


# Special Topic: Differential Privacy

Chao Lan

Does user privacy matter?

A photograph of Mark Zuckerberg on a stage, wearing a grey t-shirt. In the foreground, a hand holds a smartphone displaying a Facebook profile page. The background is a blue wall with a large, out-of-focus Facebook logo. A text overlay is positioned in the upper right area of the image.

**MARK ZUCKERBERG**

**APOLOGIZES FOR FACEBOOK USERS' PRIVACY BREACH**

# Facebook Reaches Settlement with ICO over £500,000 Data Protection Fine

Posted on November 5, 2019

POSTED IN [ENFORCEMENT](#), [INTERNATIONAL](#), [ONLINE PRIVACY](#)

On October 30, 2019, Facebook reached a settlement with the UK Information Commissioner's Office ("ICO") under which it agreed to pay (without admission of liability) the £500,000 fine imposed by the ICO in 2018 in relation to the processing and sharing of its users' personal data with Cambridge Analytica.

# Apple's 'differential privacy' still collects too much specific data, study says

By Roger Fingas

Friday, September 15, 2017, 02:49 pm PT (05:49 pm ET)

Apple's use of "differential privacy" — a method that inserts random noise into data as it's collected en masse — doesn't go far enough to protect personal information, a study suggested this week.



# Google's Differential Privacy May be Better Than Apple's



Andrew Orr  
@andrewornot

🕒 3 minute read

Sep 15th, 2017 2:51 PM EDT | Analysis

---

As it turns out, Google's version of differential privacy may be more private than Apple's implementation. Writing for *Wired*, Andy Greenberg [talks about](#) differential privacy. Specifically, about a [study](#) [PDF] that examines how Apple uses differential privacy in macOS and iOS. The researchers found that it might not be as private as Apple would have us believe.



# THE CALIFORNIA CONSUMER PRIVACY ACT (CCPA)

AN IMPLEMENTATION GUIDE





## Example GDPR Rules

Ch2.A9: “Processing of personal data revealing racial or ethnic origin, political opinions,... or sexual orientation shall be prohibited...”.

Ch3.A17: “data subject shall have the right to obtain... the erasure of personal data concerning him or her... and the controller shall... erase personal data...”



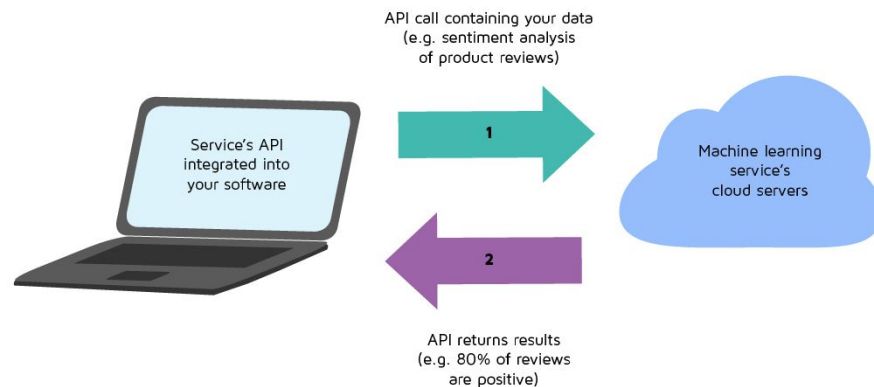
**The EU General Data Protection  
Regulation (GDPR) is the most  
important change in data privacy  
regulation in 20 years.**

**The regulation will fundamentally reshape the way in which data is handled across every sector, from healthcare to banking and beyond.**



# How is privacy related to data analytics?

## Machine Learning as a Service



Amazon ML

Microsoft Azure ML  
Studio

Google Prediction  
API

IBM Watson ML  
Model Builder

## Data Sharing



## Data Publishing



# Traditional Privacy Protection

Protect user privacy in a publicly queryable database.

			Lie, Cheat,	Behavior	Peer	Low Academic	Negative	Aggressive	Total	
Student Name	Student ID	Steal	Sneak	Problem	Rejection	Achievement	Attitude	Behavior	SRSS	GPA
Angel, Julio	2310	0	0	0	2	0	0	1	3	3.1
Akins, J'Monte	2013	0	0	0	0	0	0	0	0	4.1
Backer, Brent	2031	0	1	2	1	2	2	1	9	2.3
Boxwell, Kylie	2001	0	0	0	1	1	0	0	2	2.5
Cartright, Ashley	2152	0	1	1	1	0	1	0	4	3.2
Cox, Lucille	2002	0	0	0	0	0	0	0	0	3.9
Hankins, Erin	2017	0	0	0	0	0	2	0	2	3.7
Illio, Helen	2132	0	0	0	0	0	0	0	0	2.9
Jackson, Ronald	2003	0	1	2	2	3	2	2	12	1.7
Kemp, Patrice	2009	0	0	1	0	0	0	0	1	3.3
Parker, Stephanie	2004	0	0	0	0	1	2	0	3	2.7
Reed, Kent	2010	0	0	0	0	0	0	0	0	3.6
Sterling, Michael	2022	0	0	1	0	3	1	1	6	2.4
Thomas, James	2018	0	0	0	0	0	0	0	0	3.8
Walsh, Carter	2215	0	0	0	1	0	1	0	2	3.5

Q1: will a query reveal my identity?

