The Privacy Blanket of the Shuffle Model

Borja Balle

Joint work with J. Bell, A. Gascón and K. Nissim

Arxiv: 1903.02837

Trust vs Accuracy

Trust vs Accuracy

Trust

Accuracy

Trust		Accuracy
Central		$\Theta(1)$
Local		$\Theta(n^{1/2})$

Statistical Queries

$q : \mathbb{X} \to [0,1]$

$$F_q(x_1, \ldots, x_n) = \sum_{i=1}^{n} q(x_i)$$

Trust vs Accuracy

Trust

Accuracy

Trust	Accuracy
Central	$\Theta(1)$
Shuffle	
Local	$\Theta(n^{1/2})$

Statistical Queries

$q : \mathbb{X} \to [0,1]$

$$F_q(x_1, \ldots, x_n) = \sum_{i=1}^{n} q(x_i)$$

Trust vs Accuracy

Trust ↑ | | | Accuracy ↑

Central $\Theta(1)$

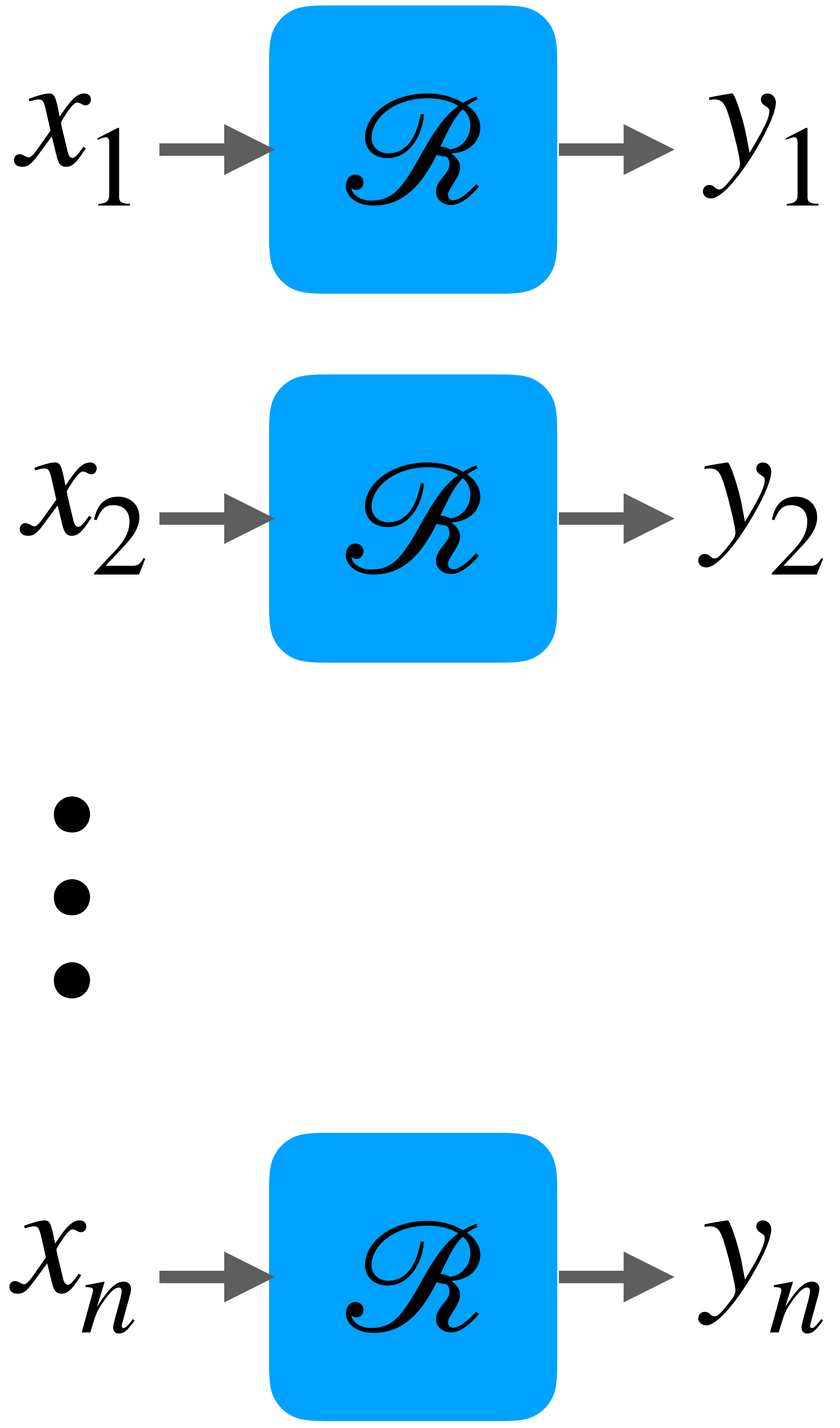
Shuffle $O(1)$ — O(√n) messages [CSUZZ19]; $\Theta(n^{1/6})$ — single message

