CS208: Applied Privacy for Data Science
DP Foundations: the Laplace Mechanism

School of Engineering & Applied Sciences
Harvard University

February 10, 2022

Pereira et al. US Broadband Coverage Data Set: A DP Data Release

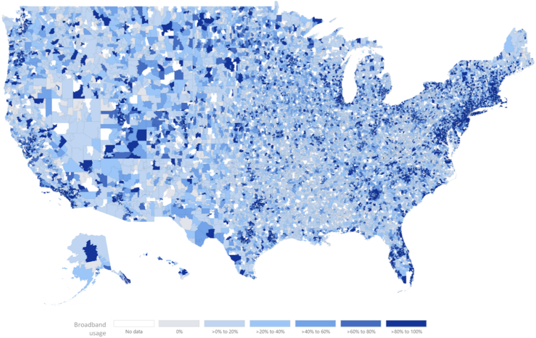


Fig. 1. Map of the United States by postal codes with indicators of broadband coverage.

From Microsoft Services, we query the following Windows telemetry data:
- L_z: Counts of devices connecting to Microsoft Services with internet speed lower than 25Mbps in zip code z.
- H_z: Counts of devices connecting to Microsoft Services with internet speed greater or equal than 25Mbps in zip code z.

Additionally, we query Microsoft Services for the following data:
- M_z: Counts of devices utilizing Microsoft Services in zip code z.
- O_z: Counts of devices not utilizing Microsoft Services in zip code z.

3.1 Privacy Loss

The privacy loss computation is a straightforward application of the parallel and sequential composition properties of differential privacy mechanisms [6]. The total privacy loss resulted from querying the internet speed telemetry is $\epsilon = 0.1$. Given that L_z^{DP} and H_z^{DP} are differentially private count queries applied to disjoint subsets of the data, from parallel composition we know that the privacy guarantee depends only on the worst of the guarantees of each analysis. The same happens when computing the privacy loss incurred from the Microsoft Services devices queries. The count queries are applied to disjoint subsets of data, resulting in an additional privacy loss of $\epsilon = 0.1$. From sequential composition we have that sequences of queries accumulate privacy costs additively. Finally, based on the post-processing immunity property described in theorem 2, the total privacy cost of the Broadband Coverage Estimates calculation is $\epsilon = 0.2$.

Differential Privacy

M is ϵ-DP if

$$Pr[M(D,q) \in T] \leq (1+\epsilon) Pr[M(D',q) \in T], \qquad \forall\, T, q.$$

Differential Privacy

M is ϵ-DP if

$$Pr[M(D,q) \in T] \leq (1+\epsilon) Pr[M(D',q) \in T], \qquad \forall\, T, q.$$

- D, D' Neighbouring datasets

Differential Privacy

M is ϵ-DP if

$$Pr[M(D,q) \in T] \leq e^{\epsilon} Pr[M(D',q) \in T], \qquad \forall\, T, q.$$

- D, D' Neighbouring datasets
- M Mechanism that Maps from data to result

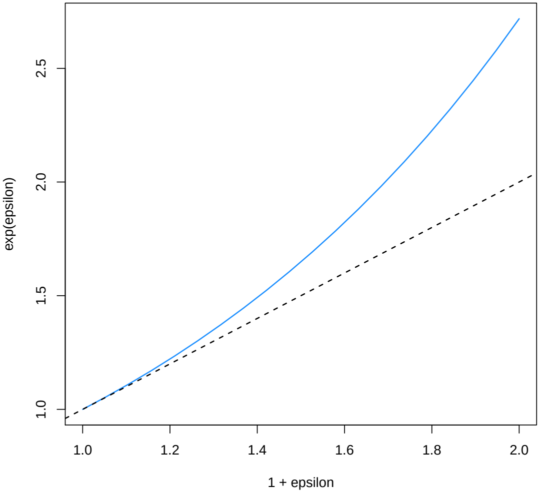
Differential Privacy

M is ϵ-DP if

$$Pr[M(D, q) \in T] \leq e^{\epsilon} Pr[M(D', q) \in T], \qquad \forall\, T, q.$$

