Advertising systems in social media (3)

Saptarshi Ghosh
and Mainack Mondal

CS 60017
Autumn 2019

The story so far …

- Social advertising systems
  - Why bother about them?
  - The curious case of Facebook ads
  - How can we leverage these systems for doing good
- Abuse of the advertising systems
  - Why is targeted advertising bad?
  - Privacy risks with PII based targeting

The story so far …

- Social advertising systems
  - Why bother about them?
  - The curious case of Facebook ads
  - How can we leverage these systems for doing good
- Abuse of the advertising systems
  - Why is targeted advertising bad?
  - Privacy risks with PII based targeting
- Now, how to prevent abuse of advertising systems and provide data privacy?

Preserving privacy of social data

- Two broad dimensions
  - Preserving privacy from the background actors, e.g., advertisers or even the social media platform
  - Preserving privacy of data from other users, e.g., your ex

Preserving privacy from background actors

What are we going to talk about?

- Mechanisms for hiding privacy sensitive attributes in databases
  - K-anonymity
  - Differential privacy

- Slides heavily borrowed from
  - Vitaly Smatikov from Cornell
  - Li Xiong from Emory

Public Data Conundrum

- Health-care datasets
  - Clinical studies, hospital discharge databases …
- Genetic datasets
  - $1000 genome, HapMap, deCode …
- Demographic datasets
  - U.S. Census Bureau, sociology studies …
- Search logs, recommender systems, social networks, blogs …
  - AOL search data, social networks of blogging sites, Netflix movie ratings, Amazon …

What About Privacy?

- First thought: anonymize the data
- How?
- Remove "personally identifying information" (PII)
  - Name, Social Security number, phone number, email, address… what else?
  - Anything that identifies the person directly
- Is this enough?

Re-identification by Linking

Microdata

ID	QID			SA
Name	Zipcode	Age	Sex	Disease
Alice	47677	29	F	Ovarian Cancer
Betty	47602	22	F	Ovarian Cancer
Charles	47678	27	M	Prostate Cancer
David	47905	43	M	Flu
Emily	47909	52	F	Heart Disease
Fred	47906	47	M	Heart Disease

Voter registration data

Name	Zipcode	Age	Sex
Alice	47677	29	F
Bob	47983	65	M
Carol	47677	22	F
Dan	47532	23	M
Ellen	46789	43	F

Latanya Sweeney's Attack (1997)

Massachusetts hospital discharge dataset

Medical Data Released as Anonymous

SSN	Name	[illegible]city	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: [illegible]e-identifying anonymous data by linking to external data

Public voter dataset

Quasi-Identifiers

- Key attributes
  - Name, address, phone number - uniquely identifying!
  - Always removed before release
- Quasi-identifiers
  - (5-digit ZIP code, birth date, gender) uniquely identify 87% of the population in the U.S.
  - Can be used for linking anonymized dataset with other datasets

Classification of Attributes

- Sensitive attributes
  - Medical records, salaries, etc.
  - These attributes is what the analysts need, so they are always released directly

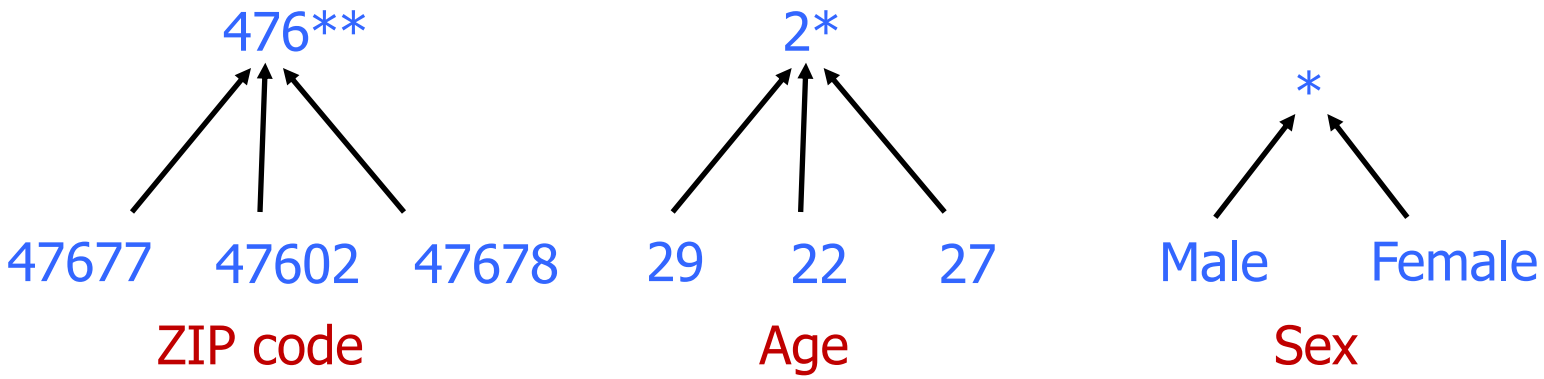
Key Attribute	Quasi-identifier			Sensitive attribute
Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

K-Anonymity: Intuition

- The information for each person contained in the released table cannot be distinguished from at least k-1 individuals whose information also appears in the release
  - Example: you try to identify a man in the released table, but the only information you have is his birth date and gender. There are k men in the table with the same birth date and gender.
- Any quasi-identifier present in the released table must appear in at least k records

Generalization

- Goal of k-Anonymity
  - Each record is indistinguishable from at least k-1 other records
  - These k records form an equivalence class
- Generalization: replace quasi-identifiers with less specific, but semantically consistent values