Local $\Theta(n^{1/2})$

Statistical Queries $q : \mathbb{X} \to [0,1]$ $F_q(x_1, \ldots, x_n) = \sum_{i=1}^{n} q(x_i)$

The Shuffle Model

$\mathscr{R} : \mathbb{X} \to \mathbb{Y}$

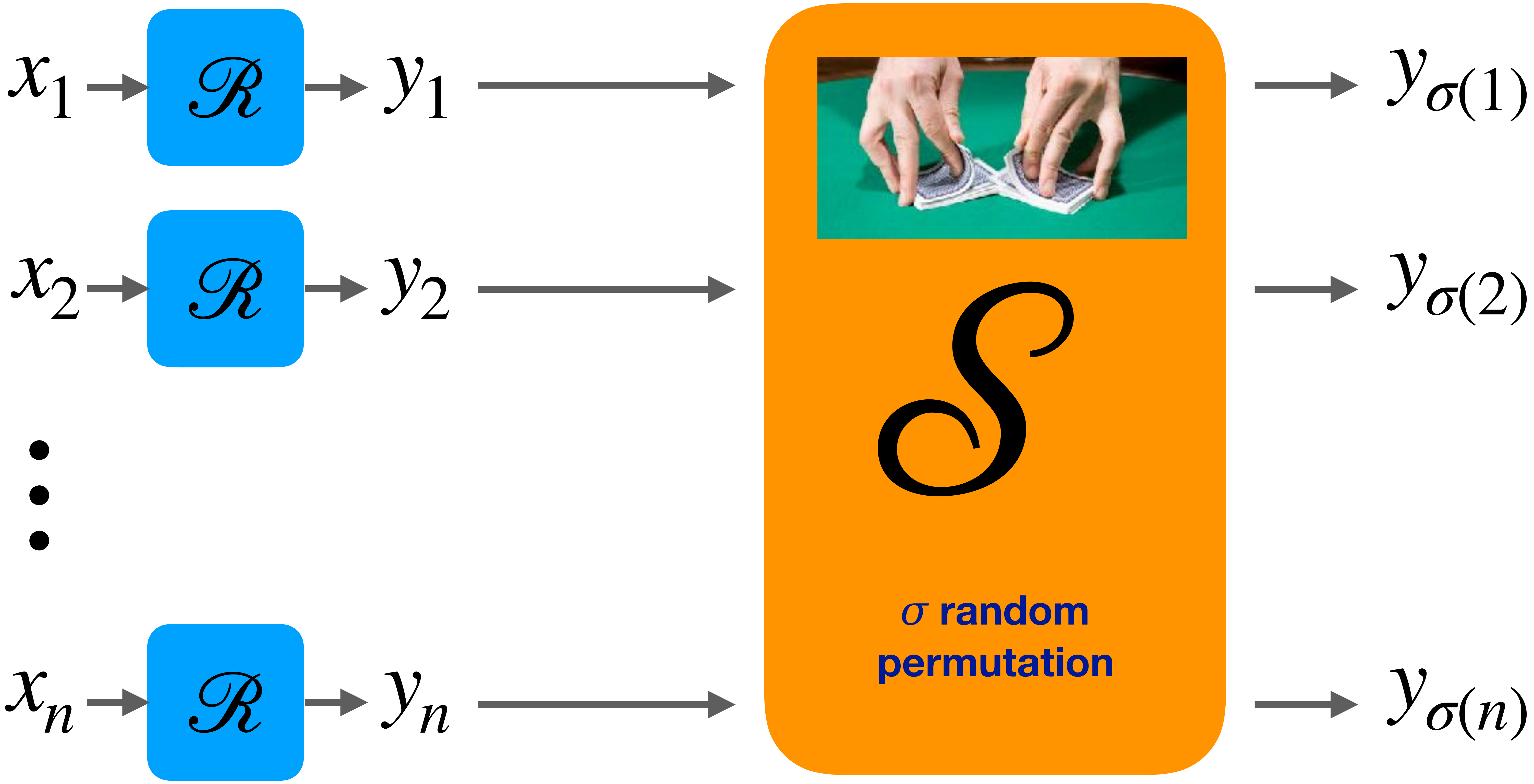


Local Randomizer

[BEMMRLRKTS17; EFMRTT19; CSUZZ19]

The Shuffle Model

$\mathcal{R} : \mathbb{X} \to \mathbb{Y}$ $\mathcal{S} : \mathbb{Y}^n \to \mathbb{Y}^n$



Local Randomizer **Trusted Shuffler**

[BEMMRLRKTS17; EFMRTT19; CSUZZ19]

The Shuffle Model

$\mathcal{R} : \mathbb{X} \to \mathbb{Y}$

$\mathcal{S} : \mathbb{Y}^n \to \mathbb{Y}^n$

$x_1 \to \mathcal{R} \to y_1 \longrightarrow$

$x_2 \to \mathcal{R} \to y_2 \longrightarrow$

⋮

$x_n \to \mathcal{R} \to y_n \longrightarrow$

$\mathcal{S}$

σ **random permutation**

$\longrightarrow y_{\sigma(1)}$

$\longrightarrow y_{\sigma(2)}$

$\longrightarrow y_{\sigma(n)}$

Local Randomizer

Trusted Shuffler

Privacy Analysis

$\mathcal{S} \circ \mathcal{R}^n$ (ε,δ)-DP

Utility Analysis

$\mathcal{A} \circ \mathcal{S} \circ \mathcal{R}^n$

[BEMMRLRKTS17; EFMRTT19; CSUZZ19]

Real Sum in the Shuffle Model

- **Problem Statement**
  - n users, each holding a number in [0,1], estimate the sum
- **Previous Work [CSUZZ, Eurocrypt 2019]**
  - One message: error $O(n^{1/2})$, communication $O(1)$
  - Multiple messages: error $O(1)$, communication $O(n^{1/2})$
- **Our Result**