			Lie, Cheat, Sneak	Behavior Problem	Peer Rejection	Low Academic Achievement	Negative Attitude	Aggressive Behavior	Total SRSS	GPA
Student Name	Student ID	Steal								
Angel, Julio	2310	0	0	0	2	0	0	1	3	3.1
Akins, J'Monte	2013	0	0	0	0	0	0	0	0	4.1
Backer, Brent	2031	0	1	2	1	2	2	1	9	2.3
Boxwell, Kylie	2001	0	0	0	1	1	0	0	2	2.5
Cartright, Ashley	2152	0	1	1	1	0	1	0	4	3.2
Cox, Lucille	2002	0	0	0	0	0	0	0	0	3.9
Hankins, Erin	2017	0	0	0	0	0	2	0	2	3.7
Illio, Helen	2132	0	0	0	0	0	0	0	0	2.9
Jackson, Ronald	2003	0	1	2	2	3	2	2	12	1.7
Kemp, Patrice	2009	0	0	1	0	0	0	0	1	3.3
Parker, Stephanie	2004	0	0	0	0	1	2	0	3	2.7
Reed, Kent	2010	0	0	0	0	0	0	0	0	3.6
Sterling, Michael	2022	0	0	1	0	3	1	1	6	2.4
Thomas, James	2018	0	0	0	0	0	0	0	0	3.8
Walsh, Carter	2215	0	0	0	1	0	1	0	2	3.5

Anonymization: hide user identity. Is it sufficient?

TABLE I. MICRODATA

ID	Attributes			
	<i>Age</i>	<i>Sex</i>	<i>Zip code</i>	<i>Disease</i>
1	26	M	83661	Headache
2	24	M	83634	Headache
3	31	M	83967	Toothache
4	39	F	83949	Cough

If we have another public database with overlapping attributes...

TABLE I. MICRODATA

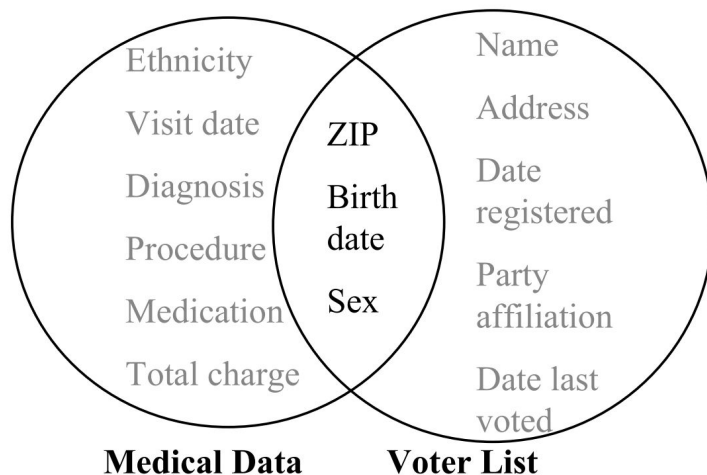
ID	Attributes			
	<i>Age</i>	<i>Sex</i>	<i>Zip code</i>	<i>Disease</i>
1	26	M	83661	Headache
2	24	M	83634	Headache
3	31	M	83967	Toothache
4	39	F	83949	Cough

TABLE II. VOTER REGISTRATION LIST

ID	Attributes			
	<i>Name</i>	<i>Age</i>	<i>Sex</i>	<i>Zip code</i>
1	Jim	26	M	83661
2	Jay	24	M	83634
3	Tom	31	M	83967
4	Lily	39	F	83949

Anonymization can be vulnerable to [linkage attack](#).

Massachusetts governor William Weld's personal health information was discovered in a supposedly anonymized public database.





# Other Examples of Linkage Attack

## 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.

Our second contribution is a very general class of

## Who's Watching?

## De-anonymization of Netflix Reviews using Amazon Reviews

Maryam Archie, Sophie Gershon, Abigail Katcoff, and Aaron Zeng  
{marchie, sgershon, akatcoff, a2z}@mit.edu

**Abstract**—Many companies' privacy policies state they can only release customer data once personal identifiable information has been removed; however it has been shown by Narayanan and Shmatikov (2008) and reinforced in this paper that removal of personal identifiable information is not enough to anonymize datasets. Herein we describe a method for de-anonymizing the Netflix Prize dataset users using publicly available Amazon review data [3], [4]. Based on the matching Amazon user profile, we can then discover more information about the supposedly anonymous Netflix user, including the user's full name and shopping habits. Even when datasets are cleaned and perturbed to protect user privacy, because of the sheer quantity of information publicly available through the Internet, it is difficult for individuals or companies like Netflix to guarantee that the data they release will not violate the privacy and anonymity of their users.

using data from the Internet Movie Database<sup>3</sup> (IMDb). They developed a formal model for privacy breaches in anonymized micro-data, e.g. recommendations. Narayanan and Shmatikov also proposed an algorithm that predicts if ratings between datasets are correlated (by date and numerical rating). Using publicly available data from IMDb, they were able to identify several users in the “anonymized” Netflix dataset and learn potentially sensitive information about them, including political affiliations [2].

