No silver bullet: De-identification still doesn't work

Arvind Narayanan
arvindn@cs.princeton.edu

Edward W. Felten
felten@cs.princeton.edu

July 9, 2014

Paul Ohm's 2009 article "Broken Promises of Privacy" spurred a debate in legal and policy circles on the appropriate response to computer science research on re-identification.[1] In this debate, the empirical research has often been misunderstood or misrepresented. A new report by Ann Cavoukian and Daniel Castro is full of such inaccuracies, despite its claims of "setting the record straight."[2]

We point out eight of our most serious points of disagreement with Cavoukian and Castro. The thrust of our arguments is that (i) there is no evidence that de-identification works either in theory or in practice[3] and (ii) attempts to quantify its efficacy are unscientific and promote a false sense of security by assuming unrealistic, artificially constrained models of what an adversary might do.

1. There is no known effective method to anonymize location data, and no evidence that it's meaningfully achievable.

A 2013 study by de Montjoye *et al*. showed that 95% of mobility traces are uniquely identifiable given four random spatio-temporal points.[4] Cavoukian and Castro downplay the privacy impact of this study on the grounds that the authors didn't actually re-identify anyone, and that obtaining four spatio-temporal locations about individuals is hard.

We disagree strongly. First, many users reveal just such information on social networks. Second, Cavoukian and Castro ignore another finding of the study, namely that over 50% of users are uniquely identifiable from just *two* randomly chosen points. The study notes that these two points are likely to correspond to the individual's home and work locations. The uniqueness of home/work pairs is corroborated by other re-identification studies.[5]

[1] Paul Ohm, *Broken promises of privacy: Responding to the surprising failure of anonymization*, UCLA L. Rev., 57, 1701 (2009).

[2] Ann Cavoukian & Daniel Castro, *Big Data and Innovation, Setting the Record Straight: De-identification Does Work* (2014), available at http://www2.itif.org/2014-big-data-deidentification.pdf

[3] At the risk of being pedantic, when we say that de-identification doesn't work we mean that it isn't effective at resisting adversarial attempts at re-identification.

[4] Yves-Alexandre de Montjoye et al., *Unique in the Crowd: The privacy bounds of human mobility,* Scientific Reports 3 (2013).

[5] Hui Zang & Jean Bolot, *Anonymization of location data does not work: A large-scale measurement study,* in Proc. 17th Intl. Conf. on Mobile Computing and Networking 145-156 (2011); Philippe Golle & Kurt Partridge, *On the anonymity of home/work location pairs*, Pervasive Computing 390-397 (2009).

You don't have to be an expert to understand how to identify a person given their home and work locations. People who know you will probably know where you live and work; and people who don't know you can buy that information from a data broker. Either way, they can find your record in a database that includes home/work location pairs.

Let's be clear about why the authors of the study didn't actually re-identify anyone: because they didn't set out to. The study addressed the issue of uniqueness of mobility patterns, which is a scientific question about human behavior. On the other hand, the percentage of individuals that are re-identifiable is not an inherent property; it is determined by the datasets that are available at a particular point in time to a particular adversary.

Cavoukian and Castro admit that "there is no known standard for de-identifying mobility data" and "it is admittedly very difficult to de-identify mobility traces, while maintaining a sufficient level of data quality necessary for most secondary purposes." Indeed, a key finding of the de Montjoye et al. study is that the main technique one might hope to use — making the data more coarse-grained — has only a minimal impact on uniqueness.

Although they chose this as their leading example, Cavoukian and Castro don't suggest how one would go about de-identifying such a data set; they don't point to any literature asserting that it can be de-identified; and they don't even claim that it is de-identifiable in principle.

2. Computing re-identification probabilities based on proof-of-concept demonstrations is silly.

Turning to the Netflix Prize re-identification study,[6] Cavoukian and Castro say: "the researchers re-identified only two out of 480,189 Netflix users, or 0.0004 per cent of users, with confidence."

