CSC 495.002 – Lecture 6
Web/Social Networks Privacy: K-anonymity

Dr. Özgür Kafalı

North Carolina State University
Department of Computer Science

Fall 2017

Targeted Advertising

- Online behavioral advertising definition
- Types of targeted advertising
- Types of cookies and how they work
- Tools to mitigate privacy concerns of targeted advertising
- People's attitudes towards private browsing tools

Problem Definition

- Data owner, e.g., hospital
- Has private dataset with user specific data
- Goal: To share a version of the dataset with researchers
  - Dataset can help researchers to train better models
  - Results can help the data owner
- Provide scientific guarantees that users in the dataset cannot be re-identified
- Data should remain practically useful

Real Problem

- For, 87% (216M of 248M) of the US population
- Uniquely identifiable based only on
  - 5-digit ZIP code
  - Gender
  - Date of birth

Sweeney. Uniqueness of Simple Demographics in the US Population, 2000

Netflix Prize

- In October 2006, Netflix offered a $1M prize for a 10% improvement in its recommendation system
- Released a training dataset for competitors to train their systems
- Disclaimer: To protect customer privacy, all personal information identifying individual customers has been removed and all customer IDs have been replaced by randomly assigned IDs

- Netflix is not the only movie-rating portal on the web
- On IMDb, individuals can rate movies “not” anonymously
- Researchers from University of Texas at Austin, linked Netflix dataset with IMDb to de-anonymize the identity of some users

Differential Privacy

- Provide guarantees for your released dataset

- Formally
  - Maximize the accuracy of queries from statistical databases
  - While minimizing the chances of identifying its records

Studies

- Look at two studies
  - Originators of k-anonymity
  - De-anonymizing the Netflix dataset

K-anonymity: A model for Protecting Privacy

k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY[1]

LATANYA SWEENEY
School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
E-mail: latanya@cs.cmu.edu

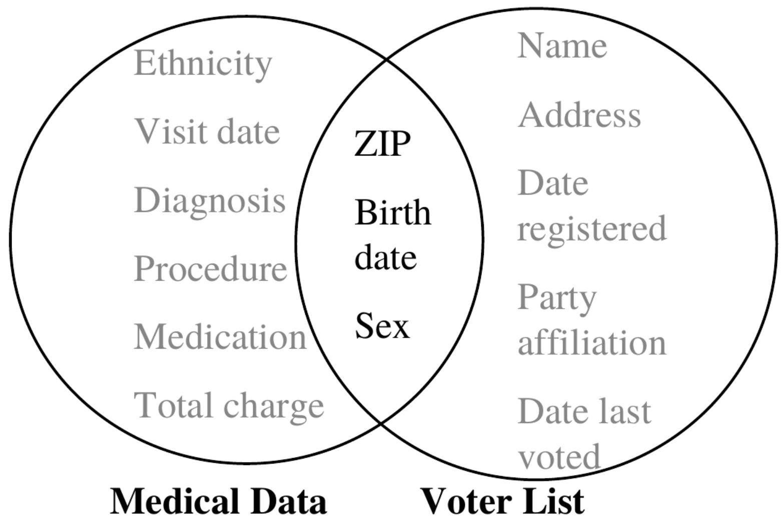
Received May 2002

Consider a data holder, such as a hospital or a bank, that has a privately held collection of person-specific, field structured data. Suppose the data holder wants to share a version of the data with researchers. How can a data holder release a version of its private data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful? The solution provided in this paper includes a formal protection model named k-anonymity and a set of accompanying policies for deployment. A release provides k-anonymity protection if the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appears in the release. This paper also examines re-identification attacks that can be realized on releases that adhere to k-anonymity unless accompanying policies are respected. The k-anonymity protection model is important because it forms the basis on which the real-world systems known as Datafly, μ-Argus and k-Similar provide guarantees of privacy protection.

Keywords: data anonymity, data privacy, re-identification, data fusion, privacy.