We aim to extend these results to show we can identify users from the “anonymized” dataset using publicly available Amazon reviews. As a result, we can learn about Netflix users' spending habits and reveal possibly private information about them.

# The 18 HIPAA Identifiers

The HIPAA privacy rule sets forth policies to protect all individually identifiable health information that is held or transmitted by a covered entity. These are the 18 HIPAA Identifiers that are considered personally identifiable information. This information can be used to identify, contact, or locate a single person or can be used with other sources to identify a single individual. When personally identifiable information is used in conjunction with one's physical or mental health or condition, health care, or one's payment for that health care, it becomes Protected Health Information (PHI).

- Name
- Address (all geographic subdivisions smaller than state, including street address, city county, and zip code)
- All elements (except years) of dates related to an individual (including birthdate, admission date, discharge date, date of death, and exact age if over 89)
- Telephone numbers
- Fax number
- Email address
- Social Security Number
- Medical record number
- Health plan beneficiary number
- Account number
- Certificate or licence number
- Any vehicle or other device serial number

Q2: will a query reveal whether I cheated or not?

			Lie, Cheat,	Behavior	Peer	Low Academic	Negative	Aggressive	Total	
Student Name	Student ID	Steal	Sneak	Problem	Rejection	Achievement	Attitude	Behavior	SRSS	GPA
Angel, Julio	2310	0	0	0	2	0	0	1	3	3.1
Akins, J'Monte	2013	0	0	0	0	0	0	0	0	4.1
Backer, Brent	2031	0	1	2	1	2	2	1	9	2.3
Boxwell, Kylie	2001	0	0	0	1	1	0	0	2	2.5
Cartright, Ashley	2152	0	1	1	1	0	1	0	4	3.2
Cox, Lucille	2002	0	0	0	0	0	0	0	0	3.9
Hankins, Erin	2017	0	0	0	0	0	2	0	2	3.7
Illio, Helen	2132	0	0	0	0	0	0	0	0	2.9
Jackson, Ronald	2003	0	1	2	2	3	2	2	12	1.7
Kemp, Patrice	2009	0	0	1	0	0	0	0	1	3.3
Parker, Stephanie	2004	0	0	0	0	1	2	0	3	2.7
Reed, Kent	2010	0	0	0	0	0	0	0	0	3.6
Sterling, Michael	2022	0	0	1	0	3	1	1	6	2.4
Thomas, James	2018	0	0	0	0	0	0	0	0	3.8
Walsh, Carter	2215	0	0	0	1	0	1	0	2	3.5

Aggregation: only show statistics. Is it sufficient?

			Lie, Cheat, Sneak	Total SRSS	
Student Name	Student ID	Steal			GPA
Angel, Julio	2310	0	0	3	3.1
Akins, J'Monte	2013	0	0	0	4.1
Backer, Brent	2031	0	1	9	2.3
Boxwell, Kylie	2001	0	0	2	2.5
Cartright, Ashley	2152	0	1	4	3.2
Cox, Lucille	2002	0	0	0	3.9
Hankins, Erin	2017	0	0	2	3.7
Illio, Helen	2132	0	0	0	2.9
Jackson, Ronald	2003	0	1	12	1.7
Kemp, Patrice	2009	0	0	1	3.3
Parker, Stephanie	2004	0	0	3	2.7
Reed, Kent	2010	0	0	0	3.6
Sterling, Michael	2022	0	0	6	2.4
Thomas, James	2018	0	0	0	3.8
Walsh, Carter	2215	0	0	2	3.5

### Example (Conditional) Statistics

# cheaters among all students

# students GPA < 3

# cheaters among students GPA<3.

.....

Q: can you infer whether Ronald cheated or not?

Student Name	Student ID	Steal	Lie, Cheat,	Total SRSS	GPA
			Sneak		
Angel, Julio	2310	0	0	3	3.1
Akins, J'Monte	2013	0	0	0	4.1
Backer, Brent	2031	0	1	9	2.3
Boxwell, Kylie	2001	0	0	2	2.5
Cartright, Ashley	2152	0	1	4	3.2
Cox, Lucille	2002	0	0	0	3.9
Hankins, Erin	2017	0	0	2	3.7
Illio, Helen	2132	0	0	0	2.9
Jackson, Ronald	2003	0	1	12	1.7
Kemp, Patrice	2009	0	0	1	3.3
Parker, Stephanie	2004	0	0	3	2.7
Reed, Kent	2010	0	0	0	3.6
Sterling, Michael	2022	0	0	6	2.4
Thomas, James	2018	0	0	0	3.8
Walsh, Carter	2215	0	0	2	3.5

Suppose we know Ronald GPA < 2.

Aggregation can be vulnerable to [inference attack](#).

