Hadamard Response: Local Private Distribution Estimation

Jayadev Acharya, Ziteng Sun, Huanyu Zhang

Cornell University

Information Theory and Applications (ITA), 2018

Based on:

https://arxiv.org/abs/1802.04705

Distribution Estimation

- $\boldsymbol{p}$: **unknown** discrete distribution over $\boldsymbol{k}$ elements
- $\boldsymbol{\alpha}$: accuracy
- Input: independent samples $X_1, X_2, \ldots, X_{\boldsymbol{n}}$ from $\boldsymbol{p}$
- Output: $\widehat{\boldsymbol{p}}$ such that w.p. at least 0.9:

$$d(\boldsymbol{p}, \widehat{\boldsymbol{p}}) \leq \boldsymbol{\alpha}$$

- We consider $\ell_{\mathbf{1}}, \ell_{\mathbf{2}}$ distances

Sample Complexity

Sample Complexity: Least $\boldsymbol{n}$ to estimate $\boldsymbol{p}$

To estimate to $\ell_{\mathbf{1}} \leq \boldsymbol{\alpha}$:

$$\Theta\left(\frac{\boldsymbol{k}}{\boldsymbol{\alpha}^{2}}\right)$$

Empirical distribution works

Distribution Estimation with Privacy

- Samples are sensitive

- Drug abuse
  - Learn underlying drug usage behavior (for policy design)
  - Maintain privacy of users

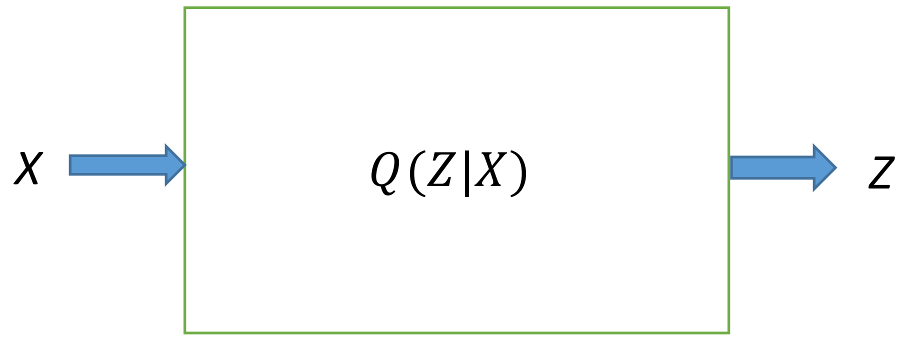
- Internet
  - Distribution of web traffic to websites
  - Maintain browsing of a particular user private

Model

- $X_1 \ldots X_{\boldsymbol{n}}$ stored over $\boldsymbol{n}$ users
- User i transmits Z_i to data collector/server
- Server has to learn $\boldsymbol{p}$
- Without privacy: send X_i

Local Differential Privacy (LDP)

- Q: a channel with input $[k]$ and output $\mathcal{Z}$



$\boldsymbol{\varepsilon}$-LDP [DuchiWainwrightJordan'12, ErlingssonPihurKorolova'14]:

$$\frac{Q(z|x)}{Q(z|x')} \le e^{\boldsymbol{\varepsilon}}$$

User i passes X_i through Q, send output Z_i

Randomized Response (RR)

[Warner'65, KairouzBonawitzRamadge'14]: $\mathcal{Z} = [\boldsymbol{k}]$

$$Q_\varepsilon(z|x) = \begin{cases} \dfrac{e^{\boldsymbol{\varepsilon}}}{e^{\boldsymbol{\varepsilon}} + \boldsymbol{k} - 1}, & z = x \\ \dfrac{1}{e^{\boldsymbol{\varepsilon}} + \boldsymbol{k} - 1}, & z \neq x \end{cases}$$

Optimal only in the low privacy regime $(\boldsymbol{\varepsilon} > \log \boldsymbol{k})$

RAPPOR

[DuchiWainwrightJordan'12, ErlingssonPihurKorolova'14]: $\mathcal{Z} = \{0,1\}^{k}$.

