TECHNOLOGY POLICY INSTITUTE
Studying the Global Information Economy

The Illusory Privacy Problem in *Sorrell v. IMS Health*

May 2011

Jane Yakowitz and Daniel Barth-Jones

The Illusory Privacy Problem in *Sorrell v. IMS Health*

Jane Yakowitz[1] and Daniel Barth-Jones[2]

[1] Visiting Assistant Professor of Law, Brooklyn Law School. B.S., Yale College; J.D., Yale Law School.
[2] Assistant Professor of Clinical Epidemiology, Mailman School of Public Health, Columbia University. M.P.H. University of Michigan; Ph.D. University of Michigan

I. Introduction

Those in the habit of looking for privacy invasions can find them everywhere. This phenomenon is on display in the recent news coverage of *Sorrell v. IMS Health Inc.*, a case currently under review by the Supreme Court. The litigation challenges a Vermont law that would limit the dissemination and use of prescription drug data for the purposes of marketing to physicians by pharmaceutical companies. The prescription data at issue identify the prescribing physician and pharmacy, but provide only limited detail about the patients (for example, the patient's age in years and gender).[3] Nevertheless, privacy organizations like the Electronic Frontier Foundation (EFF) and the Electronic Privacy Information Center (EPIC) have filed amici curiae briefs sounding distress alarms for patient privacy.[4] A recent New York Times article describes the case as one that puts the privacy interests of "little people" against the formidable powers of "Big Data."[5] The fear is that, in the information age, data subjects could be re-identified using the vast amount of auxiliary information available about each of us in commercial databases and on the internet.

Such fears have already motivated the Federal Trade Commission to abandon the distinction between personally identifiable and anonymized data in their Privacy By Design framework.[6] If the Department of Health and Human Services (HHS) were to follow suit, the result would be nothing short of a disaster for the public, since de-identified health data are the workhorse driving numerous health care systems improvements and medical research activities.

Luckily, we do not actually face a grim choice between privacy and public health. This short article describes the small but growing literature on de-anonymization—the ability to re-identify a subject in anonymized research data. When viewed rigorously, the evidence that our medical secrets are at risk of discovery and abuse is scant.

II. How Attack Algorithms Work

All de-anonymization attack algorithms are variants of one basic model. An adversary attempts to link subjects in a de-identified database to identifiable data on the entire relevant population ("population records"). The adversary links the two databases using

---

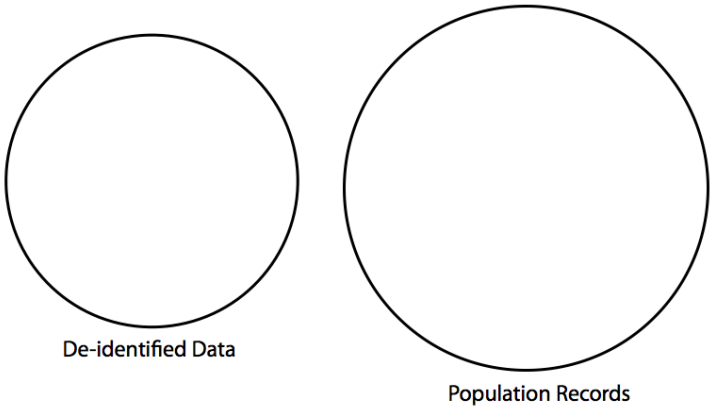
[3] The prescription data must comply with the de-identification requirements in the Health Insurance Portability and Accountability Act ("HIPAA").

[4] These briefs and all other filings in the case are available at http://www.scotusblog.com/case-files/cases/sorrell-v-ims-health-inc/.