Sweeney. K-anonymity: A model for Protecting Privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):557–570, 2002

Re-identification by Linking



Re-identification of Individuals

- William Weld: Governor of MA at the time
- His medical record in the Group Insurance Commission (GIC) data
- Lived in Cambridge, MA
- From the voter list
  - Six people with his particular birth date
  - Three of them male
  - He was the only one in his ZIP code

Statistical Databases

- Data: Person-specific information organized as a table of rows and columns
- Tuple: Corresponds to a row, describes the relationship among the set of values for a person
- Attribute: Corresponds to a column, describes a field or semantic category of information

Quasi-Identifiers

- Attributes that in combination can uniquely identify individuals
- Such as ZIP, gender, and date of birth
- Data owner should identify the quasi-identifier

Sensitive vs Nonsensitive Attributes

Zip Code	Gender	Date of Birth	Medical Condition
**	**	**	**
**	**	**	**

`nonsensitive' (at least individually)

sensitive

Exercise: Column Combinations

- Table with three columns
  - Physician
  - Patient
  - Medication
- Which combinations are sensitive?
  - R(Physician, Patient): Sensitive?
  - R(Physician, Medication): Sensitive?
  - R(Patient, Medication): Sensitive?

K-Anonymity: Formal Definition

- Informally, your information contained in the released dataset cannot be distinguished from at least k-1 other individuals whose information also appear in the dataset
- Formally,
  - Let RT($A_1, \ldots, A_n$) be a table
  - Let QI_{RT} be the quasi-identifier for RT
  - RT satisfies k-anonymity if and only if each sequence of values in RT[QI_{RT}] appears with at least k occurrences

Methods to Achieve K-anonymity

- Suppression: Values replaced with '*'
  - All or some values of a column may be replaced
  - Attributes such as "Name" or "Religion"
- Generalization: Values replaced with a broader category
  - '19' of the attribute "Age" may be replaced with '≤ 20'
  - Replace '23' with '$20 < \text{Age} \leq 30$'

Example K-Anonymous Table

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

- QI = {Race, Birth, Gender, ZIP}
- k = 2

More Examples

Race	ZIP
Asian	02138
Asian	02139
Asian	02141
Asian	02142
Black	02138
Black	02139
Black	02141
Black	02142
White	02138
White	02139
White	02141
White	02142

PT

Race	ZIP
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142

GT1

Race	ZIP
Asian	02130
Asian	02130
Asian	02140
Asian	02140
Black	02130
Black	02130
Black	02140
Black	02140
White	02130
White	02130
White	02140
White	02140

GT2

Exercise: Make This Table 4-anonymous

	Zip code	Age	Nationality	Condition
1	27609	18	Chinese	Heart Disease
2	27615	19	American	Heart Disease
3	26724	50	Indian	Cancer
4	26724	55	Chinese	Heart Disease
5	27615	21	Japanese	Viral Infection
6	26725	47	American	Viral Infection
7	27609	23	American	Viral Infection
8	27609	31	American	Cancer
9	27615	36	Japanese	Cancer
10	26725	49	American	Viral Infection
11	27609	37	Indian	Cancer
12	27615	35	American	Cancer

One Solution

	Zip code	Age	Nationality	Condition
1	276**	<30	*	Heart Disease
2	276**	<30	*	Heart Disease
3	2672*	≧40	*	Cancer
4	2672*	≧40	*	Heart Disease
5	276**	<30	*	Viral Infection
6	2672*	≧40	*	Viral Infection
7	276**	<30	*	Viral Infection
8	276**	3*	*	Cancer
9	276**	3*	*	Cancer
10	2672*	≧40	*	Viral Infection
11	276**	3*	*	Cancer
12	276**	3*	*	Cancer

L-diversity

276**	3*	*	Heart Disease
276**	3*	*	Cancer
276**	3*	*	Viral Infection
276**	3*	*	Flu

Machanavajjhala et al. L-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data, 1(1):1556–4681, 2007

