidentification, identity disclosure, inferential disclosure, association harms, group harms, and unmasking).

144. *Id*. at 13–14; *cf*. IOM STUDY, *supra* note 10.

145. Wu, *supra* note 10, at 1164.

146. Wu noted, "[t]he law tends to treat 51 percent as a magical number, or to use some other generally applicable threshold of significance. What matters with respect to privacy, however, is what effect uncertain information has, and the effect of a particular numerical level of certainty can vary widely across contexts." *Id*. (citations omitted).

All of these factors mean that a "one size fits all" standard for data release policy will not be effective. Such attempts are doomed to be either over-protective or under-protective. Data security policymakers face a similar reality. Critics of data security policy in the United States often claim they need something akin to a checklist of clearly defined rules that set out in explicit detail the steps a company must take to be compliant with the law.[147]

But like deidentification, there are too many factors to provide a consistent and detailed checklist for required data security practices. Instead, the FTC and other regulatory agencies have required "reasonable" data security, which is informed by industry standards.[148] A reasonableness approach maximizes the contextual sensitivity of a regulatory regime. Reasonableness is an established concept employed in a number of different contexts, including contracts, Fourth Amendment law, tort law, and others.[149] Because the field of deidentification advances so quickly and a determination of the risk of identification involves so many factors, deidentification policy should be contextually sensitive in a way similar to data security policy.

Risk Tolerant. The field of data security has long acknowledged that there is no such thing as perfect security.[150] As Bambauer has argued, "[s]cholars should cast out the myth of perfection, as Lucifer was cast out of heaven. In its place, we should adopt the more realistic, and helpful, conclusion that often good enough is . . . good enough."[151] Yakowitz, Wu, and even Ohm have also recognized the need to be tolerant of risk.[152]

---

147. *See generally* Gerard M. Stegmaier & Wendell Bartnick, *Psychics, Russian Roulette, and Data Security: The FTC's Hidden Data-Security Requirements*, 20 GEO. MASON L. REV. 673 (2013).

148. Press Release, Fed. Trade Comm'n, *supra* note 136.

149. LabMD, Inc., 159 F.T.C. 2145 (Jan. 16, 2014) (interlocutory order); Woodrow Hartzog & Daniel J. Solove, *The Scope and Potential of FTC Data Protection*, 83 GEO. WASH. L. REV. 2230 (2015).

150. *See, e.g.*, Derek E. Bambauer, *The Myth of Perfection*, 2 WAKE FOREST L. REV. ONLINE 22 (2012), http://wakeforestlawreview.com/2012/04/the-myth-of-perfection/ [https://perma.cc/9DKN-T2JS]; COVINGTON & BURLING, LLP, RESPONDING TO A DATA SECURITY BREACH (2014), http://www.cov.com/files/FirmService/f47dd97b-0481-4692-a3bf-36039593171f/Presentation/ceFirmServiceDocument2/Responding_to_a_Data_Security_Breach.pdf [https://perma.cc/8L3B-WW9L]; Leo Notenboom, *Security: It's a Spectrum, Not a State*, ASKLEO (Sept. 6, 2014), https://askleo.com/security-its-a-spectrum-not-a-state/ [https://perma.cc/4LKC-STWY]; Bruce Schneier, *Lessons from the Sony Hack*, SCHNEIER.COM (Dec. 19, 2014), https://www.schneier.com/blog/archives/2014/12/lessons_from_th_4.html [https://perma.cc/Z4YG-B2UE].

151. Bambauer, *supra* note 150.

152. Ohm, *supra* note 2; Wu, *supra* note 10; Yakowitz, *supra* note 6.

A risk tolerant approach to releasing data will help move us past the debate over the perfection (or lack thereof) of anonymization.[153] Because process-based regimes like the current U.S. approach to data security are agnostic about ex post harms in favor of ex ante controls, they implicitly accept that a certain number of harms will slip through the cracks.[154] By focusing on process instead of output, data release policy can aim to raise the cost of reidentification and sensitive attribute disclosure to acceptable levels without having to ensure perfect anonymization. We explore what a nuanced, process-based data release policy might look like in Part III.

C. Data Release Policy Must Be More Than Deidentification

As discussed, much of the debate surrounding anonymization is focused on the technical means for transforming data or, more narrowly, deidentification.[155] NIST acknowledged the importance of data controls such as contracts prohibiting reidentification, but it explicitly described these controls as separate from the process of deidentification.[156] NIH is among the few federal agencies to rely on a tiered access approach that combines technical measures and data controls.

We argue that the data controls are just as important as deidentification in safely releasing useful data sets. In order to bridge the previously mentioned divide between technology and policy, we recommend including both deidentification techniques and controls on data flow as part of data release policy as well as query-based methods where appropriate. While this rhetorical move might seem slight, we take the more inclusive approach in order to better emphasize the importance of a holistic approach to releasing data. This holistic approach would include not just data flow controls but also organizational structure, education, and more careful deidentification rhetoric.

Sound data release policy requires an approach that utilizes the full

---

153. *See* Shapiro, *Separating the Baby from the Bathwater*, *supra* note 134; Shapiro, *Situating Anonymization*, *supra* note 134.

154. *See* JACK GOLDSMITH & TIM WU, WHO CONTROLS THE INTERNET? ILLUSIONS OF A BORDERLESS WORLD 67 (2006) (noting that internet law "need not be *completely* effective to be *adequately* effective. All the law aims to do is to raise the costs of the activity in order to limit that activity to acceptable levels" (emphasis in original)).

155. *See supra* Section I.A and text accompanying notes 10–12 (discussing various techniques for altering quasi-identifiers).

156. NIST REPORT, *supra* note 10, at 1.

spectrum of SDL techniques—direct access, dissemination-based access, and query-based access. Some techniques may be best suited for particular contexts or best used in combination with other techniques. There is a growing consensus among scholars in the deidentification debate that access controls are critical.[157] Yianni Lagos and Jules Polonetsky proposed that administrative safeguards like contracts can be leveraged for a "reasonably good de-identification standard" as opposed to "extremely strict de-identification measures," a viewpoint aligned with others in the field.[158] A few policymakers have even recognized the importance of data controls in shaping deidentification policy. As noted above, the FTC outlined what constituted "reasonably linkable" data that triggers privacy and data security obligations from companies.[159]

The FTC's approach to deidentification is promising. We join the

---

157. Ohm has endorsed regulations grounded in trust that facilitate data access to qualified investigators. Ohm, *supra* note 2, at 1767–68 ("Regulators should try to craft mechanisms for instilling or building upon trust in people or institutions . . . . 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."). Narayanan and Felten have emphasized the need for a diverse toolkit for deidentification, including contracts limiting reidentification. NARAYANAN & FELTEN, *supra* note 45, at 8 ("Data custodians face a choice between roughly three alternatives: sticking with the old habit of de-identification and hoping for the best; turning to emerging technologies like differential privacy that involve some trade-offs in utility and convenience; and using legal agreements to limit the flow and use of sensitive data."). Barth-Jones has also endorsed the contractual approach as part of deidentification policy. *See, e.g.*, Barth-Jones, *supra* note 21.