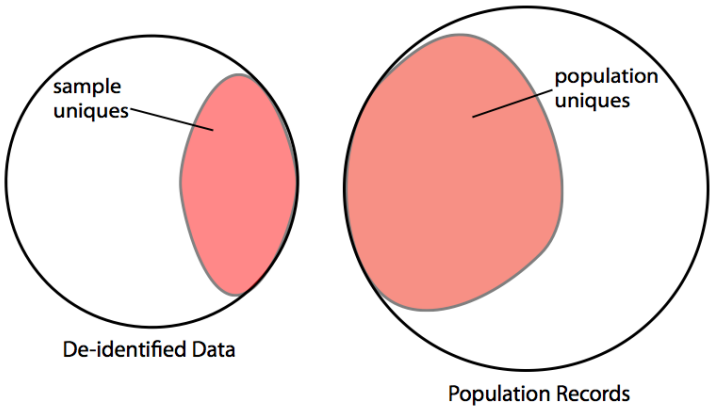
[5] Natasha Singer, "Data Privacy, Put to the Test," *New York Times*, April 30, 2011. The article downplays the fact that Vermont's law does not actually prohibit the broad dissemination of de-identified prescription data. The law restricts dissemination only for one particular use (detailing), leaving untouched the ability to sell the same data for any other purpose. The article does not address the issues *Sorrell v. IMS Health* raises about public health, cost controls, and free speech.

[6] *Protecting Consumer Privacy in an Era of Rapid Change: A Proposed Framework*, Report prepared by the Federal Trade Commission, 43 and 51-52 (2010) (available at www.ftc/gov/os/2010/12/101201privacyreport.pdf)

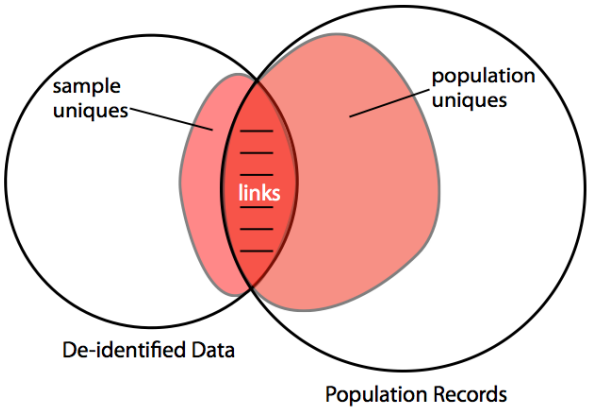
variables that the two datasets have in common. Let's call these variables "indirect identifiers." To visualize the attack, suppose the two spheres in this diagram represent the indirect identifiers in the de-identified database and the population records. Initially, these databases have no linkages:



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



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



Only a subset of the sample uniques and population uniques will be linkable because some of the sample uniques might not actually be unique in the general population, and some of the population uniques might not be in the sample of the de-identified data.[7]

The classic example of a successful matching attack was demonstrated by Latanya Sweeney, who combined de-identified Massachusetts hospital data with identifiable voter registration records to re-identify Governor Weld's medical records.[8] Because the hospital data at that time, before the passage of HIPAA, included granular detail on the patients (5-digit zip code, full birth date, and gender) many patients were unique in the hospital data and the population records.

Today, disclosure risk researchers and privacy scholars agree that this sort of "trivial de-identification" of records—the removal of only direct identifiers like names, ssns, and addresses only—is insufficient. Subjects can too easily be identified through a combination of high resolution indirect identifiers. Thus, HIPAA, like other federal privacy statutes, requires data to remove not only the obvious direct identifiers but *any* information that can be used alone or in combination with other information to identify an individual subject.[9]

---

[7] More sophisticated techniques will make matches not based on strong exact linkages but on the closeness of the matching variables and the greater distance between the best match and the second-best match. This allows an attack algorithm to make matches under more realistic conditions where databases contain measurement error, but it nevertheless requires that the adversary have access to more-or-less complete information on the general population from which the de-identified data was sampled. These methods are described more thoroughly by Josep Domingo-Ferrer et al., *Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure Risk*, Statistics and Computing 343 (2003) (available at http://arnetminer.org/viewpub.do?pid=1156009#Reference).

[8] *See* Latanya Sweeney, *Computational Disclosure Control: A Primer on Data Privacy Protection* (thesis draft of January 8, 2001) (available at http://www.eff.org/deeplinks/2009/09/what-information-personally-identifiable).

[9] HIPAA, 45 C.F.R. §164.514(b)(2)(ii). Alternatively, the disclosing entity must use the most advanced statistical methods to ensure that the risks of re-identification are "very small." 45 C.F.R. §164.514(b)(1).

