

The Failure of Anonymization and
How it will Revolutionize How we
Protect (and Regulate) Health
Privacy

Paul Ohm

Associate Professor

University of Colorado Law School

September 17, 2009

Anonymization

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

Name	Race	Birth Date	Sex	ZIP Code	Diagnosis
Sean	Black	9/20/1965	Male	02141	Short of breath
Daniel	Black	2/14/1965	Male	02141	Chest pain
Kate	Black	10/23/1965	Female	02138	Painful eye
Marion	Black	8/24/1965	Female	02138	Wheezing
Helen	Black	11/7/1964	Female	02138	Obesity
Reese	Black	12/1/1964	Female	02138	Chest pain
Forest	White	10/23/1964	Male	02138	Short of breath
Hilary	White	3/15/1965	Female	02139	Hypertension
Philip	White	8/13/1964	Male	02139	Obesity
Jamie	White	5/5/1964	Male	02139	Fever
Sean	White	2/13/1967	Male	02138	Vomiting
Adrien	White	3/21/1967	Male	02138	Back pain

Why Anonymize?

- Norms and Ethics
- The Market
- Architecture
- Law

The Law Assumes Robust Anonymization

- The Hunt for PII: “A quasi-scientific exercise in information categorization.”
- Step One: List all categories of information that might link the data to identity (PII).
- Step Two: Delete all of the information.
- Step Three: Get out of jail free.

Example: HIPAA Privacy Rule

- De-Identified Health Information (DHI) “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)

The Problem?

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

THE MASSACHUSETTS HEALTH INSURANCE STUDY

Name	Race	Birth Date	Sex	ZIP Code	Diagnosis
Sean	Black	9/20/1965	Male	02141	Short of breath
Daniel	Black	2/14/1965	Male	02141	Chest pain
Kate	Black	10/23/1965	Female	02138	Painful eye
Marion	Black	8/24/1965	Female	02138	Wheezing
Helen	Black	11/7/1964	Female	02138	Obesity
Reese	Black	12/1/1964	Female	02138	Chest pain
Forest	White	10/23/1964	Male	02138	Short of breath
Hilary	White	3/15/1965	Female	02139	Hypertension
Philip	White	8/13/1964	Male	02139	Obesity
Jamie	White	5/5/1964	Male	02139	Fever
Sean	White	2/13/1967	Male	02138	Vomiting
Adrien	White	3/21/1967	Male	02138	Back pain

Zip Code / Birthdate / Sex?

Race	Birth Date	Sex	ZIP Code	Diagnosis
Black	9/20/1965	Male	02141	Short of breath
Black	2/14/1965	Male	02141	Chest pain
Black	10/23/1965	Female	02138	Painful eye
Black	8/24/1965	Female	02138	Wheezing
Black	11/7/1964	Female	02138	Obesity
Black	12/1/1964	Female	02138	Chest pain
White	10/23/1964	Male	02138	Short of breath
White	3/15/1965	Female	02139	Hypertension
White	8/13/1964	Male	02139	Obesity
White	5/5/1964	Male	02139	Fever
White	2/13/1967	Male	02138	Vomiting
White	3/21/1967	Male	02138	Back pain

- How many people living in your ZIP code share both your birthdate and sex?
- City/Birthdate/Sex?
- County/Birthdate/Sex?

Latanya Sweeney's Results

- Percent of People in U.S. Uniquely ID'd by:
 - ZIP-5/Birthdate/Sex: 87.1%
 - City/Birthdate/Sex: 53%
 - County/Birthdate/Sex: 18%
- Governor William Weld assured Privacy

Latanya Sweeney's Results

- Governor William Weld
 - Resident of Cambridge, MA (pop. 54,000, seven zip codes)
 - \$20 for Voter Registration Rolls
 - Six with same birthdate; three men; one in same ZIP