Student Name	Student ID	Steal	Lie, Cheat, Sneak	Total SRSS	GPA
Angel, Julio	2310	0	0	3	3.1
Akins, J'Monte	2013	0	0	0	4.1
Backer, Brent	2031	0	1	9	2.3
Boxwell, Kylie	2001	0	0	2	2.5
Cartright, Ashley	2152	0	1	4	3.2
Cox, Lucille	2002	0	0	0	3.9
Hankins, Erin	2017	0	0	2	3.7
Illio, Helen	2132	0	0	0	2.9
Jackson, Ronald	2003	0	1	12	1.7
Kemp, Patrice	2009	0	0	1	3.3
Parker, Stephanie	2004	0	0	3	2.7
Reed, Kent	2010	0	0	0	3.6
Sterling, Michael	2022	0	0	6	2.4
Thomas, James	2018	0	0	0	3.8
Walsh, Carter	2215	0	0	2	3.5

Suppose we know Ronald GPA < 2.

1. # students GPA < 2 = 1

2. # cheaters whose GPA < 2 = 1



## The embarrassing dilemma between data publishing and privacy.

			Lie, Cheat, Sneak	Total SRSS	
Student Name	Student ID	Steal			GPA
Angel, Julio	2310	0	0	3	3.1
Akins, J'Monte	2013	0	0	0	4.1
Backer, Brent	2031	0	1	9	2.3
Boxwell, Kylie	2001	0	0	2	2.5
Cartright, Ashley	2152	0	1	4	3.2
Cox, Lucille	2002	0	0	0	3.9
Hankins, Erin	2017	0	0	2	3.7
Illio, Helen	2132	0	0	0	2.9
Jackson, Ronald	2003	0	1	12	1.7
Kemp, Patrice	2009	0	0	1	3.3
Parker, Stephanie	2004	0	0	3	2.7
Reed, Kent	2010	0	0	0	3.6
Sterling, Michael	2022	0	0	6	2.4
Thomas, James	2018	0	0	0	3.8
Walsh, Carter	2215	0	0	2	3.5

Need to publish some information.

Adversary can always try to infer the presence of an individual from it.

Q: can we publish useful info that does not allow accurate inference?



# Differential Privacy

Idea: whether you are in or not in, query results will be similar.

Student Name	Student ID	Steal	Lie, Cheat, Sneak	Total SRSS	GPA
Angel, Julio	2310	0	0	3	3.1
Akins, J'Monte	2013	0	0	0	4.1
Backer, Brent	2031	0	1	9	2.3
Boxwell, Kylie	2001	0	0	2	2.5
Cartright, Ashley	2152	0	1	4	3.2
Cox, Lucille	2002	0	0	0	3.9
Hankins, Erin	2017	0	0	2	3.7
Illio, Helen	2132	0	0	0	2.9
Jackson, Ronald	2003	0	1	12	1.7
Kemp, Patrice	2009	0	0	1	3.3
Parker, Stephanie	2004	0	0	3	2.7
Reed, Kent	2010	0	0	0	3.6
Sterling, Michael	2022	0	0	6	2.4
Thomas, James	2018	0	0	0	3.8
Walsh, Carter	2215	0	0	2	3.5



## Concept: Neighboring Databases

Two databases D and D' are called neighboring databases if they differ by only one record.

			Total	
Student Name	Student ID	Steal	SRSS	GPA
Angel, Julio	2310	0	3	3.1
Akins, J'Monte	2013	0	0	4.1
Backer, Brent	2031	0	9	2.3
Boxwell, Kylie	2001	0	2	2.5
Cartright, Ashley	2152	0	4	3.2
Cox, Lucille	2002	0	0	3.9
Hankins, Erin	2017	0	2	3.7
Illio, Helen	2132	0	0	2.9
Jackson, Ronald	2003	0	12	1.7
Kemp, Patrice	2009	0	1	3.3
Parker, Stephanie	2004	0	3	2.7
Reed, Kent	2010	0	0	3.6
Sterling, Michael	2022	0	6	2.4
Thomas, James	2018	0	0	3.8

			Total	
Student Name	Student ID	Steal	SRSS	GPA
Angel, Julio	2310	0	3	3.1
Akins, J'Monte	2013	0	0	4.1
Backer, Brent	2031	0	2	3.5
Boxwell, Kylie	2001	0	2	2.5
Cartright, Ashley	2152	0	4	3.2
Cox, Lucille	2002	0	0	3.9
Hankins, Erin	2017	0	2	3.7
Illio, Helen	2132	0	0	2.9
Walsh, Carter	2215	0	12	1.7
Kemp, Patrice	2009	0	1	3.3
Parker, Stephanie	2004	0	3	2.7
Reed, Kent	2010	0	0	3.6
Sterling, Michael	2022	0	6	2.4
Thomas, James	2018	0	0	3.8

## Concept: Randomized Mechanism (Query)

A randomized mechanism  $F$  is a mapping from a database  $D$  to a random variable.