While there is broad agreement on the rejection of trivial de-identification, privacy experts disagree on the efficacy of current best practices. The amici briefs of EFF and EPIC lump all anonymization techniques together and then use examples of trivial de-identification to argue that *all* de-identified data put patient privacy at risk. This is not accurate.

Data properly de-identified under the requirements of HIPAA are quite robust against re-identification attacks. In contrast to the high resolution health data that allowed Dr. Sweeney to re-identify Governor Weld fifteen years ago, Dr. Sweeney estimates that HIPAA-compliant data reporting a patient's gender, year of birth (rather than full birth date), and 3-digit zip code (rather than 5-digit) produces a re-identification risk of only 0.04 percent.[10] That is, only four people in 10,000 have a unique combination of gender, age in years, and 3-digit zip code. Moreover, because there is rarely a comprehensive population register that contains complete and accurate information on the variables that might be linked, the process for re-identifying even seemingly vulnerable data subjects is not that simple. Voter registration data, for example, cannot describe members of the population who are not registered to vote, so an adversary will over-estimate uniqueness if he does not account for this. Even if there were such a comprehensive population register, both sets of data—the de-identified data and the identifiable data—would be likely to contain errors and discrepancies, which undermine the ability to link with confidence. This is not the sort of difficulty that can be easily overcome with increased processing speed or shrewd new attack techniques; rather, it is the natural protection afforded by the inherently messy nature of data and of people.

Of the recent studies that attempt to match de-identified research data to identifiable auxiliary information, the majority use the source data that produced the de-identified data to create auxiliary information that a putative adversary might use in a re-identification attack.[11] In other words, the study authors created a perfectly clean population register which allowed re-identification to be performed error-free. Other studies (described in more detail in Part III) make over-reaching assumptions in order to facilitate the matching process without accounting for the likelihood of false matches.

Only one recent de-anonymization study was conducted under the conditions that replicate what a real adversary would face. The study, undertaken for the HHS Office of the National Coordinator for Health Information Technology ("ONC"), sheds some light on the likelihood of a successful attack on properly de-identified data.[12] The ONC put together a team of statistical experts to assess whether data properly de-identified under

---

For a discussion of the same savings clause language in other federal privacy statutes, *see* Jane Yakowitz, *Tragedy of the Data Commons*, 125 Harv. J. L. & Tech. __ (forthcoming, 2011) (available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1789749.)

[10] National Committee on Vital and Health Statistics Report to the Secretary of the U.S. Department of Health and Human Services, "Enhanced Protections for Uses of Health Data: A Stewardship Framework for "Secondary Uses" of Electronically Collected and Transmitted Health Data," at 36 (December 19, 2007)(*available at* www.ncvhs.hhs.gov/071221lt.pdf).

[11] *See*, e.g., Josep Domingo-Ferrer et al., *supra* note 5.

[12] Deborah Lafky, *The Safe Harbor Method of De-Identification*, ONC Presentation, October 9, 2009 (*available at* www.ehcca.com/presentations/HIPAAWest4/lafky_2.pdf).

HIPAA can be combined with readily available outside data to re-identify patients. The study was performed under realistic conditions and verified the re-identifications—important aspects missing from many other studies of this sort. The team began with a set of approximately 15,000 patient records that had been de-identified in accordance with HIPAA. Next they sought to match the de-identified records with identifiable records in a commercially available data repository and conducted manual searches through external sources (InfoUSA database) to determine whether any of the records in the identified commercial data would align with anyone in the de-identified data set. The team determined that it was able to accurately re-identify two of the fifteen thousand individuals, for a match rate of 0.013%. In other words, the risk, even after extraordinary effort, was very small.[13]

The data at issue in *Sorrell v. IMS Health* are also governed by HIPAA, and are also required to have very small re-identification risks. Consider the risk imposed when prescriber data includes only the patient's age and gender: Even if the prescription data allow us to know that "a 50-year-old woman who lives in Central Vermont; has prescriptions filled in Monpelier; [and] is a patient of Dr. Jones in Montpelier… regularly takes an antidepressant," an example put forward in the dissenting opinion from the 2nd Circuit, an adversary would not have the means to determine which 50-year-old woman it is. Knowing only age and gender for certain, the adversary is not able to link this woman to a population unique.[14]

