Differential Privacy

Amrita Roy Chowdhury

Slides modified from Vitaly Schmatikov, Katrina Ligett

Insta

hulu

NETFLIX

amazon®

23andMe

deCODE genetics

ZEDO

exelate

bluekai

PubMatic

patientslikeme®

What about my privacy?

"We do not collect **personally identifiable information**"

National Institute of
Standards and Technology
U.S. Department of Commerce

Special Publication 800-122

Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)

Recommendations of the National Institute of Standards and Technology

Erika McCallister
Tim Grance
Karen Scarfone

HIPAA

Health Insurance Portability and Accountability Act

- HIPAA "Safe Harbor" De-Identification of Medical Record Information

- Remove 18 specified PIIs from data

NIH
National Institute
of Mental Health

United States
Census
Bureau

Aol.

x_1	…	
x_2	…	
x_3		
…		
x_n		

Dataset

"De-identification"
"Anonymization"

Remove PII

Privacy

Anonymity!

- Personal Identifiable Information – PII
- Quasi Identifiers
- Identifiers vs Sensitive Attribute

PII	QID			SA
Name	Zipcode	Age	Sex	Disease
Alice	47627	59	F	Prostate Cancer
Bob	47621	52	M	Ovarian Cancer
Charles	47624	35	M	Flu
Dave	47630	43	M	Heart Disease
Eve	47650	37	F	Heart Disease

PII

- NIST - “any information about an individual maintained by an agency,… that can be used to distinguish or trace an individual's identity…”
- Name, SSN, Credit Card, Full address, Phone Number
- Legal concept- not a technical one

Quasi Identifiers

- Attributes that may not be uniquely identifying on their own, any attribute can be potentially identifying in combination with others
- Age, Gender, 5 digit Zipcode

Sensitive Attributes

- Medical records, salaries, etc
- These attributes is what the researchers need, so they are released unmodified

PII	QID			SA
Name	Zipcode	Age	Sex	Disease
Alice	47627	59	F	Ovarian Cancer
Bruce	47621	52	M	Prostate Cancer
Charles	47624	35	M	Flu
Dave	47630	43	M	Heart Disease
Eve	47650	37	F	Heart Disease

Not Private

PII	QID			SA
Name	Zipcode	Age	Sex	Disease
Alice	47627	59	F	Ovarian Cancer
Bruce	47621	52	M	Heart Disease
Charles	47624	35	M	Flu
Dave	47630	43	M	Heart Disease
Eve	47650	37	F	Heart Disease

Private

QID			SA
Zipcode	Age	Sex	Disease
47627	59	F	Ovarian Cancer
47621	52	M	Heart Disease
47624	35	M	Flu
47630	43	M	Heart Disease
47650	37	F	Heart Disease


PRIVACY

Some Privacy Disasters

Netflix Settles Privacy Lawsuit, Cancels Prize Sequel

Taylor Buley Contributor
The Firewall Contributor Group
News developer, in all senses of the phrase

On Friday, Netflix announced on its corporate blog that it has settled a lawsuit related to its Netflix Prize, a $1 million contest that challenged machine learning experts to use Netflix's data to produce better recommendations than the movie giant could serve up themselves.

The Privacy Report Blog Contributors About Contact

OCTOBER 28, 2009

Back to the Future: NIH to Revisit Genomic Data-Sharing Policy

By: Dan Vorhaus
Category: Genomics
Topic: Database of Genotypes and Phenotypes, dbGaP, genetic privacy, GenomeWeb, GWAS, Kaiser Permanente, NHLBI, NIH, Personal Genome Project, PLoS Genetics, WGS

TECHNOLOGY

Harvard Researchers Accused of Breaching Students' Privacy

Social-network project shows promise and peril of doing social science online

AOL Proudly Releases Massive Amounts of Private Data

Michael Arrington @arrington?lang=en / 13 years ago Comment

Yet Another Update: AOL: "This was a screw up"

Microdata

QID			SA
Zipcode	Age	Sex	Disease
47627	59	F	Ovarian Cancer
47621	52	M	Prostate Cancer
47624	35	M	Flu
47630	43	M	Heart Disease
47650	37	F	Heart Disease