158. Yianni Lagos & Jules Polonetsky, *Public vs. Nonpublic Data: The Benefits of Administrative Controls*, 66 STAN. L. REV. ONLINE 103, 104 (2013), http://www.stanfordlawreview.org/sites/default/files/online/topics/66_StanLRevOnline_103_LagosPolonetsky.pdf [https://perma.cc/HX4F-YZ6N]. Omer Tene and Christopher Wolf asserted in a white paper for the Future of Privacy Forum that administrative safeguards and legal controls were critical in defining what constitutes "personal data." OMER TENE & CHRISTOPHER WOLF, FUTURE OF PRIVACY FORUM, THE DEFINITION OF PERSONAL DATA: SEEING THE COMPLETE SPECTRUM (2013), http://www.futureofprivacy.org/wp-content/uploads/FINAL-Future-of-Privacy-Forum-White-Paper-on-De-Id-January-201311.pdf [http://perma.cc/E6JB-HCX9]. Deven McGraw has proposed the use of data controls to make individuals and entities accountable for unauthorized reidentification. Deven McGraw, *Building Public Trust in Uses of Health Insurance Portability and Accountability Act*, 20 J. AM. MED. INFORMATICS ASS'N 29, 31 (2013) ("Accountability for unauthorized re-identification can be accomplished in the following two ways: (1) through legislation prohibiting recipients of de-identified data from unauthorized re-identification of the information; and (2) by requiring HIPAA-covered entities (and business associates) to obtain agreements with recipients of de-identified data that prohibit the information from being re-identified without authorization."). Peter Swire has asserted that organizational controls such as data separation within organizations and contractual prohibitions on reidentification are crucial but underappreciated aspects of deidentification. Peter Swire, Comments to the FCC on Broadband Consumer Privacy (Apr. 28, 2015), https://transition.fcc.gov/cgb/outreach/FCC-testimony-CPNI-broadband.pdf [https://perma.cc/E5XA-4SK6].

159. *See* FED. TRADE COMM'N, *supra* note 9, at iv, 20–21.

growing chorus of voices calling for an increased focus on data controls in the deidentification debate.[160] But rather than commit to one particular data control, such as contracts, qualified investigators, or enclaves, we argue that the full range of control options should be utilized in conjunction with data treatment techniques, organizational support, and mindful framing to establish a sound deidentification regime.

But if risk, access, and control are to become central in data release policy, then a harsh truth is revealed: many kinds of public releases of data must be curtailed. It is much more difficult to assess the risk of reidentification when those who share data lose control over it. There are simply too many factors that cannot be accounted for or even reliably estimated. Therefore, we argue that sound process-based policy minimizes or eliminates "release-and-forget" deidentification as an acceptable strategy. At the very least, the data release process should require DUAs from data recipients promising to refrain from reidentification, to keep an audit trail, and to perpetuate deidentification protections.

Of course, the release-and-forget model has its advantages, but with respect to deidentified data, the benefits of being free from data controls do not outweigh the cost of relinquishing control and protection. To begin with, release-and-forget deidentification fuels the paranoia surrounding anonymization. The best-known reidentification attacks all involve release-and-forget data sets.

Additionally, if properly drafted and executed, DUAs should not be overly burdensome for data recipients. Contracts are ubiquitous. Consumers and organizations enter into tens if not hundreds of complex, less-justifiable contracts every week in the form of End User License Agreements (EULAs), terms of service, and other standard-form contracts, to say nothing of the contemplated, bargained-for contracts for negotiated goods and services.

By contrast, DUAs governing the release of deidentified data can be workable. Privacy researcher Robert Gellman suggested that data recipients should agree not to attempt reidentification, take reasonable steps to keep related parties from reidentifying data, and keep potentially identifiable data confidential unless the recipient agrees to the same reidentification restrictions.[161] These terms represent a holistic approach

160. *See supra* Sections I.A.3–.4.

161. Robert Gellman, *The Deidentification Dilemma: A Legislative and Contractual Proposal*, 21 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 33, 51–52 (2010). Gellman also suggested that data recipients implement reasonable administrative, technical, and physical data safeguards and be transparent to others regarding all such data agreements the recipient is subject to. *Id*.

designed to mitigate the failures of technical deidentification through data treatment. Likewise, they reflect a "chain-link" data protection approach by imposing substantive protections, requiring that future recipients of data be similarly vetted and bound, and that the contractual chain will be perpetuated.[162] In addition, terms regarding record keeping, audit trails, and other transparency mechanisms could be added to ensure compliance.[163] Yakowitz suggested that obligations on the data recipient not to engage in reidentification could be backed by criminal penalties for violations.[164] Of course, any such statutory prohibition would need robust exemptions for security research into deidentification and related topics.[165]

But not every data use agreement need be equally robust. As previously mentioned, we envision an inverse ratio relationship between data treatment and data controls, whereby technical and legal controls can be adjusted according to context. Yet *some* form of data control seems necessary in most situations. Even many presumptively "open" data sets require assent to terms of use agreements.[166]

We envision deidentification policy that adopts a sort of inverse-ratio rule for data treatment and data controls. Controlling for other factors, the more rigorous and robust the data treatment, the less potent the data controls need to be. The more protective data controls become, the less thorough data treatment needs to be.[167]

Because sound deidentification is dependent upon many factors, companies should be allowed some flexibility in choosing which data controls are reasonable in a given context. However, as we will describe below, some data controls, like contractual prohibitions on reidentification, might be indispensable in all but the most benign circumstances.

---

162. *See* Woodrow Hartzog, *Chain-Link Confidentiality*, 46 GA. L. REV. 657, 659, 660–61 (2012) (advocating a "chain-link confidentiality" approach to protecting online privacy).

163. *See supra* note 82.

164. *See supra* note 82.

165. Gellman, *supra* note 161. Gellman's model bill would make it a felony to engage in "knowing and willful reidentification or attempted reidentification" and a major felony with the possibility of imprisonment where there is "intent to sell, transfer, or use personal information for commercial advantage, personal gain, or malicious harm." *Id*. at 53. Yakowitz also advocated criminalizing attempts at reidentification but only when "an adversary discloses the identity and a piece of non-public information to one other person who is not the data producer." Yakowitz, *supra* note 6, at 48–49. This approach seeks to avoid "unintentionally criminalizing disclosure-risk research." *Id*.

