

# Data and Society

## Anonymity – Lecture 6

2/11/21

# Today (2/11/21)

- *Briefing Instructions*
- Lecture / Discussion – Anonymity
- Student Presentations

# Briefing Instructions – Team Project

- Briefing topic: Your team should choose a **data-related bill currently in Congress**. Describe its content, status, and potential implications for enforcement for your reader – your boss / elected official.
- Briefings are informational pieces. Everything you need to know should be summarized and explained within the briefing. Your briefing should summarize key aspects of the bill (see next slide) and provide a recommendation.
- **Briefing is due by 11:59 p.m., 2/26/21**
- **Format:** Two pages, .docx, 11 pt. font or larger, cite references in text as needed

**Briefing is worth 15 points. How your team be graded:**

**Each team member will get the same briefing grade.**

- **Content: 8 points**
  - Does the briefing address the questions?
  - Is it clear and self-contained?
  - Is the briefing interesting to read?
  - Did you provide a recommendation?
- **Writing: 7 points**
  - Is the writing compelling, concise and informative
  - Does the piece read well? Is it informative?
  - Is the spelling and grammar correct? Is the piece appropriately referenced?

# Briefing Structure

- A briefing is **informational**, often prepared for a decision-maker, who may need to make hard choices about topics that they do not understand well or don't have time to research in-depth.
- A briefing fills in key details your decision-maker needs to know. Your briefing will also **propose a recommendation** on whether to vote for the bill or not.
- **Your briefing should address the following questions**
  - **What bill are you describing, who introduced it, and when?**
  - **What does the bill do?** (significant aspects)
  - **Who will the bill impact and how?**
  - **What are its limitations?** (legal limitations, what it does not cover)
  - **How will it be enforced?**
  - **What is your recommendation?** (should your stakeholder vote for or against this bill and why)
  - References as needed (not counted in page count)

# Useful resources

- **Focus on data-related bills in the 116<sup>th</sup> and 117<sup>th</sup> Congress that HAVE NOT been passed**
  - Privacy bills in the 116<sup>th</sup> Congress:  
<https://crsreports.congress.gov/product/pdf/LSB/LSB10441>
  - <https://www.congress.gov/browse>

# Briefing Teams – email [bermaf@rpi.edu](mailto:bermaf@rpi.edu) if you don't have your partner's email

- Justin C. and Nicholas J.
- Jeff H. and Isaac L.
- Jin H. and Ishita P.
- Nathan S. and Eric X.
- Davis E. and Hannah L.
- Adam M. and Sola S.
- Angelina M. and Liam M.
- Grant B. and Justin O.
- Julian C. and Chris P.
- Josh M. and Greg S.

# Reading / Speaker for February 18


- Reading for next time: “**Ben Wizner: Pull back to reveal**”, Guernica, <https://www.guernicamag.com/pull-back-to-reveal/>

AMERICAN EMPIRES: POWER AND ITS DISCONTENTS | INTERVIEW | October 1, 2014

## Ben Wizner: Pull Back to Reveal

*The privacy advocate and legal advisor to Edward Snowden on today's surveillance empire.*

By Henry Peck



Between 2011 and 2013, a gargantuan structure took shape in the Utah desert. Covering one million square feet and costing over a billion dollars, the facility is the largest data center of the National Security Agency to date. Designed to hold what the *Washington Post* called “oceans of bulk data,” it’s a tangible manifestation of an American empire based in virtual space, a modern-day watchtower for electronic surveillance.

The scope of US surveillance had been far less conspicuous than the NSA’s architecture (which itself is off limits to the public) until the release in 2013 of top-secret NSA documents by the computer analyst Edward Snowden. These revelations were particularly significant for Ben Wizner, director of the American Civil Liberties Union’s Speech, Privacy & Technology Project. Wizner had for years been bringing cases challenging the legality of surveillance programs, only to see them dismissed due to