- One hot encoding: $x \to e_x$ (basis vector with xth entry 1)
- Flip each entry in e_x with probability $\frac{1}{e^{\varepsilon/2}+1}$

$e_x, e_{x'}$ differ in at most two positions

- Optimal only for $\varepsilon \lesssim 1$, and $\varepsilon > 2 \log k$

Subset Selection (SS)

[WangHuangWangNieXuYangLiQiao'16, YeBarg'17]:

$\mathcal{Z}$: strings in $\{0,1\}^{\boldsymbol{k}}$ with Hamming weight $\left\lceil \frac{\boldsymbol{k}}{e^{\boldsymbol{\varepsilon}}+\mathbf{1}} \right\rceil$

Optimal in all regimes

Sample Complexity

ε	RR	RAPPOR	SS	HR
$(0,1)$	$\frac{k^3}{\varepsilon^2\alpha^2}$	$\frac{k^2}{\varepsilon^2\alpha^2}$	$\frac{k^2}{\varepsilon^2\alpha^2}$	$\frac{k^2}{\varepsilon^2\alpha^2}$
$(1, \log k)$	$\frac{k^3}{e^{2\varepsilon}\alpha^2}$	$\frac{k^2}{e^{\varepsilon/2}\alpha^2}$	$\frac{k^2}{e^{\varepsilon}\alpha^2}$	$\frac{k^2}{e^{\varepsilon}\alpha^2}$

For constant ε, say $\varepsilon = 1$,

$$\frac{k}{\alpha^2} \rightarrow \frac{k^2}{\alpha^2}$$

Other Resources

Computational Complexity:

What is the encoding/decoding time?

Impractical if high running time, even if sample optimal

Communication Complexity:

How much communication to server?

Many papers considering these resources, including today on both!

Resources for $\varepsilon \in (0, 1)$

	RR	RAPPOR	SS	HR
Communication	$\log k$	k	k	$\log k$
Decoding time	n	$n \cdot k$	$n \cdot k$	n
Samples	$\frac{k^3}{\varepsilon^2 \alpha^2}$	$\frac{k^2}{\varepsilon^2 \alpha^2}$	$\frac{k^2}{\varepsilon^2 \alpha^2}$	$\frac{k^2}{\varepsilon^2 \alpha^2}$

How to claim bounds on time and communication?

Faithful implementation:

- Communication $\geq H(Z)$ bits.
- Decoding Time $\geq n \cdot H(Z)$

Communication requirements

	RR	RAPPOR	SS	HR
Communication	$\log \boldsymbol{k}$	$\log \boldsymbol{k} + \frac{\boldsymbol{k}}{\boldsymbol{e}^{\boldsymbol{\varepsilon}/2}}$	$\log \boldsymbol{k} + \frac{\boldsymbol{k}}{\boldsymbol{e}^{\boldsymbol{\varepsilon}}}$	$\log \boldsymbol{k}$

All these are entropy bounds!!

Other Resources

Large domain:
- Browsing patterns of internet users
- Distribution of product purchases of Target

Communication:
- Handheld devices with low uplink capacity
- Low battery power, 4G data

General encoding matrices

M: ± 1 matrix of size $\boldsymbol{k} \times \boldsymbol{K}$

h: #1's in each row

$$Q_\varepsilon(z|x) = \begin{cases} \dfrac{e^{\boldsymbol{\varepsilon}}}{e^{\boldsymbol{\varepsilon}} + \boldsymbol{K} - h}, & M(x,z) = +1 \\ \dfrac{1}{e^{\boldsymbol{\varepsilon}} + \boldsymbol{K} - h}, & M(x,z) = -1 \end{cases}$$

$$\begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{matrix} \quad \textbf{OR} \quad \begin{matrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{matrix}$$

Hadamard Matrix

H_m: $m \times m$ matrix

$H_1 = [1]$, and for other m:

$$H_m = \begin{bmatrix} H_{m/2} & H_{m/2} \\ H_{m/2} & -H_{m/2} \end{bmatrix}.$$

- The first row & column has m ‘1’s
- Every other row & column has $\frac{m}{2}$ ‘1’s
- Hamming distance between any two rows is $\frac{m}{2}$
- Matrix vector multiplication real fast!

