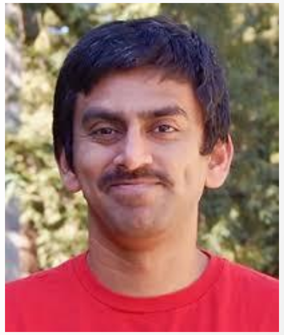
Amplification by Shuffling:
From Local to Central Differential Privacy via Anonymity

Vitaly Feldman

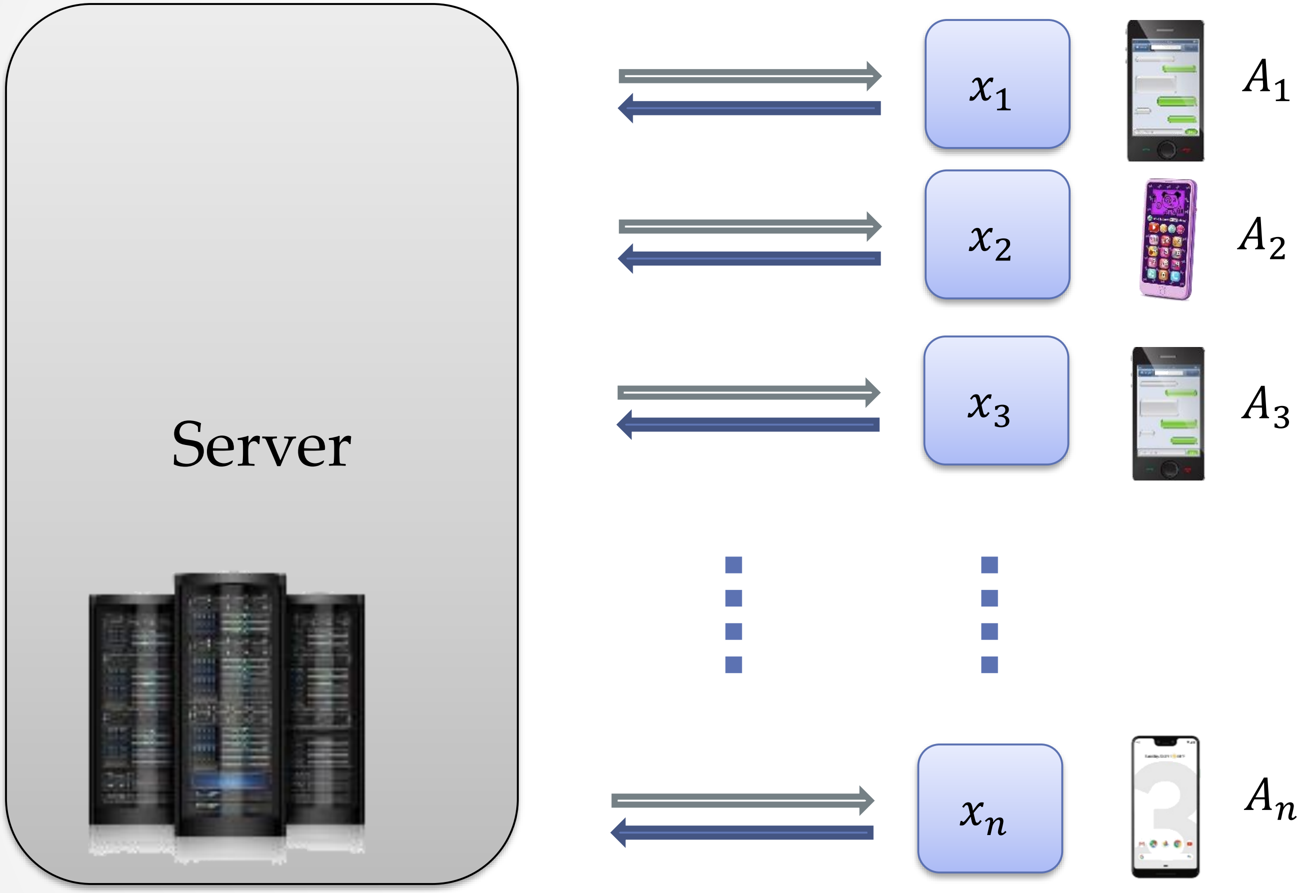
Ulfar Erlingsson Ilya Mironov Ananth Raghunathan Kunal Talwar Abhradeep Thakurta