Achieving k-Anonymity

- Generalization
  - Replace specific quasi-identifiers with less specific values until get k identical values
  - Partition ordered-value domains into intervals

Example of a 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

Figure 2 Example of *k*-anonymity, where *k*=2 and QI={*Race*, *Birth*, *Gender*, *ZIP*}

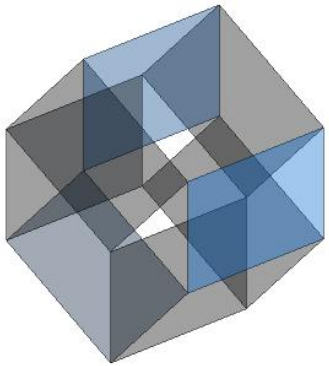
At least two people
With same attributes

QI = quasi identifier tuple

Curse of Dimensionality

[Aggarwal VLDB '05]

- Generalization fundamentally relies on spatial locality
  - Each record must have k close neighbors
- Real-world datasets are very sparse
  - Many attributes (dimensions)
    - Amazon customer records: several million dimensions
  - Not possible to create k close neighbors
- Projection to low dimensions loses all info ⇒ k-anonymized datasets are useless



Two (and a Half) Interpretations

- Membership disclosure: Attacker cannot tell that a given person in the dataset
- Sensitive attribute disclosure: Attacker cannot tell that a given person has a certain sensitive attribute
- Identity disclosure: Attacker cannot tell which record corresponds to a given person

This interpretation is correct, assuming the attacker does not know anything other than quasi-identifiers

But this does not imply any privacy!

Example: k clinical records, all HIV+

Attacks on k-Anonymity

- k-Anonymity does not provide privacy if
  - Sensitive values in an equivalence class lack diversity
  - The attacker has background knowledge

Homogeneity attack

Bob	
Zipcode	***Age***
47678	27

Background knowledge attack

Yoshiko		
Zipcode	***Age***	***Race***
47673	36	Japanese

Low chance of heart disease

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

k-Anonymity Considered Harmful

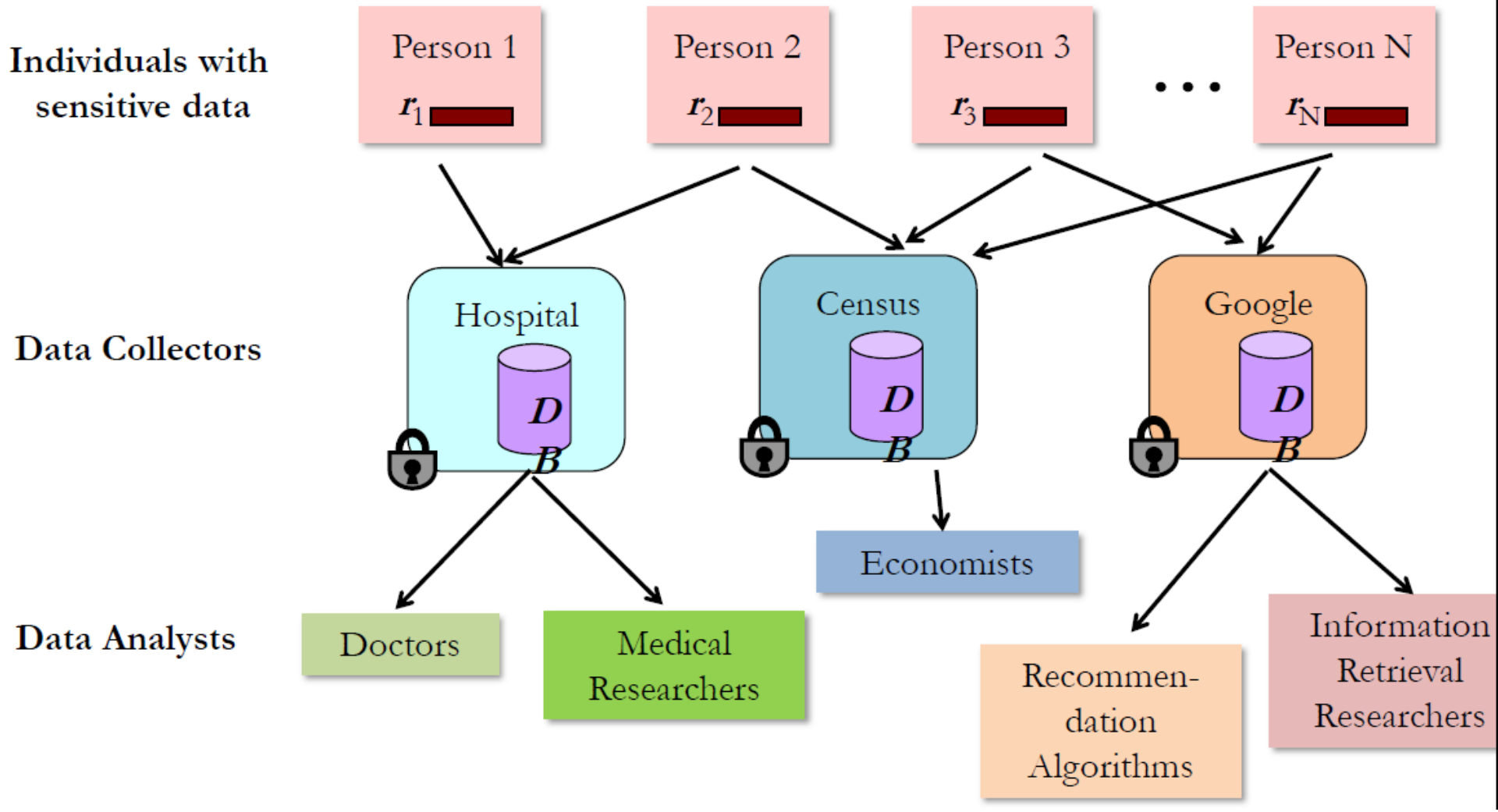
- Syntactic
  - Focuses on data transformation, not on what can be learned from the anonymized dataset
  - "k-anonymous" dataset can leak sensitive information
- "Quasi-identifier" fallacy
  - Assumes a priori that attacker will not know certain information about his target
- Relies on locality
  - Destroys utility of many real-world datasets

What are we going to talk about?