(b, B)-Hadamard Matrix

b, B: powers of 2, and $K = b \cdot B$

$$H_K^b = \begin{pmatrix} H_b & P_b & \cdots & P_b \\ P_b & H_b & & P_b \\ \vdots & & \ddots & \vdots \\ P_b & P_b & \cdots & H_b \end{pmatrix}$$

P_b: $b \times b$ matrix with all entries '-1'

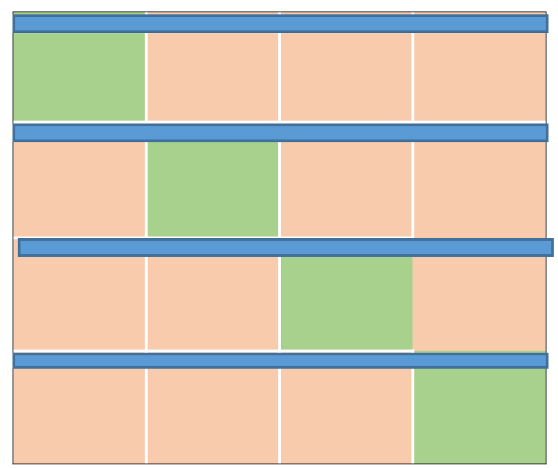
$B = 1,\ H_K^b = H_b$

$b = 1, H_K^b =$ Identity matrix

Encoding Matrix

Rows of H_K^b have different number of 1's

- Delete the first row of each embedded H_b
- The first $\boldsymbol{k}$ rows is the encoding matrix M



Selecting the parameters

B: largest power of 2 less than $\min\{e^{\varepsilon}, 2\boldsymbol{k}\}$

b: smallest power of 2 larger than $\left\lceil \frac{\boldsymbol{k}}{B} \right\rceil$

$$\boldsymbol{K} = B \cdot b \leq 4\boldsymbol{k}$$

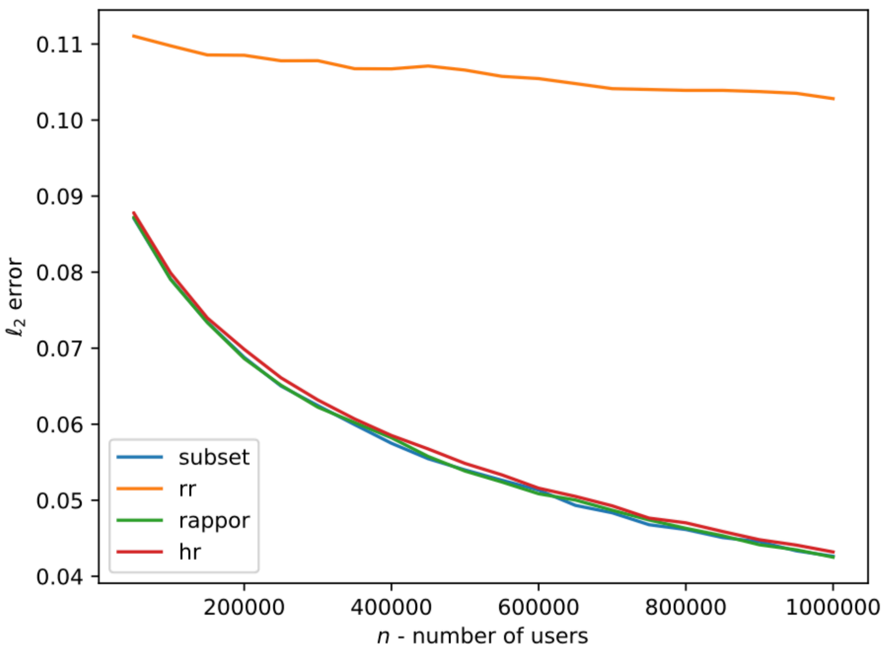
Communication: $\log \boldsymbol{K} \leq \log \boldsymbol{k} + 2$ bits.