166. *See*, *e.g*., *The D4D Challenge Is a Great Success!*, *supra* note 120.

167. *See* IOM STUDY, *supra* note 10.

D. Seven Risk Factors

Perhaps the best way to build upon the FTC's framework is to identify the different risk vectors to be balanced in determining how protective a company must be when releasing a data set. There are at least seven variables to consider, many of which have been identified in reports by NIST and others[168]:

Volume of Data: The FTC's approach does not discriminate based upon the nature of the data. But the volume of the data can affect the risk of reidentification and sensitive attribute disclosure. Some large data sets have a high degree of unicity, which makes it easier to launch reidentification attacks.[169]

Sensitivity of the Data: Some information, like health and financial information, is more sensitive and thus more likely to be targeted by attackers. As Ohm argues in a recent paper, sensitive information is important because it is strongly connected to privacy harms affecting individuals.[170] It also lends itself to a threat modeling approach for assessing the risk of such harms.[171]

Recipient of the Data: There are at least three different types of recipients of data, each increasingly risky: (1) internal recipients, (2) trusted recipients, and (3) the general public. Internal recipients are in most respects the least problematic, though how "internal" is conceptualized is important. Trusted recipients are riskier, but should be an integral part of any data release policy. De Montjoye and his colleagues have argued that data sharing regimes should facilitate more sharing among trusted data recipients "with strong processes, data security, audit, and access control mechanisms in place. For example, trusted third parties at research universities might warrant access to richer, less anonymized data for research purposes and be relied on not to try to re-identify individuals or to use the data inappropriately."[172] There might exist several tiers of trusted recipients, with increasing protections tied to less-trustworthy recipients. Data sharing with

168. *See* ICO CODE, *supra* note 140; *supra* note 10.

169. *See* de Montjoye et al., *supra* note 40.

170. *See* Paul Ohm, *Sensitive Information*, 88 S. CAL. L. REV. 1125, 1125–28 (2015).

171. *Id*.

172. DE MONTJOYE ET AL., *supra* note 121, at 4–5.

recipients at the lowest tier would be treated as the equivalent of public release. Finally, as we discuss below, public releases should be seen as inherently problematic and require the greatest amount of protections, all other variables being equal.

One way to assign organizations to these categories is by evaluating their privacy programs. Does the organization collect and store data in a way that minimizes the risk of reidentification and sensitive attribute disclosure? Does it offer privacy training to its employees, segment the virtual and physical storage of data, implement company policies regarding deidentification, and set a tone within the organization regarding data minimization and anonymization as important privacy protections?

These structures are crucial not only to ensure that data treatment techniques and controls are consistently and correctly executed, but also to protect against the insider threat to deidentified data. Wu drew a distinction between "insider" or "outsider" threats. He wrote that "[p]rivacy 'insiders' are those whose relationship to a particular individual allows them to know significantly more about that individual than the general public does. Family and friends are examples."[173] Wu noted that co-workers and service providers at the corporate and employee levels could also be insiders, "for example, employees at a communications service provider, or workers at a health care facility."[174] Insider attacks present a range of different problems for deidentification. Wu noted, "[i]n security threat modeling, analysts regard insider attacks as 'exceedingly difficult to counter,' in part because of the 'trust relationship . . . that genuine insiders have.'"[175]

Use of the Data: Some uses of data are riskier, less necessary, or more problematic than others. Will the data be used for routine, administrative purposes like record keeping, website development, or customer service? Or will it be used for commercial or discriminatory purposes? Will certain uses of data create a motivation for attackers to attempt reidentification? Information that is to be used for more problematic purposes likely must be better protected given the potential harm and motivations by attackers to identify people or sensitive attributes. Some have also argued that protections should be lowered if

---

173. Wu, *supra* note 10, at 1154.

174. *Id*.

175. *Id*. (quoting SUSAN LANDAU, SURVEILLANCE OR SECURITY?: THE RISKS POSED BY NEW WIRETAPPING TECHNOLOGIES 162–63 (2010)).

the data is to be used for a significant public good or to help people avoid serious harm.[176]

Data Treatment Techniques: Risk varies according to the ways data is manipulated through the use of deidentification and SDL techniques to protect data subjects. Data values are suppressed, generalized, substituted, diluted with noise, and hashed to protect identities and sensitive attributes.[177] Sometimes entirely new data sets that do not map to actual individuals are synthetically created as safer surrogates than authentic data. Query-based systems provide another form of treatment, whereby only parts of a database are made available to recipients in response to specific queries. Such controls can leverage techniques like differential privacy to protect the identity and attributes of users.

Data Access Controls: Risk is also contingent upon the way data is released. When SDL and other access controls are utilized to limit who can access data and how they can access it, this lowers the risk of reidentification or sensitive data disclosure. Companies can choose to release data only to internal staff or trusted recipients, provided they contractually agree to protect the data and refrain from attempting reidentification. Recipient controls can be combined with distribution controls. Furthermore, they can make data available only via on-site terminals or secure portals.

Data Subject's Consent or Expectations: People are told that their data is often collected only for specific purposes. These representations are made in permission forms, privacy policies, marketing materials, orally, and as part of an app or website's design. Meaningful, properly obtained consent can mitigate the need to offer robust protections. Also, as we discuss below, in order to avoid being deceptive, protections should meet or exceed consumer expectations created by a company's statements or omissions.

E. Data Release Policy Should Embrace Industry Standards

In order to be effective and sustainable, data release policy must be nimble, which in turn requires a relative lack of specificity. The more

---

176. DE MONTJOYE ET AL., *supra* note 121, at 4 ("Special consideration should be given to cases where the data will be used for significant public good or to avoid serious harm to people.").

177. *See* NIST REPORT, *supra* note 10.

detailed data release law becomes, the quicker it becomes outdated. Laws are difficult to amend. The better alternative to regulatory specificity is to tether obligations of reasonable conduct to industry standards.