Voter Registration Data

Name	Zipcode	Age	Sex
Alice	47627	59	F
Bruce	43756	35	M
Carol	47677	42	F
Dan	47632	47	M
Ellen	42789	23	F

Latanya Sweeney's Attack (1997)

Massachusetts hospital discharge dataset

Medical Data Released as Anonymous

SSN	Name	Ethnicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
			09/27/64	female	02139	divorced	hypertension
			09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party	....
....	....	....	....	....	....	....	
....	....	....	....	....	....	....	
Sue J. Carlson	1459 Main St.	Cambridge	*02142*	*9/15/61*	*female*	democrat	....
....	....	....	....	....	....	....	

Figure 1: Re-identifying anonymous data by linking to external data

Public voter dataset


NETFLIX
Netflix Prize
Home Rules Leaderboard Register Update Submit Download
NETFLIX
Browse Recommendations Friends Queue Buy DVDs
Movies For You
You really liked it...
The Big One
Welco
The Netflix Prize seeks to substantially improve the accuracy of predictions about how much someone is going to love a movie based on their movie preferences. Improve it enough and you win one (or more) Prizes. Winning the Netflix Prize improves our ability to connect people to the movies they love.
Read the Rules to see what is required to win the Prizes. If you are interested in joining the quest, you should register a team.
You should also read the frequently-asked questions about the Prize. And check out how various teams are doing on the Leaderboard.
Good luck and thanks for helping!


DATA BREACH
DATA BREAC
DATA BREACH
© CanStockPhoto.com - csp48100125

AOL User 4417749

- AOL query logs have the form
- <AnonID, Query, Query Time, ItemRank, ClickURL<truncatedURL>
- Sample queries of user with AnonID 4417749: – “numb fingers”, “60 single men”, “dog that urinates on everything”, “landscapers in Lilburn, GA”, several people with the last name Arnold •
- Only 14 citizens with the last name Arnold near Lilburn
- NYT contacted the 14 citizens, found out AOL User 4417749 is 62-year-old Thelma Arnold

Lesson Learnt
PII is technically meaningless

PII is info “with respect to which there is a reasonable basis to believe the information can be used to identify the individual.”

HIPAA
Health Insurance Portability and Accountability Act

- Any piece of data can be used for re-identification

Narayanan, Shmatikov
CACM column, 2010

What analysis can we do?

- Limit to analysis on large population
- Aggregate Statistics
- Reveal ordinary facts
- All of the above susceptible to leakage

What should we guarantee?

- Output should not reveal anything about an individual that could not have been learnt without access to the input

Is this possible?

Privacy/Utility Tradeoff

What can we guarantee?

- Output should not reveal anything significantly more about an individual than what could have been learned from the same analysis by omitting the individual’s data from the input database

What can we guarantee?

- Think of the output to be randomized
- Promise to individual – if you leave the database the output does not change by much
- Incentive for individual data owners – since output does not change by much whether you participate, might as well give your data

Statistical Database Model

- X = Set of all possible rows for a person
- Database x is a set of rows in $\mathbb{N}^{|X|}$, i.e., a histogram representation

Analysts Objective

- Wants to compute some statistics on D $\in \mathbb{N}^{|X|}$
- Preserve privacy of individuals
- Find a randomized mapping from D to some output space such that it masks small changes in D

What is Differential Privacy?

Neighboring Datasets

Two datasets D_1 and D_2 are defined to be neighboring datasets if they differ in a single row

$$||D_1 - D_2|| \leq 1$$

$$D_1, D_2 \in \mathbb{N}^{|X|}$$

What is Differential Privacy?