Key arguments

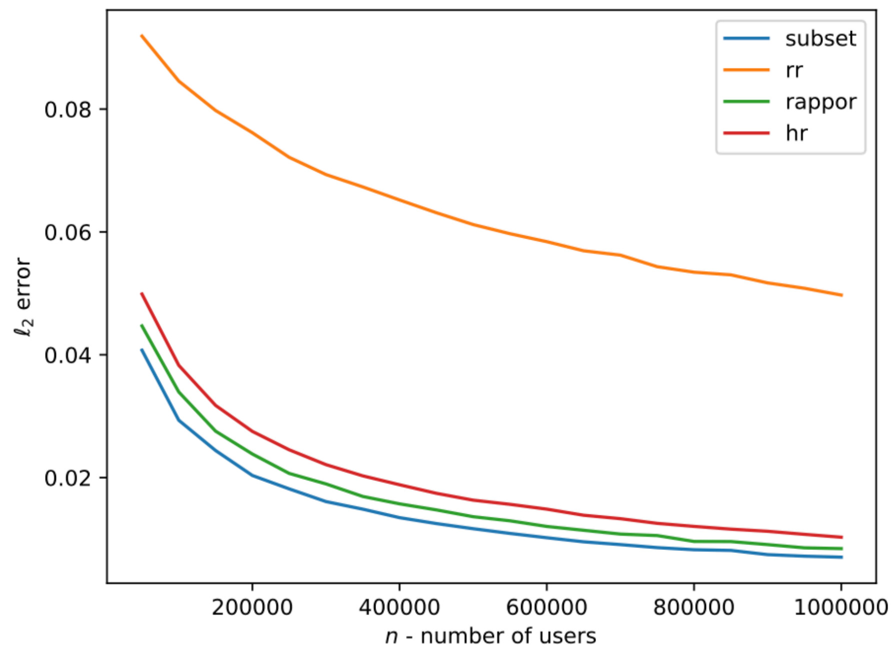
Large Hamming distance -> Sample Optimality

Fast Hadamard Transform -> Fast Decoding

L2 error plots ($k = 1000$, Geo(0.8))

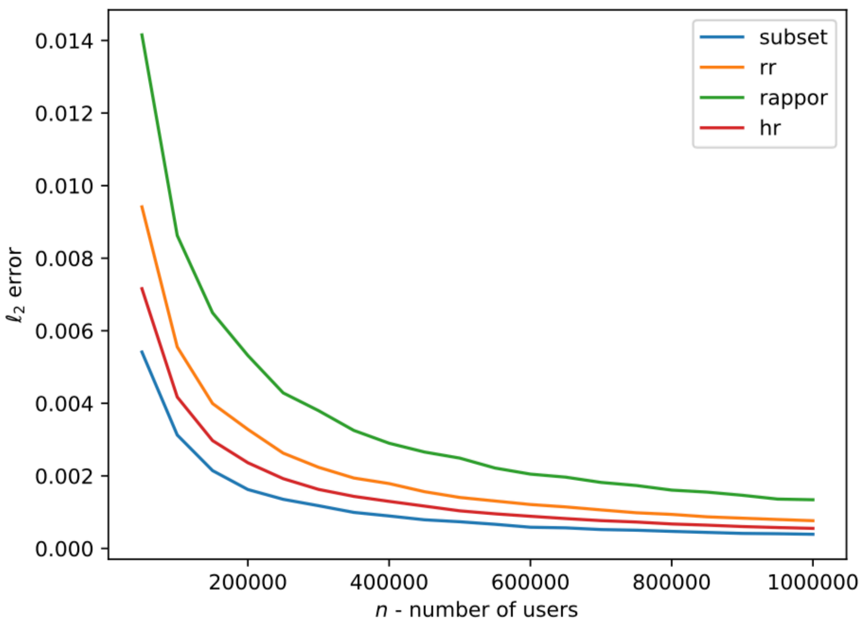


(a) $\varepsilon = 0.5$

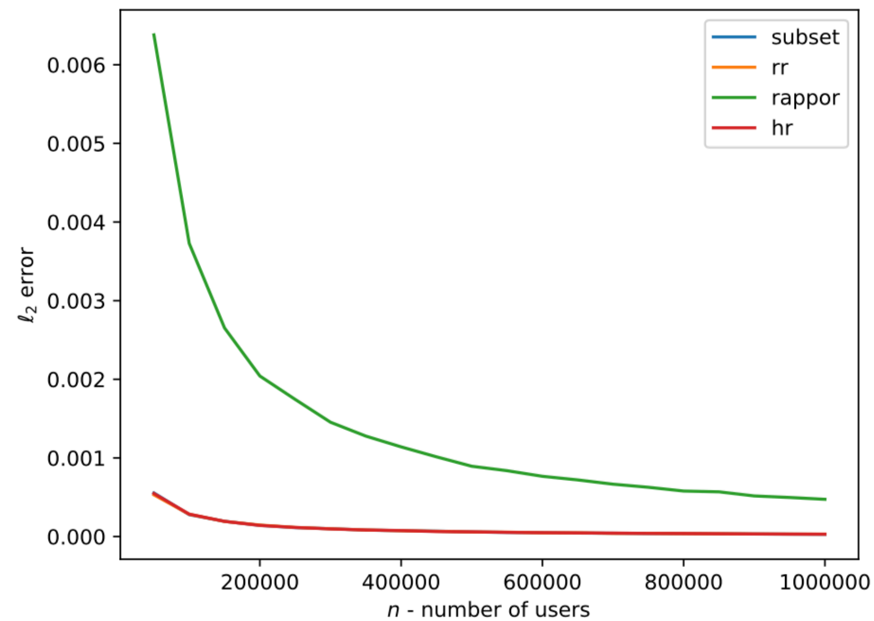


(b) $\varepsilon = 2$

L2 error plots ($k = 1000$, Geo(0.8))