  - One message: error $\Theta(n^{1/6})$, communication $O(\log n)$

Privacy Amplification by Shuffling

- **Problem Statement**
  - Characterize the privacy of shuffled mechanisms in terms of the privacy of its local randomizers
- **Previous Work [EFMRTT, SODA 2019]**
  - Shuffle-then-randomize (with adaptativity):

$$\varepsilon = O\left(\varepsilon_0\sqrt{\log(1/\delta)/n}\right)$$

for $\varepsilon_0 = O(1)$

- **Our Result**
  - Randomize-then-shuffle (one randomizer):

$$\varepsilon = O\left((\varepsilon_0 \wedge 1)e^{\varepsilon_0}\sqrt{\log(1/\delta)/n}\right)$$

for $\varepsilon_0 <= 0.5 \log(n) + O(1)$

Real Summation Protocol

- Discretize [0,1] into k+1 bins of equal length

$$\mathscr{R} : [0,1] \rightarrow \left\{ 0, \frac{1}{k}, \frac{2}{k}, \ldots, 1 \right\}$$

- Apply randomized rounding and randomized response with prob. γ

$$\mathscr{R}(x_i) = \begin{cases} Round(x_i) & \sim_{wp} 1 - \gamma \\ Uniform & \sim_{wp} \gamma \end{cases}$$

- After shuffling, add all the messages and remove the bias

$$\mathscr{A}(\overrightarrow{y}) = deBias\left(\sum_{i=1}^{n} y_i \right)$$

Analysis Overview

- Bound the MSE (ie. variance) of the protocol (as a function of k and γ)

$$\mathbb{E}\left[\left(\mathscr{A} \circ \mathscr{R}^n(\overrightarrow{x}) - \sum_i x_i\right)^2\right] = O\left(\frac{n}{k^2}\right) + O\left(\gamma n\right)$$

