THE UNIVERSITY OF MELBOURNE

COMP20008 Elements of Data Processing

Semester 2 2018

Lecture 22: Differential Privacy – Local and Global

THE UNIVERSITY OF MELBOURNE

Announcements

- Exam consultation sessions:
  – Monday 22/10/2018 Room 07.02 Doug McDonell 10:00am-12:00pm
  – Thursday 25/10/2018 Room 07.02 Doug McDonell 10:00am-12:00pm
- Phase 3 marks will be released next Tuesday 16/10/2018 7pm
- Final update of exam guide
  – Available next Friday 19/10/2018
- Reminder: Subject Experience Survey (SES) provides valuable feedback to the University and to your subject coordinators for subject improvements for future cohorts
  – Log in directly from your SES notification email, the SES login page (ses.unimelb.edu.au), or from your LMS homepage.
  – We value your feedback about the subject

THE UNIVERSITY OF MELBOURNE

Plan today

- Recap of *k*-anonymity and *l*-diversity
  – Concept
  – Homogeneity and background attack
  – Location/trajectory privacy
- An introduction to differential privacy

THE UNIVERSITY OF MELBOURNE

k-anonymity recap

- Data owner determines quasi identifier(s)
- Data owner or individuals choose parameter *k*

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Figure 1. Inpatient Microdata

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Figure 2. 4-anonymous Inpatient Microdata

l-Diversity: Privacy Beyond k-Anonymity. Machanavajjhala, Gehrke, Kifer and Venkitasubramaniam, 2007

l-diversity

- To protect privacy against
  - Homogeneity attack
  - Background knowledge attack

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	*	Heart Disease
4	1305*	≤ 40	*	Viral Infection
9	1305*	≤ 40	*	Cancer
10	1305*	≤ 40	*	Cancer
5	1485*	> 40	*	Cancer
6	1485*	> 40	*	Heart Disease
7	1485*	> 40	*	Viral Infection
8	1485*	> 40	*	Viral Infection
2	1306*	≤ 40	*	Heart Disease
3	1306*	≤ 40	*	Viral Infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

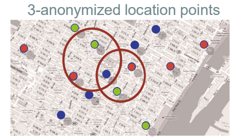
THE UNIVERSITY OF MELBOURNE

Overview of Location Privacy Models

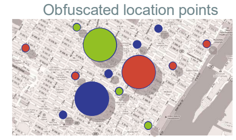
- Location privacy
  - *k*-anonymity (cloaking)
    - If individuals' location information cannot be distinguished from *k*−1 other individuals
  - Obfuscation
    - The greater the imperfect knowledge about a user's location, the greater the user's privacy

Exact location points

3-anonymized location points



Obfuscated location points



THE UNIVERSITY OF MELBOURNE

Summary

- To reduce risk of re-identification of individuals in released datasets
  - Choose value of *k*
  - Manipulate data to make it *k*-anonymous, either
    - Replace categories by broader categories
    - Suppress attributes with a * (limited utility)
  - Further manipulate data to make it *l*-diverse
    - Ensure there are at least *l* different values of the sensitive attribute in each group
- Privacy is difficult to maintain in high-dimensional datasets like trajectory datasets
  - Cloaking provides spatial *k*-anonymity
  - Obfuscation ensures location imprecision

THE UNIVERSITY OF MELBOURNE

Question: Could we release this dataset publicly?

Anonymous ID	Gender	Subject	Grade
a0b76	Male	COMP20008	89
539a2	Male	COMP20008	99
32435	Male	COMP20008	70
ae545	Male	COMP20008	63
ea6f5	Female	COMP20008	88
56acc	Female	COMP20008	90
9103b	Female	COMP20008	52
9a99a	Female	COMP20008	78
….	….	….	….
539a2	Male	COMP20003	31
..	…	..	..

- **Student 539a2 tweets that "I got 99 for COMP20008!"**

What to do?

"The future of privacy is lying"

– (April 10 2013, Matt Buchanan, New Yorker)

Differential privacy: Local and global

- **Global**: We have a sensitive dataset, a trusted data owner Alice and a researcher Bob. Alice does analysis on the raw data, adds noise to the answers, and reports the (noisy) answers to Bob
- **Local**: Each person is responsible for adding noise to their own data. Classic survey example each person has to answer question "Do you use drugs?"
  - They flip a coin in secret and answer "Yes" if it comes up heads, but tell the truth otherwise.
  - Plausible deniability about a "Yes" answer
- We will next be looking further at the global case (global systems generally more accurate, and less noise is needed)

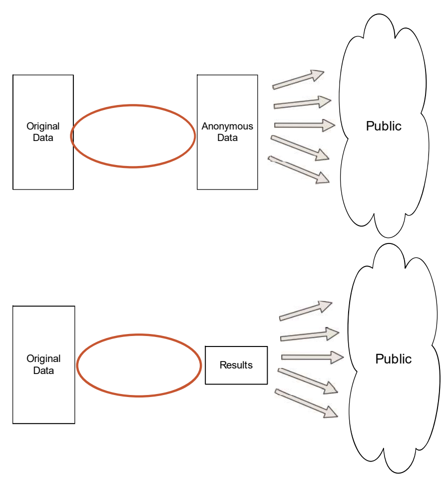
Differential privacy: Where?