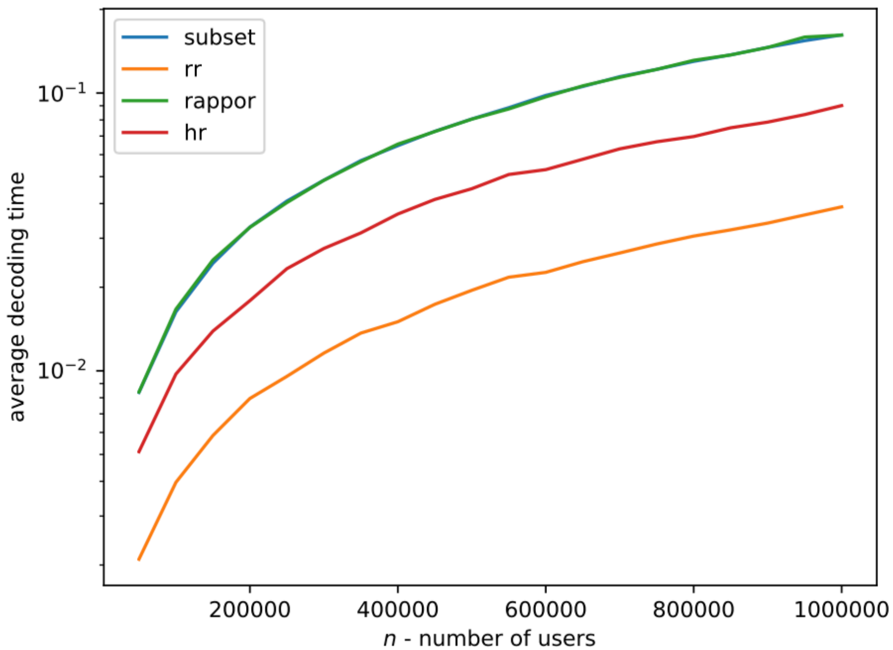


(c) $\varepsilon = 5$

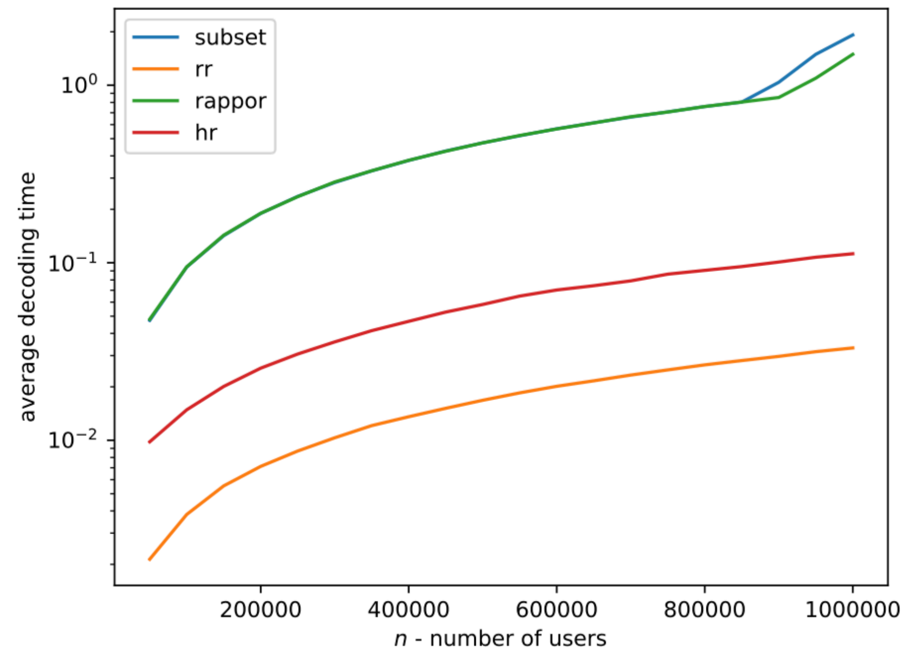


(d) $\varepsilon = 7$

Running time Geo(0.8)