- Mechanisms for hiding privacy sensitive attributes in databases
  - K-anonymity
  - Differential privacy

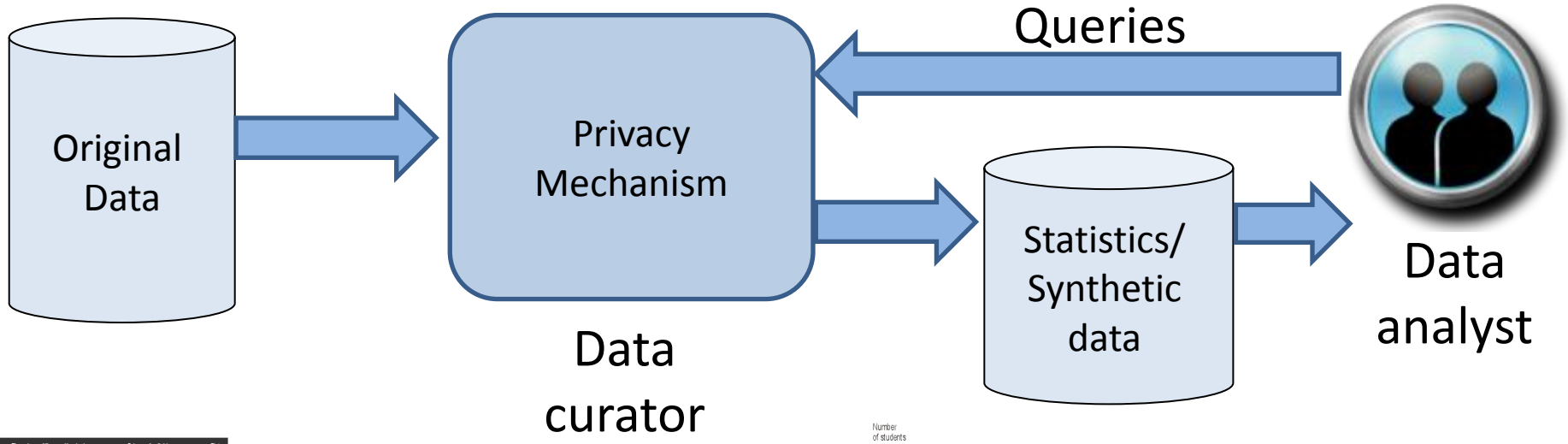
- Slides heavily borrowed from
  - Vitaly Smatikov from Cornell
  - Li Xiong from Emory

Statistical Databases

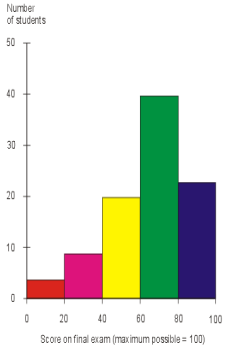


Statistical Data Privacy

- Non-interactive vs interactive
- Privacy goal: individual is protected
- Utility goal: statistical information useful for analysis



Name	Thread pitch (mm)	Minor diameter tolerance	Nominal diameter (mm)	Head shape	Price for 50 screws	Available at factory outlet?	Number in stock	Flat or Phillips head?
M4	0.7	4g	4	Pan	$10.08	Yes	276	Flat
M5	0.8	4g	5	Round	$13.89	Yes	183	Both
M6	1	5g	6	Button	$10.42	Yes	1043	Flat
M8	1.25	5g	8	Pan	$11.98	No	298	Phillips
M10	1.5	6g	10	Round	$16.74	Yes	488	Phillips
M12	1.75	7g	12	Pan	$18.26	No	998	Flat
M14	2	7g	14	Round	$21.19	No	235	Phillips
M16	2	8g	16	Button	$23.57	Yes	292	Both
M18	2.1	8g	18	Button	$25.87	No	664	Both
M20	2.4	8g	20	Pan	$29.09	Yes	486	Both
M24	2.55	9g	24	Round	$33.01	Yes	982	Phillips
M28	2.7	10g	28	Button	$35.66	No	1067	Phillips
M36	3.2	12g	36	Pan	$41.32	No	434	Both
M50	4.5	15g	50	Pan	$44.72	No	740	Flat



