A Taiwanese embassy worker in Haiti is browsing Mandarin Wikipedia at work, and the Chinese government is trying to spy on them…

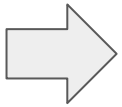
Date	Zh Wikipedia Views in Haiti
4 Jan 2022	32
5 Jan 2022	42
6 Jan 2022	48
7 Jan 2022	39
8 Jan 2022	33
9 Jan 2022	68
10 Jan 2022	39
11 Jan 2022	20
Average views per day	**40.125**

Let’s say that there are 26 pages. Our embassy worker looks at sensitive page Z (a rare medical condition, let’s say) one time…

Country	Language	Page	Views
Haiti	Mandarin	Safe page A	19
Haiti	Mandarin	Safe page B	8
…	…	…	…
Haiti	Mandarin	Safe page Y	0
Haiti	Mandarin	Sensitive page Z	1

K-anonymity/bucketing

Country	Language	Page	Views
Haiti	Mandarin	Safe page A	19
Haiti	Mandarin	Safe page B	8
…	…	…	…
Haiti	Mandarin	Safe page Y	0
Haiti	Mandarin	Sensitive page Z	1



Country	Language	Page	Views
Haiti	Mandarin	Safe page A	11-20
Haiti	Mandarin	Safe page B	1-10
…	…	…	…
Haiti	Mandarin	Safe page Y	0
Haiti	Mandarin	Sensitive page Z	1-10

K-anonymity/bucketing: What could good faith users learn here?

- Page A is the most popular
- Page Y is least popular
- Pages B and Z are around equally popular in the middle
    - This is actually *wrong* — page B is significantly more popular than page Z

Country	Language	Page	Views
Haiti	Mandarin	Safe page A	11-20
Haiti	Mandarin	Safe page B	1-10
…	…	…	…
Haiti	Mandarin	Safe page Y	0
Haiti	Mandarin	Sensitive page Z	1-10

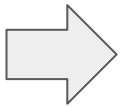
K-anonymity/bucketing: What could the government learn here?

- 100% certainty: There is a person in Haiti who looked at the Mandarin wikipedia page for disease Z that day
- If you combine that with some outside information:
  - Evidence that the embassy worker recently went to the doctor
  - Access logs to the embassy building so you know that person had internet access that day
  - IP request logs from the embassy buildings so you know someone was on Wikipedia
- You could be almost certain that that the embassy worker has disease Z

Country	Language	Page	Views
Haiti	Mandarin	Safe page A	11-20
Haiti	Mandarin	Safe page B	1-10
…	…	…	…
Haiti	Mandarin	Safe page Y	0
Haiti	Mandarin	Sensitive page Z	1-10

Differential Privacy

Country	Language	Page	Views
Haiti	Mandarin	Safe page A	19
Haiti	Mandarin	Safe page B	8
…	…	…	…
Haiti	Mandarin	Safe page Y	0
Haiti	Mandarin	Sensitive page Z	1



Country	Language	Page	Views
Haiti	Mandarin	Safe page A	17
Haiti	Mandarin	Safe page B	11
…	…	…	…
Haiti	Mandarin	Safe page Y	2
Haiti	Mandarin	Sensitive page Z	0

Differential Privacy: What could good faith users learn here?

- Page A is the most popular
- Page B has middling popularity
- Pages B and Z are around equally unpopular

More accurate picture than K-anonymity

Country	Language	Page	Views
Haiti	Mandarin	Safe page A	17
Haiti	Mandarin	Safe page B	11
…	…	…	…
Haiti	Mandarin	Safe page Y	2
Haiti	Mandarin	Sensitive page Z	0

Differential Privacy: What could the government learn here?

- Noise added (some positive, some negative) is small and averages to 0 over the entire dataset
- Overall statistics of the dataset are unchanged, but a small amount of error is added, similar to sampling error in polls
- However, noise protects against any kind of attack *that is reliant on the dataset*
  - Impossible to know if the embassy worker looked at disease Z on Wikipedia
  - Impossible to know if the embassy worker is even contained in the dataset
- Importantly, the government can still deduce things from other data sources, but the dataset we are releasing is (1) anonymized and (2) provides no extra information about the people in it

Country	Language	Page	Views
Haiti	Mandarin	Safe page A	17
Haiti	Mandarin	Safe page B	11
…	…	…	…
Haiti	Mandarin	Safe page Y	2
Haiti	Mandarin	Sensitive page Z	0

Summary

- K-anonymity injects more error and less specificity into our data, is vulnerable to reidentification and linkage attacks from outside data sources, and has no measurable privacy metric

- Differential privacy is generally more accurate, is invulnerable to reidentification and linkage attacks, and has a measurable (and by extension, governable) privacy metric