- D, D' Neighbouring datasets
- M Mechanism that Maps from data to result
- q Query
- T Set providing a decision rule

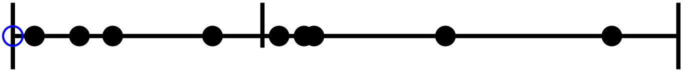
e^{ϵ} vs. $(1+\epsilon)$

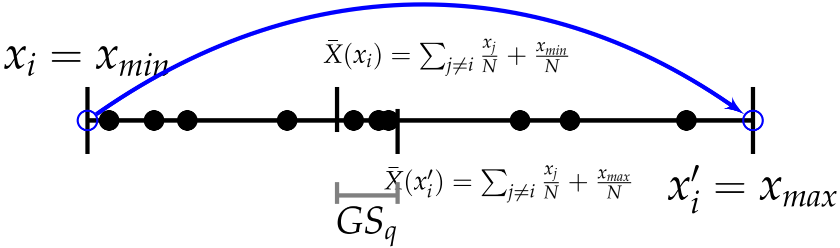


see `expEpsilon.r`

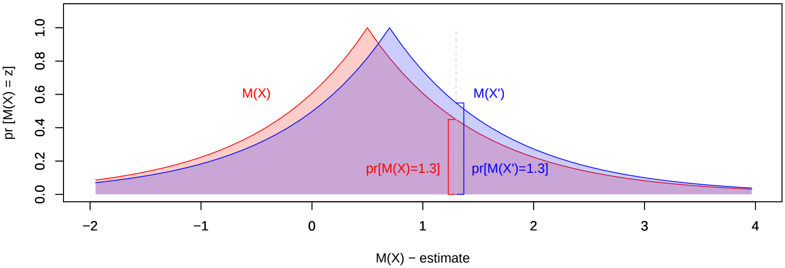
$x_i = x_{min}$

$\bar{X}(x_i) = \sum_{j \neq i} \frac{x_j}{N} + \frac{x_{min}}{N}$





$x_i = x_{min}$
$\bar{X}(x_i) = \sum_{j \neq i} \frac{x_j}{N} + \frac{x_{min}}{N}$
$\bar{X}(x'_i) = \sum_{j \neq i} \frac{x_j}{N} + \frac{x_{max}}{N}$
$x'_i = x_{max}$
GS_q



M(X)
M(X')
pr[M(X)=1.3]
pr[M(X')=1.3]
pr [M(X) = z]
0.0
0.2
0.4
0.6
0.8
1.0
−2
−1
0
1
2
3
4
M(X) − estimate

The Laplace has density over y:

$$f_{Laplace}(y|s,\mu) = \mathrm{Lap}(s,\mu) = \frac{1}{2s}\exp\Big(-\frac{|y-\mu|}{s}\Big)$$

The Laplace has density over y:

$$f_{Laplace}(y|s) = \text{Lap}(s) = \frac{1}{2s}\exp\Big(-\frac{|y|}{s}\Big)$$

We were given the theorem:

$$M(x, q) = q(x) + \text{Lap}(GS_q/\epsilon)$$

The Laplace has density over y:

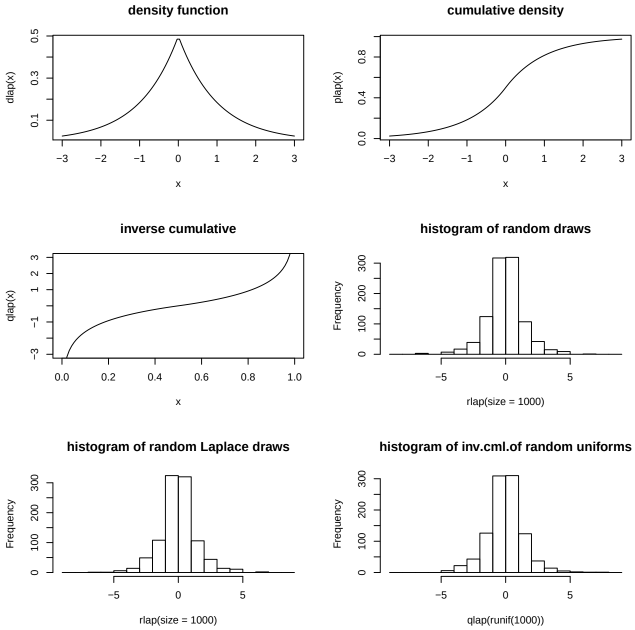
$$f_{Laplace}(y|s) = \text{Lap}(s) = \frac{1}{2s} \exp\Big(-\frac{|y|}{s} \Big)$$

We were given the theorem:

$$M(x, q) = q(x) + \text{Lap}(GS_q/\epsilon)$$

So our differentially private mean, $M(X)$, which combines the "true" sample mean with Laplace noise, becomes:

$$M(x) = \bar{x} + Z; \qquad Z \sim \text{Lap}(s = GS_q/\epsilon)$$

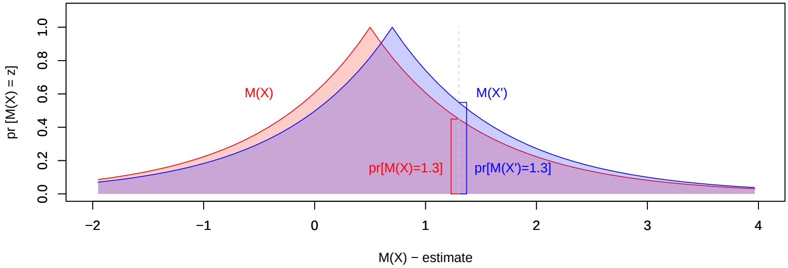


see `showLaplaceDistributions.r`

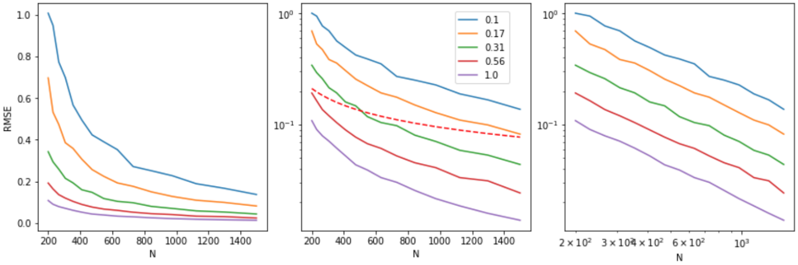
$$\frac{pr[M(x)=t]}{pr[M(x')=t]} = \frac{e^{\frac{-\epsilon|\bar{x}-t|}{GS_q}}}{e^{\frac{-\epsilon|\bar{x}'-t|}{GS_q}}} = e^{\frac{\epsilon|\bar{x}'-t|-\epsilon|\bar{X}-t|}{GS_q}} = e^{\frac{\epsilon|\bar{x}'-\bar{x}|}{GS_q}} \leq e^{\epsilon}$$

since we know $GS_q \geq |\bar{x}' - \bar{x}|$ by the def. of sensitivity. Thus we meet the original definition:

$$Pr[M(x)=t] \leq e^{\epsilon} Pr[M(x')=t]$$



Two Laplace distributions, for two adjacent datasets x and x'. The definition of ϵ-differential privacy requires the ratio of $M(x)/M(x')$ is not greater than e^ϵ for all points along the x-axis. Thus for any realized output z (for example here, $z = 1.3$) we can not determine that x or x' were more likely to have produced z.



see `laplace_mechanism_and_opendp.ipynb`