Industry standards are attractive for regulatory purposes because they are regularly updated. They are also, by definition, feasible and have the support of an industry's majority. The key to data security law in the U.S. is a reasonable adherence to industry standards.[178] This approach has kept data security standards fluid, negotiable based upon context and resources, and ascertainable to those responsible for securing data. Rather than looking to the law for specific data security practices to follow, data security professionals look to state-of-the-art standards from industry and international standards organizations and then reasonably follow along.[179]

This approach provides a good deal of breathing space to organizations where it is difficult to prescribe with precision the optimal protections in a given context. It also helps ensure that rules surrounding such a highly technical field as data security remain grounded in reality and up-to-date. For example, Vadhan and his colleagues have proposed that regulatory agencies maintain a safe harbor list of data-sharing mechanisms appropriate for different contexts that can be maintained and regularly updated with the input of experts and stakeholders.[180]

Deferring to industry standards is not without risk. Certain minimal protections for people must be ensured. Simply because a practice is standard does not ensure that it is sufficient. Thus, regulators must ensure a co-regulatory approach (like Vadhan's or otherwise) that helps shape minimum industry standards and steps in when industry standards

---

178. *See* Hartzog & Solove, *supra* note 149; Kristina Rozan, *How Do Industry Standards for Data Security Match Up with the FTC's Implied "Reasonable" Standards—and What Might This Mean for Liability Avoidance?*, PRIVACY ADVISOR (Nov. 25, 2014), https://privacyassociation.org/news/a/how-do-industry-standards-for-data-security-match-up-with-the-ftcs-implied-reasonable-standards-and-what-might-this-mean-for-liability-avoidance/ [https://perma.cc/YW6L-BKWB].

179. *See supra* note 177.

180. Vadhan et al., *supra* note 70. In particular, they propose that each entry in this list would:

> [S]pecify a class of data sources (e.g. electronic health records that do not include any genomic data), a class of data-sharing methods (e.g. HIPAA-style de-identification by the removal of certain fields, or interactive mechanisms that achieve a given level of differential privacy), a class of informed consent mechanisms, and a class of potential recipients. Together, these components of an entry specify a set of contexts in which a safe harbor would apply, and case-by-case IRB [Institutional Review Board] review could be avoided. In the long term, one can hope for this list to be sufficiently comprehensive so that the vast majority of research projects can proceed without IRB review of informational harms.

Id. at 7. We believe this proposal has much merit.

fail to deliver that minimum standard of care. Yet, generally speaking, deference to industry standards has proven workable if not fruitful in the field of data security.[181]

Data release policy should also be tethered to international data security standards, some of which already address deidentification and data release. There are at least five popular data security standards that have helped shaped policy, two of which (NIST 800-53[182] and ISO 27001[183]) enjoy widespread support.[184] There is substantial overlap between these standards as well.[185]

Some of these standards have begun to address deidentification and data release, though their guidance needs to become more specific. Appendix J of the popular NIST 800-53 standard simply identifies anonymization and deidentification as techniques that support the fair information principle of data minimization.[186] Even the specific publication on protecting the confidentiality on PII only includes a small Section on deidentifying and anonymizing information that provides little guidance to companies.[187]

Yet industry and international standards are on their way, as demonstrated by the NIST Draft Report and the UK's Information Commissioner's Office (ICO) report.[188] If developed correctly, standards will bring with them both a common vocabulary and consensus on process. Even though the NIST Draft Report has yet to offer advice on proper process, it is a remarkably concise and useful summary of the problem and articulation of common terms.

There are a number of other possible standards that could set the bar for deidentification policy. For example, the Article 29 Data Protection

---

181. *Id*.

182. KELLEY DEMPSEY ET AL., NIST COMPUT. SEC. DIV., SUMMARY OF NIST SP 800-53 REVISION 4, SECURITY AND PRIVACY CONTROLS FOR FEDERAL INFORMATION SYSTEMS AND ORGANIZATIONS (2014), http://csrc.nist.gov/publications/nistpubs/800-53-rev4/sp800-53r4_summary.pdf [https://perma.cc/MM6F-J23U].

183. INT'L ORG. FOR STANDARDIZATION, ISO/IEC 27001:2013 INFORMATION TECHNOLOGY—SECURITY TECHNIQUES—INFORMATION SECURITY MANAGEMENT SYSTEMS—REQUIREMENTS (2013), http://www.iso.org/iso/catalogue_detail?csnumber=54534 [https://perma.cc/5BYD-LL4Y].

184. Rozan, *supra* note 178.

185. *Id*.

186. DEMPSEY ET AL., *supra* note 182, at J-2, J-14.

187. ERIKA MCCALLISTER ET AL., NAT'L INST. OF STANDARDS & TECH., GUIDE TO PROTECTING THE CONFIDENTIALITY OF PERSONALLY IDENTIFIABLE INFORMATION: RECOMMENDATIONS OF THE NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY 4-3, 4-4 (2010), http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf [https://perma.cc/2QWW-2TW6].

188. *See* ICO CODE, *supra* note 140.

Working Party recently published an opinion laying out the strengths and weaknesses of the main anonymization techniques as well as the common mistakes related to their use.[189] While this opinion offers much useful guidance, it never quite resolves a tension in European data protection law between the legal implications of anonymization[190] and a reasonableness standard for determining whether a person is identifiable.[191]

Some of the most promising guidance capable of being standardized by industry is a 2012 anonymization code of practice issued by the United Kingdom's ICO.[192] The ICO Code is focused on identifying risks when anonymizing information and articulating specific practices to minimize them. Most importantly, the Code is risk tolerant and focused on process rather than output.[193] Thus, notwithstanding its use of the term anonymization, it is a good candidate for policymakers to borrow from when creating a process-based deidentification policy.

* * *

In this Part, we have outlined the three core aspects of a process-based approach to mitigating the risk of releasing data. Borrowing from data security, data release policy should be broader than just deidentification techniques. It should also incorporate SDL techniques like query-based access and other data controls to protect against many different kinds of threats. Finally, by fostering and relying upon industry standards similar to data security policy, data release policy can become

---

189. *See Opinion 05/2014 on Anonymisation Techniques by the Working Party on the Protection of Individuals with Regard to the Processing of Personal Data*, 0829/14/EN, WP 216 [hereinafter *Opinion on Anonymisation Techniques*], http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf [https://perma.cc/C46F-3GV9].

190. For example, *Opinion on Anonymisation Techniques*, *supra* note 189, states that "principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable," which amounts to a perfect anonymization requirement. *Id*. at 5 (quoting Council Directive 95/46/EC, Recital 26, 1995 O.J. (L 281) 31, 31 (EC) [hereinafter Directive 95/46/EC]).

191. In contrast, Directive 95/46/EC, *supra* note 190, states that account should be taken of all the "means likely reasonably to be used" to identify a person. *Opinion on Anonymisation Techniques*, supra note 189. Although the Working Party struggles to split the difference between these two competing conceptions, it achieves limited success. *See id*. at 8 (referring to an earlier opinion in which it "clarified that the 'means . . . reasonably to be used' test is suggested by the Directive as a criterion to be applied in order to assess whether the anonymisation process is sufficiently robust, i.e. whether identification has become 'reasonably' impossible"). But "reasonably impossible" is clearly a self-contradictory notion.