This is an unfortunate misrepresentation of the results considering that the Netflix paper explicitly warns against this: "Our results should thus be viewed as a proof of concept. They do not imply anything about the percentage of IMDb users who can be identified in the Netflix Prize dataset."

Cavoukian and Castro seem to fundamentally miss the point of proof-of-concept demonstrations. By analogy, if someone made a video showing that a particular car security system could be hacked, it would be an error to claim that there is nothing to worry about because only one out of 1,000,000 such cars had been compromised.

The IMDb re-identification was not even the most important result of the Netflix study. The study shows in detail that if someone knows just a little bit about the movie preferences of a user in the Netflix dataset (say, from Facebook or a water-cooler conversation), there's an upwards of

[6] Arvind Narayanan & Vitaly Shmatikov, *Robust de-anonymization of large sparse datasets*, in Proc. 2008 IEEE Symp. on Security and Privacy 111-125 (2008).

80% chance of identifying that user's record in the dataset. Predictably, Cavoukian and Castro ignore all this.

Again, Cavoukian and Castro don't suggest or cite any method for de-identifying the Netflix dataset or even claim that it is de-identifiable in principle.

3. Cavoukian and Castro ignore many realistic threats by focusing narrowly on a particular model of re-identification.

Cavoukian and Castro have an implicit model of re-identification in mind that drives their viewpoint. They don't explicitly state it, but it is important to make it clear. First, they limit themselves to re-identification studies that use only free, *already existing*, *publicly available* auxiliary data. They mostly ignore the possibility of re-identification by a spouse, friend, nosey neighbor, or investigator based on specific knowledge about the victim, as well as a data-broker applying re-identification based on their existing datasets to enrich their dossiers.[7]

Second, they consider only *large-scale* re-identification that has a *high probability* of identifying each person, as opposed to, say, deanonymization of a targeted individual by a rival or political opponent, where the analyst may follow up on leads generated by re-identification, or opportunistic re-identification by an adversary who will take advantage of whatever portion of the population they can re-identify.

We invite you to judge for yourself whether it is worth worrying about attacks that can re-identify targeted individuals, or that can re-identify (say) only 1% of the U.S. population — over 3 million people.

Re-identification of Governor Weld's Health Record

A good example of the authors' narrow, unrealistic threat model is their discussion of Latanya Sweeney's famous demonstration that she could re-identify the medical record of the then-governor of Massachusetts, William Weld. Sweeney started with a released medical database that included each patient's gender, date of birth and home ZIP code, then used a public dataset of registered voters to get the governor's gender, date of birth, and ZIP code, to extract the governor's record. She confirmed the result by comparing the medical record to publicly known facts about the governor.

In a long paragraph on pp. 7-8, Cavoukian and Castro make a series of incorrect claims about Sweeney's study. First, they claim that Sweeney's method had a "fatal flaw" because Sweeney "assume[d] that the population register used to re-identify individuals is a complete and accurate representation of the true population." In fact, Sweeney made no such assumption, nor did she need to. All she needed was the governor's gender, date of birth, and ZIP code, which were

[7] The authors claim that data brokers' databases "are often incomplete, making it difficult to positively identify someone with a high degree of confidence." This is cold comfort to a person who is re-identified because they do appear in the database. And it doesn't consider that a realistic adversary often can just buy access to another database if the first one doesn't meet their needs.

accurate in the voter database and in any case could have been learned from another source if, for some reason, the governor was not registered to vote. (For example, a minute spent on Wikipedia will reveal that current Massachusetts Governor Deval Patrick, who is male, was born on July 31, 1956 and lives in ZIP code 02186.)

Cavoukian and Castro claim further that "there was zero risk for individuals not appearing in the voter database." This also is incorrect, because such individuals would be at risk of re-identification by anyone who could learn their gender, date of birth, and ZIP code, which are not exactly state secrets. For example, if the governor had not been in the voter database, his record could still have been re-identified based on public information.