Race	Birth Date	Sex	ZIP Code	Diagnosis
Black	9/20/1965	Male	02141	Short of breath
Black	2/14/1965	Male	02141	Chest pain
Black	10/23/1965	Female	02138	Painful eye
Black	8/24/1965	Female	02138	Wheezing
Black	11/7/1964	Female	02138	Obesity
Black	12/1/1964	Female	02138	Chest pain
White	10/23/1964	Male	02138	Short of breath
White	3/15/1965	Female	02139	Hypertension
White	8/13/1964	Male	02139	Obesity
White	5/5/1964	Male	02139	Fever
White	2/13/1967	Male	02138	Vomiting
White	3/21/1967	Male	02138	Back pain

Race	Birth Date	Sex	ZIP Code	Diagnosis
Black	9/20/1965	Male	02141	Short of breath
Black	2/14/1965	Male	02141	Chest pain
Black	10/23/1965	Female	02138	Painful eye
Black	8/24/1965	Female	02138	Wheezing
Black	11/7/1964	Female	02138	Obesity
Black	12/1/1964	Female	02138	Chest pain
White	10/23/1964	Male	02138	Short of breath
White	3/15/1965	Female	02139	Hypertension
White	8/13/1964	Male	02139	Obesity
White	5/5/1964	Male	02139	Fever
White	2/13/1967	Male	02138	Vomiting
White	3/21/1967	Male	02138	Back pain

THE NETFLIX PRIZE STUDY

The Netflix Prize

User	Titanic	Dark Knight	Star Wars	Shrek 2	E.T.	Phantom Menace	Dead Man's Chest	Spider-Man	Revenge of the Sith
1	1	4	3	2	5	3	2	5	3
2	3	5	2	3	5	3	2	1	4
3	3	2	2	3	2	5	5	2	3
4	5	3	2	5	5	5	5	5	5
5	3	5	2	3	5	5	2	4	3
6	1	1	1	1	1	1	1	1	1
7				3	1		3		
8	5			4					
9	3	1	1	4	3	3	2	5	2

The Netflix Prize

User	Zyzyx Road	Death Defying Acts	Blonde Ambition	Scorched	Harold	Head	Molly	Unknown	Shade
1	3								
2			4				2		
3		1							
4				5	5				
5						3		2	
6					3			3	
7			1						1
8						1			2
9				5					

The Netflix Prize



The Netflix Prize

- 100 Million Ratings
- 480,000 Users
- 18,000 Movies



Results

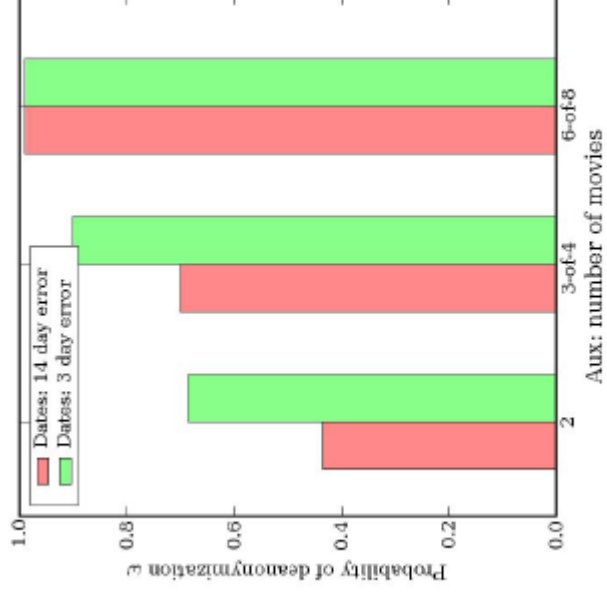


Figure 4. Adversary knows exact ratings and approximate dates.