192. *See* ICO CODE, *supra* note 140.

193. The report avoids absolute framing and instead focuses on language like "mitigating," not eliminating, risk. *Id*. at 18.

more specific, flexible, and tethered to reality and the state of the art. In the next Part, we will discuss how process-based data release policy might work in practice.

III. IMPLEMENTING PROCESS-BASED DATA RELEASE POLICY

Let us recap what we have covered so far. In Part I, we reviewed the anonymization debate and stagnant policy. In Part II, we proposed that data release policy should be focused on the process of minimizing risk. Drawing from data security law, we developed a process-based data release policy as a holistic, contextual and risk tolerant approach. In this Part, we propose several legal reforms to safely release data.

Data release policy is not hopelessly broken. It regularly works quite well. However, many current laws and policies should be updated given the uncertainty surrounding reidentification and sensitive attribute risk. Policymakers could incorporate process-based data release rules without dramatic upheaval to relevant privacy regimes. Process-based data release can be implemented in increments and serve as an additional protective strategy as well as a replacement to output-based regimes in some contexts. In this Part, we review a few areas where the law could be changed to focus more on process rather than output or use more accurate rhetoric to better shape expectations.

A. From Output to Process

There are a number of deidentificaiton and data release laws that depend on outputs related to the data itself. For example, common conceptualizations of PII hinge upon whether an individual is or can be ultimately identified from a data set.[194] The EU Data Protection Directive includes personal data within its scope on similar grounds and excludes "data rendered anonymous in such a way that the data subject is no longer identifiable."[195] The HIPAA deidentification regime turns on whether data lacks certain attributes or whether an expert finds a threshold level of risk has been crossed with respect to the data set.

These regimes could be modified to focus on ensuring a process to protect information was followed, rather than looking to the state of the data itself. Like data security law, HIPAA could simply require the

194. *See* Paul M. Schwartz & Daniel J. Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 N.Y.U. L. REV. 1814, 1828–36 (2011).

195. Directive 95/46/EC, *supra* note 190.

implementation of "reasonable data release protections."

What does this mean? Again, the best place to turn for guidance is the law of data security. The FTC requires that companies collecting personal information provide "reasonable data security."[196] A combination of the FTC's complaints, its statement issued in conjunction with its fiftieth data security complaint, and a guide on data security reveals that there are four major components of "reasonable data security": (1) assessment of data and risk; (2) data minimization; (3) implementation of physical, technical, and administrative safeguards; and (4) development and implementation of a breach response plan.[197]

We propose that these four tenets of reasonable data security can be modified to establish a general requirement that businesses provide "reasonable data release protections." The tenets of reasonable process-based data release protections would look similar to those of data security:

1) Assess data to be shared and risk of disclosure;
2) Minimize data to be released;
3) Implement reasonable (proportional) deidentification and/or additional data control techniques as appropriate;
4) Develop a monitoring, accountability, and breach response plan.

These requirements would be informed by the nascent industry standards, including accepted deidentification and SDL techniques as well as a consideration of the seven risk vectors described above. This approach is context-sensitive and would allow companies to tailor their obligations to the risk.

Notions of reasonable, process-based data release protections could be implemented in various privacy regimes. The HIPAA Privacy Rule currently outlines two paths for deidentifying health data sets, the Safe Harbor method and expert determinations.[198] Both have long been subjects of criticism.[199] HIPAA could move closer to process-based data releases in several different ways. First, the Safe Harbor method could be modified to require technological, organizational, and contractual mechanisms for limiting access to deidentified data sets as well as deidentification. Additionally, experts might be asked to certify

196. *See, e.g.*, Press Release, Fed. Trade Comm'n, *supra* note 136.

197. *Id*. The FTC added specifics to these general tenets in its guide to data security for businesses with ten general rules of thumb. *Id*.

198. *See supra* note 115.

199. *See* McGraw, *supra* note 158.

processes along the lines described by El Emam and Malin[200] and Shapiro,[201] rather than assess risk. Companies seeking to be certified as HIPAA compliant would be asked to demonstrate that they have implemented a comprehensive data release program analogous to the comprehensive privacy and security programs articulated in FTC consent orders.[202] This would include performing a threat analysis, identifying mitigating controls, and documenting the methods and results of this analysis (as required by the expert determination method).[203] Although these approaches have their own drawbacks,[204] they would better incentivize robust data release protections and mitigate the inherent difficulty of assessing reidentification and sensitive attribute disclosure risk.

More generally and regardless of whether HIPAA applies, any company seeking to fortify data releases should implement procedures to minimize risk. Instead of mechanically removing a pre-set list of identifiers, threat modeling should be used to calculate risk as soundly and accurately as possible. These threat models would then guide companies toward the implementation of deidentification safeguards or use of other SDL methods, including direct access methods and query-based access methods such as differential privacy.

Using reasonable data release protections as a regulatory trigger would have several advantages over output-based risk thresholds. Companies would be incentivized to embrace the full spectrum of SDL methods and to combine deidentification techniques with access controls to protect data. Data release policy would create efficiencies by becoming aligned with data security law. A co-regulatory approach would drive the development of industry standards and safe-harbor lists, which would keep data release laws feasible and grounded. As discussed above, process-based approaches grounded in a reasonableness standard are nimble, contextual, and risk tolerant. Using risk analysis to inform process rather than ultimately determine regulatory application also

---

200. *See supra* note 10.

201. *See supra* note 153.

202. *See, e.g.*, Daniel Solove & Woodrow Hartzog, *Snapchat and FTC Privacy and Security Consent Orders*, LINKEDIN (May 12, 2014), https://www.linkedin.com/pulse/20140512053224-2259773-the-anatomy-of-an-ftc-privacy-and-data-security-consent-order [https://perma.cc/9EL2-LWUG].

203. For a related suggestion, see McGraw, *supra* note 158, at 32 (advocating that HHS explore certifying or accrediting entities that regularly deidentify data or evaluate reidentification risk).

204. *Id*. (discussing the prospects for eliminating or modifying deidentification methods under the Privacy Rule).

provides some wiggle room for an inexact exercise.