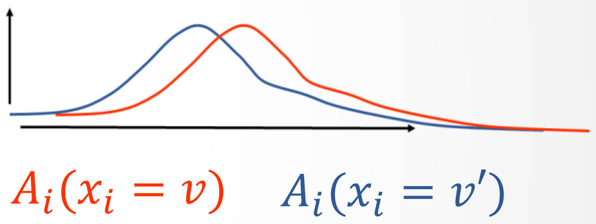


Local Differential Privacy (LDP)



Compute (approximately)
$f(x_1, x_2, \ldots, x_n)$

For all i, A_i is a local ϵ-DP randomizer:
for all $v, v' \in X$



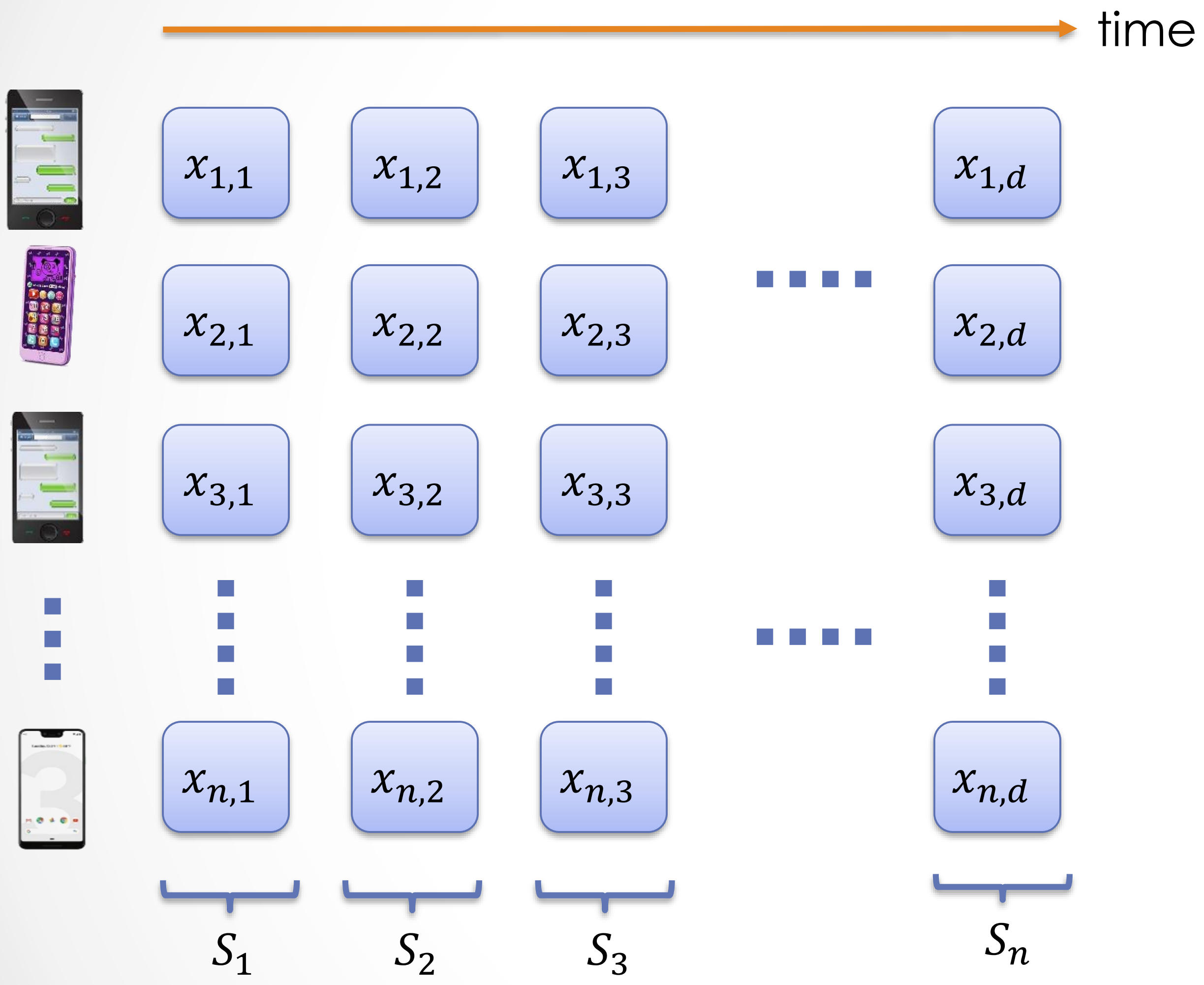
[Warner '65; EGS '03; KLNRS '08]