Finally, Cavoukain and Castro claim that "such attacks are 'no better than the flip of a coin' if there is only one other person with the same [gender, date of birth, and ZIP code]." This is not true either. The records at issue contained more than just gender, date of birth, and ZIP code: they also contained detailed medical records for the individual. There might be two people whose identifiers match a record, but if the record shows treatment for a medical condition with visible symptoms, these will match (or not) a targeted individual, letting an analyst who knows or can see the individual more information about whether they match the rest of the record.

The subtext in the authors' discussion of Sweeney's research is that it would somehow be cheating for an attacker to take any steps beyond what Sweeney did. But Sweeney didn't stop when she did because she couldn't think of more steps to try. She stopped because she had already shown that she could re-identify the governor's record. Cavoukian and Castro's strawman attacker might just copy Sweeney's steps exactly and then stop. But a real attacker would carry out the extra easy steps needed to re-identify the record(s) they wanted.

4. Cavoukian and Castro concede that de-identification is inadequate for high-dimensional data. But nowadays most interesting datasets are high-dimensional.

High-dimensional data consists of numerous data points about each individual, enough that every individual's record is likely to be unique, and not even similar to other records. Cavoukian and Castro admit that: "In the case of high-dimensional data, additional arrangements may need to be pursued, such as making the data available to researchers only under tightly restricted legal agreements." Of course, restrictive legal agreements are not a form of de-identification. Rather, they are a necessary protection in cases where de-identification is unavailable or is likely to fail.

But what Cavoukian and Castro don't seem to recognize is that high-dimensional data is now the norm, not the exception. Low-dimensional datasets may have been common in the past because of the difficulty of generating a large number of observations about a single individual, but these days it is rare for useful, interesting datasets to be low-dimensional.

In the Netflix Prize dataset, the average customer has 213 movie ratings and dates associated with them — that's 426 dimensions. The cell phone location study mentioned above used a dataset that recorded the user's location every two hours — that's hundreds of dimensions per month of data. Social network data is also high-dimensional: the median Facebook user in 2011 had about 100 friends — that's at least 100 dimensions.[8] Genetic data — the kind you get by spitting in a tube for $99 — is one-million-dimensional.

Medical records are arguably high-dimensional, depending on how many data points about each individual are recorded. If we include those, then every single dataset discussed by Cavoukian and Castro is high-dimensional.

5. Penetrate-and-patch is not an option.

Security professionals rightly criticize the penetrate-and-patch approach to security, which focuses on remediating past attacks rather than designing systems from the ground up to resist future attacks. Systems built on a penetrate-and-patch principle tend to fail repeatedly, not only because attackers are always discovering new tricks, but also because such systems are built on a foundation of sand.

Yet Cavoukian and Castro advocate a penetrate-and-patch mindset, in which we ask whether a de-identification method can resist certain past attacks, rather than insisting on affirmative evidence that the method cannot leak information regardless of what the attacker does.

Penetrate-and-patch is a poor approach to securing software, but it is even worse for de-identification. When a security bug in software is found, a patch that fixes the bug can be distributed to end users, who will eagerly install it in order to protect their own systems. But if a dataset that has been distributed is found to be re-identifiable, there is no way to patch it. Anyone who already has the dataset will be able to re-identify it, and we cannot force them to replace their re-identifiable version with a patched version that is strictly less useful to them. When patching isn't an option, penetrate-and-patch becomes simply penetrate.