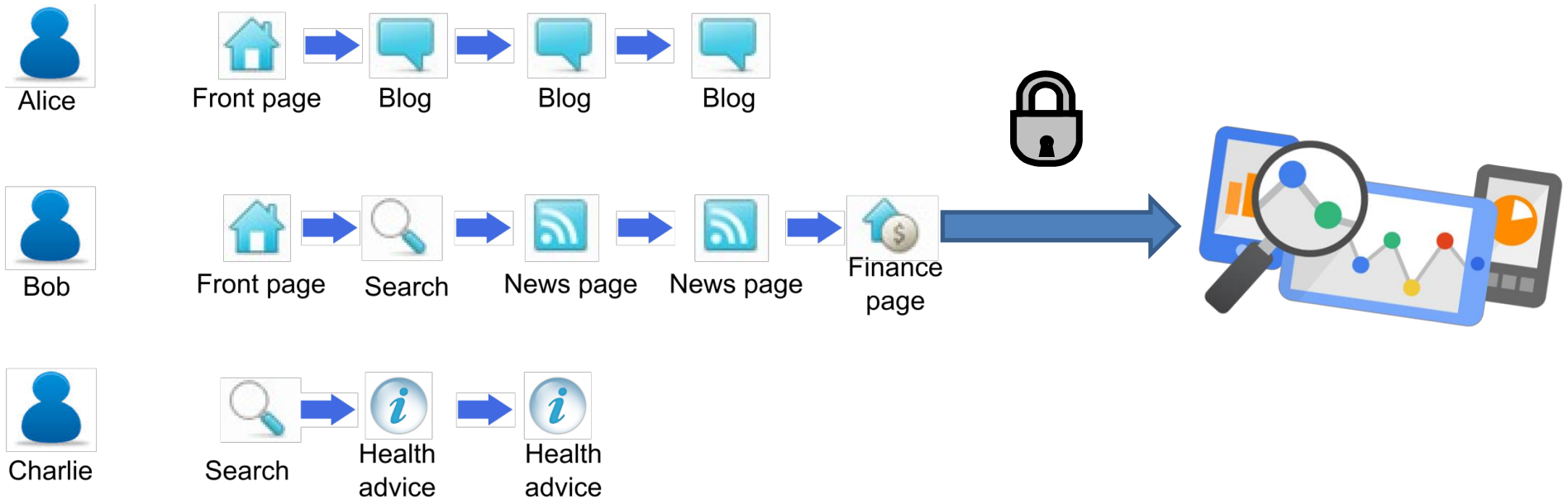
Name	Thread pitch (mm)	Minor diameter tolerance	Nominal diameter (mm)	Head shape	Price for 50 screws	Available at factory outlet?	Number in stock	Flat or Phillips head?
M4	0.7	4g	4	Pan	$10.08	Yes	276	Flat
M5	0.8	4g	5	Round	$13.89	Yes	183	Both
M6	1	5g	6	Button	$10.42	Yes	1043	Flat
M8	1.25	5g	8	Pan	$11.98	No	298	Phillips
M10	1.5	6g	10	Round	$16.74	Yes	488	Phillips
M12	1.75	7g	12	Pan	$18.26	No	998	Flat
M14	2	7g	14	Round	$21.19	No	235	Phillips
M16	2	8g	16	Button	$23.57	Yes	292	Both
M18	2.1	8g	18	Button	$25.87	No	664	Both
M20	2.4	8g	20	Pan	$29.09	Yes	486	Both
M24	2.55	9g	24	Round	$33.01	Yes	982	Phillips
M28	2.7	10g	28	Button	$35.66	No	1067	Phillips
M36	3.2	12g	36	Pan	$41.32	No	434	Both
M50	4.5	15g	50	Pan	$44.72	No	740	Flat

Differential Privacy

- Promise: an individual will not be affected, adversely or otherwise, by allowing his/her data to be used in any study or analysis, no matter what other studies, datasets, or information sources, are available”
- Paradox: learning nothing about an individual while learning useful statistical information about a population

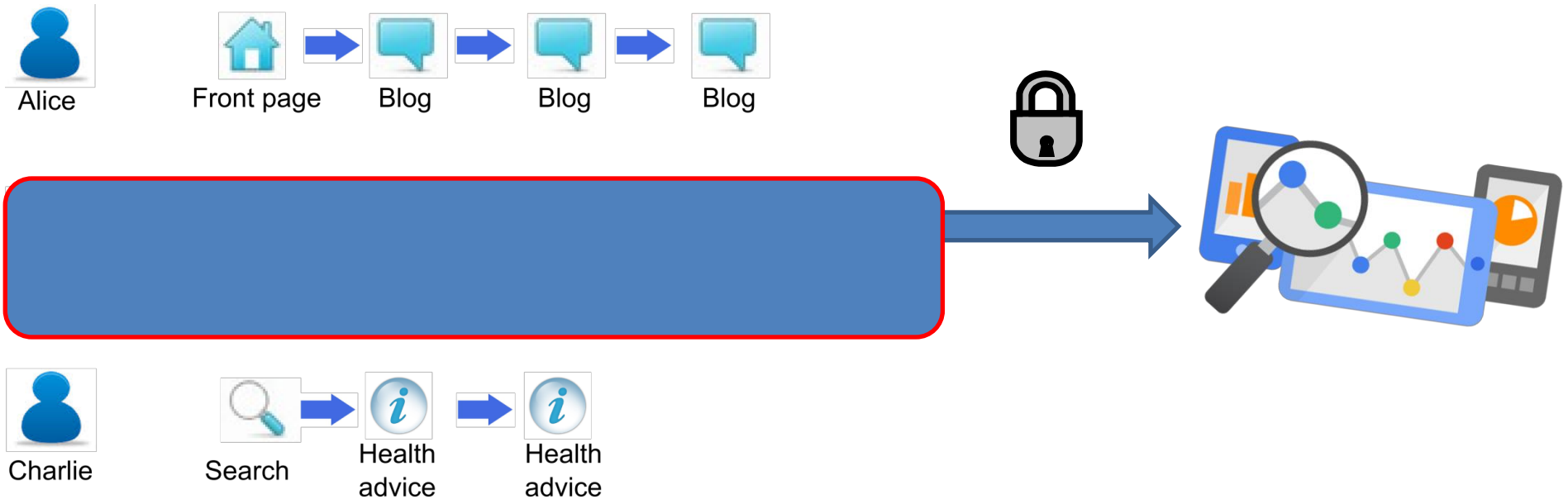
Differential Privacy

- Statistical outcome is indistinguishable regardless whether a particular user (record) is included in the data