Outline

Online monitoring with LDP

Benefits of anonymity:
privacy amplification by shuffling

Online monitoring



$x_{i,j} \in \{0,1\}$

Status of user i on day j

Assume that each user's status changes at most k times

- only for utility

Estimate the daily counts $S_j = \sum_{i=1}^{n} x_{i,j}$ for all $j \in [d]$

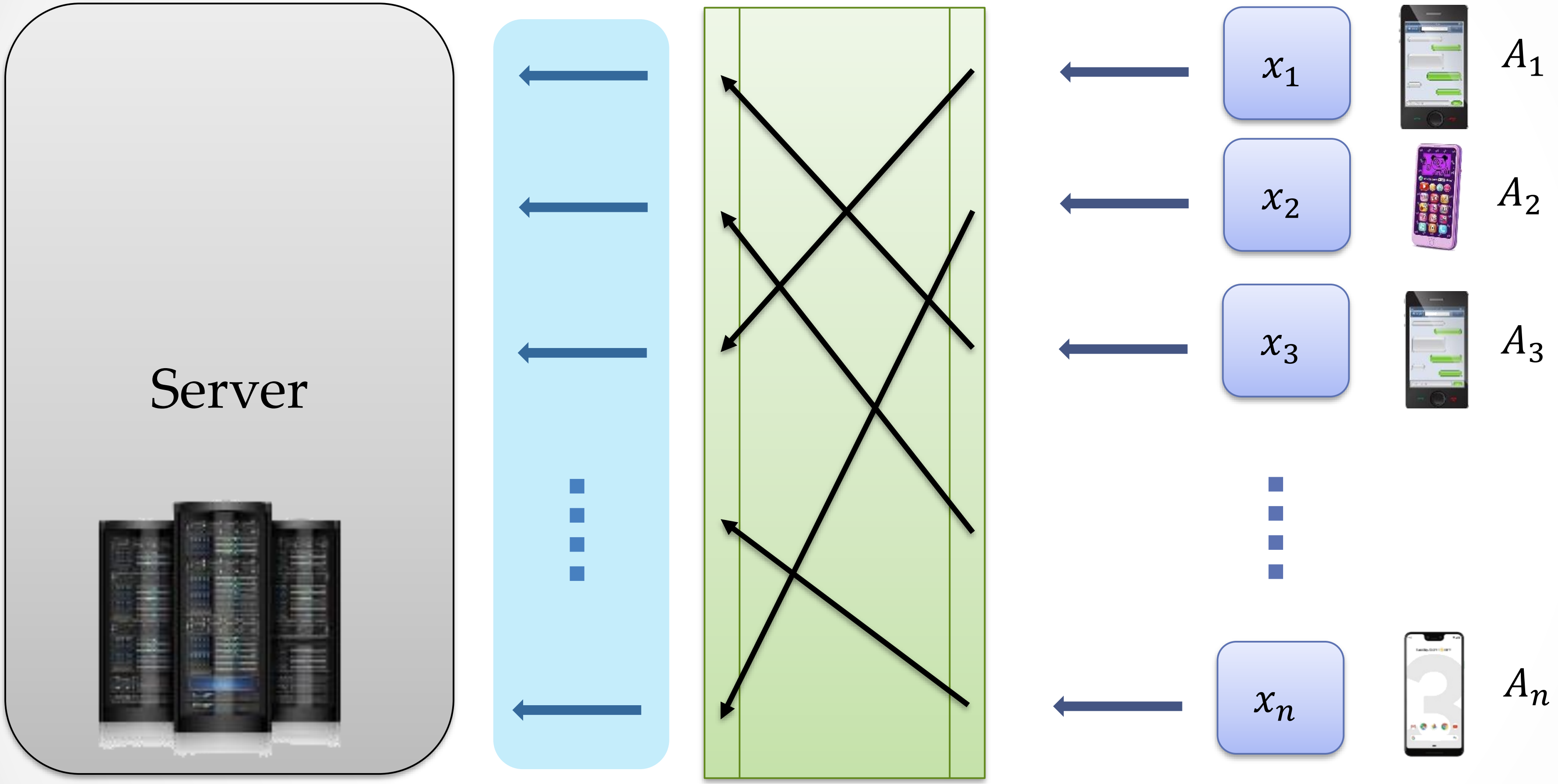
Monitoring with LDP

There exists an ϵ-LDP algorithm that constructs estimates $\hat{S}_1, \hat{S}_2, \dots, \hat{S}_d$ such that with high prob. for all $j \in [d]$,

$$|S_j - \hat{S}_j| = O\left(\frac{\sqrt{n}k\,(\log d)^2}{\epsilon}\right)$$

- Report the status changes (only first k)
- Maintains a tree of counters each over an interval of time
- Based on [DNPR '10; CSS '11]

Encode-Shuffle-Analyze (ESA) [Bittau et al. '17]



Shuffle and anonymize

Privacy amplification by shuffling