An example is a recently released dataset of New York City taxi logs, which provided information about roughly 173 million New York taxi trips, including "anonymized" identifiers for the driver and the cab. When the driver and cab identities were re-identified, the leak could not be undone.[9]

Could the NYC taxi dataset have been de-identified reliably? Nobody knows. We do know that the driver and cab identifiers were "anonymized" using a simple hash function, which is known to be a flawed approach[10] (though some non-experts still advocate it). If we were stuck in a penetrate-and-patch mindset, we might fall into the trap of thinking the dataset would be safe to release if we just improved the handling of driver and cab IDs by replacing the hash function with something better. But it's not enough to prevent one specific re-identification method. This

[8] Johan Ugander, Brian Karrer, Lars Backstrom, & Cameron Marlow. *The anatomy of the Facebook social graph*. arXiv Preprint arXiv:1111.4503 (2011).

[9] Vijay Pandurangan, *On Taxis and Rainbows: Lessons from NYC's improperly anonymized taxi logs* (2014), available at https://medium.com/@vijayp/of-taxis-and-rainbows-f6bc289679a1

[10] Edward Felten, *Does Hashing Make Data "Anonymous"?*, Tech @FTC blog (2012), available at http://techatftc.wordpress.com/2012/04/22/does-hashing-make-data-anonymous/

is still a large, complex, high-dimensional dataset that probably leaks information about people in other ways. All we can do now is hope for the best, because the dataset can't be unreleased.

6. Computer science knowledge is relevant and highly available.

Cavoukian and Castro dismiss the research literature because "All of the above examples of primary literature are research-based articles within the highly specialized field of computer science".

This is a strange claim in several ways. First, it seems odd to dismiss the role of computer science knowledge when the question at issue is whether computers can be used to re-identify computerized data. Second, although computer science skills might be rare in some circles, there are tens of millions of software developers in the world today and computer science enrollments are soaring.

Most "anonymized" datasets require no more skill than programming and basic statistics to de-anonymize. Indeed, only a small fraction of re-identification research is publishable — the rest never gets written up as papers (although it occasionally makes for nice blog posts)[11] or gets rejected during peer review for being too simple (yes, this happens regularly). Re-identifying the NYC taxi data would make a good undergraduate homework assignment, but is nowhere near publishable. Moreover, even published research on de-anonymization has a tendency to look obvious a decade later; such is the nature of progress in algorithms and computer science.

7. Cavoukian and Castro apply different standards to big data and re-identification techniques.

Consider this paragraph from Cavoukian and Castro's paper: "Advances in computing technology are unlocking opportunities to collect and analyze data in ways never before imagined. The analysis of data may lead to important insights and innovations that will benefit not only individual consumers, but society at large. While data is typically collected for a single purpose, increasingly it is the many different secondary uses of the data wherein tremendous economic and social value lies."

We agree! In particular, just as big data techniques keep advancing, so do re-identification techniques. In fact, re-identification can help unlock many of the secondary uses that Cavoukian and Castro allude to, although the benefit to individuals or society at large is dubious.

Moreover, the putative benefits of big data come from machine-learning techniques that allow indirect, probabilistic inferences about people. For example, the excitement around personal genomics is about being able to predict certain diseases slightly better, and not with anything approaching certainty. In this context, the authors' insistence on only worrying about privacy-breaching inferences that have a high degree of accuracy is a clear double standard. They sum-

[11] Arvind Narayanan, *Lendingclub.com: A De-anonymization Walkthrough,* 33 Bits of Entropy blog (2008), available at http://33bits.org/2008/11/12/57/

marily dismiss inference of medical conditions that are only as accurate as a “flip of a coin,” and yet an equivalent ability to predict disease risks would be considered an extraordinary success.

8. Quantification of re-identification probabilities, which permeates Cavoukian and Castro’s arguments, is a fundamentally meaningless exercise.

Cavoukian and Castro endorse the idea of estimating re-identification probabilities. Indeed, this is the only prescription they make that goes beyond simply hoping for the best. Yet this is a meaningless exercise. The reason is simple: it always depends on arbitrary and fragile assumptions about what auxiliary datasets and general knowledge are available to the adversary.[12]

Their paper itself contains evidence of this. Consider this text that motivates the fact that careful de-identification is better than a superficial version: “the risk of unique identification drops off sharply when given slightly more abstract data. For instance, if an individual’s date of birth is replaced with only the month and year of birth, the percentage of those uniquely identifiable drops [from 63 per cent] to 4.2 per cent.”

