Jane Yakowitz.
UCLA School of Law

Anonymization Is Not Dead

As an empirical researcher who relies on anonymous data, I am concerned that the Federal Trade Commission is considering regulations that would inhibit the dissemination of anonymous data. The "Technology and Privacy" panelists at the recent FTC Roundtable uniformly agreed that the concepts of anonymization and PII are dead. This view, which suggests that extreme and drastic adjustments are needed, is motivated by a few recent studies that have become detached from the law and the science. I will use the Netflix Prize database as the compelling example. Other databases that have been held up as proof that anonymization does not work (such as the AOL database and the Massachusetts medical dataset from the 1990s) are not relevant because these databases were never properly anonymized in the first place. The former included last names and neighborhood identifiers[1] and the latter included the obvious overlapping quasi-identifiers of zip codes, age, and gender.[2]

The Netflix Illustration

a. Privacy Is Not Breached When Re-Identification Requires Special Information

The Netflix Prize database is at the heart of Narayanan's and Shmatikov's research on de-anonymizing sparse datasets.[3] The authors illustrated how their attack algorithms could be used to re-identify Netflix users if an adversary has some information about the target. To situate this work in the current legal framework, it's important to understand that there's a discrepancy between the type of privacy that is protected by law (via various statutes and the common law) and the breaches that Narayanan and Shmatikov study. Narayanan and Shmatikov examine how auxiliary information learned through any source, even at the water cooler, could be used to identify a target. But special information gleaned through private channels is not part of the legal calculus of Personally Identifiable Information. If public policy had embraced this expansive definition of privacy—that privacy is breached if somebody in the database could be reidentified by anybody else using special information— dissemination of data would never have been possible. Instead, privacy law in its various forms requires data producers to watch out for traceable characteristics that are, or foreseeably could be, in

---

[1] Michael Barbaro & Tom Zeller Jr., *A Face Is Exposed for AOL Searcher No. 4417749*, N.Y. TIMES, Aug. 9, 2006.
[2] *Recommendations to Identify and Combat Privacy Problems in the Commonwealth Before the H. Select Comm. On Information Security* (Penn. 2005) (statement of Latanya Sweeney)
[3] 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.

the public domain.[4] These are the quasi-identifiers that can lead to re-identification, even by a stranger.

Before the Information Age, the main sources of public information about private individuals were found in collections of public records disseminated by third parties. The internet has changed this; now everybody is their own self-publisher. The law has not resolved what to do with information that is nonpublic for the majority of people, but is voluntarily broadcast publicly by some people (e.g. IMDb users who rate movies using their real names). Narayanan and Shmatikov and other privacy advocates have assumed that these characteristics must be treated as quasi-identifiers, but this treatment is much too simplistic to be the foundation of sound policy. If I blog about a hospital visit, should my action render California's entire public hospital admissions database (relied on by epidemiologists and health policy advocates) violative of privacy law? Do we really want the bounds of information flow to be determined by the behavior of the most extroverted among us? We should take great care in reframing privacy expectations so broadly.

b. The Harms of the Netflix Prize Data Are Illusory

The distress calls of privacy advocates in response to the Netflix database have little proportionality to the plausible harms. Even for the subgroup of data subjects that have publicized some of their movie reviews in other fora (like IMDb),[5] the potential for harm caused by the Netflix database is very limited.

Narayanan and Shmatikov have made clear that matches between IMDb and the Netflix database work best when the movies reviewed on IMDb are less popular films.[6] The authors identified two (out of twelve) IMDb users in the Netflix dataset, but while they describe the movies rated in the Netflix database in vivid detail, they provide no information on the movies that the targets had freely chosen to rate publicly. This information is crucial for understanding the marginal risks associated with the Netflix database because the inferences that some privacy advocates are drawing from the Netflix ratings—that they reveal political affiliation or sexual orientation or, as the complaint for a recent lawsuit against Netflix alleges, "personal struggles with issues such as domestic violence, adultery, alcoholism, or substance abuse"[7]— can be drawn just as easily from the set of movies that the target had publicly rated in the first place. Stated another way, if the proverbial adversary already knows 5 or 6 movies that the target has watched, *that*

---

[4] For example, the Federal Family Education Privacy and Rights Act ("FERPA") defines PII as "information that, alone or in combination, is linked or linkable to a specific student that would allow a reasonable person in the school community, **who does not have personal knowledge of the relevant circumstances**, to identify the student with reasonable certainty." 34 CFR 99.3 (emphasis added). Likewise, "At a minimum, each statistical agency must assure that the risk of disclosure from the released data when combined with **other relevant publicly available data** is very low." Federal Committee on Statistical Methodology, *Statistical Policy Working Paper 22 (Second version, 2005)*, December 2005, at 3.