Student Name	Student ID	Steal	Total SRSS	GPA
Angel, Julio	2310	0	3	3.1
Akins, J'Monte	2013	0	0	4.1
Backer, Brent	2031	0	9	2.3
Boxwell, Kylie	2001	0	2	2.5
Cartright, Ashley	2152	0	4	3.2
Cox, Lucille	2002	0	0	3.9
Hankins, Erin	2017	0	2	3.7
Illio, Helen	2132	0	0	2.9
Jackson, Ronald	2003	0	12	1.7
Kemp, Patrice	2009	0	1	3.3
Parker, Stephanie	2004	0	3	2.7
Reed, Kent	2010	0	0	3.6
Sterling, Michael	2022	0	6	2.4
Thomas, James	2018	0	0	3.8

### Example Random Mechanisms

$$F(D) = \# \text{ cheaters} + \eta$$

$$F(D) = \text{mean GPA} \propto N(3.11, \sigma^2).$$

## Concepts: Mechanism **Range** and its **Subset**

Let  $R(F)$  be the range of a randomized mechanism  $F$ . Let  $S$  be any subset of  $R(F)$ .

Student Name	Student ID	Steal	Total SRSS	GPA
Angel, Julio	2310	0	3	3.1
Akins, J'Monte	2013	0	0	4.1
Backer, Brent	2031	0	9	2.3
Boxwell, Kylie	2001	0	2	2.5
Cartright, Ashley	2152	0	4	3.2
Cox, Lucille	2002	0	0	3.9
Hankins, Erin	2017	0	2	3.7
Illio, Helen	2132	0	0	2.9
Jackson, Ronald	2003	0	12	1.7
Kemp, Patrice	2009	0	1	3.3
Parker, Stephanie	2004	0	3	2.7
Reed, Kent	2010	0	0	3.6
Sterling, Michael	2022	0	6	2.4
Thomas, James	2018	0	0	3.8

### Example Range and Subset

$R(F) = \{0, 1, 2, \dots, n\}$  # cheaters

$S = \{2, 3\}$ , or  $\{4, 7, 8\}$ , or ... subset

$R(F) = [0, 4]$  GPA range

$S = [0, 2]$  or  $[2.5, 3.5]$  or ... subset

## Definition: $\epsilon$ -Differential Privacy ( $\epsilon > 0$ )

We say a randomized mechanism  $F$  satisfies  $\epsilon$ -differential privacy if, for any neighboring databases  $D$  and  $D'$ , and any subset  $S$  of  $R(F)$ , there is

$$e^{-\epsilon} \leq \frac{\Pr\{F(D) \in S\}}{\Pr\{F(D') \in S\}} \leq e^{\epsilon}$$

Here, smaller  $\epsilon$  gives stronger privacy guarantee.

## Understanding the two probabilities.

$$\Pr\{F(D) \in S\}$$

$$\Pr\{D: \# \text{ cheaters} + \varepsilon \in \{2,3\}\}$$

Student Name	Student ID	Steal	Total SRSS	GPA
Angel, Julio	2310	0	3	3.1
Akins, J'Monte	2013	0	0	4.1
Backer, Brent	2031	0	9	2.3
Boxwell, Kyle	2001	0	2	2.5
Cartright, Ashley	2152	0	4	3.2
Cox, Lucille	2002	0	0	3.9
Hankins, Erin	2017	0	2	3.7
Illio, Helen	2132	0	0	2.9
Jackson, Ronald	2003	0	12	1.7
Kemp, Patrice	2009	0	1	3.3

$$\Pr\{F(D') \in S\}$$

$$\Pr\{D': \# \text{ cheater} + \varepsilon \in \{2,3\}\}$$

Student Name	Student ID	Steal	Total SRSS	GPA
Angel, Julio	2310	0	3	3.1
Akins, J'Monte	2013	0	0	4.1
Backer, Brent	2031	0	9	2.3
Boxwell, Kyle	2001	0	2	2.5
Cartright, Ashley	2152	0	4	3.2
Cox, Lucille	2002	0	0	3.9
Hankins, Erin	2017	0	2	3.7
Illio, Helen	2132	0	0	2.9
Walsh, Carter	2215	0	2	3.5
Kemp, Patrice	2009	0	1	3.3

Here, D and D' are arbitrary, and S is arbitrary.



These two probabilities should be similar.

Smaller  $\epsilon$  implies the two probabilities are similar, thus we have stronger privacy guarantee.

$$\boxed{e^{-\epsilon}} \leq \frac{\Pr\{F(D) \in S\}}{\Pr\{F(D') \in S\}} \leq \boxed{e^{\epsilon}}$$

Q: is this randomized mechanisms  $\epsilon$ -differential private?

Student Name	Student ID	Steal	Lie, Cheat, Sneak	Total SRSS	GPA
Angel, Julio	2310	0	0	3	3.1
Akins, J'Monte	2013	0	0	0	4.1
Backer, Brent	2031	0	1	9	2.3
Boxwell, Kylie	2001	0	0	2	2.5
Cartright, Ashley	2152	0	1	4	3.2
Cox, Lucille	2002	0	0	0	3.9
Hankins, Erin	2017	0	0	2	3.7
Illio, Helen	2132	0	0	0	2.9
Jackson, Ronald	2003	0	1	12	1.7
Kemp, Patrice	2009	0	0	1	3.3
Parker, Stephanie	2004	0	0	3	2.7
Reed, Kent	2010	0	0	0	3.6
Sterling, Michael	2022	0	0	6	2.4
Thomas, James	2018	0	0	0	3.8
Walsh, Carter	2215	0	0	2	3.5

Suppose we know Ronald GPA < 2.