L-diversity Solution

276**	3*	*
276**	3*	*
276**	3*	*
276**	3*	*

Exercise: L-diversity

	Zip code	Age	Nationality	Condition
1	276**	<30	*	Cancer
2	276**	<30	*	Cancer
3	2672*	≧40	*	Flu
4	2672*	≧40	*	Heart Disease
5	276**	<30	*	Heart Disease
6	2672*	≧40	*	Heart Disease
7	276**	<30	*	Heart Disease
8	276**	3*	*	Flu
9	276**	3*	*	Heart Disease
10	2672*	≧40	*	Flu
11	276*	3*	*	Flu
12	276**	3*	*	Heart Disease

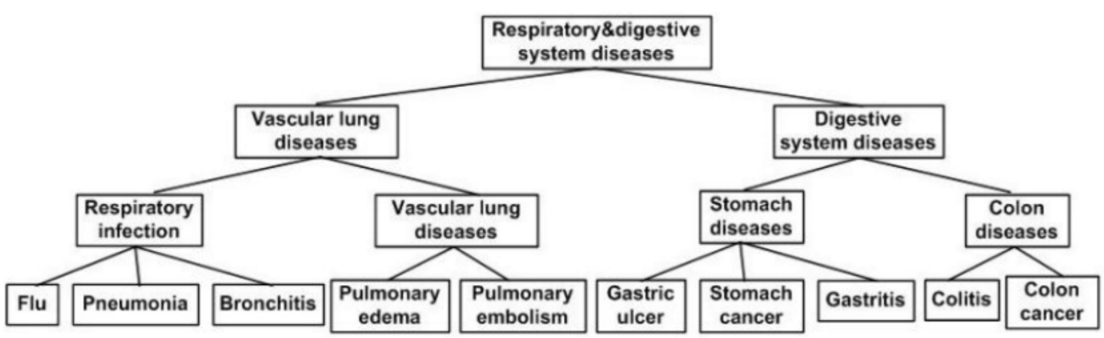
L-diversity Blocks

	Zip code	Age	Nationality	Condition
1	276**	<30	*	Cancer
2	276**	<30	*	Cancer
7	276**	<30	*	Heart Disease
5	276**	<30	*	Heart Disease
3	2672*	≧40	*	Flu
4	2672*	≧40	*	Heart Disease
6	2672*	≧40	*	Heart Disease
10	2672*	≧40	*	Flu
8	276**	3*	*	Flu
9	276**	3*	*	Heart Disease
11	276*	3*	*	Flu
12	276**	3*	*	Heart Disease

L-diversity Concerns

- Some medical conditions are more sensitive than others
- Some medical conditions may indicate same disease

T-closeness



- Measure semantic distance between concepts

Li et al. T-closeness: Privacy beyond k-anonymity and l-diversity. International Conference on Data Engineering, pages 106–115, 2007

Example T-closeness Table

Zip code	Age	Disease
4767*	<40	Gastric ulcer
4767*	<40	Stomach cancer
4767*	<40	Pneumonia
4790*	>39	Gastritis
4790*	>39	Flu
4790*	>39	Bronchitis
2760*	<40	Gastritis
2760*	<40	Bronchitis
2760*	<40	Stomach cancer

Attacks against K-anonymity

- 3 common attacks

- Unsorted matching attack
- Complementary release attack
- Temporal inference attack

Unsorted Matching Attack

- Based on the order of rows in the released datasets
- This problem is often ignored in real-world use
- Easy to correct by randomly sorting the rows

Exercise: Unsorted Matching Attack

Race	ZIP
Asian	02138
Asian	02139
Asian	02141
Asian	02142
Black	02138
Black	02139
Black	02141
Black	02142
White	02138
White	02139
White	02141
White	02142

PT

Race	ZIP
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142

GT1

Race	ZIP
Asian	02130
Asian	02130
Asian	02140
Asian	02140
Black	02130
Black	02130
Black	02140
Black	02140
White	02130
White	02130
White	02140
White	02140