Date	Topic	Speaker	Date	Topic	Speaker
1-25	Introduction	Fran	1-28	The Data-driven World	Fran
2-1	Data and COVID-19	Fran	2-4	Data and Privacy -- Intro	Fran
2-8	Data and Privacy – Differential Privacy	Fran	2-11	Data and Privacy – Anonymity / Briefing Instructions	Fran
2-15	NO CLASS / PRESIDENT’S DAY		2-18	Data and Privacy – Law	Ben Wizner
2-22	Digital rights in the EU and China	Fran	2-25	Data and Discrimination 1	Fran
3-1	Data and Discrimination 2	Fran	3-4	Data and Elections 1	Fran
3-8	Data and Elections 2	Fran	3-11	NO CLASS / WRITING DAY	
3-15	Data and Astronomy	Alyssa Goodman	3-18	Data Science	Fran
3-22	Digital Humanities	Brett Bobley	3-25	Data Stewardship and Preservation	Fran
3-29	Data and the IoT	Fran	4-1	Data and Smart Farms	Rich Wolski
4-5	Data and Self-Driving Cars	Fran	4-8	Data and Ethics 1	Fran
4-12	Data and Ethics 2	Fran	4-15	Cybersecurity	Fran
4-19	Data and Dating	Fran	4-22	Data and Social Media	Fran
4-26	Tech in the News	Fran	4-29	Wrap-up / Discussion	Fran
5-3	NO CLASS				



# Lecture – Anonymity


- Anonymity (Sweeney)
- Netflix Competition (Narayanan and Shmatikov)

# Privacy and Anonymity

- **Privacy:** *The state of being free from being observed or disturbed by other people; the state of being free from public attention.*
- **Anonymity:** *Lack of outstanding, individual, or unusual features; impersonality*

# Anonymity: Can you keep your data private by removing explicit identifiers?

- **Latanya Sweeney et al:**  
Removing / Changing explicit identifiers will *not* get you anonymity



**DATA PRIVACY LAB**  
De-identification Project

## Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression

by Pierangela Samarati and [Latanya Sweeney, Ph.D.](#)

**Abstract**

Today's globally networked society places great demand on the dissemination and sharing of person-specific data. Situations where aggregate statistical information was once the reporting norm now rely heavily on the transfer of microscopically detailed transaction and encounter information. This happens at a time when more and more historically public information is also electronically available. When these data are linked together, they provide an electronic shadow of a person or organization that is as identifying and personal as a fingerprint, even when the sources of the information contains no explicit identifiers, such as name and phone number. In order to protect the anonymity of individuals to whom released data refer, data holders often remove or encrypt explicit identifiers such as names, addresses and phone numbers. However, other distinctive data, which we term *quasi-identifiers*, often combine uniquely and can be linked to publicly available information to re-identify individuals.

In this paper we address the problem of releasing person-specific data while, at the same time, safeguarding the anonymity of individuals to whom the data refer. The approach is based on the definition of *k-anonymity*. A table provides *k-anonymity* if attempts to link explicitly identifying information to its contents ambiguously map the information to at least *k* entities. We illustrate how *k-anonymity* can be provided by using generalization and suppression techniques. We introduce the concept of minimal generalization, which captures the property of the release process not to distort the data more than needed to achieve *k-anonymity*. We illustrate possible preference policies to choose among different minimal generalizations. Finally, we present an algorithm and experimental results when an implementation of the algorithm was used to produce releases of real medical information. We also report on the quality of the released data by measuring precision and completeness of the results for different values of *k*.

# Key Definitions (informal))

- **Anonymous data:** Data that cannot be manipulated or linked to confidently identify the entity that is the subject of the data.
- **Explicit identifier:** Set of data elements (e.g. {name, address} or {name, phone number} for which *with no additional information*, the designated person can be directly and uniquely ascertained.
- **Quasi-identifier:** Set of data elements that in combination can be used to identify an entity uniquely or almost uniquely.
- **De-identified data:** Data with explicit identifiers removed, generalized, or replaced with a made-up alternative.

# Re-identifying data may be straightforward

- Most states collect **hospital discharge data**, which is distributed to researchers, sold to industry, and often made publicly available.
- When **coupled with census data** or **voter registration data**, combinations of characteristics can identify individuals, **even if the data has been de-identified**.
  - {Zip, gender, month and year of birth}
  - {Zip, gender, age}
  - {County, gender, date of birth}
  - {County, gender, age}
  - Etc.

