

Protecting Privacy After the Failure of Anonymisation

Associate Professor Paul Ohm
University of Colorado Law School

UK Information Commissioner's Office
30 March 2011

The Paper

- Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA LAW REVIEW 1701 (2010)
- Available for download: <http://paulohm.com>

Anonymisation

- Manipulation of information in a database to try to make it difficult to identify the subjects of the data.

Anonymisation is:

- Ubiquitous
- Trusted
- Rewarded by Law

Example: HIPAA Privacy Rule

- De-Identified Health Information (DHI)
- “Statistical Standard”
- “Safe Harbor Standard”: Health information without (partial list):
 - Names, geographic subdivisions smaller than state (except some partial ZIP codes), most dates, telephone numbers, fax numbers, e-mail addresses, social security numbers, medical record numbers, health plan beneficiary numbers, account numbers, certificate/license numbers, device identifiers and serial numbers, URLs, IP addresses, biometric identifiers, full face photos.
 - 45 C.F.R. § 164.514(b)(1)
- Every other privacy statute, too.

The Problem?

- “Data can be either useful or perfectly anonymous but never both.”

Netflix Prize Study

- Public Release of Anonymised Database with 100 Million Ratings for 480,000 Users rating 18,000 Movies
- Each record containing:
 - Date
 - Movie
 - Rating
 - Anonymised, Cross-Session ID

Netflix Prize (Narayanan & Shmatikov)

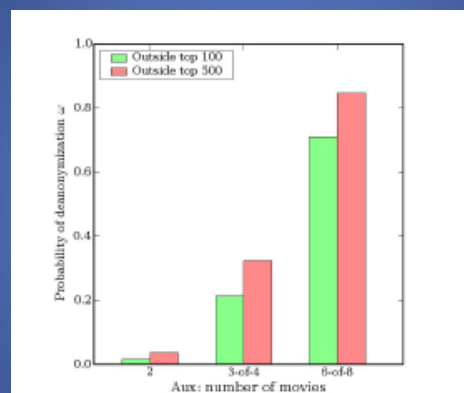


Figure 8. Adversary knows exact ratings but does not know dates at all.

Netflix Prize (Narayanan & Shmatikov)

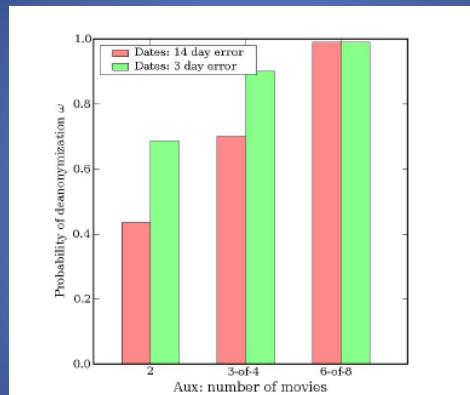


Figure 4. Adversary knows exact ratings and approximate dates.

The Trend

- Series of Studies challenging intuitions about robustness of anonymisation
- Even experts often seem surprised by how:
 - Easy
 - Cheap (\$, CPU)
 - Fast
- Five years from now, none of this will surprise us.
- For now: We're facing an intuition gap

The Objections

- Examples of bad anonymisation.
- “Who Cares?”
- The Myth of the Superuser

**(HOW) SHOULD POLICYMAKERS
RESPOND?**

The End of PII

- Should we expand HIPAA to include:
 - Movie ratings
 - Search query search strings
 - Social network friends' relationships graph?
- No!
 - Regulatory Whack-a-Mole
 - Privacy Theater

Why Technology Won't Save Us

- The Privacy/Utility Relationship
 - The Impossibility Result
 - The Inverse Relationship
 - The Imbalanced Relationship

The Way Forward

- Contextual Risk Assessment, considering:
 - 1. Data Handling Techniques
 - 2. Private versus Public Release
 - 3. Quantity
 - 4. Motive
 - 5. Trust
- Balance benefits of access to information against risk of privacy harm

New Privacy Regulation Principles

- Shift:
 - From silver bullets to risk assessments
 - From promises to good faith efforts
 - From math to sociology
 - From requirements to accountability
 - From personal information to unjustifiably risky collections of information

Risky Data Collections

- Special attention for databases satisfying any of these:
 - Numerosity
 - Rarity
 - Diversity
- Irrespective of presence of personal data, PII, identifiable data, etc.

Data.gov.uk

- Benefits
 - Potentially immense
 - But speculative
- Costs
 - Risk of harm
 - Varies widely by database
 - Will increase with time

Revising the Data Protection Directive

- “personal data”
 - (a) 'personal data' shall mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity;

Revising the Data Protection Directive

- Recital 26
 - (26) Whereas the principles of protection must apply to any information concerning an identified or identifiable person; whereas, to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person; whereas the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable; whereas codes of conduct within the meaning of Article 27 may be a useful instrument for providing guidance as to the ways in which data may be rendered anonymous and retained in a form in which identification of the data subject is no longer possible;

THANK YOU
<http://paulohm.com>