GT2

Complementary Release Attack

Race	BirthDate	Gender	ZIP	Problem
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

PT

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
person	1965	female	0213*	painful eye
person	1965	female	0213*	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	0213*	short of breath
person	1965	female	0213*	hypertension
white	1964	male	0213*	obesity
white	1964	male	0213*	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

GT1

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1960-69	male	02138	short of breath
white	1960-69	human	02139	hypertension
white	1960-69	human	02139	obesity
white	1960-69	human	02139	fever
white	1960-69	male	02138	vomiting
white	1960-69	male	02138	back pain

GT3

Complementary Release Attack: Linked Table

Race	BirthDate	Gender	ZIP	Problem
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

PT

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	02138	short of breath
white	1965	female	02139	hypertension
white	1964	male	02139	obesity
white	1964	male	02139	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

LT

- LT no longer satisfies the k-anonymity requirement

Exercise: Protection for Complementary Releases

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
person	1965	female	0213*	painful eye
person	1965	female	0213*	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	0213*	short of breath
person	1965	female	0213*	hypertension
white	1964	male	0213*	obesity
white	1964	male	0213*	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

GT1

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1960-69	male	02138	short of breath
white	1960-69	human	02139	hypertension
white	1960-69	human	02139	obesity
white	1960-69	human	02139	fever
white	1960-69	male	02138	vomiting
white	1960-69	male	02138	back pain

GT3

- How can you protect against this type of attack?
- $QI_{GT3} = QI \cup \{Problem\}$
- GT1 is the basis of GT3

Temporal Inference Attack

- Data collections are dynamic
- Rows are added, removed, and updated
- Similar to the previous problem of consequent releases

- Let original table be T_0 at time t = 0
- RT_0 is released for T_0 satisfying k-anonymity
- Assume some rows are added to T_0 at time t (becomes T_t)
- RT_t is released for T_t
- Linking RT_0 and RT_t might cause problems

Example: Temporal Inference Attack

Race	BirthDate	Gender	ZIP	Problem
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

PT

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
person	1965	female	0213*	painful eye
person	1965	female	0213*	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	0213*	short of breath
person	1965	female	0213*	hypertension
white	1964	male	0213*	obesity
white	1964	male	0213*	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

GT1

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1960-69	male	02138	short of breath
white	1960-69	human	02139	hypertension
white	1960-69	human	02139	obesity
white	1960-69	human	02139	fever
white	1960-69	male	02138	vomiting
white	1960-69	male	02138	back pain

GT3

Robust De-anonymization of Large Sparse Datasets

Robust De-anonymization of Large Sparse Datasets

Arvind Narayanan and Vitaly Shmatikov
The University of Texas at Austin

Abstract

We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary's background knowledge.

We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world's largest online movie rental service. We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset. Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.

and sparsity. Each record contains many attributes (*i.e.*, columns in a database schema), which can be viewed as dimensions. Sparsity means that for the average record, there are no "similar" records in the multi-dimensional space defined by the attributes. This sparsity is empirically well-established [7, 4, 19] and related to the "fat tail" phenomenon: individual transaction and preference records tend to include statistically rare attributes.

Our contributions. Our first contribution is a formal model for privacy breaches in anonymized micro-data (section 3). We present two definitions, one based on the probability of successful de-anonymization, the other on the amount of information recovered about the target. Unlike previous work [25], we do not assume *a priori* that the adversary's knowledge is limited to a fixed set of "quasi-identifier" attributes. Our model thus encompasses a much broader class of de-anonymization attacks than simple cross-database correlation.

