Applied Cryptography and Network Security

Adam J. Lee
adamlee@cs.pitt.edu
6111 Sennott Square

Lecture #27: Data Privacy
April 15, 2014

Announcements

Exam: Saturday, April 26 from 10:00 – 11:50 AM

This Thursday: Exam review

- I'll give a short course wrap up
- You will provide questions for us to discuss

I will hold office hours next week as planned

Outline

What is data privacy? Why should I care?

Models for data privacy

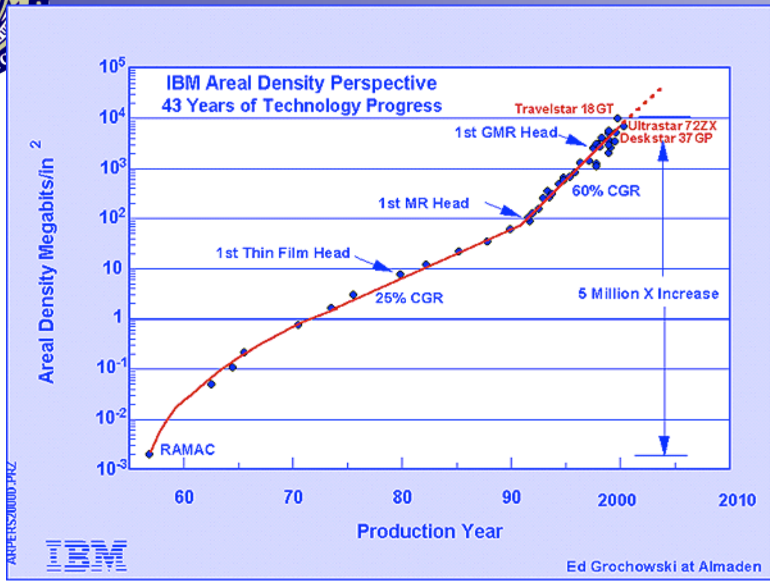
- Anonymize and release
- Mediated query processing
- Outsourced data management

Case study: k-Anonymity

- How does it work?
- Why doesn't it work?

Future directions

Data, data everywhere!



Hard drive sizes are absurd!

- Capacity increasing
- Cost decreasing
- Example: 1TB backup drive costs ~ $100

Thought: Why delete anything? Just as easy to keep it all…

Result: Our whole lives are on disk!

These days, "data" means more than "documents"

- Electronic health records
- Pay as you go car insurance
- Browsing/shopping histories
- Location-based services
- Social networking blunders
- …

Result: Compromise can hurt more than productivity

We can learn a lot from this data

Google: Advertising and search

- Why are Google's services free?
- Because they use your information to intelligently place ads!
- Portions of this data is also available to you (cf. Analytics)

Walmart: Marketing experts

- Over 580 TB in 2006, hosted on 1000 processor system
- Data used to predict/control inventory, coordinate with suppliers, and adjust to local trends

Medical data and imaging

- Medical data mining
- Google flu trends (http://www.google.org/flutrends/)
- Drug and prosthesis design
- …

Widespread data availability is not always a good thing, though…

August 2006: AOL releases search data

- 20,000,000 search keywords
- Over 650,000 users
- 3 months worth of records

Intended use: Learning about search patterns

Result: Records for individual users were recovered!

October 2006: Netflix releases movie rating data

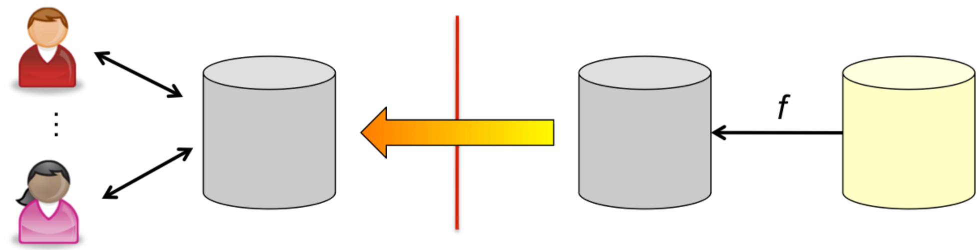
- 100,480,507 ratings that 480,189 users gave to 17,770 movies
- ⟨User, Movie, Date, Rating⟩ tuples

Intended use: Developing and testing new collaborative filtering algorithms

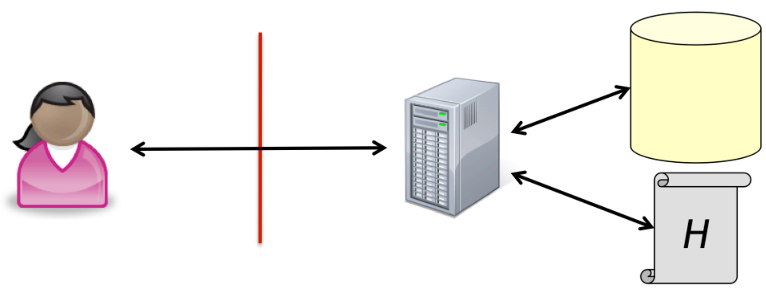
Result: Records for individual users were recovered!