Rounding Uniform

- Analyze privacy of the protocol (as a function of k and γ)

$$\gamma = O\left(\frac{k \log(1/\delta)}{n\varepsilon^2}\right)$$

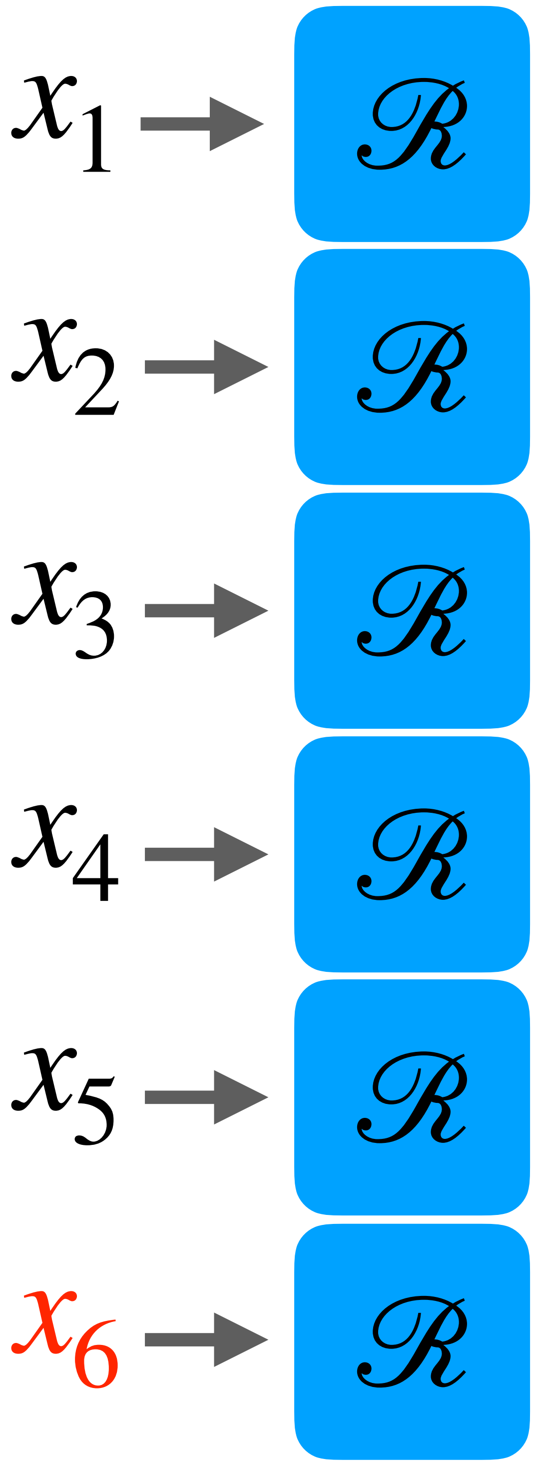
(ε,δ)-DP

- Optimize over k to minimize error

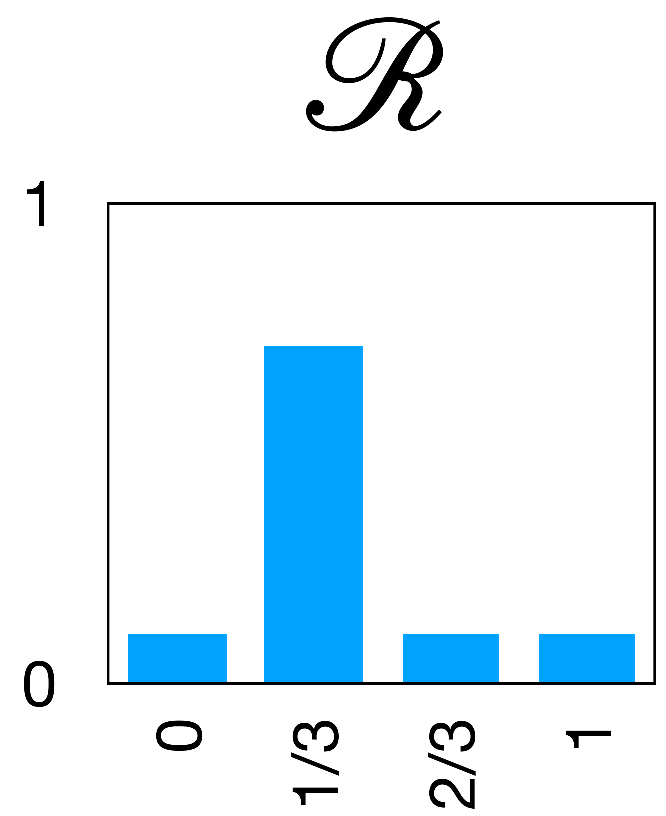
$$\text{MSE}(\mathscr{A} \circ \mathscr{R}^n) = O\left(\frac{n^{1/3} \log^{2/3}(1/\delta)}{\varepsilon^{4/3}}\right)$$