Randomized mechanism: # cheaters +  $\eta$

# GPA < 2 +  $\eta$  = 3 (not 1)

# cheaters with GPA < 2 +  $\eta$  = 2 (not 1)

# Differential Privacy Techniques

## General Process of Achieving Differential Privacy

We have an ideal deterministic query mechanism  $f$ .

- e.g.,  $f(D) = \text{average GPA}$

We want to randomize  $f$  to obtain a randomized mechanism  $F$

- e.g.,  $F(D) = f(D) + \eta$
- if randomization is proper,  $F$  can satisfy  $\epsilon$ -differential privacy

Concept: **Sensitivity** of  $f$  on fixed-sized databases.

Assume the size of database is fixed, e.g., it always contains  $n$  students.

Sensitivity of  $f$  is its maximum output change on any two neighboring databases.

$$s(f) = \max_{D, D'} ||f(D) - f(D')||$$

Q: if  $f(D)$  = average GPA,  $|D| = 15$ , what is  $s(f)$ ?

Student Name	Student ID	GPA
Angel, Julio	2310	3.1
Akins, J'Monte	2013	4.1
Backer, Brent	2031	2.3
Boxwell, Kylie	2001	2.5
Cartright, Ashley	2152	3.2
Cox, Lucille	2002	3.9
Hankins, Erin	2017	3.7
Illio, Helen	2132	2.9
Jackson, Ronald	2003	1.7
Kemp, Patrice	2009	3.3
Parker, Stephanie	2004	2.7
Reed, Kent	2010	3.6
Sterling, Michael	2022	2.4
Thomas, James	2018	3.8
Walsh, Carter	2215	3.5

Assume GPA in  $[0,4]$ .

Q: intuitively, what should we do if  $s(f)$  is large?

Suppose we want to add proper noise  $\eta$  so that  $F(D) = f(D) + \eta$  is differentially private.

Sensitivity of  $f$  is its maximum output change on any two neighboring databases.

$$s(f) = \max_{D, D'} ||f(D) - f(D')||$$



## Two Strategies to Achieve $\epsilon$ -Differential Privacy.

Laplacian Mechanism

Exponential Mechanism

# Laplacian Mechanism

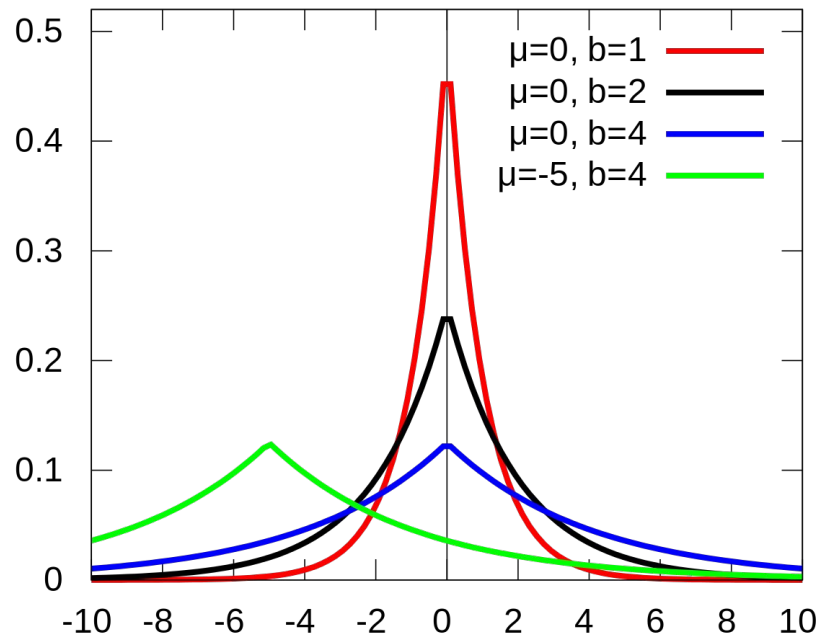
Idea: add Laplacian noise to output.

$$F(D) = f(D) + \eta$$

where  $\eta \sim \text{Lap}(0, s(f)/\epsilon)$ .

Here  $\mu = 0$  and  $b = s(f)/\epsilon$ .

PDF of Laplacian Variable



Q: how do  $\epsilon$  and  $s(f)$  affect the noise?