There is a need to balance privacy and availability when releasing data

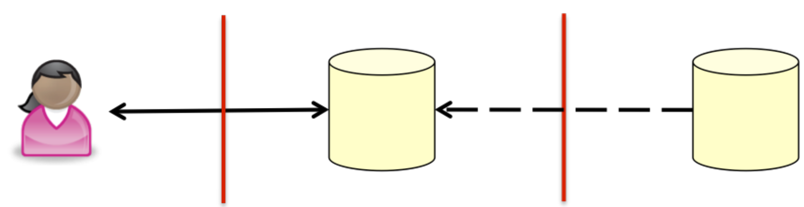
Today, we'll talk about three privacy models for data



Anonymize and Release

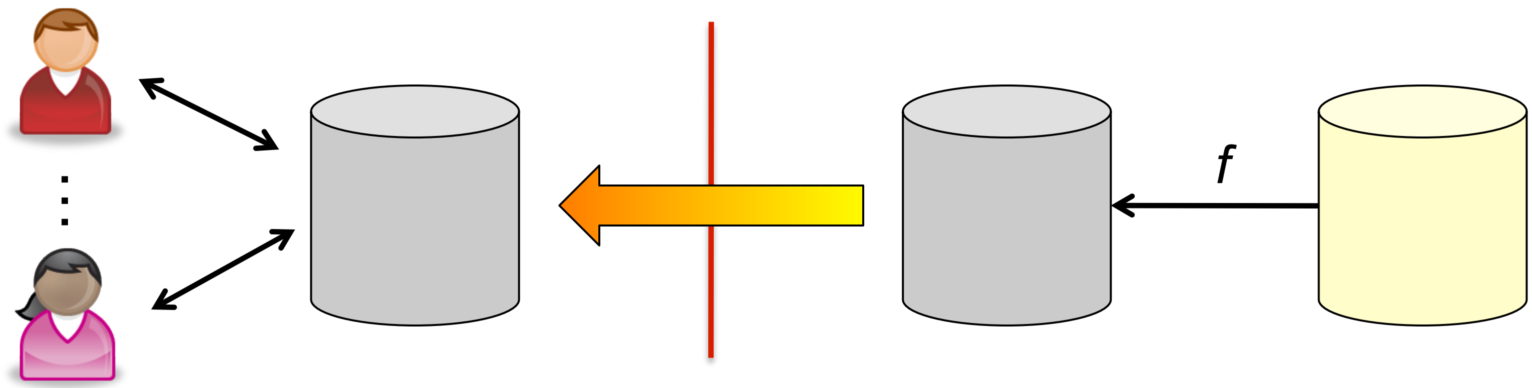


Mediated Query Processing



Outsourced Data Hosting

Anonymize and Release



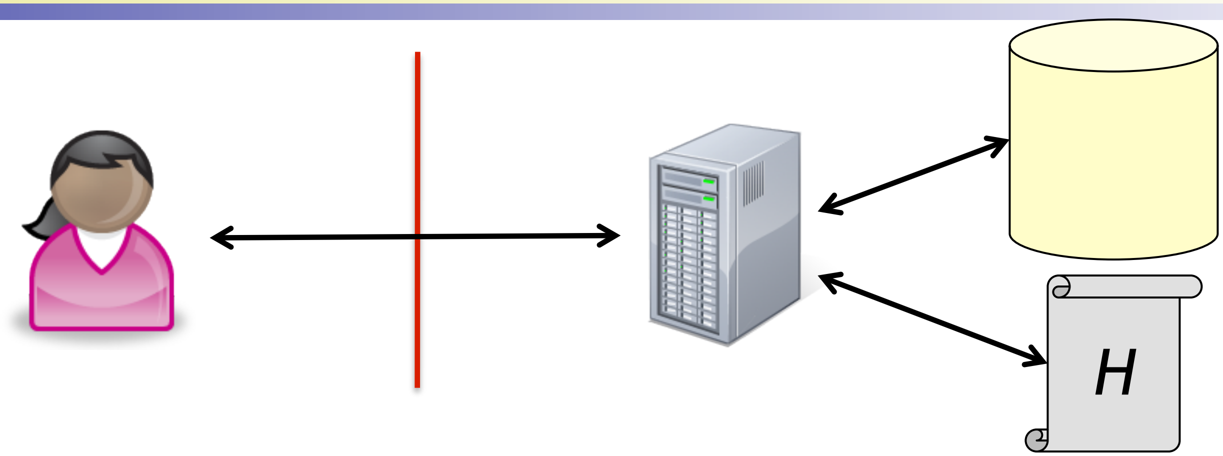
Rather than releasing the original dataset, data providers release a modified version of the dataset to the public/analysts