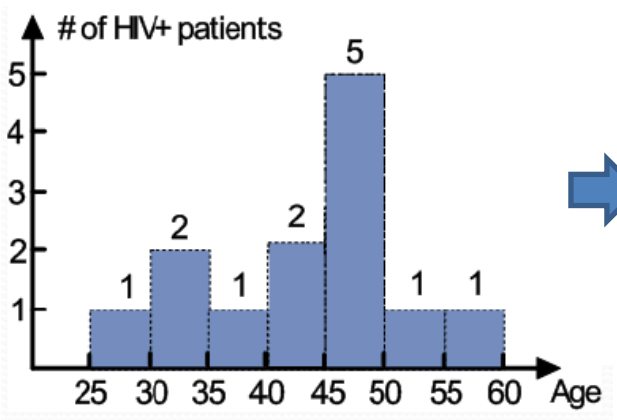
Differential Privacy

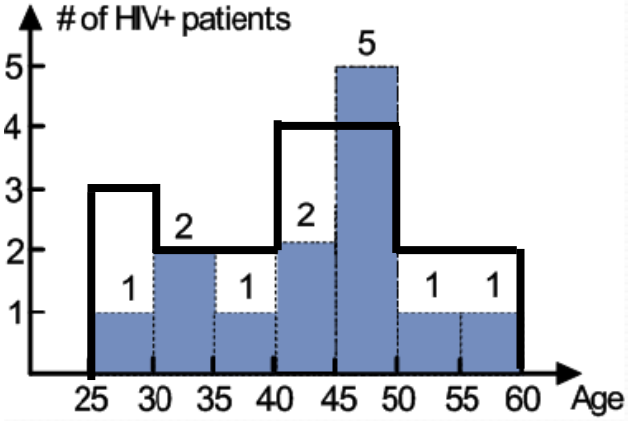
- Statistical outcome is indistinguishable regardless whether a particular user (record) is included in the data



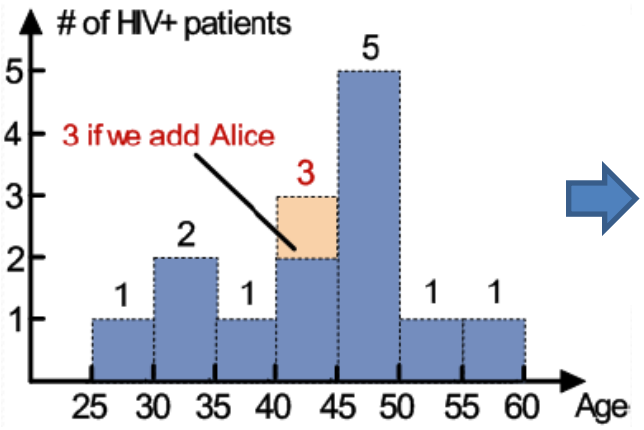
Differential privacy: an example

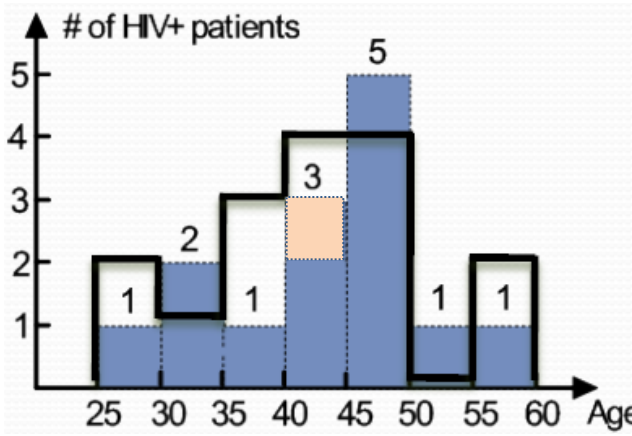
Name	Age	HIV+
Frank	42	Y
Bob	31	Y
Mary	28	Y
Dave	43	N
…	…	…





Name	Age	HIV+
Alice	43	Y
Frank	42	Y
Bob	31	Y
Mary	28	Y
Dave	43	N
…	…	…





Original records

Original histogram

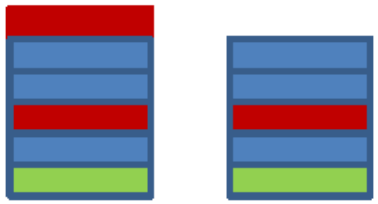
Perturbed histogram with differential privacy

Differential Privacy

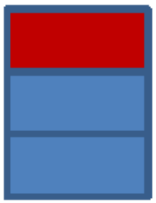
[Dwork ICALP 2006]

For every pair of inputs that differ in one row

For every output …



D_1 D_2



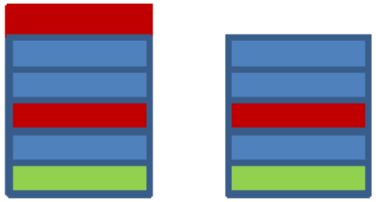
O

Adversary should not be able to distinguish between any D_1 and D_2 based on any O

$$\log\left(\frac{\Pr[A(D_1) = O]}{\Pr[A(D_2) = O]}\right) < \varepsilon \quad (\varepsilon > 0)$$

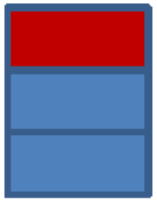
Why *all* pairs of datasets …?