Narayanan and Shmatikov. Robust De-anonymization of Large Sparse Datasets. IEEE Symposium on Security and Privacy, pages 111–125, 2008

Problem: Linking Databases

- De-anonymization attacks
- Linking datasets (public or private) together to gain additional information about users
- Even if sensitive attributes are not contained in the dataset, they can be inferred with high accuracy

AOL Search Data

- In 2006, AOL released 20 million web queries from 650,000 users over a 3 month period
- User names were removed, but there were still connections to user accounts
- New York Times journalists identified some individuals from the search records by cross referencing them with phonebook listings

- Reputation: Incident made it to “101 Dumbest Moments in Business”
- Violation: Lawsuit filed against AOL after a month

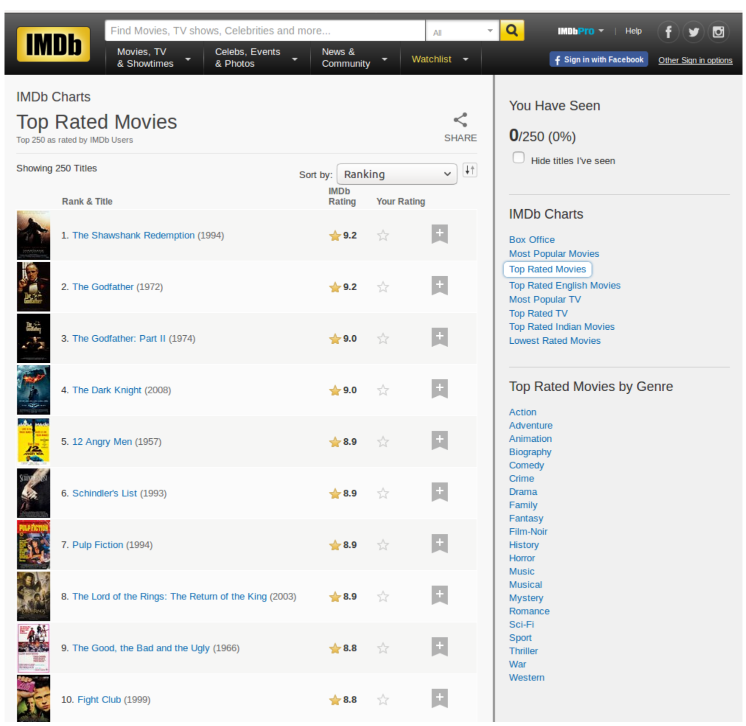
https://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/
http://www.nytimes.com/2006/08/09/technology/09aol.html?_r=0
http://money.cnn.com/magazines/business2/101dumbest/2007/full_list/index.html

Netflix Dataset

- “Anonymous” movie ratings of 480,189 subscribers of Netflix
- 100,480,507 movie ratings
- Between 1999 and 2005
- Less than 1/10 of the entire 2005 database

- Sparsity: Individual rows in the dataset include statistically rare attributes
- Is sparsity enough to identify individual rows?

Public IMDb Ratings



Research Questions

- If the adversary knows a few movies that the user watched, can the adversary learn all the movies that the user watched?
- Can the adversary still identify if only a subset of the original dataset is released?
- Can the adversary still identify if some rows are perturbed?
- Can the adversary still identify in the existence of wrong knowledge about the user?

Assumptions

- Adversary needs some background knowledge about the user
- Movie ratings (only approximately)
- Dates when ratings are entered (with a 14-day error margin)
- Some of that knowledge can be completely wrong
- Develop a “robust” algorithm uniquely identifies a user with high accuracy

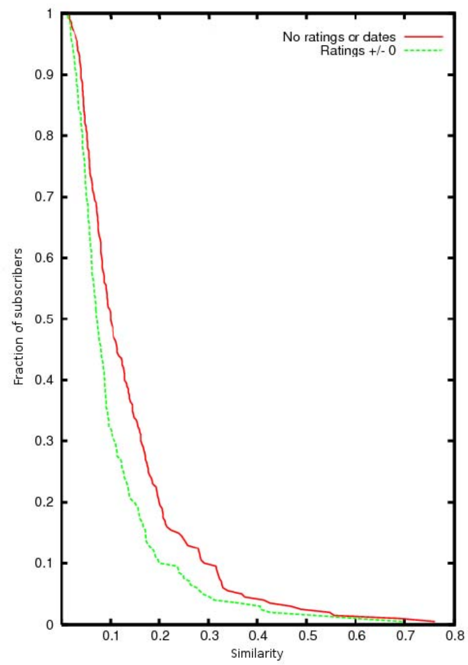
Sparsity

- Each individual row contains values for a tiny fraction of the attributes
- For example, shopping online on Amazon
- Or, rating movies on Netflix