Data analysts often prefer this model of data release (Why?)

Common operations performed on data include:

- Stripping out names and other identifiers (Suppression)
- Grouping data values into less precise buckets (Generalization)
- Adding noise to records or groups of records (Perturbation)

Mediated Query Processing



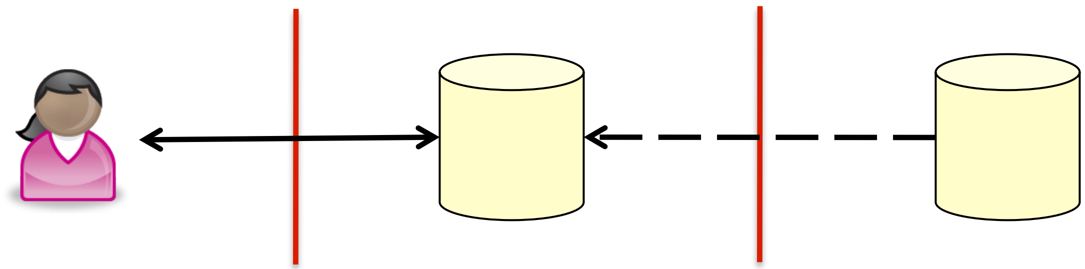
Critical point: Data is not released by the data owner

Since data is retained by the owner, they also retain control

- Do I think that this query is safe to answer?
- What other questions has this querier asked? Should this affect my answer?
- What type of perturbation is needed to make answering this question safe?

This data model is used by the US Census Bureau

Outsourced Data Hosting



Scenario: Let's pay someone else to host our data

- Became popular with the increased prevalence of the web
- Increasingly interesting as cloud computing becomes a reality

Potential uses include offsite backups and outsourced DB management

Depending on the reasons behind outsourcing, a variety of questions deserve some attention:

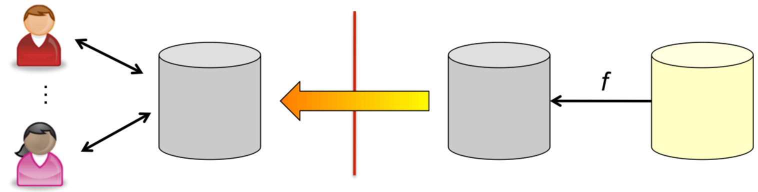
- Should the data host be able to read the data?
- Should the data host be able to learn about the organization of the data?
- Should queries be revealed to the data host?
- Is the data that is claimed to be hosted actually available?

This is currently a very active area of academic research

Question

What do you think are the strengths and weaknesses of each of these three data management scenarios?

Each of these data models has various pros and cons

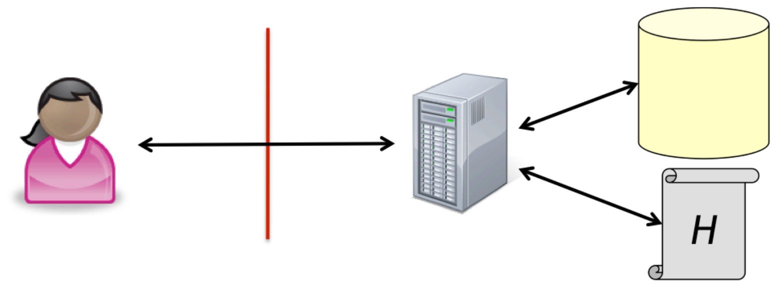


Strengths

- Analysts get (nearly) complete access to data
  - Can explore data in novel/ unpredicted ways
  - Can ask any questions they want
- Providers do not need to host data locally

Weaknesses

- When is data "safe" to release?
- Balancing privacy versus utility?
- Quantifying anonymization?