For every pair of inputs that differ in one row



D_1 D_2

For every output …

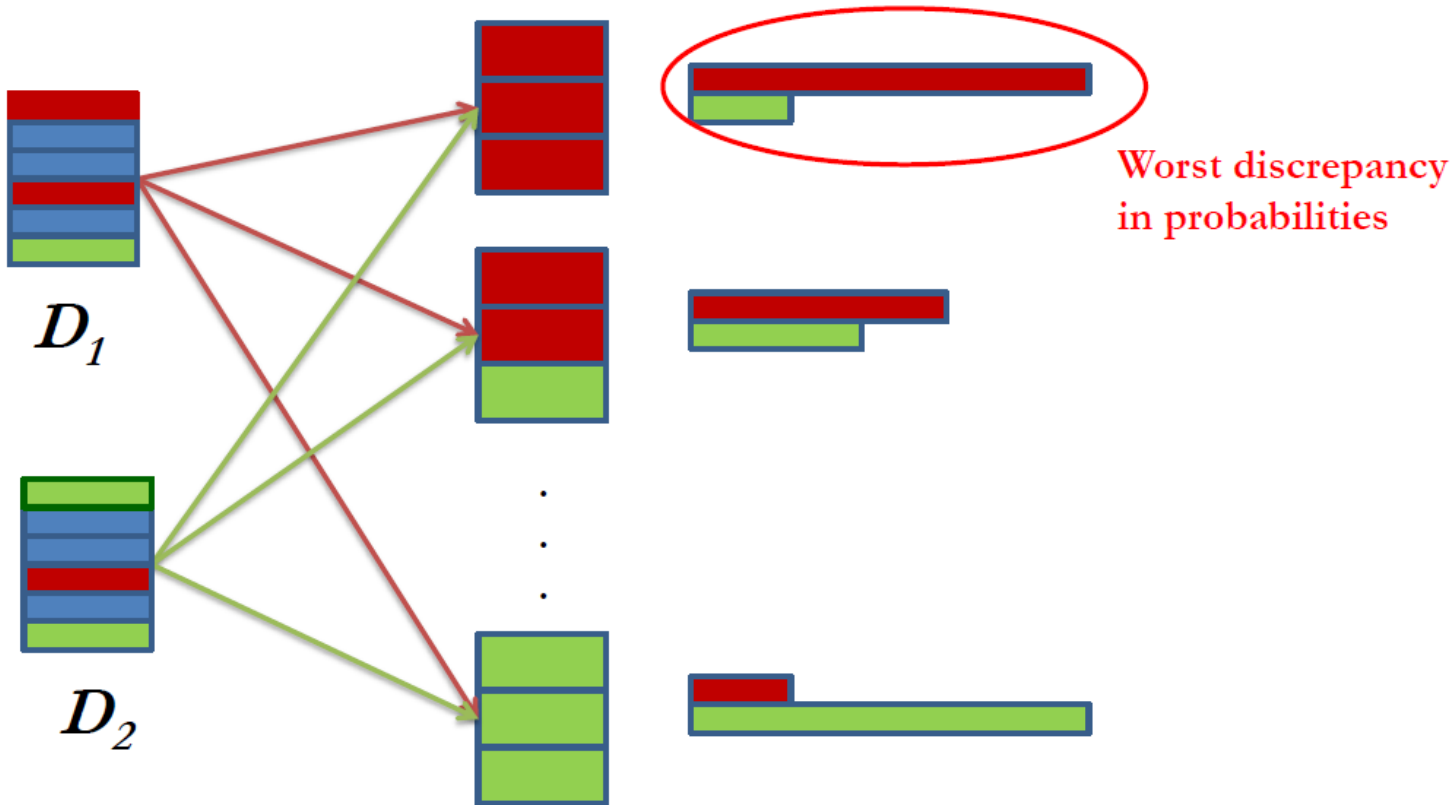


O

Guarantee holds no matter what the other records are.

Why *all* outputs?

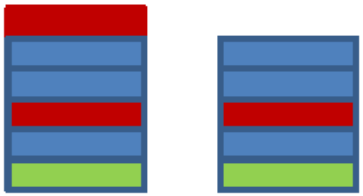
Should not be able to distinguish whether input was D_1 or D_2 no matter what the output