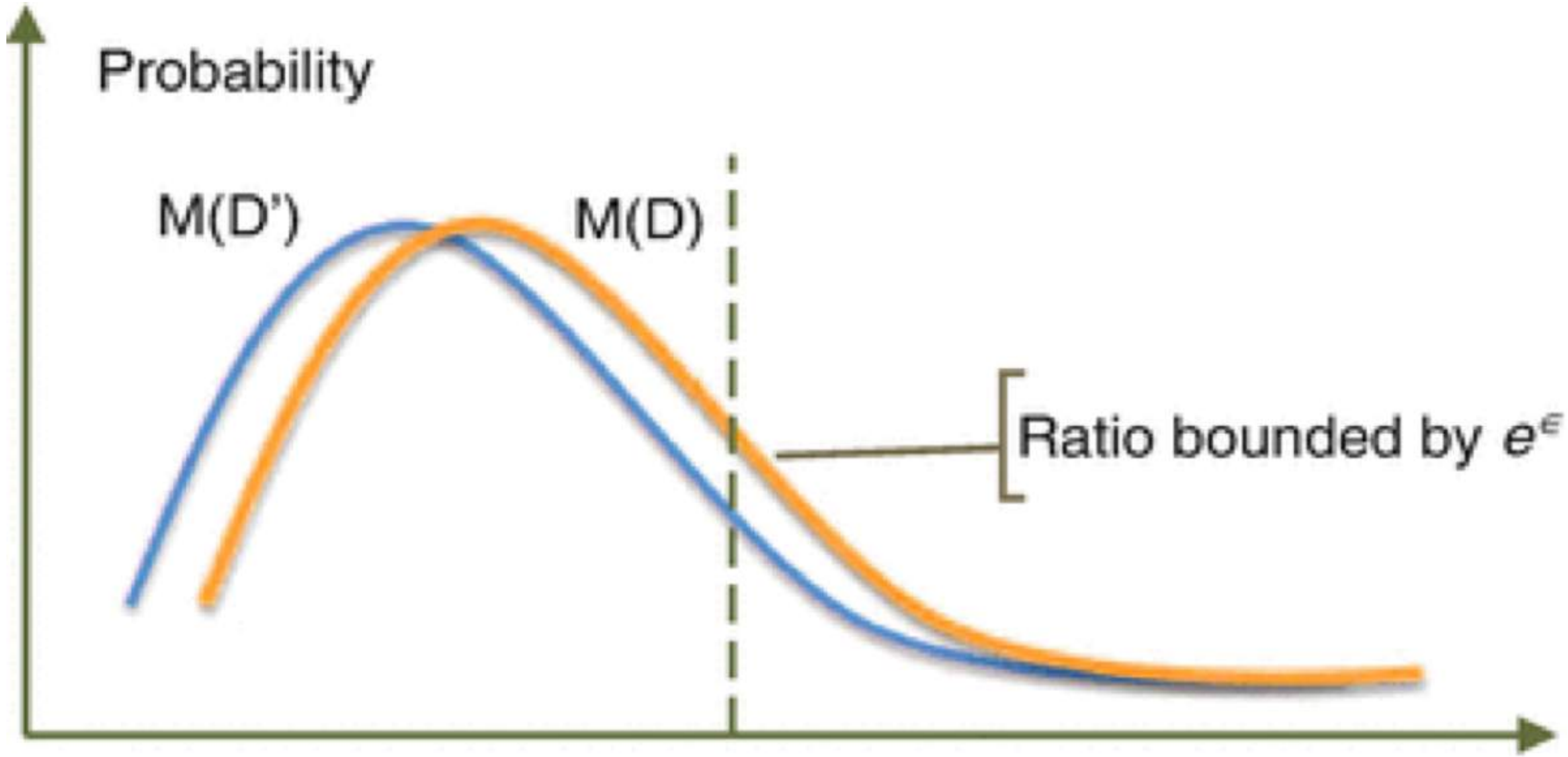
Algorithm $\mathcal{A}$ is ε - differentially private if, for all output $\mathcal{S} \subseteq Range(\mathcal{A})$ and two databases D_1 and D_2 such that they differ only in a single row

$$Prob\big(\mathcal{S} \in \mathcal{A}(D_1)\big) \leq e^{\varepsilon}\, Prob\big(\mathcal{S} \in \mathcal{A}(D_2)\big)$$

$$e^{\varepsilon} \sim (1 + \varepsilon)$$

What is Differential Privacy?