Strengths

- Analysts can ask many types of queries to the data store
- Providers can see all access to data and can adjust as needed

Weaknesses

- Potentially need to store LOTS of query history
- How should data be perturbed?

Case Study: k-Anonymity

L. Sweeney, "k-anonymity: A Model for Protecting Privacy," International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

The state of the art for protecting privacy in the early 1990s was simply removing "identifiers"

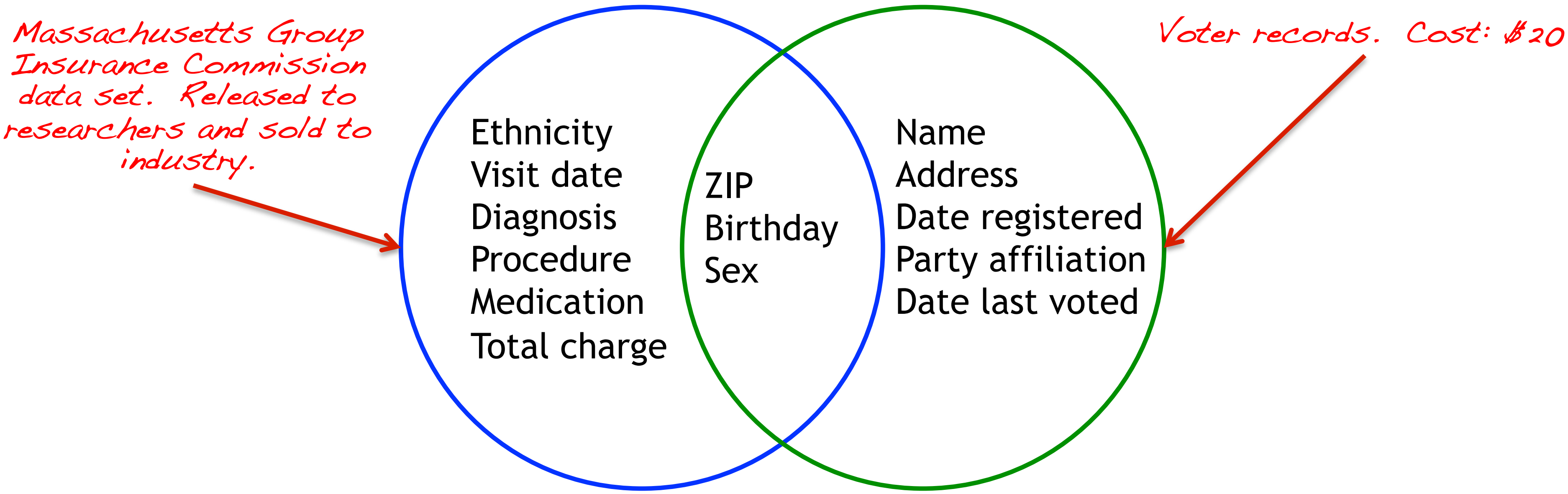
Name	*Race*	*Birth*	*Gender*	*ZIP*	*Problem*
Aaron	Black	1965	M	02145	Short breath
Bob	Black	1965	M	02143	Chest pain
Christina	Black	1965	F	02133	Hypertension
Danielle	Black	1965	F	02137	Hypertension
Eve	Black	1964	F	02137	Obesity
Francine	Black	1964	F	02134	Chest pain
George	White	1964	M	02138	Chest pain
Harry	White	1964	M	02138	Obesity
Ian	White	1964	M	02134	Short breath
James	White	1967	M	02133	Chest pain
Kevin	White	1967	M	02133	Chest pain

Question: Who can see a problem with this?

Answer: Your name is not your only unique identifier!

One example: The triple (City, Birthday, Sex) is a unique identifier for 53% of the population, while (County, Birthday, Sex) identifies 18%

Interesting attack: Reidentifying medical records



After joining these two datasets, Sweeny was able to recover the medical records of William Weld, the (then) governor of Massachusetts!

Some terminology…

Quasi-identifiers

Explicit identifier

Sensitive attribute(s)

Name	*Race*	*Birth*	*Gender*	*ZIP*	*Problem*
Aaron	Black	1965	M	02145	Short breath
Bob	Black	1965	M	02143	Chest pain

…

Steps for "anonymize and release" data processing:

1. Remove all explicit identifiers
2. Manipulate rows to ensure that quasi identifiers cannot be used to map specific individuals to sensitive attributes

Seems easy, right?

k-Anonymity was one of the first rigorously studied anonymization methods

High-level goal: Each unique quasi-identifier should appear at least *k* times in the released data set

This provides a sort of plausible deniability…

How can we accomplish this goal?