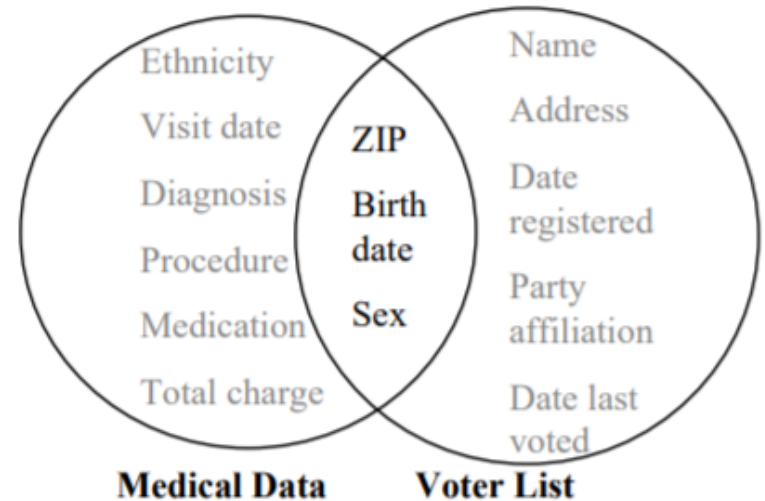


Figure 1 Linking to re-identify data

# Sweeney/2000: Simple quasi-identifiers can be used to identify entities

- Sweeney showed that {zip, gender, birthdate} (quasi-identifier) could be used to identify most people (87%) in the U.S.
  - Methodology used public data (1990 U.S. Census) and other publicly and semi-publicly available health data

Race	Birth	Gender	ZIP	Problem
Black	09/20/65	m	02141	short of breath
Black	02/14/65	m	02141	chest pain
Black	10/23/65	f	02138	hypertension
Black	08/24/65	f	02138	hypertension
Black	11/07/64	f	02138	obesity
Black	12/01/64	f	02138	chest pain
White	10/23/64	m	02138	chest pain
White	03/15/65	f	02139	hypertension
White	08/13/64	m	02139	obesity
White	05/05/64	m	02139	short of breath
White	02/13/67	m	02138	chest pain
White	03/21/67	m	02138	chest pain

Figure 6 De-identified data

# K-anonymity (Sweeney and Samarati)

- **K-anonymity** is a property of a data set, usually used in order to describe the **data set's level of anonymity**.
- A data set is said to be **k-anonymous** if the information for each person contained in the set cannot be distinguished from at least  $k-1$  other individuals whose information also appear in the data-set

# Ways to increase k in k-anonymity: Suppression and Generalization

- **Suppression:** Some values of the attributes are replaced by \*
- **Generalization:** Individual values of the attributes are replaced by a broader category
- **Upper table** has 1-anonymity because Ramsha can be identified uniquely by age.
- **Lower table** has 2-anonymity wrt {age, gender, state of domicile}

Name	Age	Gender	State of domicile	Religion	Disease
Ramsha	30	Female	Tamil Nadu	Hindu	Cancer
Yadu	24	Female	Kerala	Hindu	Viral infection
Salima	28	Female	Tamil Nadu	Muslim	TB
Sunny	27	Male	Karnataka	Parsi	No illness
Joan	24	Female	Kerala	Christian	Heart-related
Bahuksana	23	Male	Karnataka	Buddhist	TB
Rambha	19	Male	Kerala	Hindu	Cancer
Kishor	29	Male	Karnataka	Hindu	Heart-related
Johnson	17	Male	Kerala	Christian	Heart-related
John	19	Male	Kerala	Christian	Viral infection

Name	Age	Gender	State of domicile	Religion	Disease
*	20 < Age ≤ 30	Female	Tamil Nadu	*	Cancer
*	20 < Age ≤ 30	Female	Kerala	*	Viral infection
*	20 < Age ≤ 30	Female	Tamil Nadu	*	TB
*	20 < Age ≤ 30	Male	Karnataka	*	No illness
*	20 < Age ≤ 30	Female	Kerala	*	Heart-related
*	20 < Age ≤ 30	Male	Karnataka	*	TB
*	Age ≤ 20	Male	Kerala	*	Cancer
*	20 < Age ≤ 30	Male	Karnataka	*	Heart-related
*	Age ≤ 20	Male	Kerala	*	Heart-related
*	Age ≤ 20	Male	Kerala	*	Viral infection



# Caveats

## K-anonymity is susceptible to attacks:

- K-anonymity fails in high-dimensional data sets and most real-world datasets of individual recommendations and purchases.
- When background knowledge is available to an attacker, such attacks become even more effective.