III. The Privacy Organizations Cite To Flawed Evidence of Re-Identification Risk

The amici brief of EPIC cites to Dr. Sweeney's study of pre-HIPAA hospital discharge data to support the claim that prescriber data is risky for patients, arguing that "almost all medical patients can be re-identified using the zip code, date of birth, and gender categories in de-identified data."[15] The comparison is misleading. While 63 percent of the U.S. population have a unique combination of gender, 5-digit zip code, and full date of birth, only 0.2 percent are unique when the exact date of birth is replaced by age in years.[16] The percentage drops further still when geographic information is blurred, as is the case with the prescription data.

---

[13] These findings are consistent with an earlier study that examined re-identification attacks under realistic conditions. *See* Walter Muller, et al., *Disclosure Risk for Microdata Stemming from Official Statistics*, 46 Statistica Neerlandica 69 (1992).

[14] Latanya Sweeney has studied the re-identification of prescription data under the assumption that the pharmacy's zip code is identical to the patient's zip code. Even with this strong assumption, which will surely be right sometimes but will just as surely be wrong some times, the prescription data could rarely be linked to population registers—2.3 percent of the time in New York, 1.24 percent of the time in Arizona, and less than .04 percent of the time in the other seven states studied. Latanya Sweeney, *Patient Identifiability in Pharmaceutical Marketing Data* (*available at* dataprivacylab.org/projects/identifiability/pharma1.pdf).

[15] Amici curiae brief for EPIC at 26.

[16] Philippe Golle, *Revisiting the Uniqueness of Simple Deomgraphics in the US Population*, Proceedings of the 5th Association for Computing Machinery Workshop on Privacy in Electronic Society, at 2 (2006)(*available at* http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.4147&rep=rep1&type=pdf.)

A de-anonymization study of Netflix data has greatly influenced the privacy policy debate, but the applicability of this study is much more limited than the privacy advocates suggest. In 2006, Netflix released an anonymized dataset to the public consisting of movie reviews of 500,000 of its members. Arvind Narayanan and Vitaly Shmatikov used information from user reviews on the Internet Movie Database (IMDb) to re-identify subjects in the Netflix Prize dataset.[17] The authors take advantage of the fact that each of our movie-viewing histories makes a unique fingerprint. The idea is, if an adversary knows enough of the movies that a subject in the Netflix database has watched, then that person can be identified since he or she will be unique in the Netflix sample.

Narayanan and Shmatikov did not have a "population register" of the movies viewed by every American, so their attack faced a hurdle: how do we know that somebody who looks unique in the Netflix database based on reviewing, say, a string of 6 movies is unique not just in the Netflix data, but also among the American population? To overcome this hurdle, Narayanan and Shmatikov do something clever: they use the Netflix data themselves to predict the likelihood that somebody else in the general population might have watched the same set of movies. Since the Netflix data are a more-or-less random sample, the authors could use a measure of "uniqueness" for a set of movie reviews in Netflix to predict the likelihood that somebody outside the sample would have seen the same set of movies. That way, they do not need to have access a population register. Instead, they are able to predict with confidence when a unique in the Netflix database would also be a unique among all Americans.

However, the Netflix dataset does not report all the movies that the subjects in the Netflix database have watched. Indeed, it does not even report all of the movies that the subjects have watched from Netflix. Instead, the Netflix database reports only the movies that the subjects chose to review online. Thus, a string of six movies that looks unique in the Netflix database *might not be unique even among the subjects in the Netflix database.* Unless an adversary accounts for the incompleteness of the Netflix database, the algorithm will underestimate the likelihood that a match made between IMDb and Netflix is a false match.

Narayanan and Shmatikov performed a proof of concept attack and were able to match two IMDb profiles (out of a few dozen) to the Netflix database, but the authors do not report having verified the matches.[18] In any event, EPIC's claim that the researchers "successfully identified 99 percent of the people in the Netflix database" is patently false.[19]

---

[17] Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)*, in proc. of *29th IEEE Symposium on Security and Privacy*, Oakland, CA, May 2008, pp. 111-125. IEEE Computer Society, 2008.