- Attribute generalization
- Attribute suppression
- Attribute perturbation

Race	***Birth***	***Gender***	***ZIP***	***Problem***
Black	1965	M	02145	Short breath

Black	1967	*	021**	Short breath

Let's see an example…

This is a 2-anonymous version of our hospital data table example

Race	*Birth*	*Gender*	*ZIP*	*Problem*
Black	1965	M	0214*	Short breath
Black	1965	M	0214*	Chest pain
Black	1965	F	0213*	Hypertension
Black	1965	F	0213*	Hypertension
Black	1964	F	0213*	Obesity
Black	1964	F	0213*	Chest pain
White	1964	M	0213*	Chest pain
White	1964	M	0213*	Obesity
White	1964	M	0213*	Short breath
White	1967	M	0213*	Chest pain
White	1967	M	0213*	Chest pain

Question: Why is this table 2-anonymous?

- Each quasi-identifier appears (at least) 2 times

Question: Is the following table 3-anonymous?

Race	*Birth*	*Gender*	*ZIP*	*Problem*
Black	1965	M	0214*	Short breath
Black	1965	M	0214*	Chest pain
Black	1965	M	0214*	Hypertension
Black	1965	F	0213*	Hypertension
Black	1964	F	0213*	Obesity
Black	1964	F	0213*	Chest pain
White	1964	M	0213*	Chest pain
White	1964	M	0213*	Obesity
White	1964	M	0213*	Short breath

k-Anonymity sounds great! So the data anonymization problem is solved, right?

Problem 1: Solving the anonymization quality/efficiency trade-off

This table is 5-anonymous, but useless!

Race	*Birth*	*Gender*	*ZIP*	*Problem*
*	19**	*	*	Short breath
*	19**	*	*	Chest pain
*	19**	*	*	Hypertension
*	19**	*	*	Hypertension
*	19**	*	*	Obesity

Question: How can we define the "goodness" of a dataset?

- Less suppression/generalization/perturbation → better quality

Fact: Finding an optimal k-anonymization is an NP-Hard problem

Fortunately, heuristic methods do a pretty good job of this with fairly low overheads (see work by LeFevre et al.)

Efficiency is solved, but what other problems are there?

Problem 2: How do we choose the value of k to use?

Essentially, there is no good answer to this question…

- How much better is 3-anonymity than 2-anonymity?
- Is the same value of k reasonable for all individuals in the dataset?
- How much does adjusting k impact the quality of the released data?

More problems still…

Scenario: Bob has a record in the dataset, was born in the 1960s, and lives in the 15260 ZIP code

Race	*Birth*	*Gender*	*ZIP*	*Problem*
Black	196*	M	15260	Brain cancer
Black	196*	M	15260	Brain cancer
Black	196*	M	15260	Brain cancer
Black	196*	M	15260	Brain cancer

This is a 4-anonymous table, but Bob has brain cancer…

Race	*Birth*	*Gender*	*ZIP*	*Problem*
Black	196*	M	15260	Brain cancer
Black	196*	M	15260	Lung cancer
Black	196*	M	15260	Leukemia
Black	196*	M	15260	Bone cancer

This is a 4-anonymous table, but Bob has cancer…

A generalization of this problem…

These problems emerge because the data set was not diverse

- All entries for a quasi-identifier map to the same sensitive attribute
- All entries for a quasi-identifier map to related sensitive attributes

Follow on work addresses this, but is subject to attacks of its own!

The bigger problem is that this class of solutions does not adequately model the knowledge of the attacker

- I know that Bob visited the hospital and should be in this data set
- I know that Bob has some type of cancer
- …

Recent work on differential privacy works for any attacker, but uses the mediated query model

In short, this is still a very active research area

Conclusions

Today we talked about three types of models for managing private data

- Anonymize and release
- Mediated query processing
- Outsourced data hosting

k-Anonymity is one solution in the “anonymize and release” model

- Strip out explicit identifiers
- Be sure that each quasi-identifier appears at least k times

How do we manage diversity and model attacker knowledge?

- Recent work does this with limited success

Take away point: Data anonymization is hard. “Anonymization” is probably a flawed term, as it is hard to quantify…

Next time: Wrap up!