- Since its introduction in 2006:
  - US Census Bureau in 2012: *On The Map* project
    - Where people are employed and where they live
  - Apple in 2016: iOS 10
    - User data collection, e.g. for emoji suggestions
    - https://images.apple.com/privacy/docs/Differential_Privacy_Overview.pdf
  - NSW Department of Transport open release of 2016 Opal ticketing system data
    - https://opendata.transport.nsw.gov.au/sites/default/files/resources/Open%20Opal%20Data%20Documentation%20170728.pdf

Global differential privacy: Our focus

k-anonymity
l-diversity

Differential privacy



What is being protected?

- Imagine a survey is asking you:
  - How old are you?
    - Result: Number of individuals >40 will be reported
  - What is your gender?
    - Result: Number of females will be reported
  - Are you a smoker?
    - Result: Number of smokers will be reported

ID	Age	Gender	Smoker
sdhj5vbg	20	Male	False
wu234u4	25	Female	True
hi384yrh	17	Female	False
po92okwj	50	Male	False

- Would you take part in it?

THE UNIVERSITY OF MELBOURNE

What is being protected?

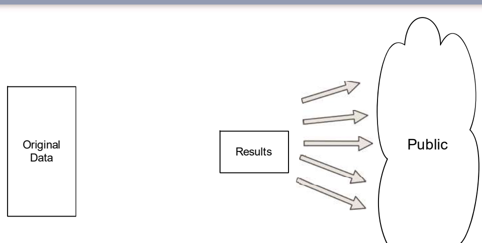
I would feel safe submitting the survey if:

I know the chance that the privatized result would be R was nearly the same, whether or not I take part in the survey.

- Does this mean that an individual's answer has no impact on the release result?

THE UNIVERSITY OF MELBOURNE

Overview of the process: Global differential privacy



- The privatized analysis comprises two steps:
  - Query the data and obtain the real result, e.g., how many female students are in the survey?
  - Add random noise to hide the presence/absence of any individual. Release noisy result to the user.

$D_{original} \xrightarrow{\text{Query}} R_{real} + \text{Noise} \longrightarrow R_{released}$

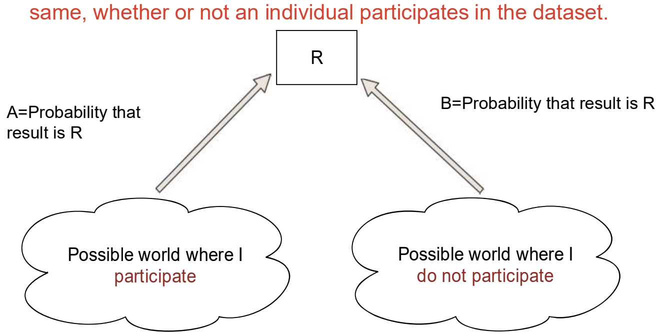
THE UNIVERSITY OF MELBOURNE

The released results will be different each time (different amount of noised added)

- Query: How many females in the dataset? (true result = 32)
- Generate some random values, according to a distribution with mean value 0: {1,2,-2,-1,0,-3,1,0}, add to true result and release
  - 1st query: Released result=33 (32+1)
  - 2nd query: Released result=34 (32+2)
  - 3rd query: Released result=30 (32-2)
  - 4th query: Released result=31 (32-1)
  - 5th query: Released result=32 (32+0)
  - 6th query: Released result=29 (32-3)
  - 7th query: Released result=33 (32+1)
  - 8th query: Released result=32 (32,0)
  - …
- On average, the released result will be 32, but observing a single released result doesn't give the adversary exact knowledge

The promise of differential privacy

- The chance that the noisy released result will be R is nearly the same, whether or not an individual participates in the dataset.



- If we can guarantee A≅B (A is very close to B), then no one can guess which possible world resulted in R.

THE UNIVERSITY OF MELBOURNE

The promise of differential privacy

- Does this mean that the attacker cannot learn anything sensitive about individuals from the released results?

THE UNIVERSITY OF MELBOURNE

Differential privacy: How?

- How much noise should we add to the result? This depends on
  - **Privacy loss budget:** How private we want the result to be (how hard for the attacker to guess the true result)
  - **Global sensitivity:** How much difference the presence or absence of an individual could make to the result.