[18] One of the two matches uses review dates as well as movie titles, which suggests that the match is very likely to be correct. The other match uses movie titles without matching the dates.

[19] Misreadings of the Netflix study are common in the media and among litigants. *See* Ryan Singel, *Netflix Spilled Your Brokeback Mountain Secret, Lawsuit Claims*, WIRED December 17, 2009, http://www.wired.com/threatlevel/2009/12/netflix-privacy-lawsuit/; Seth Schoen, *What Information is "Personally Identifiable?"*, Electronic Freedom Foundation (available at

IV. Re-identification Risk Has Never Materialized

One fact often gets lost in the discussion of data disclosure risks: as of today, there is no evidence that re-identification by a true adversary (somebody other than a researcher or journalist interested in the efficacy of privacy protections) has actually happened. This is not particularly surprising when one considers the skill and effort required to launch a de-anonymization attack on a properly anonymized dataset. Indeed, the EFF notes that the mathematics involved in the Netflix attack algorithm is complex.[20] Moreover, de-anonymization attacks do not scale well because of the challenges of determining the characteristics of the general population. Each attack must be customized to the particular de-identified database and to the population as it existed at the time of the data collection. This is likely to be feasible only for small populations under unusual conditions.

The EPIC and EFF amici briefs drift from re-identification risk to the harms of commercial data accretion in a way that might seem to imply re-identification is already a common practice employed by commercial data aggregators. In fact commercial databases collect and combine *identifiable* information. The collection and dissemination of this sort of commercial data is at the center of a live debate—the Stearns and Kerry/McCain consumer privacy bills directly address these issues. Whatever their merits, these issues are wholly separate from the alleged privacy risks of properly de-identified data.

V. Unintended Consequences

If the Supreme Court or U.S. policymakers were to abandon the distinction between personally identifiable information and anonymized data, the public would lose one of its strongest policy and social justice tools. The anonymized prescription data in *Sorrell* have been used to monitor national trends in prescription drug usage and to answer a variety of important public health questions including, rather ironically, whether various marketing practices have an impact on prescriber behavior.[21] Moreover, it's not just prescription data but *all* anonymized research data that hang in the balance of the debate over re-identification risk—housing data used to evaluate racial segregation, education data used to monitor teacher effectiveness, and data used to examine trends in U.S. income distributions. The data subject consent model advocated by EFF would result in great selection bias that would frustrate research that advances not only public health and

---

http://www.eff.org/deeplinks/2009/09/what-information-personally-identifiable); Electronic Privacy Information Center's "Re-identification" page (available at http://epic.org/privacy/reidentification/). Several recent lawsuits have made the argument that there is no longer a tenable difference between anonymized information and personally identifiable information. *See*, e.g., Gaos v. Google Inc., Case No. 2010cv04809 (Cal., filed 2010); Doe v. Netflix, C09 05903 (filed December 17, 2009) (petition available at http://www.wired.com/images_blogs/threatlevel/2009/12/doe-v-netflix.pdf); Elinor Mills, *AOL Sued Over Web Search Data Release*, cnet News (September 25, 2006)(available at http://news.cnet.com/8301-10784_3-6119218-7.html)

[20] Brief of Electronic Frontier Foundation in Support of Petitioners, at 10

[21] *See* Brief of Academic Research Scientists as Amici Curiae in Support of Respondents.

social policy, but data privacy improvements as well. This would be devastating for the public.[22] And for this loss we would get only the illusion of privacy protection in return.

[22] EFF's amici brief states that direct patient consent would be a preferable model for the dissemination of de-identified data. EFF brief, at 7. In contrast to the patient consent model, Marc Rodwin argues that medical data constitute a public good and should be required to be supplied in de-identified form to the federal government and made available to researchers. Marc A. Rodwin, *Patient Data: Property, Privacy, & the Public Interest*, 36 Am. J. L. & Med. 586 (2010).