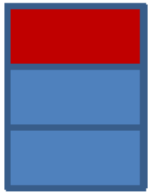
Privacy Parameter ε

For every pair of inputs that differ in one row

For every output …



D_1 D_2



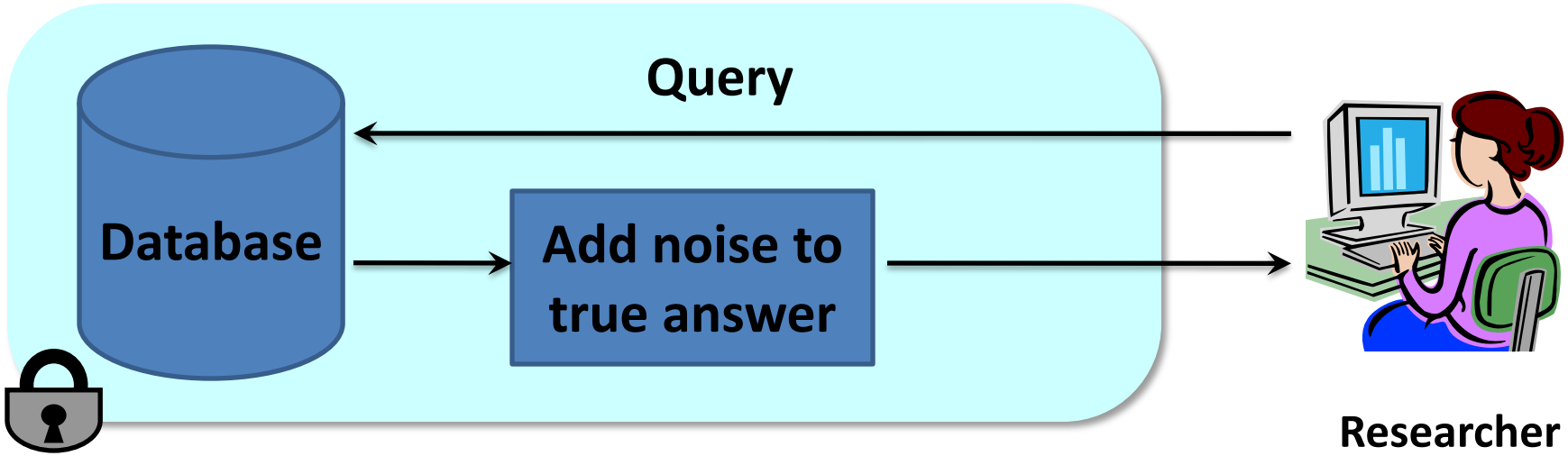
O

$$\Pr[A(D_1) = O] \leq e^{\varepsilon} \Pr[A(D_2) = O]$$

Controls the degree to which D_1 and D_2 can be distinguished.
Smaller the ε more the privacy (and better the utility)

Can deterministic algorithms satisfy differential privacy?

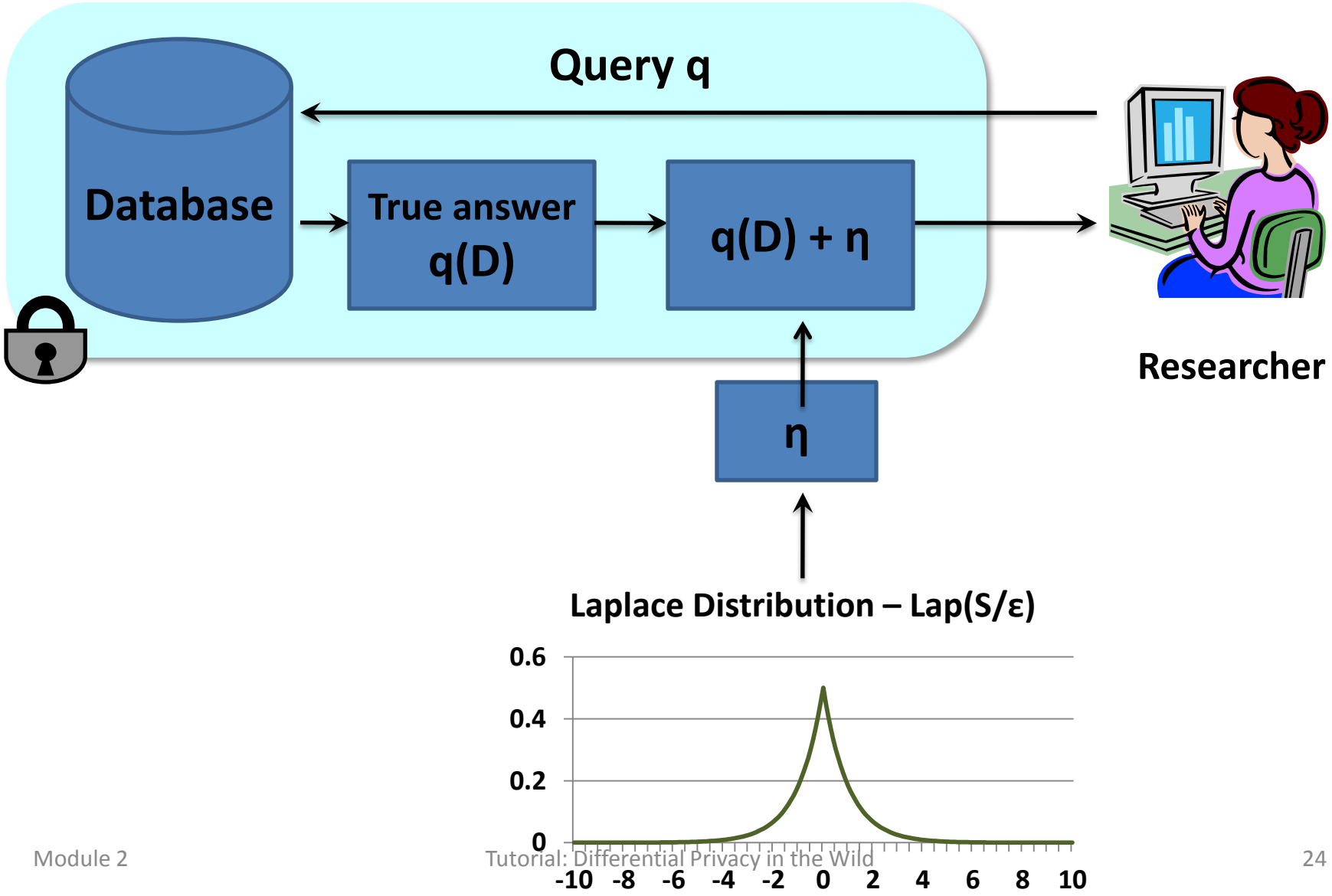
Output Randomization



- Add noise to answers such that:
  - Each answer does not leak too much information about the database.
  - Noisy answers are close to the original answers.