Similarity

- Map a pair of rows (users) to an interval [0, 1]
- supp(r): Support of a row (the set of non-null attributes in a row)
- Sim(r_1, r_2) = $\frac{\sum Sim(r_{1i}, r_{1i})}{|supp(r_1) \cup supp(r_2)|}$
- You can also define similarity for attributes in a similar way, e.g., to compute similarity of a pair of movies

Netflix Dataset Sparsity



De-anonymization

- Adversary model:
  - Sample a row r randomly
  - Give background knowledge to adversary related to r
  - Subset of the supp(r): Might be perturbed or simply wrong
  - Background knowledge chosen arbitrarily
- Adversary objective: Gain as much information about the user's attributes that isn't already known

De-anonymization Algorithm: Inputs and Outputs

- Input: Released subset D′ of database D
- Input: Row r of interest
- Input: Background knowledge Aux related to r

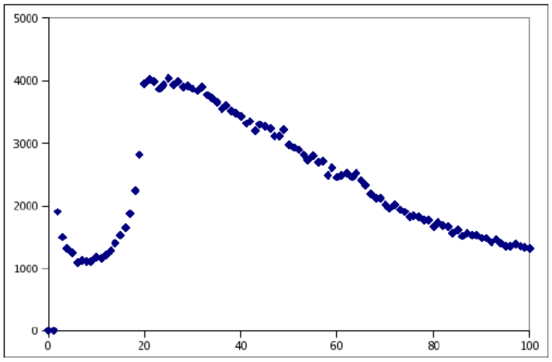
- Output: A row r′, or
- Output: A set of candidate rows with an associated probability distribution

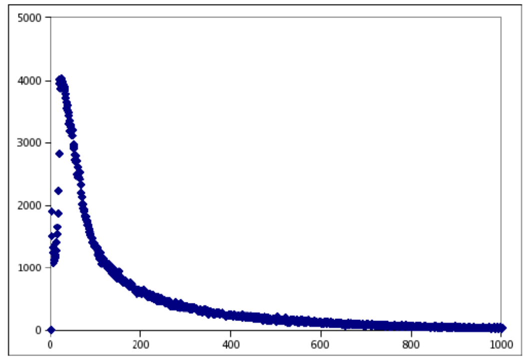
NC STATE UNIVERSITY
TECHNIQUES & STUDIES

De-anonymization Algorithm: Steps

1. Scoring function: Assigns a numerical score to each row in D′ based on how well it matches Aux:
   Compute Score(Aux, r′) for each r′ $\in$ D′
   Score(Aux, r′) = $\min_{i \in supp(Aux)} Sim(Aux_i, r'_i)$
2. Matching criterion: Determine the matching set of rows:
   M = $\{$r′ $\in$ D′: Score(Aux, r′) $> \alpha\}$
3. Row selection: Select one best-guess row or a set of candidates

Netflix Dataset Characteristics





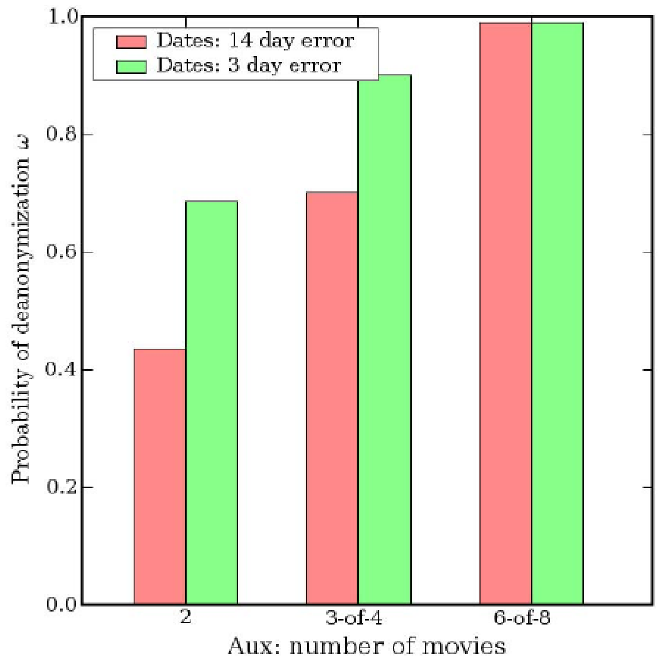
- Number of users with X ratings: X $\leq$ 100, X $\leq$ 1000