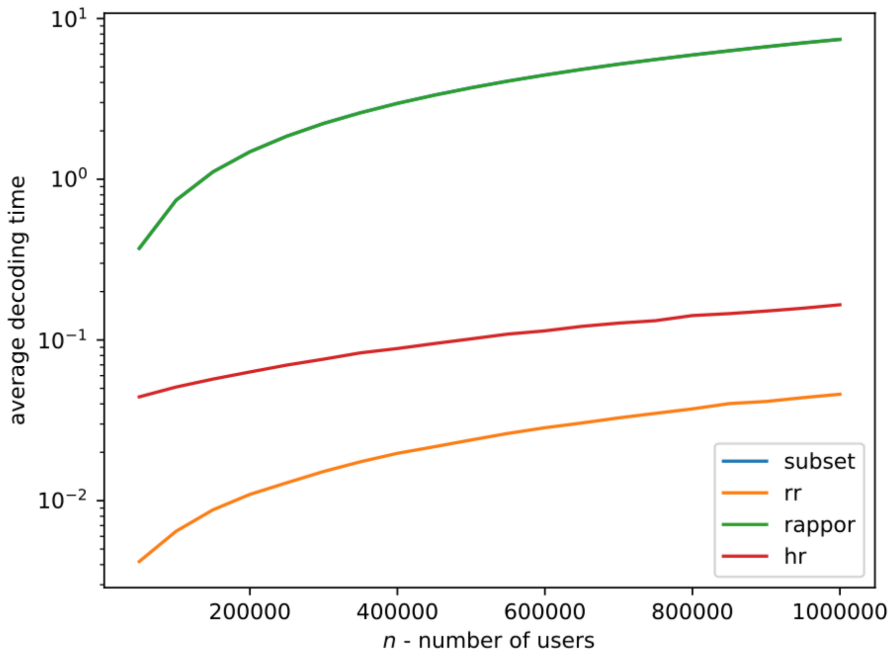


(a) $k = 100$

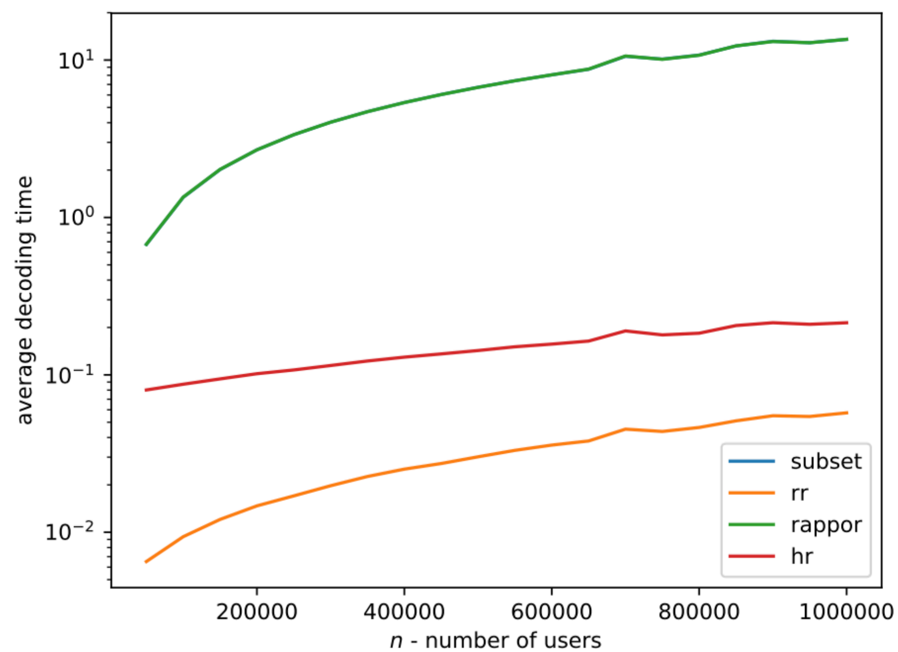


(b) $k = 1000$

Running time Geo(0.8)



(c) $k = 5000$



(d) $k = 10000$

Thank You

Details in the paper online!

https://arxiv.org/abs/1802.04705