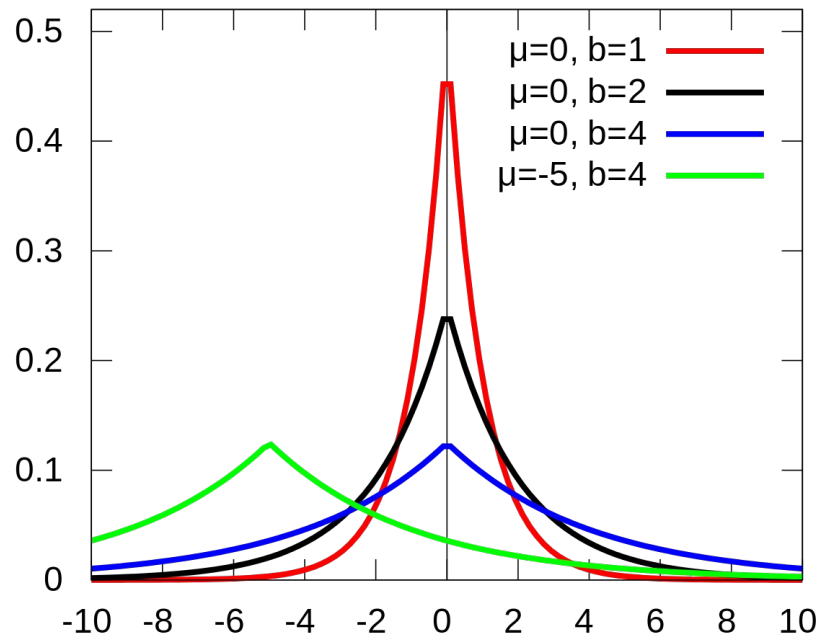
Idea: add Laplacian noise to output.

$$F(D) = f(D) + \eta$$

where  $\eta \sim \text{Lap}(0, s(f)/\epsilon)$ .

Here  $\mu = 0$  and  $b = s(f)/\epsilon$ .

PDF of Laplacian Variable



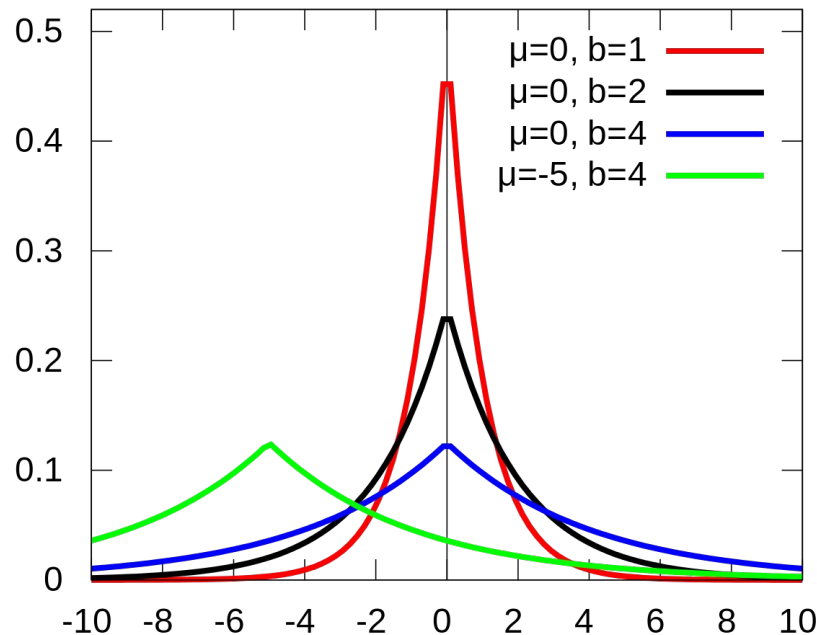
## Q: what is the scale of Laplacian noise?

We know  $f(D) = \text{ave.GPA}$  and  $s(f) = 4/15$ .

If we want to design  $F(D) = f(D) + \eta$  which satisfies 0.1-differential privacy, we need

$$\eta \propto \text{Lap}(0, s(f)/\epsilon) = \text{Lap}(0, \underline{\hspace{1cm}}).$$

PDF of Laplacian Variable



# Exponential Mechanism

Idea: randomly return a result  $y$ , but return matching ones with higher probabilities.

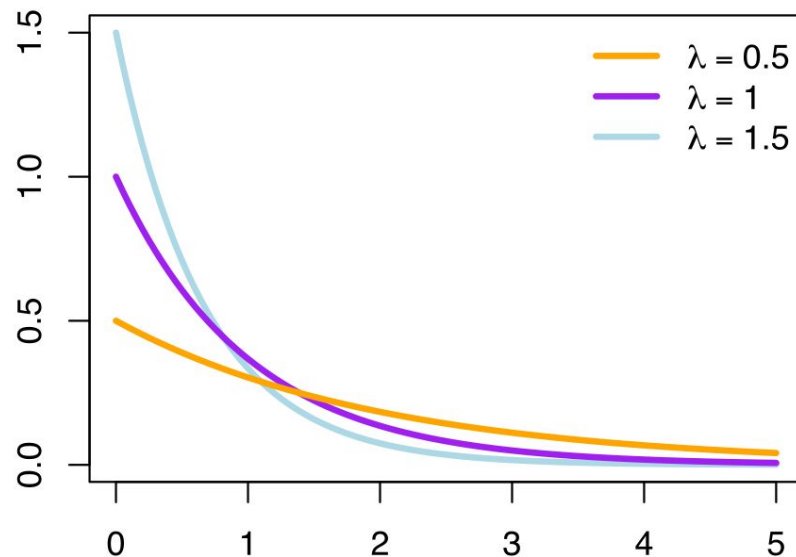
$$\Pr(y|D) \propto \exp( \epsilon * u(y|D) / 2s(u) )$$

-  $u(y|D)$  is **utility** of  $y$  in  $D$  (self-defined)

$$- s(u) = \max_{y,D,D'} |u(y|D) - u(y|D')|$$

Here  $\lambda = \epsilon * H(y|D) / 2s(H)$ .