# The 2006 Netflix Competition

- **Netflix 2006 competition:** Netflix offered \$1M prize for improving their movie recommendation service.
  - Dataset with 100M movie ratings created by 480K Netflix subscribers between 1995 and 2005 provided.
  - Ratings did not appear to have been perturbed significantly
- NetFlix Prize dataset did not provide user names. In answer to the question **“Is there any customer information in the database that should be kept private?”**, Netflix said (FAQ):
  - *“No, all customer identifying information has been removed; all that remains are ratings and dates. This follows our privacy policy, which you can review [here](#). Even if, for example, you knew all your own ratings and their dates you probably couldn’t identify them reliably in the data because only a small sample [of data] was included (less than one-tenth of our complete dataset) and that data was subject to perturbation. Of course, since you know all your own ratings that really isn’t a privacy problem, is it?”*
- **Narayanan and Shmatikov showed that the data set could be de-anonymized**

# Netflix data set: removing identifying information was not sufficient for anonymity

- Researchers used an “adversary approach” to de-anonymize users, *even when some of the auxiliary information was imprecise*
- *Used additional information* to identify individual subscribers:
  - Private records of Netflix subscribers they knew.
  - Public IMDB ratings
- Narayanan and Shmatikov studied the question “**How much does the adversary need to know about a Netflix subscriber in order to identify her record in the data set, and thus learn her complete movie viewing history**”.

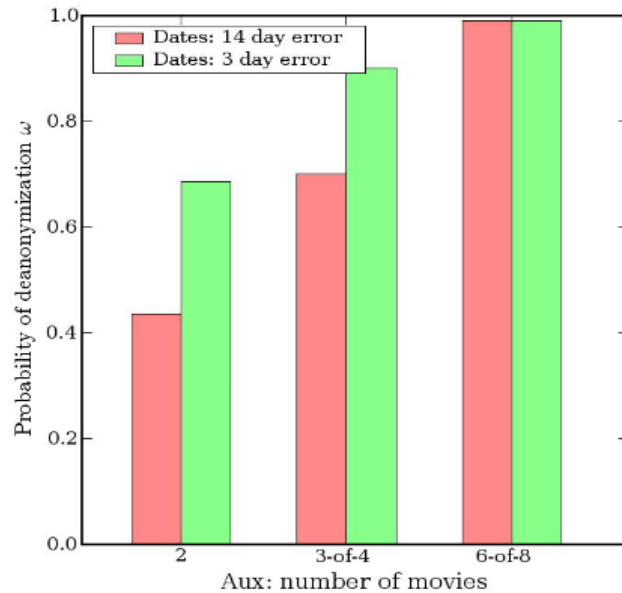
# Methodology

- Adversary's goal is to de-anonymize an anonymous record R from the public database.
  - Adversary has a small bit of **auxiliary information or background knowledge** related to R (restricted to a subset of R's attributes)
  - The auxiliary information may be imprecise or incorrect
- Designate Netflix users to be “**similar**” when two subscribers create **ratings that are close with respect to date** (within 3 days, within 14 days, within infinity days [no date given])) **and value** (same or within 1)
- **De-anonymization approach:**
  - Assign a numerical score to each record based on how well it matches some auxiliary/outside information
  - Use matching criteria used to see if there is a match between records in the database and auxiliary information
  - Select “best guess” candidate records with highest score(s)

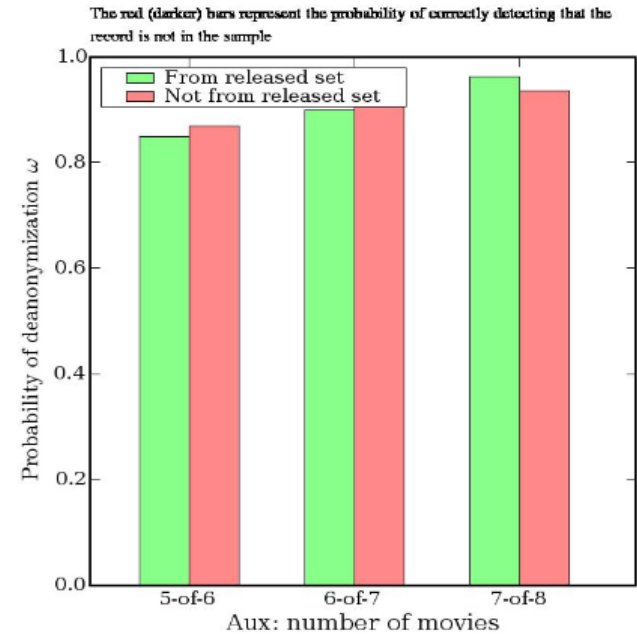
# Additional information

- **How does an adversary get auxiliary information?**
  - Conversation or overheard information
  - Personal blogs and Google searches
  - Public IMDB ratings (likely strong correlation with Netflix ratings)
- Researchers used a few dozen IMDB users to breach data set
- **Sparsity of information in dataset increases the probability that the adversary strategy succeeds** in de-anonymizing the data and decreases the amount of auxiliary information needed.
  - True of Netflix dataset
  - **Many real-world datasets** containing individual transactions, preferences, etc. **are sparse**