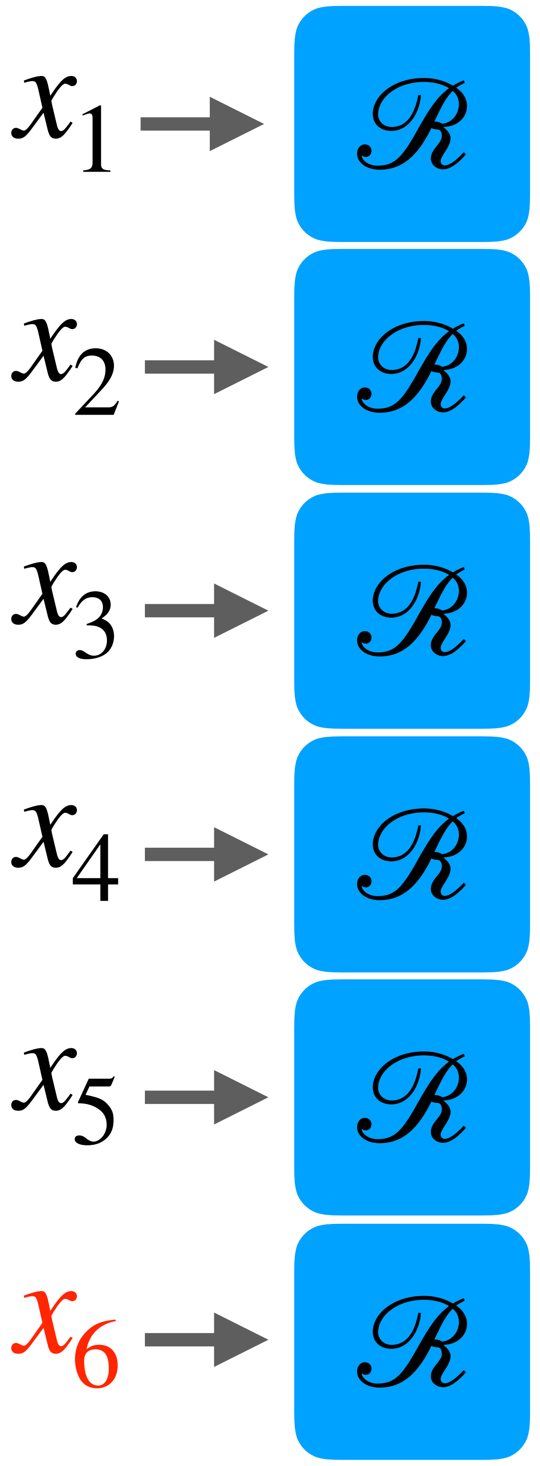
$$k = \tilde{O}(\varepsilon^{2/3} n^{1/3})$$