The FTC could extend data release policy to all data sets via section 5 of the FTC Act.[205] In addition to its proposed jurisdictional test, the agency could regulate unreasonably protected releases of data sets as an unfair trade practice. If process-based data release protection proves workable, it could even be used in a new baseline privacy law that discouraged release-and-forget anonymization, encouraged data use agreements, and regulated both data release procedures as well as reidentification attempts.[206]

The transition to a risk-based process also begins to resolve several lingering problems in the contemporary anonymization debate. First, it mitigates Ohm's "broken promises" objection by treating deidentification not as a jurisdictional test in privacy law but rather as one of several possible approaches to sharing data using the full gamut of SDL methods. As previously noted, following a risk-based approach relaxes certain privacy requirements but not others.[207] It follows that no one has to make "breakable promises" regarding (perfect) anonymity. Rather, organizations will offer appropriate assurances based on reasonable security measures.

Second, it suggests a possible workaround to the auxiliary information problem. Ohm correctly noted that solving this problem via regulation quickly turns into a game of "whack-a-mole."[208] While it may be impossible to limit auxiliary information, the use of trusted recipients and direct access methods to deprive most adversaries of access to protected data sets is much less challenging. This may seem cumbersome and may discourage some researchers from engaging in important work and yet it reflects current thinking about the costs and benefits of open data.[209]

B. Deceptive Deidentification

The way companies and the media talk about deidentified data matters, and data holders regularly play fast and loose with the concept of anonymity. The terms "anonymous" and "anonymization" simply over-promise. They create expectations of near-perfection and lull

---

205. It could do so either as an unfair or deceptive trade practice, depending on context. *See* Solove & Hartzog, *Common Law of Privacy*, *supra* note 137.

206. *See* Gellman, *supra* note 161.

207. *See supra* Section II.D.

208. Ohm, *supra* note 2, at 1742.

209. *See supra* Section I.A.4.

people into a false sense of security. It is no wonder that the media keep proclaiming the death of anonymity—we keep expecting the impossible.

In previous work, one of us has noted:

> The resolution of a debate often hinges on how the problem being debated is presented. In communication, sociology, psychology, and related disciplines, this method of issue presentation is known as framing. Framing theory holds that even small changes in the presentation of an issue or event can produce significant changes of opinion. For example, people are more willing to tolerate rallies by controversial hate groups when such rallies are framed as free speech issues, rather than disruptions of the public order.[210]

So it goes for the deidentification debate. In the same way that there is no such thing as perfect data security, there is no such thing as perfect deidentification. Our policy and rhetoric should reflect this fact.

Ohm makes a similar point, suggesting that we "abolish the word anonymize" and replace it with a word like "scrub" that "conjures effort, not achievement."[211] We agree with Ohm that rhetoric is a key aspect of this debate, and the terms "anonymous" and "anonymization" should be used very sparingly and with due attention to precision. They are counterproductive because they create unrealistic consumer expectations. We view terms such as "pseudonymous" as often more technically accurate.[212] However, we disagree with Ohm's suggestion that we also abandon the term "deidentification," which we find a useful umbrella term to incorporate data transformation as well as data controls. Rather than jettisoning deidentificaiton, we should clarify its meaning as a broad, general term referring to *the process by which data custodians treat and control data to make it harder for users of the data to determine the identities of the data subjects*.

While "anonymization" has far too much baggage to be useful anymore, "deidentification" is a more responsible and useful way to refer to the process by which a data custodian uses a combination of data

---

210. Woodrow Hartzog, *The Fight to Frame Privacy*, 111 MICH. L. REV. 1021, 1021 (2013) (citing Thomas E. Nelson et al., *Media Framing of a Civil Liberties Conflict and Its Effect on Tolerance*, 91 AM. POL. SCI. REV. 567 (1997)).

211. Ohm, supra note 2, at 1744.

212. *See, e.g.*, *Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individual with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection), Compromise Amendments on Articles 30–91*, at Arts. 33, 81, COM (2012) 11 (Oct. 17, 2013), http://www.computerundrecht.de/comp_am_art_30-91en.pdf [http://perma.cc/GEM8-SL2A] (distinguishing personal data from pseudonyms).

alteration and removal techniques and sharing and mining controls to make it harder or more unlikely for users of the data to determine the identities of the data subjects.

In previous research, one of us has developed the concept of "obscurity" as the preferable way of conceptualizing notions of privacy in shared information.[213] When information is obscure, that is, unlikely to be found or understood, it is, to a certain degree, safe. NIST correctly notes the efficacy of obscured, deidentified data.[214] But even "anonymized" data (which NIST sees as ensuring that previously identified data cannot be reidentified) exists along a continuum of obscurity. "Anonymization" just makes it harder, but not impossible, to find out someone's identity. NIST's obscurity framing for deidentified data is thus the most accurate, even for "anonymized" information.

Getting the framing for the deidentification debate right is critical to setting people's expectations regarding how their data will be protected. If companies do not promise perfection and people do not expect it, then deidentification policy will be more likely to reflect reality. Risk tolerant rules become politically palatable and consumers can better sense the extent to which their disclosures make them vulnerable.

There is great benefit to improving the accuracy of consumer expectations. Consider an "anonymous social network"[215] app called Whisper, which was the subject of a series of articles by *The Guardian* in fall 2014, asserting that the app might be less than anonymous.[216] Whisper has sold itself as the "safest place" on the internet.[217] However, its terms of use have evolved to tell a more realistic and less bulletproof

---

213. Evan Selinger & Woodrow Hartzog, *Obscurity and Privacy*, *in* ROUTLEDGE COMPANION TO PHILOSOPHY OF TECHNOLOGY (Joseph Pitt & Ashley Shew eds., forthcoming 2016); Woodrow Hartzog & Frederic Stutzman, *The Case for Online Obscurity*, 101 CALIF. L. REV. 1 (2013); Woodrow Hartzog & Frederic Stutzman, *Obscurity by Design*, 88 WASH. L. REV. 385 (2013).

214. *See* NIST REPORT, *supra* note 10.

215. *Whisper*, GOOGLE PLAY, https://play.google.com/store/apps/details?id=sh.whisper&hl=en [https://perma.cc/7MY7-44AT] (last visited Apr. 24, 2016).

216. Paul Lewis & Dominic Rushe, *Revealed: How Whisper App Tracks 'Anonymous' Users*, THE GUARDIAN (Oct. 16, 2014), http://www.theguardian.com/world/2014/oct/16/-sp-revealed-whisper-app-tracking-users [https://perma.cc/NQ6E-FGAU]. *But see* Corrections and Clarifications Column Editor, *Whisper—A Clarification*, THE GUARDIAN (Mar. 11, 2015), http://www.theguardian.com/
news/2015/mar/11/corrections-and-clarifications [https://perma.cc/T4LQ-8R4H].