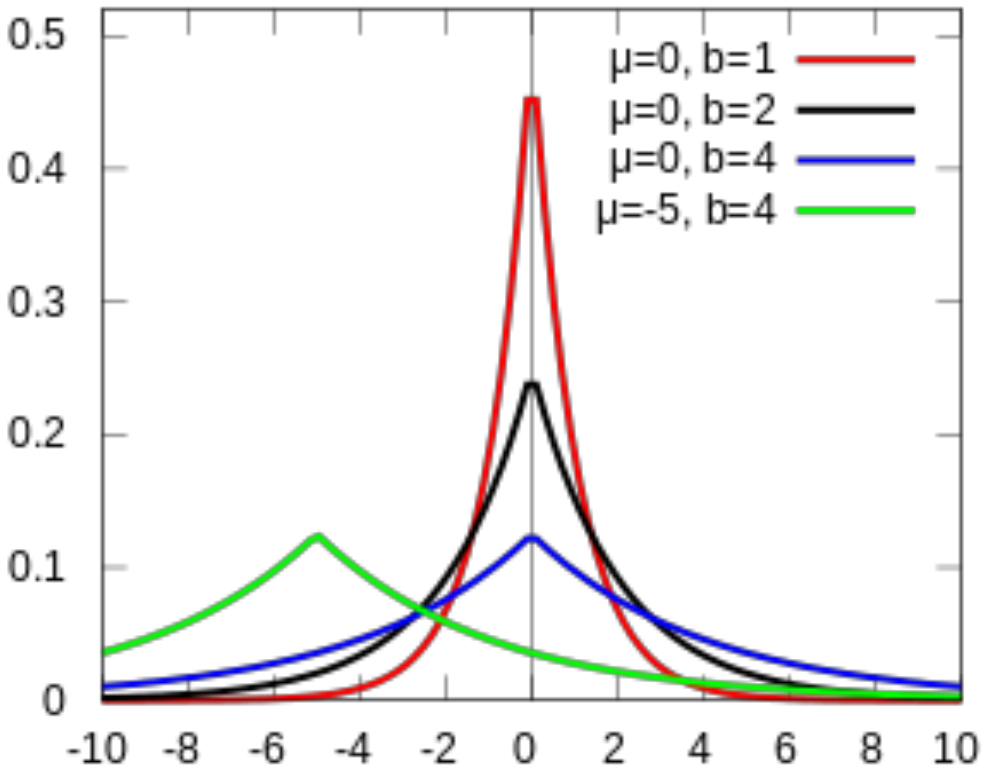
[DMNS 06]

Laplace Mechanism



Laplace Distribution

- PDF: $f(x \mid \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$
- Denoted as Lap(b) when u=0
- Mean u
- Variance $2b^2$



How much noise for privacy?

[Dwork et al., TCC 2006]

Sensitivity: Consider a query q: $I \rightarrow R$. S(q) is the smallest number s.t. for any neighboring tables D, D',

$$| q(D) - q(D') | \leq S(q)$$

Theorem: If **sensitivity** of the query is **S**, then the algorithm $A(D) = q(D) + Lap(S(q)/\varepsilon)$ guarantees ε-differential privacy

Sensitivity

- Semantically Sensitivity is
  - Given a query, what the maximum amount that the output will change by adding a row?

Example 1

- Let's consider a simple count query
  - Number of people clicking on an ad / having a disease?
  - What is the sensitivity?

Example: COUNT query

- Number of people having disease
- Sensitivity = 1

- Solution: 3 + η,
  where η is drawn from Lap(1/ε)
  - Mean = 0
  - Variance = $2/\varepsilon^2$

D

Disease (Y/N)
Y
Y
N
Y
N
N

Example 2

- Let's consider another count query
  - Number of people clicking on an ad / having a disease rounded to nearest multiple of 10?
  - What is the sensitivity?

Privacy of Laplace Mechanism

- Consider neighboring databases D and D'
- Consider some output O

$$\frac{\Pr[A(D) = O]}{\Pr[A(D') = O]} = \frac{\Pr[q(D) + \eta = O]}{\Pr[q(D') + \eta = O]}$$

λ = variance = S(q)/ε

$$= \frac{e^{-|O-q(D)|/\lambda}}{e^{-|O-q(D')|/\lambda}}$$

$$\leq e^{|q(D)-q(D')|/\lambda} \leq e^{S(q)/\lambda} = e^{\varepsilon}$$

Utility of Laplace Mechanism

- Laplace mechanism works for **any function** that returns a real number
- Error: $E(\text{true answer} - \text{noisy answer})^2$

$$= \text{Var}(\text{Lap}(S(q)/\varepsilon))$$

$$= 2*S(q)^2 / \varepsilon^2$$

- Where is there room for improvement?
  - The Laplace mechanism adds independent noise to every coordinate…
  - What happens if the user asks (essentially) the same question in every coordinate?
  - Read [Dinur,Nissim03]: a computationally efficient attack that gives blatant non-privacy for a mechanism that adds noise bounded by $o(\sqrt{n})$