- Blue Line – Probability to receive certain output t given D’
- Orange Line - Probability to receive certain output t given D
- D and D’ are neighboring datasets



What is Differential Privacy?

- Is a statistical property of the mechanism
- Many ways to implement it with same privacy guarantee but different utility
- Independent of the adversary’s computational power
- Unaffected by any auxiliary information

Approximate Differential Privacy

Algorithm $\mathcal{A}$ is (ε, δ) - differentially private if, for all output $\mathcal{S} \subseteq Range(\mathcal{A})$ and two databases D_1 and D_2 such that they differ only in a single row

$$Prob\big(\mathcal{S} \in \mathcal{A}(D_1)\big) \leq e^{\varepsilon} Prob\big(\mathcal{S} \in \mathcal{A}(D_2)\big) + \delta$$

chrome


We always do the same thing.
Fine. You wanna go skydiving?
No way.
You're not easy to please.
I just want to veg out and watch movies.
Like we always do. I love us.
Beats
Cancel
PINTEREST
Show More
Beats Pill+ Speaker - White
Beats Solo2 Wireless On-Ear Headphones - Black
Beats Studio Wireless Over-Ear Headphones - Red
MUSIC
Search in App
I'm coming to get you now.
Hey. Are you thinking what I'm thinking?
I'm starving. Let's get
Differential privacy


Apple will not
see your data

United States Census 2020

The U.S. Census Bureau Adopts Differential Privacy

John M. Abowd, *U.S. Census Bureau*

Publication Date

8-2018

Abstract

The U.S. Census Bureau announced, via its Scientific Advisory Comm
that it would protect the publications of the 2018 End-to-End Census
(E2E) using differential privacy. The E2E test is a dress rehearsal for
2020 Census, the constitutionally mandated enumeration of the popu
used to reapportion the House of Representatives and redraw every
legislative district in the country. Systems that perform successfully i

About Epsilon and Delta

- Does higher delta mean better privacy?
- Does lower epsilon mean better privacy?

Randomized Response

- Q: Have you ever broken the law?
- A: Yes / No

- Randomize the response

Randomized Response Cntd

- Flip a coin
- If it is a head, then report truthfully
- Else, flip a second coin – responds “Yes” if Head , “No” if Tail

Randomized Response Cntd

Claim – Randomized Response is (ln3,0) – DP

Proof –

$$\frac{\Pr[\text{Response=YES}\,|\,\text{Truth=YES}]}{\Pr[\text{Response=YES}\,|\,\text{Truth=NO}]}$$

$$=\frac{3/4}{1/4}$$

$$=\frac{\Pr[\text{Response=NO}\,|\,\text{Truth=NO}]}{\Pr[\text{Response=NO}\,|\,\text{Truth=YES}]}$$