217. Dana Goodyear, *Open Secrets*, NEW YORKER (Dec. 8, 2014), http://www.newyorker.com/magazine/2014/12/08/open-secrets-5 [https://perma.cc/L2MD-KXD4]; Stephen Loeb, *Heyward: Whisper Is "the Safest Place on the Internet,"* VATORNEWS (Oct. 4, 2014), http://vator.tv/news/2014-10-04-heyward-whisper-is-the-safest-place-on-the-internet [https://perma.cc/3HT7-MKH2].

story about anonymity.[218] At one point, Whisper's privacy policy stated:

> We collect very little information that could be used to identify you personally. . . . Our goal is to provide you with a tool that allows you to express yourself while remaining anonymous to the community. However, please keep in mind that your whispers will be publicly viewable, so if you want to preserve your anonymity you should not include any personal information in your whispers. . . . [E]ven if you do not include personal information in your whispers, your use of the Services *may still allow others, over time, to make a determination as to your identity* based on the content of your whispers as well as your general location. . . . [W]e encourage you to be careful and avoid including details that may be used by others to identify you.[219]

Note the explicit emphasis on the fallibility of anonymity. Such accuracy is desirable, though it may accomplish little for consumers who do not and cannot be expected to read the fine print.[220] Users are much more likely to read the app's marketing description as "anonymous" and proceed accordingly. Such practices breed deception and confusion and frustrate sound deidentification policy.

Yet the rhetoric of anonymity remains effective for one simple purpose: convincing people to trust data recipients with their personal information. To be anonymous online is to be safe. Companies that promise anonymity gain the benefit of people's trust even when there is a notable risk of reidentification from poorly anonymized data sets.

The FTC should continue to use its authority under section 5 of the FTC Act to ensure that promises of anonymity are not deceptive. Put simply, companies cannot guarantee anonymity. However, companies can promise that they have assessed the risk of harm from the use and release of data and have implemented appropriate protections according to industry standards. Tempering the language of anonymization and deidentification will help appropriately set consumer expectations.

---

218. Paul Lewis & Dominic Rushe, *Whisper App Has Published Its New Terms of Service and Privacy Policy*, THE GUARDIAN (Oct. 16, 2014), http://www.theguardian.com/world/2014/oct/16/-sp-whisper-privacy-policy-terms-of-service [https://perma.cc/M5NR-4AYZ].

219. *Privacy Policy*, WHISPER, https://whisper.sh/privacy (last visited Apr. 30, 2015) (emphasis added).

220. Woodrow Hartzog, *The Problems and Promise with Terms of Use as the Chaperone of the Social Web*, CONCURRING OPINIONS (June 11, 2013), http://concurringopinions.com/archives/2013/06/the-problems-and-promise-with-terms-of-use-as-the-chaperone-of-the-social-web.html [https://perma.cc/PR84-ZWGJ].

Promising process rather than perfection and output will also force companies to actively embrace data release protections rather than passively benefit from speculative risk calculations.

Truthful deidentification rhetoric can also be encouraged in ethical engineering principles and in business-to-business contracts and communications. Data release policy should focus on education efforts for people, companies, and, critically, the media. Like privacy, the rumors of deidentification's death have been greatly exaggerated. Yet media coverage of successful reidentification attempts remains a critical component of understanding the limits of deidentification and the larger world of protections for the disclosure of data. A better dialogue between journalists, scholars, and policymakers would benefit all.

C. Data Release Policy and PII

As noted above, PII typically defines the scope and boundaries of privacy law.[221] Although there are several different approaches to defining PII,[222] the key point is that many privacy laws associate privacy harm with PII and leave non-PII unregulated.[223] Thus, many organizations devise a compliance strategy premised on this distinction and take steps to transform PII into non-PII with the goal of limiting or eliminating their obligations under applicable privacy statutes and regulations.

By now the problems associated with this deidentification strategy are familiar. First, a lack of relevant deidentification standards means that many organizations do a poor job "anonymizing" data, yet claim its unregulated status. Second, while deidentification reduces risk, it never achieves perfection. Thus, even organizations that follow best practices may not be wholly successful in protecting the privacy of data subjects. Finally, release-and-forget methods exacerbate these problems by creating publicly available data sets over which organizations are incapable of imposing controls.

In a path-breaking article, Schwartz and Solove argue that despite these and other problems, privacy law should not abandon the concept of PII but rather develop a new model using a standards-based approach.[224] According to their revitalized standard, PII should be conceptualized in

---

221. *See supra* Section I.A.2.

222. *See* Schwartz & Solove, *supra* note 194, at 1828–36 (describing three main approaches).

223. *Id*.

224. *Id*. at 1870–72.

terms of a risk-based continuum, with three categories: information that refers to (1) an identified person, (2) an identifiable person, and (3) a non-identifiable person.[225] A person is identified when her identity is "ascertained" or he or she can be "distinguished" from a group; a person is identifiable when specific identification is "not a significantly probable event" (i.e., the risk is low to moderate); while non-identifiable information carries only a "remote" risk of identification.[226] Moreover, Schwartz and Solove argue that the applicability of the FIPPs turns on these categories. Thus, while all of the FIPPs generally should apply to information that refers to an identified person, only some of the FIPPs—data quality, transparency, and security (but not notice, access, and correction rights)—should apply to identifiable data.[227]

This reconceptualization of PII complements our risk-based approach to deidentification as proposed above. The tripartite approach requires an ex ante assessment of whether a given data set should be treated as falling into category 1 (and accorded protection under the full FIPPs), category 2 (partial FIPPs apply) or category 3 (no FIPPs apply). According to Schwartz and Solove, this assessment must look at "the means likely to be used by parties with current or probable access to the information, as well as the additional data upon which they can draw" as well as additional contextual factors such as "the lifetime for which information is to be stored, the likelihood of future development of relevant technology, and parties' incentives to link identifiable data to a specific person."[228] We agree. While Schwartz and Solove might be overly optimistic about the availability of "practical tools" for assessing the risk of identification,[229] their approach—with one important modification—presents a clear path for overcoming the regulatory problems noted above. The modification is to treat public release of data sets as an overriding factor in assigning data sets to categories 1, 2, or 3.

Under this modified version of PII 2.0 (call it PII 2.1), regulators should create a default presumption that publicly released data sets are identifiable, even if the data custodian deidentifies the data set by removing common identifiers. This presumption could be overcome by determining that the data custodian meets process-based data release

225. *Id*. at 1877–79.

226. *Id*.

227. *Id*. at 1879–83. The authors are silent on the remaining FIPPs.

228. *Id*. at 1878.

229. *Id*. at 1879. They do not factor in the auxiliary information problem or respond to criticisms based on the lack of mathematical rigor in assessing the risk of reidentification. *Id*.

requirements as described below. Obviously, this would require changes to the HIPAA Privacy Rule.