For any $\epsilon = O(1)$ and any sequence of ϵ-LDP algorithms $(A_1, \ldots, A_n)$, let

$$A_{\text{shuffle}}(x_1, \ldots, x_n) = A_1\left(x_{\pi(1)}\right), A_2\left(x_{\pi(2)}\right), \ldots, A_n\left(x_{\pi(n)}\right)$$

for a random and uniform permutation $\pi: [n] \to [n]$

Then A_{shuffle} is (ϵ', δ)-DP in the central model for $\epsilon' = O\left(\frac{\epsilon\sqrt{\log(1/\delta)}}{\sqrt{n}}\right)$

Holds for adaptive case: A_i may depend on outputs of $A_1, \ldots, A_{i-1}$

Comparison with subsampling

Running ϵ-DP algorithm on random q-fraction of elements is $\approx q\epsilon$-DP ($\epsilon \leq 1$) **[KLNRS '08]**

Shuffling includes all elements so $q = 1$

Output $A_1(x_{i_1}), A_2(x_{i_2}), \ldots, A_n(x_{i_n})$

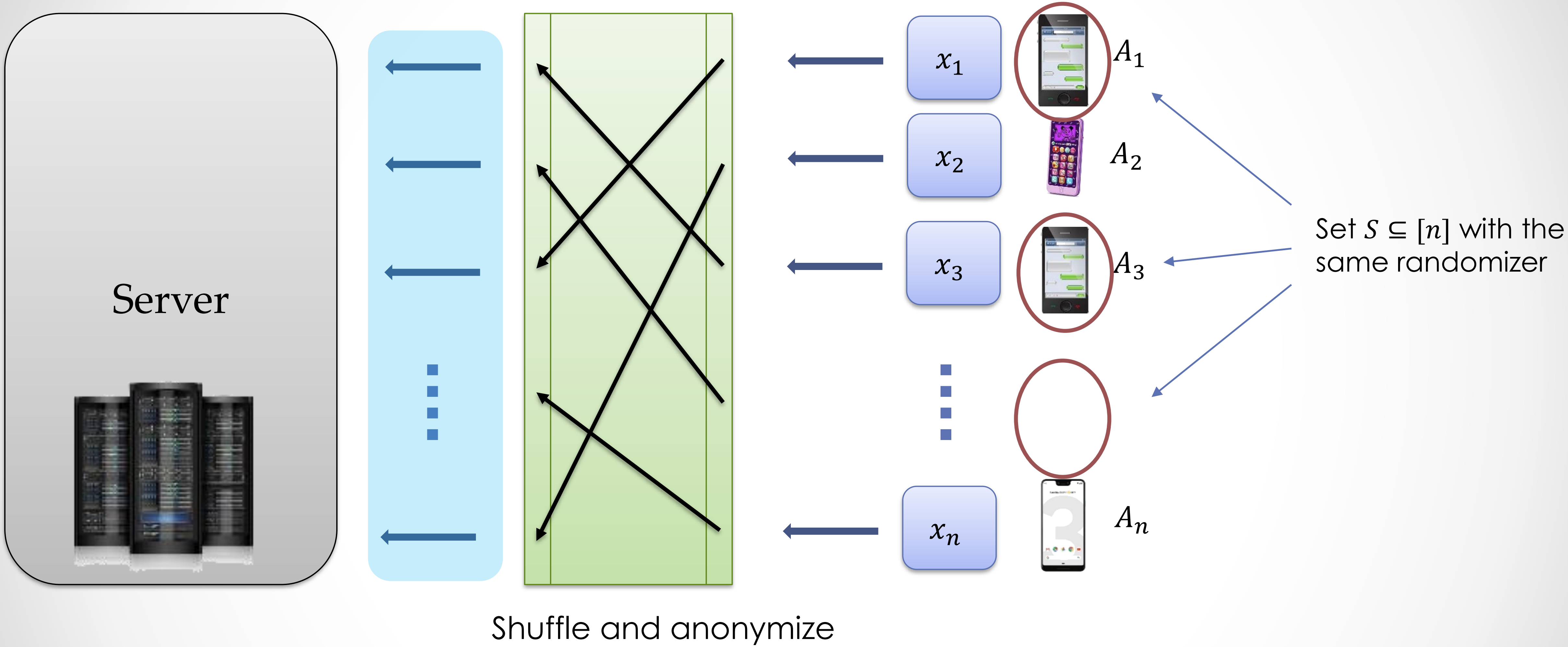
where $i_1, i_2, \ldots, i_n \sim [n]$ (independently) is (ϵ', δ)-DP for $\epsilon' = O\left(\frac{\epsilon\sqrt{\log(1/\delta)}}{\sqrt{n}}\right)$