Privacy Analysis

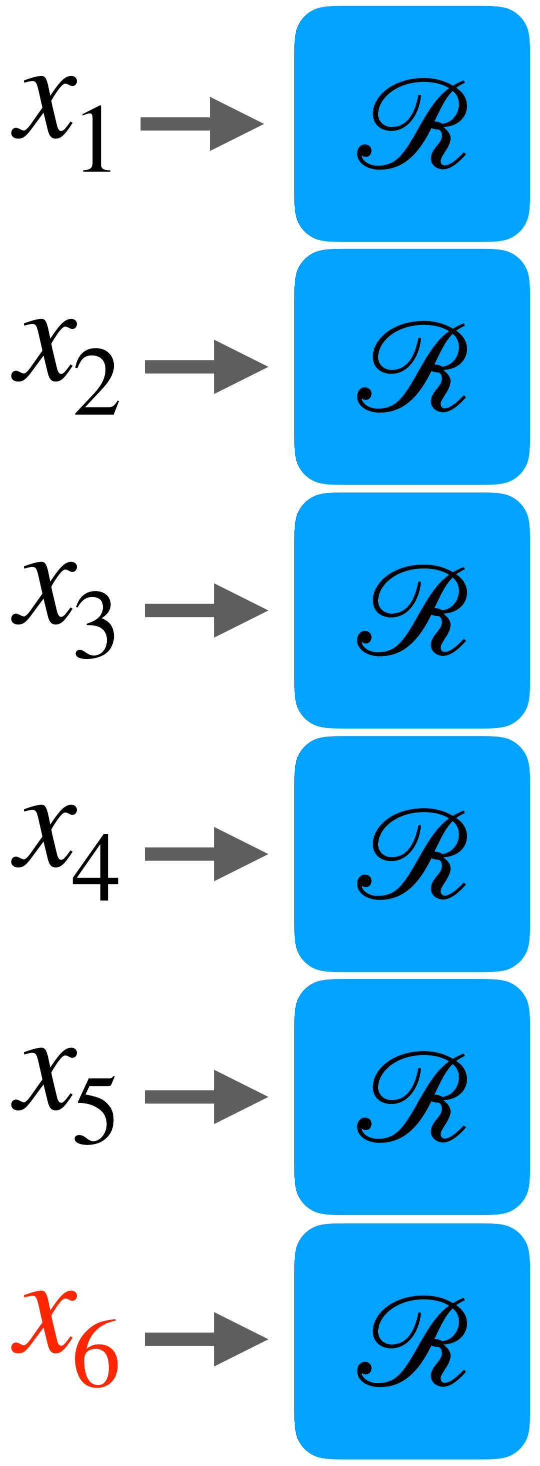


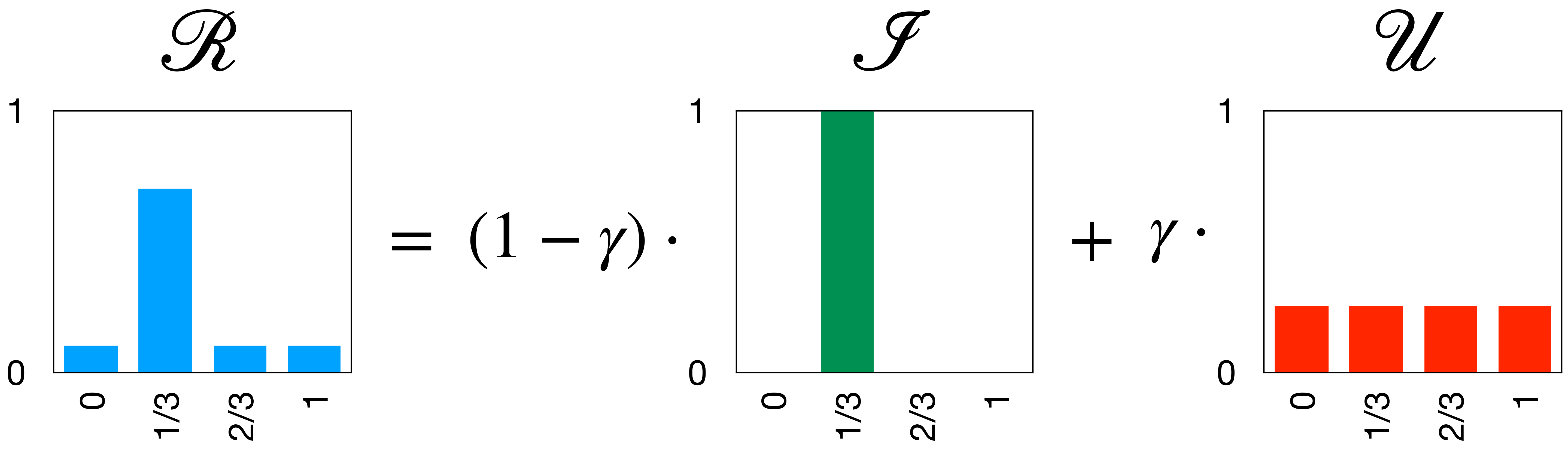