While Cavoukian and Castro recognize that making the released data more abstract greatly affects re-identification probabilities, they fail to realize that making the adversary’s auxiliary dataset more *specific* has the equal and opposite impact! Of course, with high-dimensional datasets, there are strong limits to how much the data can be generalized without destroying utility, whereas auxiliary information has the tendency to get more specific, accurate, and complete with each passing year.

In fact, the example that they hold up as the way to do de-identification correctly serves as the perfect illustration of our point. The dataset in question comes from the Heritage Health Prize, released for a USD 3MM data-mining competition to predict future health outcomes based on past hospitalization (insurance claims) data. The dataset was de-identified by Khaled El Emam and his team.[13] Cavoukian and Castro approvingly note, “Based on this empirical evaluation, it was estimated that the probability of re-identifying an individual was .0084”.

We are in a position to say something about this claim, because it so happens that the Heritage Health Prize organizers contacted Narayanan to perform an adversarial (“red-team”) analysis of the de-identified data. (In the rest of this section, the first person singular refers to Narayanan.)

Happily for the patients in the dataset, *large-scale* auxiliary databases of medical information (hospital visits, etc.) that could be used for re-identification don’t appear to be available *publicly* at the *present time*. All things considered, if the data were going to be released publicly, I believe that El Emam and his team did the best job of de-identification that one could.

[12] For probabilistic guarantees that don’t depend on assumptions about the adversary, we must look outside de-identification, specifically, to differential privacy.

[13] Khaled El Emam et al., *De-identification methods for open health data: the case of the Heritage Health Prize claims dataset,* Journal of Medical Internet Research 14(1) (2012).

Where I differ sharply with him, however, is in the estimates of re-identification probabilities. His estimates were derived based on a specific, somewhat arbitrary set of assumptions. Here's just one: "we assume that the adversary does not know the order of the quasi-identifier values. For example, the adversary would not know that the heart attack occurred before the broken arm…" I cannot fathom the reason for this. If the auxiliary information comes from online review sites or from personal knowledge of the subject, detailed timeline information is very much available.

In my report to the Heritage Health Prize organizers, to show just how arbitrary El Emam's estimates are, I performed my own re-identification analysis with a slightly different set of assumptions, one that seems at least as realistic to me as the original. In particular, I assumed that the adversary knows the year but not the month or day of each visit. I was then able to show that one would derive dramatically different re-identification probabilities — up to 12.5% of members.[14]

It is very tempting to look for an assurance that (say) only 1% of individuals in a dataset can be re-identified. But there is simply no scientific basis for interpreting re-identification probabilities of de-identified high-dimensional datasets as anything more than (weak) lower bounds, and we urge the reader to be wary of false promises of security.

Conclusion: no silver bullet

We've explained why de-identification fails to resist inference of sensitive information either in theory or in practice and why attempts to quantify its efficacy are unscientific and promote a false sense of security. While we did this by looking at the inaccuracies and misrepresentations in Cavoukian and Castro's piece, our arguments apply equally well against other attempts to defend de-identification.

Data privacy is a hard problem. 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. These solutions aren't fully satisfactory, either individually or in combination, nor is any one approach the best in all circumstances.

Change is difficult. When faced with the challenge of fostering data science while preventing privacy risks, the urge to preserve the status quo is understandable. However, this is incompatible with the reality of re-identification science. If a "best of both worlds" solution exists, de-identification is certainly not that solution. Instead of looking for a silver bullet, policy makers must confront hard choices.

[14] Arvind Narayanan, *An Adversarial Analysis of the Reidentifiability of the Heritage Health Prize Dataset,* Manuscript (2011).