e.g. **[BST '14]**

Advantages of shuffling:

- does not affect the statistics of the dataset
- does not increase LDP cost

Implications for ESA



For every $i \in S$, the output is $\left(O\left(\frac{\epsilon\sqrt{\log(1/\delta)}}{\sqrt{|S|}}\right), \delta\right)$-DP for element at position i

Special case: binary randomized response

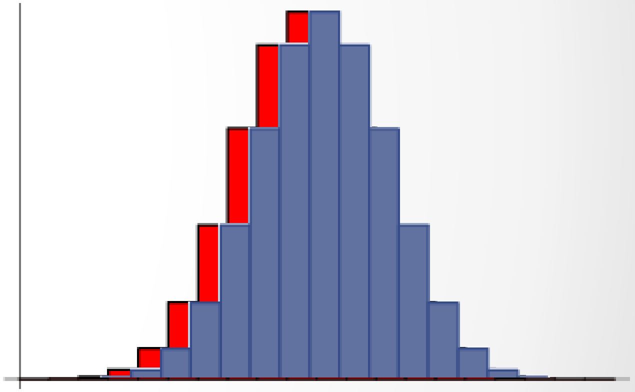
RR: For $x \in \{0,1\}$, return x flipped with probability $1/3$. Satisfies $(\log 2)$-LDP

Output distribution is determined by $m = \#_1(\mathrm{RR}(x_1), \ldots, \mathrm{RR}(x_n))$

$m \sim \mathrm{Bin}\left(k, \frac{2}{3}\right) + \mathrm{Bin}\left(n-k, \frac{1}{3}\right)$, where $k = \#_1(x_1, \ldots, x_n)$

For a neighboring dataset: $k' = k \pm 1$

$$\mathrm{Bin}\left(k, \frac{2}{3}\right) + \mathrm{Bin}\left(n-k, \frac{1}{3}\right) \approx_{\left(\sqrt{\frac{\log(1/\delta)}{n}}, \delta\right)} \mathrm{Bin}\left(k+1, \frac{2}{3}\right) + \mathrm{Bin}\left(n-k-1, \frac{1}{3}\right)$$



[DKMMN ‘06]

Also given in [Cheu,Smith,Ullman,Zeber,Zhilyaev ‘18] (independently)

Conclusions

- Monitoring with LDP and log dependence on time
- General privacy amplification technique
  - Match state of the art in the central model
  - Can be used to derive lower bounds for LDP
- Provable benefits of anonymity for ESA-like architectures
- To appear in SODA 2019
- arxiv.org/abs/1811.12469