$$= 3$$

Sensitivity

$$\Delta f = \text{max}_{\text{D1,D2}} || f(D_1) - f(D_2) ||_1$$

Measures how much a single record can affect the output

Sensitivity Cntd

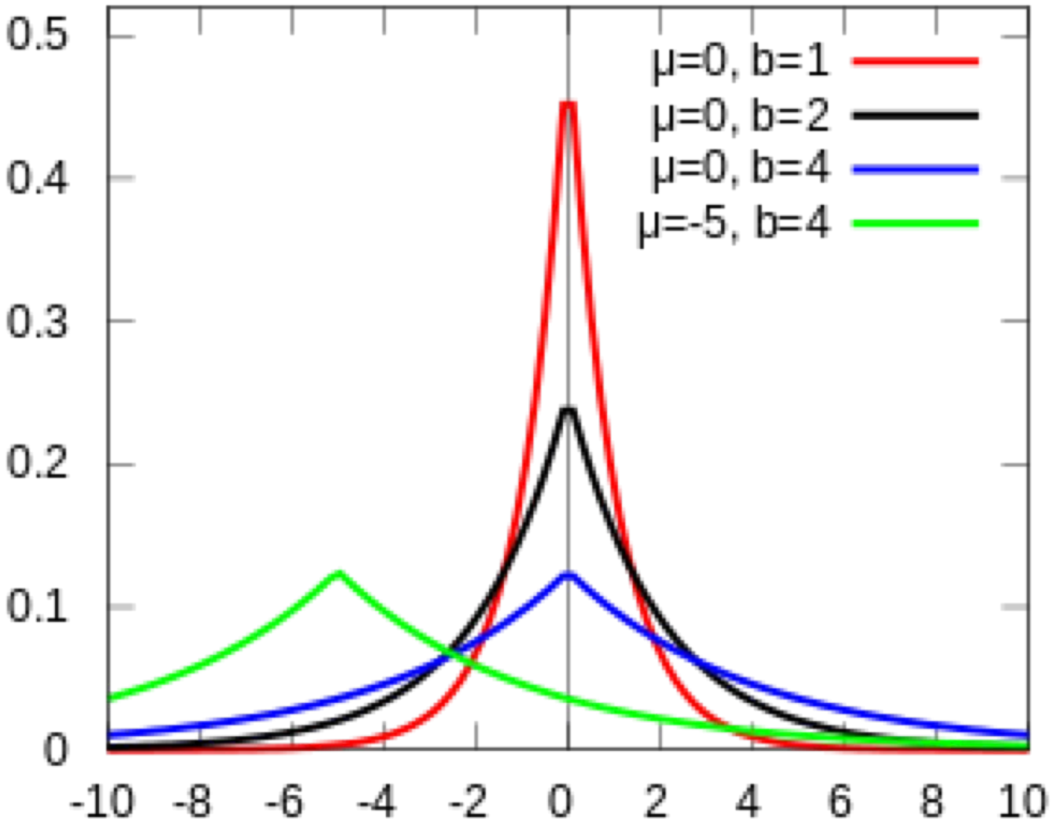
- Counting Queries
- Number of people in the database satisfying a predicate P
- Sensitivity = 1

- Sum Query
- Find the sum of the ages of the people in the database where Age [1,100]
- Sensitivity = 100

- Histogram Query
- Output the Age histogram
- Sensitivity = 1

Laplace Distribution

- Double exponentian
- Two parameters μ and b
- PDF(x)=$\frac{1}{2b}$exp(|x-μ|)
- Variance = $2b^2$
- Y$\sim Lap$(b), Pr[Y$\geq$ bt]=exp(-t)



Laplace Mechanism

Given $f: D \rightarrow Rk$, a $\epsilon -$ differentially private mechanism $\mathcal{M}$ publishes

$$f(D) + [Lap(\frac{\Delta f}{\epsilon})]^{\mathrm{k}}$$

Examples

Linear Query

- How many people with Age in [40,50] who watch Powerpuff Girls?
- Sensitivity = 1
- Add noise from $Lap(\frac{1}{\epsilon})$

Group Privacy

- Thm: Any $(\varepsilon, 0)$ – DP algorithm $\mathcal{A}$ is also $(\mathrm{k}\varepsilon, 0)$ – DP for groups of size k, i.e., for all

$$||D_1\text{-}D_2|| \leq \mathrm{k}$$

$$D_1, D_2 \in \mathbb{N}^{|X|}$$

And for all $\mathcal{S} \subseteq Range(\mathcal{A})$

$$Prob\big(\mathcal{S} \in \mathcal{A}(D_1)\big) \leq \mathrm{e}^{k\varepsilon} Prob\big(\mathcal{S} \in \mathcal{A}(D_2)\big)$$

Post Processing

- Thm: Let $\mathcal{A}$: $\mathbb{N}^{X|} \rightarrow \mathbb{R}$ be a ε – DP algorithm. Let f: $\mathbb{R} \rightarrow \mathbb{R}'$ be a randomized mapping. Then f $\circ$ $\mathcal{A}$ also satisfies ε – DP .

Composition

Thm- For i $\in [k]$, let $\mathcal{A}_{\mathrm{i}}$: $\mathbb{N}^{X|} \rightarrow \mathbb{R}_{\mathrm{i}}$ be ε_i –DP. Then the mechanism ($\mathcal{A}_1$(D),…, $\mathcal{A}_{\mathrm{k}}$(D)) is $\sum_i \varepsilon_i$- DP.

"Advanced" version available too