**(HOW) SHOULD POLICYMAKERS
RESPOND?**

What Easy Reidentification Does to

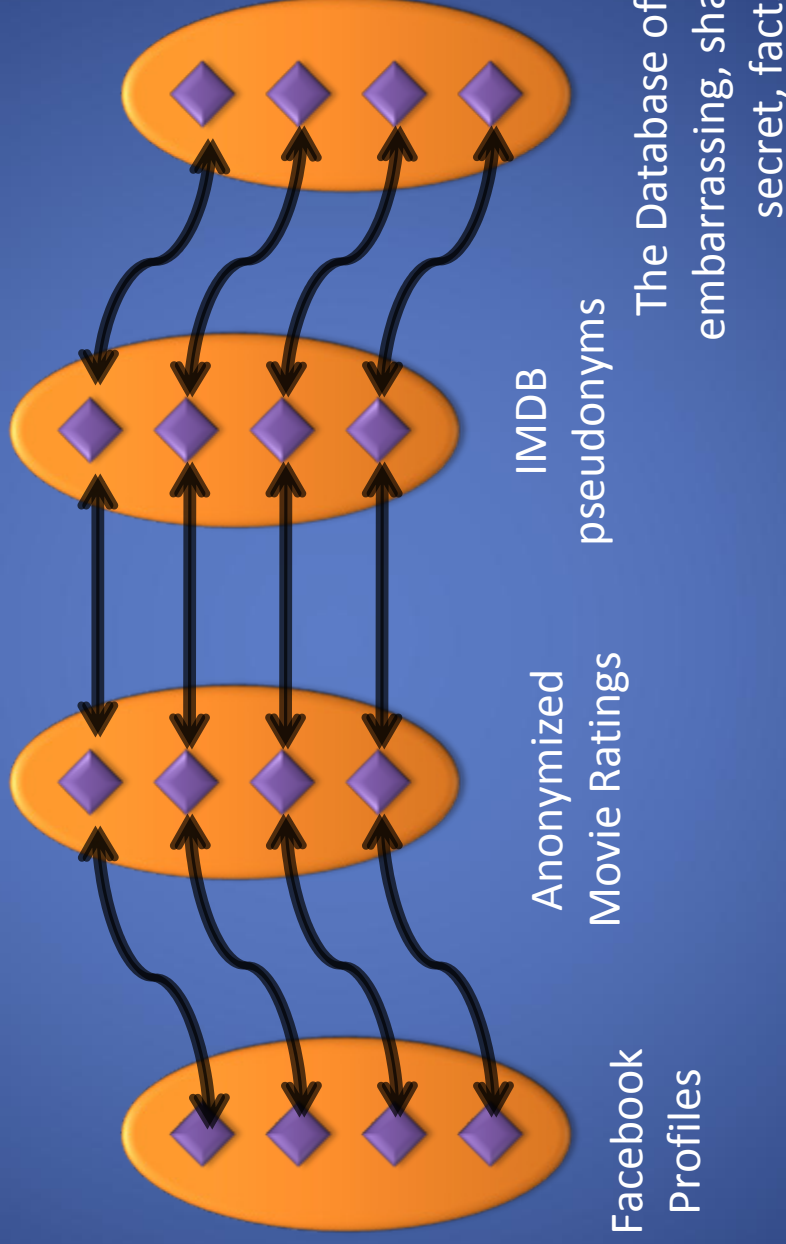
HIPAA

- De-Identified Health Information may contain:
 - Hospital Name, diagnosis, year of visit, patient's age, partial zip codes.
- An adversary with rich outside information can use these to defeat anonymity.
- Result: Regulatory Whack-a-Mole; “Privacy Theater”

Half-Measures and False Starts

- 1. Punish Those Who Harm Strictly

Response: The Accretion Problem



Half-Measures and False Starts

- 1. Punish Those Who Harm Strictly
- 2. Wait for Technology to Save Us

Response: Hard to Achieve Balance

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

The Way Forward

- Risk Assessment
 - What is the Risk of Reidentification in this particular context?
 - From Math to Sociology
 - Think Comprehensively and Contextually

Risk Assessment Factors

- 1. Data Handling Techniques
- 2. Private versus Public Release
- 3. Quantity
- 4. Motive
- 5. Trust

A New HIPAA

- Trust the technology less; trust the people more.
- 1. Formalize and codify rules of trust.
- 2. Free up sharing between trusted parties.
- 3. Prevent abuse through safeguards and accountability mechanisms
 - Vary based on sensitivity.
 - Very sensitive: NSA-inspired clearances; researchers go to the data, not the other way around.
 - Less sensitive: Relax protections

Result of this new HIPAA?

- Stifle research?
- Expand research?

THANK YOU

<http://paulohm.com>