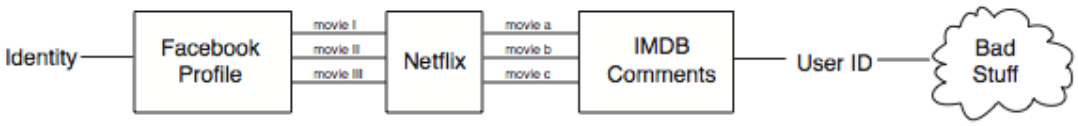
[5] In analyzing marginal risks, it would be helpful to know approximately how many of the subjects in the Netflix dataset have commented on the IMDb site using traceable usernames. Though IMDB has 17 million user ids (according to its Wikipedia entry), the majority post comments using non-traceable usernames. My quick and rough estimation (using the comments on the first couple pages of Avatar) is that only 1/3 of all users have traceable usernames.

[6] Narayanan & Shmatikov, *supra* note 3, at 7.

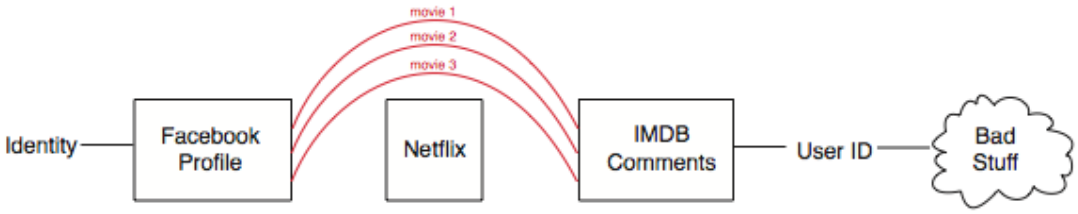
[7] Doe v. Netflix, [cite] (petition available at http://www.wired.com/images_blogs/threatlevel/2009/12/doe-v-netflix.pdf)

knowledge can go a long way toward pigeonholing and making assumptions (often wrong) about the target.[8]

Another concern is that the Netflix Prize Database might form a link, or an inner join, of a long series of information that connects identities on one end to truly sensitive information on the other. For example, the identity of a target in the Netflix database might be learned by matching to a Facebook profile, and then the Netflix data could be used to find his (up until now) anonymous username on IMDb. That formerly anonymous username can then be traced all over the internet.



This too overstates the marginal risks. If the target is listing movies on facebook and also reviewing movies on IMDB, these movies are likely to overlap.



In other words, the Netflix Prize dataset can be taken out of the stream altogether, and the target will still be at similar risk.[9] The inner join concept might apply to other databases,

---

[8] Privacy policy should not aspire to regulate these wrong-headed inferences; plenty of straight people enjoyed "Brokeback Mountain", and plenty of liberals dislike Michael Moore. But even if movie reviews are windows to the soul, the marginal information gained by re-identifying somebody in the Netflix dataset is likely to be small.

[9] Sophisticated followers of the computer science literature might notice this oversimplifies how the Netflix database is used to match to other resources. Narayanan's and Shmatikov's algorithm uses all information in the Netflix database not only to match records, but also to insure against the possibility of a false match. It's true that my diagrams oversimplify the process because without the Netflix dataset, assumptions would have to be made about the likelihood of a false positive. However, this highlights a tacit simplification that Narayanan and Shmatikov make in their attack models: They assume that the Netflix ratings are a complete set of movies viewed by each subject in the data. This is problematic. Even the most enthusiastic Netflix rater has seen far more movies than they bother to rate in Netflix. While the authors' models account for sample incompleteness, they do not adequately account for attribute incompleteness. The set of movies a person has seen and might comment on is much larger than the set of movies he has rated in Netflix. And likewise, far more people—even Netflix users in the prize database—have watched a given movie than have rated it. Thus, while the true movie-viewing behavior and preferences are no doubt very sparse, they are much less sparse than the Netflix database. Narayanan's and Shmatikov's algorithms are susceptible to false negatives and, most importantly, false positives, since the algorithm uses the Netflix dataset and not "true life" to measure the probability that two people might have viewed the same set of movies and enjoyed them the same amount. In more technical terms, the Netflix dataset D' is a subset of the real (nonexistent) complete database of movie viewing, dates, and preferences D_{TRUE} such that for all individuals i, $r_i \in r_{TRUEi}$ but r_i is not equal to r_{TRUEi} Then $Sim(r_1, r_2)$ is less than or equal to Sim_{TRUE} (r_{TRUE1}, r_{TRUE2}). It follows that D' can be (ε, δ)-sparse with respect to Sim even if $Pr(Sim_{TRUE}$ (r_{TRUEi}, r'_{TRUEi})$> \varepsilon$ for all r_{TRUEi} not= r'_{TRUEi}) $> \delta$. This is not fully corrected by the authors' allowance for one or two auxiliary attributes to be completely wrong, because (a) the similarity scores are still wholly dependent on the Netflix

but the conditions are so particular that this is no more than a theoretical risk at the moment.[10]

Anonymous Data is Useful

At the same time that the computer science literature has become detached from the legal definitions of privacy, the legal literature has misunderstood the findings of computer scientists. Paul Ohm has asserted that if data is useful to researchers, it is also, by definition, re-identifiable, and this assertion has been repeated in the national media.[11] This statement is false on its face. A database with just one quasi-identifying variable (e.g. gender) tied to non-public information (such as pharmaceutical purchases) can be tremendously valuable for a *specific* research question (e.g. "Do women purchase drugs in proportion to the national rates of diagnosis?") without the remotest possibility of revealing identities. Ohm and the media outlets were thrown way off because the technical studies use a definition of data-mining utility that encompassed all possible research questions that could be probed by the original database.[12] So, for example, if race and geographic indicators are removed from the database, the utility for *all possible* research questions plummets even though its utility for the specific research question I've suggested stays intact. For specific research questions, utility and anonymity can and often do coexist.