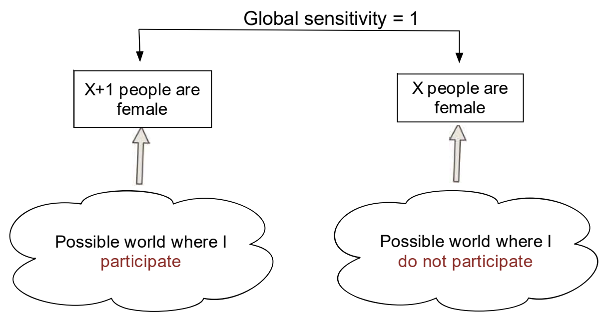
THE UNIVERSITY OF MELBOURNE

Global sensitivity

- Global sensitivity of a query Q is the maximum difference in answers that adding or removing any individual from the dataset can cause (maximum effect of an individual)
- Intuitively, we want to consider the worst case scenario
- If asking multiple queries, global sensitivity is equal to the sum of the differences

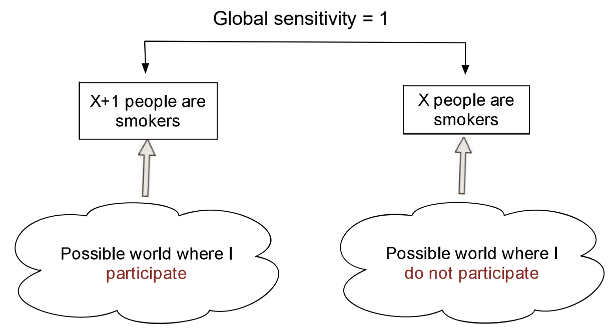
Global sensitivity

- QUERY: How many people in the dataset are female?



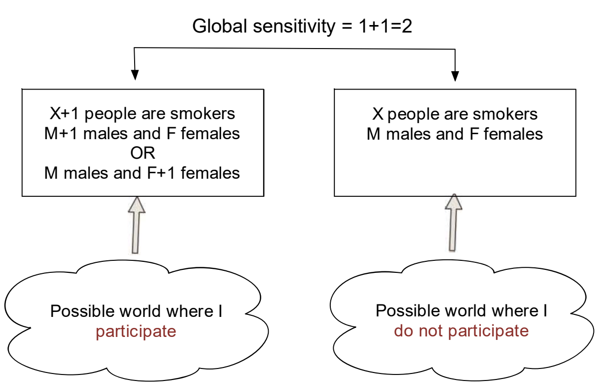
Global sensitivity

- QUERY: How many people in the dataset are smokers?



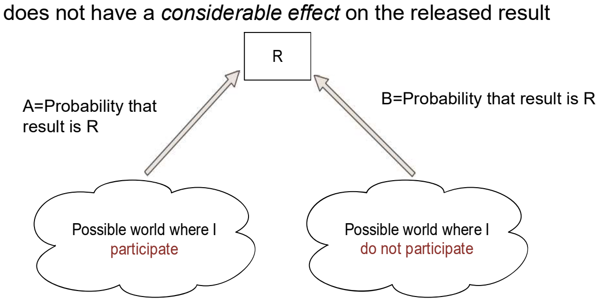
Global sensitivity

- QUERY: How many people in the dataset are female? And how many people are smokers?



Privacy loss budget = k

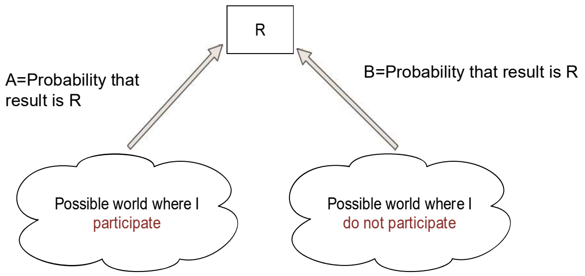
- We want that the presence or absence of a user in the dataset does not have a *considerable effect* on the released result



Privacy loss budget = k ($k \geq 0$)

Choose k to guarantee that $A \leq 2^k \times B$

Privacy loss budget = k



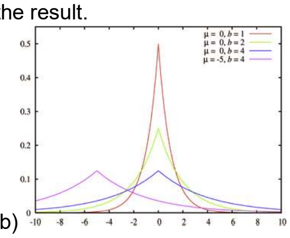
Privacy loss budget=k (k ≥0)

Choose k to guarantee that A ≤ 2^k × B

- k=0: No privacy loss (A=B), low utility
- k=high: Larger privacy loss, higher utility
- k=low: Low privacy loss, lower utility

Differential privacy: How?

- How much noise should we add to the result? This depends on
  - **Privacy loss budget (k):** How private we want the result to be (how hard for the attacker to guess the true result)
  - **Global sensitivity (G):** How much difference the presence of absence of an individual could make to the result.
- Strategy: Add noise to the result according to
  - Released result = True result + noise
    - Where noise is a number randomly sampled from a distribution having
      - average value = 0 (μ)
      - standard deviation (spread)= G/k (b)
    - Details about the distribution are beyond the scope of our study (it is called the Laplace distribution)



Summary

- Differential privacy guarantees that the presence or absence of a user cannot be revealed after releasing the query result
  - It does not prevent attackers from drawing conclusions about individuals from the aggregate results over the population
- We need to determine the budget and global sensitivity to know what is the scale of the noise to be added

Further reading

- **Protecting unit-record level personal information:** The limitations of de-identification and the implications for the *Privacy and Data Protection Act 2014*
  - https://www.cpdp.vic.gov.au/images/content/pdf/privacy_papers/20180503-De-identification-report-OVIC-V1.pdf

Final two lectures

- Ethics (Wednesday)
- Wrap up (Friday)