Results

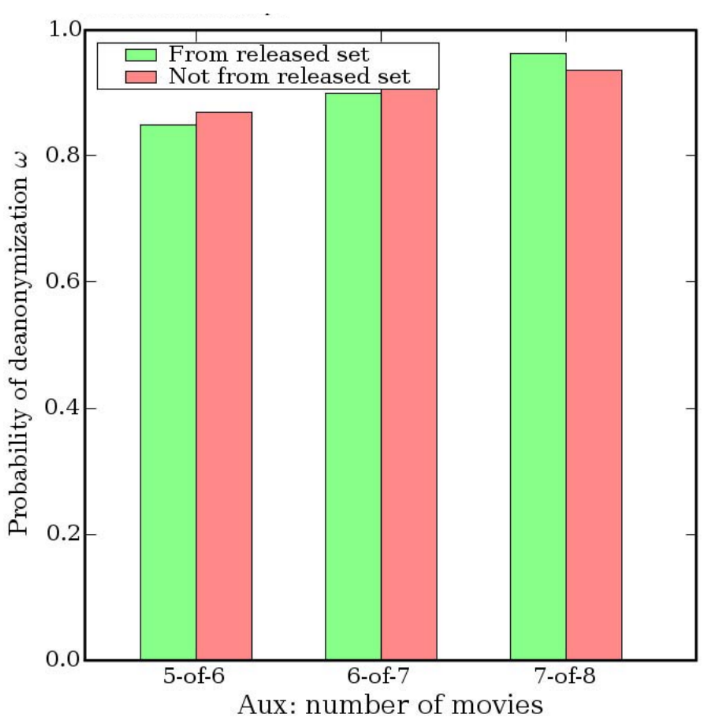
- With 8 movie ratings known (2 of them might be completely wrong)
- And, dates having a 14-day error margin
- 99% of users can be uniquely identified