# Results



**Figure 4. Adversary knows exact ratings and approximate dates.**



**Figure 5. Same parameters as Fig. 4, but the adversary must also detect when the target record is not in the sample.**

[http://www.cs.utexas.edu/~shmat/shmat\\_oak08netflix.pdf](http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf)

# Why does de-anonymizing the Netflix dataset matter?

- Cross-correlation methods can reveal other non-public personal information (viewing habits may indicate political, sexual, religious or other preferences)
- Privacy breaches can endanger “future privacy” – private information in future sessions
- General methodology can be used with other similar sparse datasets (e.g. those focusing on social relationships)
- Could violate data privacy policy that claims that Netflix customer data will be shared and used anonymously

# Lecture 8 References (not already on slides)

- **“Simple Demographics Often Identify People Uniquely”**, Latanya, Sweeney, CMU Working Paper,  
<https://dataprivacylab.org/projects/identifiability/paper1.pdf>
- **K-anonymity**, <https://www.quora.com/What-is-k-anonymity>
- **“Robust de-anonymization of sparse data sets”**, Arvind Narayanan and Vitaly Shmatikov,  
[http://www.cs.utexas.edu/~shmat/shmat\\_oak08netflix.pdf](http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf)
- **“Have I Been Pwned — which tells you if passwords were breached — is going open source”**, The Verge,  
<https://www.theverge.com/2020/8/7/21359191/password-breach-have-i-been-pwned-open-source-troy-hunt>



# Presentations



# Upcoming Presentations

## February 18

- **“Analysis: California privacy reboot puts rights in spotlight”**, Bloomberg Law, <https://news.bloomberglaw.com/bloomberg-law-analysis/analysis-california-privacy-reboot-puts-rights-in-spotlight>
- **“To fix social media now, focus on privacy, not platforms”**, The Hill, <https://thehill.com/opinion/technology/535824-to-fix-social-media-now-focus-on-privacy-not-platforms>

## February 22

- **“Grindr on the hook for 10M euro violations over GDPR consent violations”**, TechCrunch, [https://techcrunch.com/2021/01/26/grindr-on-the-hook-for-e10m-over-gdpr-consent-violations/?guccounter=1&guce\\_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce\\_referrer\\_sig=AQAAABSWJHh2KAv6uq4Ncippikhw3Ce8GDoEKFMOcJPFJy1kTbj1Fn\\_rpur6O7sq1LYNhqv1HzwQ7AVNLVUCIRMG9wPBBVXTIxLK2WDqIMMtpFc68TjvPWzjrF0U4sqHCzns0wFJoubxi4WMIloTy6bswMgd-YBJC xvHYwuGyB9scWgeT](https://techcrunch.com/2021/01/26/grindr-on-the-hook-for-e10m-over-gdpr-consent-violations/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAABSWJHh2KAv6uq4Ncippikhw3Ce8GDoEKFMOcJPFJy1kTbj1Fn_rpur6O7sq1LYNhqv1HzwQ7AVNLVUCIRMG9wPBBVXTIxLK2WDqIMMtpFc68TjvPWzjrF0U4sqHCzns0wFJoubxi4WMIloTy6bswMgd-YBJC xvHYwuGyB9scWgeT)
- **“How the West got China’s social credit system wrong,”** Wired, <https://www.wired.com/story/china-social-credit-score-system/>

# Need Volunteers

## February 25 (Vaccines and discrimination)

- **“Where do the vaccine doses go and who gets them? The algorithms decide.”**, New York times, <https://www.nytimes.com/2021/02/07/technology/vaccine-algorithms.html?referringSource=articleShare> (Nicholas J.)
- **“Getting a Covid vaccine can be required by your boss. Why that's a good thing — and a danger”**, NBC News, <https://www.nbcnews.com/think/opinion/getting-covid-vaccine-can-be-required-your-boss-why-s-ncna1256389> (Julian C.)

# Presentations for February 11

- **“We’re banning facial recognition. We’re missing the point.”** New York Times, <https://www.nytimes.com/2020/01/20/opinion/facial-recognition-ban-privacy.html> (Josh)
- **“This site published every face from Parler’s Capitol riot videos”**, Wired, <https://www.wired.com/story/faces-> (Nate)