

Anonymous does not stay anonymous.

Four decades of re-identification

Anonymisation is a claim about the technique available when the data was released. It is not a property the data carries with it into the future.

Each case below involved data that was genuinely believed to be anonymous by competent people acting in good faith, and each was undone by someone with a public dataset and some patience.

2006: AOL, and the woman in Lilburn

On 4 August 2006, AOL's research division published about 20 million search queries from more than 650,000 users, covering three months. User names were replaced with numbers. It was intended as a gift to researchers.

Five days later, the New York Times published an article identifying user 4417749 as Thelma Arnold, a 62-year-old widow in Lilburn, Georgia. They found her by reading her searches: a run of queries about her town, about people sharing her surname, and about ailments common to older women. No hacking was required. The searches described her.

AOL pulled the file within three days of release, but it had already been mirrored. Two employees were dismissed and the chief technology officer resigned later that month.

2008: the Netflix Prize

Netflix released a dataset of just over 100 million film ratings from about 480,000 subscribers, with identities removed, as the basis of a public competition to improve its recommendation engine.

Arvind Narayanan and Vitaly Shmatikov showed that this was not enough. In *Robust De-anonymization of Large Sparse Datasets*, presented at the IEEE Symposium on Security and Privacy in 2008, they matched entries in the anonymised set against publicly posted ratings on a film database, and recovered identities. The insight is uncomfortable and general: your particular combination of tastes is close to unique, so a handful of ratings with dates functions as a fingerprint.

Three ordinary facts

In a Carnegie Mellon working paper published in 2000, Latanya Sweeney reported that 87% of the United States population, based on 1990 census data, had a combination of five-digit ZIP code, sex and date of birth that was likely to be unique to them.

In fairness we should give the correction as prominently as the claim. In 2006, Philippe Golle revisited the question with 2000 census data and arrived at 63%, and 61% for the 1990 data Sweeney had used, and he stated plainly that he could not explain the discrepancy because he lacked detail about the earlier method. So the honest position is that this is a failure to replicate rather than a demonstrated error, and the defensible statement is that a majority of Americans, though not 87% of them, are uniquely identified by three facts that nobody would think twice about giving out.

Either figure is enough to make the point. There is no such thing as an identifying field and a harmless field. There are only combinations.

The joins are where the danger is

The legal scholar Paul Ohm gathered these cases into an argument in *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization* (UCLA Law Review, 2010). His conclusion was that reidentification would keep improving, and would draw each of us closer to what he called our personal "databases of ruin": the point at which enough separate records have been linked that there exists, somewhere, a record containing at least one fact about you whose exposure would cause serious concrete harm.

He returned to it in the Harvard Business Review in 2012 with a warning aimed at the people building these systems, and with the observation that matters most here: once such a database has been assembled, it is unlikely we will ever be able to take it apart again.

2018: the heatmap and the military bases

A last case, because it shows that aggregation is not a defence either.

In November 2017 the fitness company Strava published a global heatmap drawn from about 700 million activities and 1.4 trillion GPS points. No names, no individual routes, nothing but accumulated brightness on a world map.

In January 2018 an Australian student, Nathan Ruser, noticed that in parts of Afghanistan and Syria where the surrounding terrain was dark, there were small

bright loops. They were military personnel jogging the perimeters of bases. The aggregate had disclosed the thing none of its individual records did.

Nobody was identified. That was not necessary. The pattern was the disclosure.

Where to read more

- [The Smaller Kingdom](#): the whole argument, and what to do about it.
- [Consent does not travel](#): what happens when the company holding your data is sold.
- [Records outlive regimes](#): the Amsterdam population register, and what was done with it.
- [The philosophy](#): why we build software that cannot collect anything.

Frequently asked

Is anonymised data really anonymous?

Not permanently. Anonymisation reflects the technique available at the time of release. AOL search logs in 2006, Netflix ratings in 2008 and demographic records have all been re-identified after publication as anonymous data.

What is re-identification?

Matching an anonymised record back to a named person, usually by combining it with another dataset that is public. Narayanan and Shmatikov re-identified Netflix subscribers by matching anonymised ratings against publicly posted film ratings.

What is a database of ruin?

A phrase from the legal scholar Paul Ohm's 2010 article on the failure of anonymisation. As separate records accumulate and link, there will eventually exist for each person a database containing at least one closely guarded fact whose exposure would cause serious harm.

Can aggregated data reveal individuals?

It can reveal patterns that matter even without naming anyone. Strava's 2017 global heatmap contained no identities, but the shape of the aggregate exposed the perimeters of military bases in Afghanistan and Syria when it was examined in January 2018.

How many facts does it take to identify someone?

Very few. A 2000 Carnegie Mellon working paper put the figure at 87% of Americans uniquely identified by ZIP code, sex and date of birth; a 2006 replication

using later census data arrived at 63%. Either way, a majority of people are identified by three unremarkable facts.

From Offline.Ltd, Amsterdam. The web version of this page lives at <https://offline.ltd/smaller-kingdom/anonymous-does-not-stay-anonymous> and is kept current there.