Privacy Concerns About Anonymous Data are Driven by Unjustified Fear and Confusion

Most databases—even large ones—can be properly anonymized even in our information age. Popular arguments that sound scientific and are leveraged to increase our anxiety about anonymous data don't hold up with deeper inspection. Take, for example, a comment made by Peter Eckersley of the Electronic Frontier Foundation at the 2nd Privacy roundtable. Peter said that every person on earth can be identified through 33 bits of information. He and others have arrived at this fact because each bit has two values (0 and 1), and 2 to the 33rd power is well over the population of earth—in fact it comes to about 8.6 billion.

The fact that we can all be uniquely described using 33 pieces of information sounds startling until we consider the conditions that must be met for this to be so. First, those thirty-three bits must each split the population in half. Other than gender, geographic

---

database rather than on the true likelihoods of overlap, luring a hacker into thinking that a false positive is less likely than it really is, and (b) in any event, there's no reason to think that the target bothered to rate up to 75% of the movies he has seen and might discuss with friends on Netflix without external evidence that this is so. These criticisms are inapplicable if the auxiliary information is about movies that the attacker *knows* the target would have rated in Netflix.

[10] To use an inner join for de-anonymization one must be certain (or at least quite sure) that the target is a member of all three information sets. Databases rarely cover the same populations since data producers have long been wary of overlapping disclosures on the same sample population. This is discussed in some detail in *Statistical Policy Working Paper 22*, *supra* note 4, at 82.

[11] Ryan Singel, *Netflix Spilled Your Brokeback Mountain Secret, Lawsuit Claims*, WIRED December 17, 2009, http://www.wired.com/threatlevel/2009/12/netflix-privacy-lawsuit/

[12] Justin Brickell & Vitaly Shmatikov, *The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing*, in proc. of *14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, Las Vegas, NV, August 2008, pp. 70-78.

divisions, and age divisions, it's hard to think of simple attributes that do this. Also, these attributes must all be independent from every other attribute. If gender is included, education level cannot be included (or if it is, you'll need more bits to overcome the covariance between those two variables.) It's possible to devise such a network of bits, but these wouldn't be the sorts of attributes included in anonymous datasets. Finally, all of the attributes have to be traceable somehow to the individual in order to matter from a privacy perspective. It does no good for an adversary to know that his target has a deformed molar, and drinks his tea with milk.

Of course, this entire exercise is quite remote to privacy law since a database will violate *somebody's* privacy long before it violates everybody's. The general detachment of the enterprise from common sense is the relevant point. It generates unnecessary public distrust. So far as I am aware, no anonymous dataset has been reverse engineered for unscrupulous reasons, and indirect evidence of deceitful practices is notably lacking. The Federal Trade Commission's own statistics show a declining incidence of identity theft.[13] Until anonymization techniques have failed us in some appreciable way, barriers to data dissemination should be implemented only with caution and deliberate care.

Abandoning the Concept of Anonymization Will Cause Significant Harm to the Public

If data anonymity is presumed not to exist, the future of public use datasets and all of the social utility that flows from them will be thrown into question. The uses of anonymous data are not limited to behavioral marketing researchers and academics. Nearly every recent public policy debate has benefited from mass dissemination of anonymous data. For example, anonymous data provided by the Federal Financial Institutions Examination Council has informed the subprime home mortgage debate, and eventually led to the passage of legislation prohibiting certain predatory lending practices. Similarly, research performed by health economists and epidemiologists using Medicare and Medicaid claims data are now central to discussions about health care reform. When data can be shared freely, it creates a research synergy that cannot be imitated through restricted data and license agreements. The Economist Magazine recognized the unmatchable public value of freely accessible data and crowdsourcing in its recent article *Of Governments and Geeks*.[14] A determination by the Federal Trade Commission that every piece of data is PII will strike a great blow to the data commons, and to innovations in the field of Information Technology. Please do not overlook these important factors in your attempts to strike the right balance for the digital age. I urge the FTC to consult with disclosure risk analysis experts[15] before disposing of the concept of data anonymity.

---

[13] See Thomas M. Lenard and Paul H. Rubin, In Defense of Data, Technology Policy Institute (2009) (available at ssrn.com/abstract=1407731)

[14] *Of Governments and Geeks*, The Economist, February 6, 2010. *See also* Chris Soghoian, *AOL, Netflix and the End of Open Access to Research Data*, CNET November 30, 2007, http://news.cnet.com/8301-13739_3-9826608-46.html.

[15] For example, The University of Michigan's Interuniversity Consortium for Political and Social Research (ICPSR) would be a great resource. ICPSR has archived half a million datafiles, and the vast majority are anonymous public-use files with no restrictions on access. If anonymous data has led to re-identification abuse, ICPSR would know.