- With 2 ratings and 3-day error dates, 68% of users

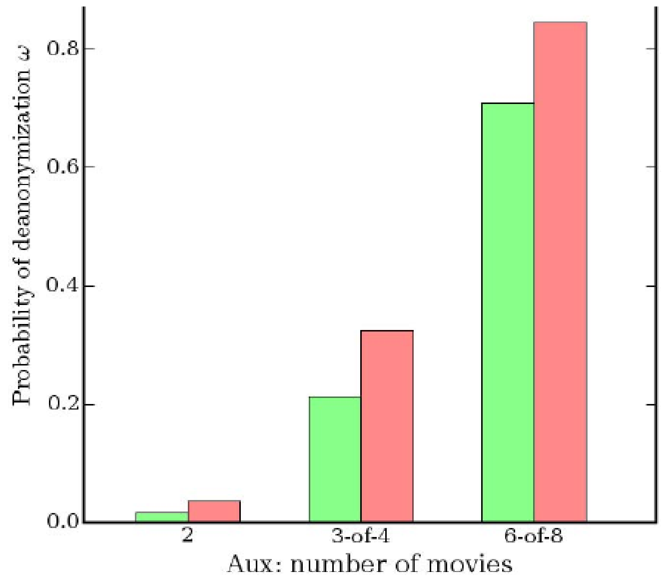
Results: Adversary Knows Exact Ratings



Results: Adversary Must Detect User is Not Present

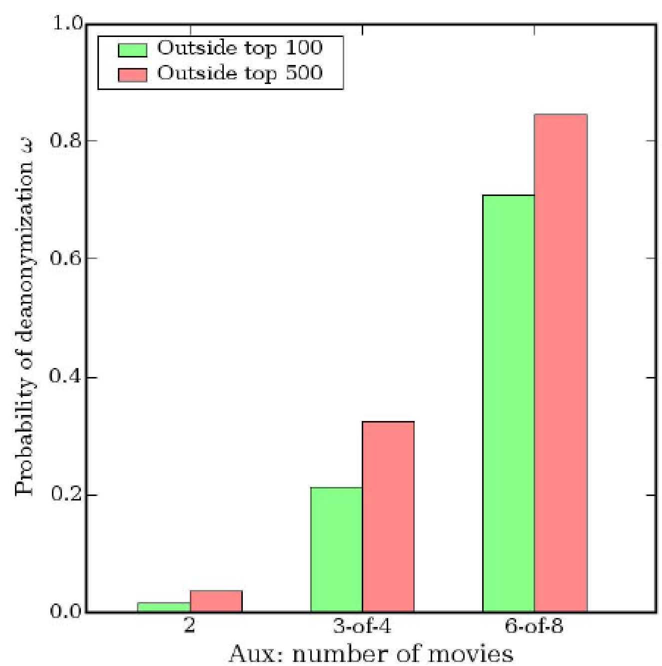


Exercise: What are the Red and Green Bars?

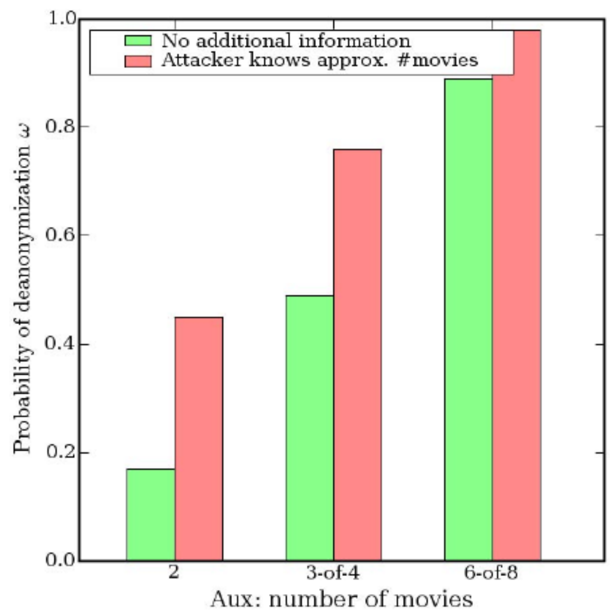


- Adversary only knows the movie ratings

Results: Movie Popularity



Results: Adversary Knows Number of Movies



Implications

- Why would someone who (not anonymously) rates movies on IMDb care about privacy of Netflix ratings?

- Extract entire movie viewing history from Netflix
- Infer political orientation
- Infer religious views

Hulu and Quora Disclosures

- Hulu news article: http://www.reuters.com/article/2013/12/23/us-hulu-privacy-lawsuit-idUSBRE9BM0OJ20131223
- Quora news article: http://techcrunch.com/2012/08/14/after-privacy-uproar-quora-backpedals-and-will-no-longer-show-data-on-what-other-users-have-viewed/
- Links are also on the course website

Things to Look For

- What are the similarities and differences between the two incidents?
- Mitigation (using methods we have seen): Prevention, detection, recovery
- Take 10 minutes to look at the incidents on your own

- Now discuss with your neighbor
- Also take a look at the summary reports
  - Hulu: https://drive.google.com/file/d/0B3m-I0YVAv0EWWhfR2t2YzlDQ1E/view
  - Quora: https://drive.google.com/file/d/0B3m-I0YVAv0EVW4tZjBUdXBHUjA/view