Privacy Analysis



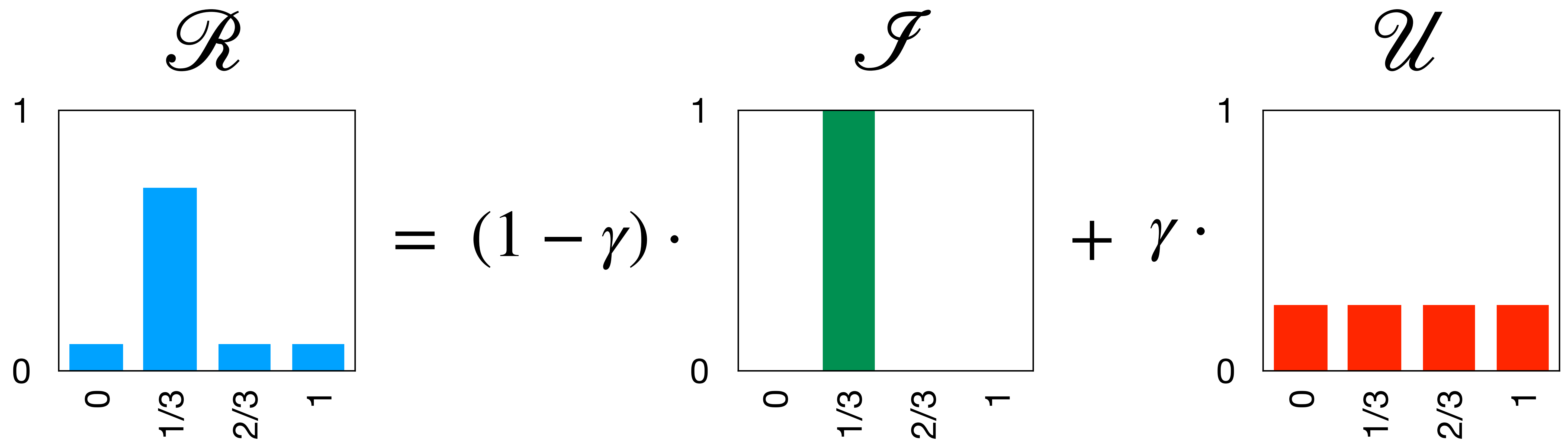


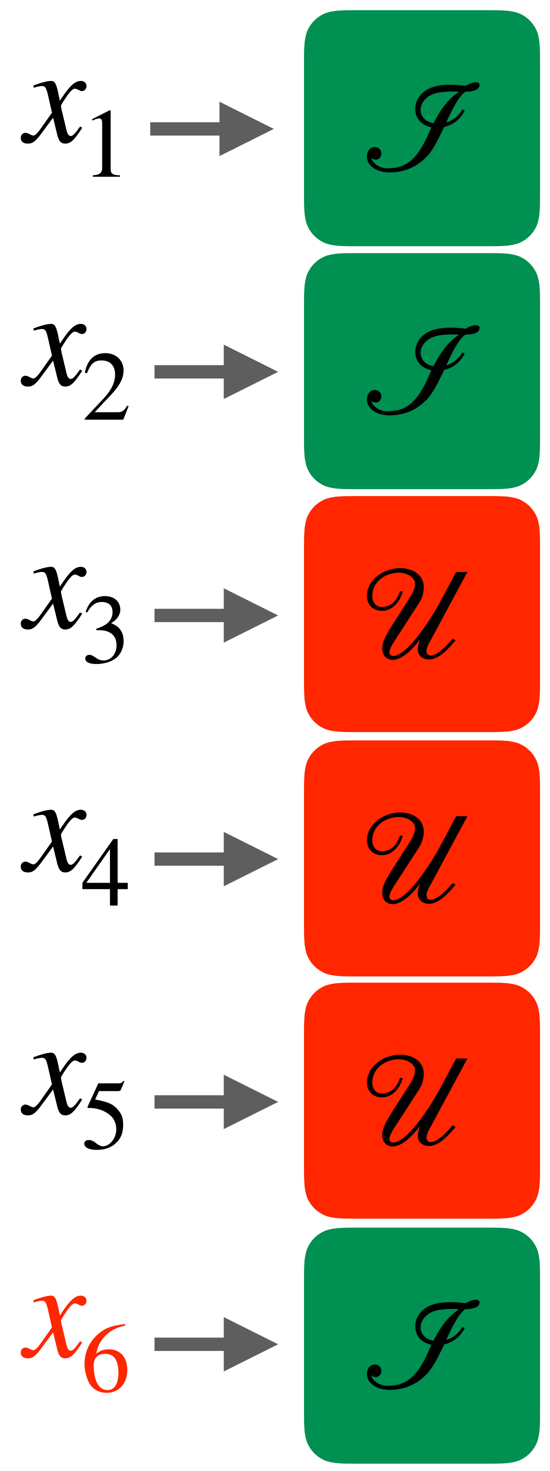