PDF of Exponential Variable



Q: how do  $\epsilon$ ,  $H(y|D)$  and  $s(H)$  affect the random selection?

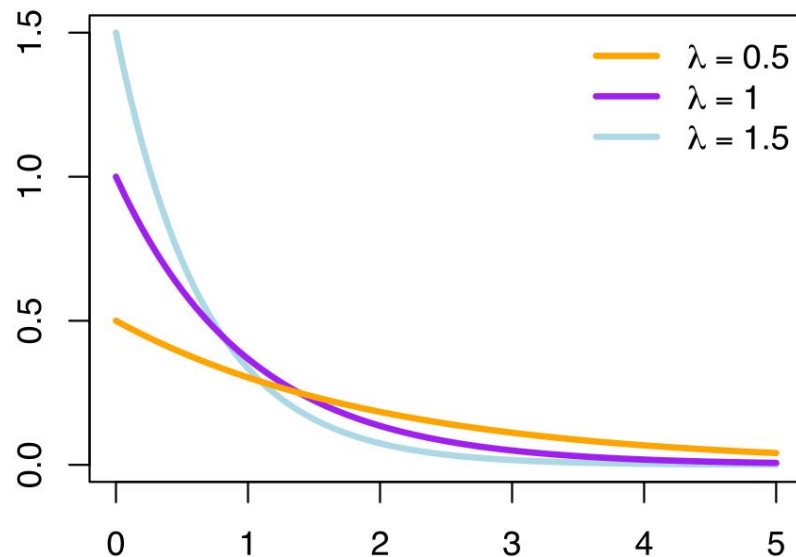
Idea: randomly return a result  $y$ , but return matching ones with higher probabilities.

$$\Pr(y|D) \propto \exp( \epsilon * H(y|D) / 2s(H) )$$

- $H(y|D)$  is **utility** of  $y$  in  $D$  (self-defined)
- $s(H) = \max_{D, D'} |H(D) - H(D')|$  is sensitivity

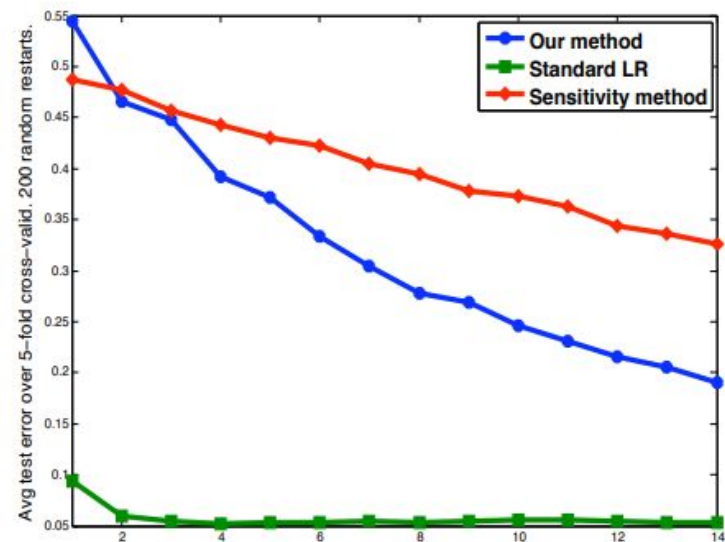
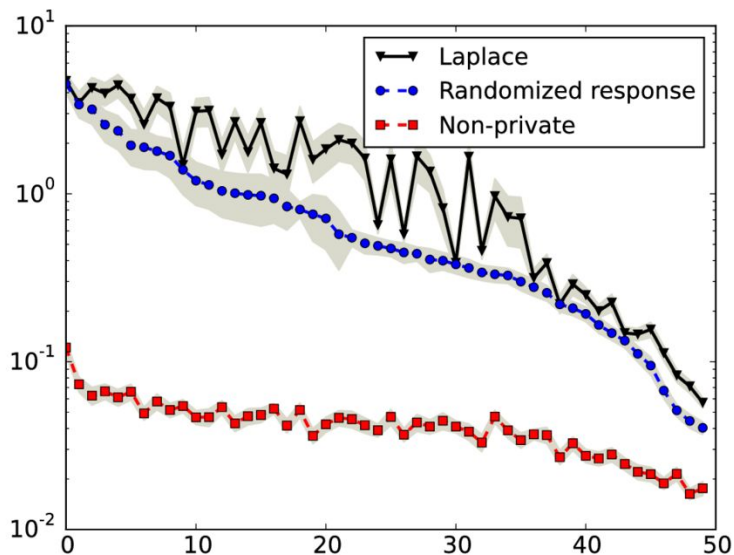
Here  $\lambda = \epsilon * H(y|D) / 2s(H)$ .

PDF of Exponential Variable



Both strategies have been applied to make machine learning algorithms differentially private...

Both strategies have accuracy-privacy tradeoff.





Let's see how to prove differential privacy.