Our proposal will operate similarly to the FTC's deidentification framework, which acts as a threshold PII test as well. Recall that the FTC uses a "reasonable linkability" standard for determining the scope of its privacy framework.[230] While "reasonable linkability" seems output-based, it is mainly a process requirement. Obtain contracts, promise to protect the data, and scrub the data to a sufficient degree, and the information is excluded from the framework. While the scrubbing of data is output-based, it need not be. Our proposal for process-based data release policy could be similarly repurposed, such that proper data release protections meeting a reasonableness standard and/or utilizing a data-sharing mechanism on a safe-harbor list in the appropriate context would exempt companies from additional privacy restrictions because the risk of harm to data subjects has likely been sufficiently mitigated.

CONCLUSION

The debate about the failure of anonymization illustrates what we will call the first law of privacy policy: there is no silver bullet. Neither technologists nor policymakers alone can protect us. But we have been missing the big picture. We should think of reidentification as a data release problem. Sound data release policy requires a careful equilibrium on multiple fronts: law and technology, data treatment and data controls, privacy and utility.

It is important to keep data release policy and the surrounding debate from becoming parochial and separated from other parts of privacy and data security law. Hacking, surveillance, and inducement to breach confidentiality are all alternatives to reidentification attacks. Additionally, identification and sensitive attribute disclosure are just a few of many modern privacy problems, alongside occasionally related but technically separate issues like discrimination and manipulation.

Yet if data release policy becomes too ambitious, it will become intractable and ineffective. The variables affecting the likelihood of reidentification and sensitive attribute disclosure are vexing enough. Thus, we have argued the locus of data release policy should be the process of mitigating these risks.

Process-based data release protections are the best way to develop policy in the wake of the perceived and real failures of anonymization.

---

230. *See* FED. TRADE COMM'N, *supra* note 9, at 20.

Such an approach is driven by policies balancing protection with data utility. It is holistic and integrated. Perhaps most importantly, it is flexible and can evolve alongside the relevant science and the lessons of implementation.

The vigorous and productive debate over how to protect the identity of data subjects has matured. Even though there are sharp disagreements, there is more consensus than at first appears. The next step is to develop policy from our lessons learned. Anonymization is dead. Long live the safe release of data.

APPENDIX

Anonynmization and Risk: A Glossary of Terms

Auxiliary information (background information; outside information): information outside of a data set. Auxiliary information can be used in an attempt to identify individuals in a data set. [Page 711.]

Data use agreement (DUA): a contract that conditions access to, and use of, a data set on agreement to specific terms. A DUA may include such terms as refraining from reidentifying subjects in the data set, maintaining an audit trail, and perpetuating deidentification protections. [Page 739–40.]

Deidentificaiton: the process by which data custodians remove the association between identifying data and the data subject. [Page 754.]

Direct access: a form of statistical disclosure limitation (SDL) that encompasses both licensed data, which allows users who click-through the applicable licensing terms to perform any data query and receive full results, and authorized access to research data centers, which also allows any query but only returns vetted results. Direct access imposes the fewest restrictions on data but limits data access to qualified investigators who must agree to licensing terms or execute a DUA, which may also stipulate security measures and prohibit redistribution of the data sets or attempts to reidentify or contact data subjects. [Page 717–18.]

Direct identifier: data that directly identifies a unique individual, such as name or social security number. [Page 710.]

Dissemination-based access: a form of SDL that refers to the practice of publicly releasing reduced, altered, or synthetic data (i.e., hypothetical data that have similar characteristics to the real data). A researcher using dissemination-based access can view full results to any query in a data set. The data custodian applies various techniques to construct the transformed data set before publicly releasing it. This form of access combines public release of data with masking of data sets by methods including generalization and suppression. Deidentification is a form of dissemination-based access. [Page 718.]

K-Anonymity: a process that requires the data administrator to ensure that, given what the adversary already knows, the adversary does not

reduce the set of potential target records to fewer than *k* records in the released data. A weakness in this approach is that *k*-anonymity assumes that only a small number of attributes may be used as quasi-identifiers for purposes of a linkages attack. Several researchers have taken issue with this claim. [Pages 712–13.]

Linkage attack: an attempt to reidentify individuals in a data set by linking the deidentified data set with additional information. The term "attack" is borrowed from computer security literature, hence the individual carrying out the attack is called an "adversary." The additional information is called "outside," "auxiliary," or "background" information. [Page 711 & 734.]

Open data: information that is accessible to everyone, machine readable, and offered online at zero cost, and has no limits on reuse and redistribution. [Page 719–20.]

Personally identifiable information (PII): includes a range of information that can be used to identify an individual; some kinds of information can more readily identify an individual than others. Privacy laws focus on the collection, use, and disclosure of PII, and privacy harm depends in part on whether disclosed information is PII. However, as Schwartz and Solove have shown, there is no uniform definition of PII in United States privacy law. [Page 755–55.]

Pseudonymization: a form of deidentification that uses a replacement value (like a pseudonym or number) for the identity of data subjects. [Pages 711, 753–54.]

Quasi-identifier: data that does not itself identify a specific individual but can be aggregated and linked with information in other data sets to identify data subjects. Examples include birthday, ZIP code, and gender. [Page 711–12.]

Query-based access: a form of SDL that allows users to interact with the data by posing queries, typically over a secure internet connection. There are several sub-categories of query-based access. (1) Remote analysis servers allow researchers to analyze confidential data without ever seeing the underlying data, although both the queries they can pose and the results they can obtain may be subject to limitations. (2) Differential privacy is a set of techniques whereby query results are altered, often by adding noise, so that released information does not

reveal any person's data with certainty. In query-based access, data analysis uses statistical queries without direct access to underlying data sets. [Page 718–19.]

Reidentification: the process of attempting to determine the identities of the data subjects whose identifiers have been removed from the data set. [Page 710–11.]

Release and forget: a term used by Paul Ohm to describe when a data administrator releases deidentified records without restrictions or tracking what happens to the records after release. [Page 725.]

Statistical disclosure limitation (SDL): comprises the principles and techniques that researchers have developed for disseminating official statistics and other data for research purposes while protecting the privacy and confidentiality of data subjects. Satkartar Kinney divides SDL into three major forms: direct access, dissemination-based access, and query-based access. [Page 717.]

Unicity: a concept used to quantify how much outside information one would need, on average, to reidentify a specific and known user in a simply anonymized data set. The higher a data set's unicity, the easier it is to reidentify data subjects in the anonymized data. Mobile phone metadata is highly unique and therefore can be reidentified using little outside information. The same is roughly true of credit card data. Unicity was coined by Yves-Alexandre de Montjoye et al. [Page 714.]