Privacy Analysis



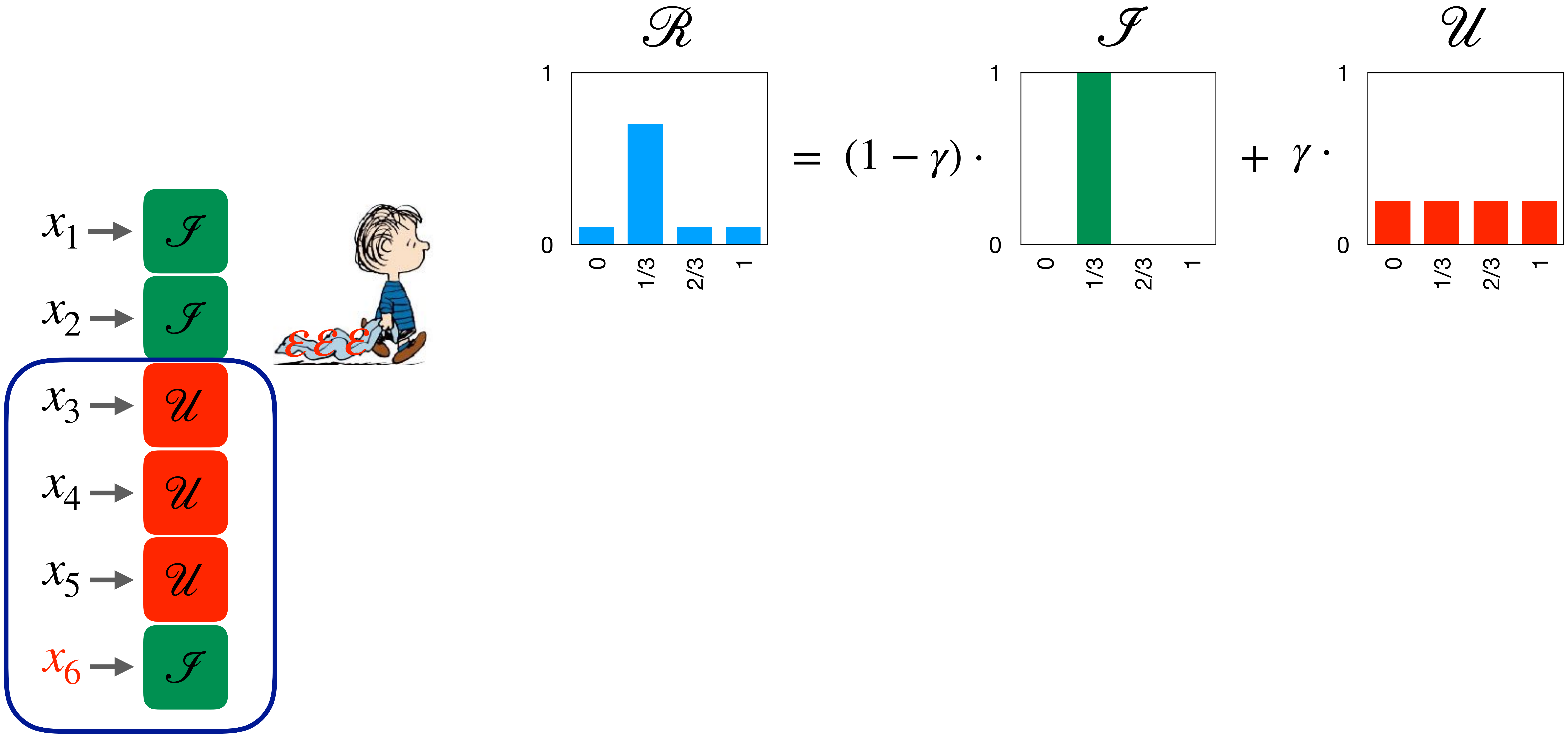


Privacy Analysis

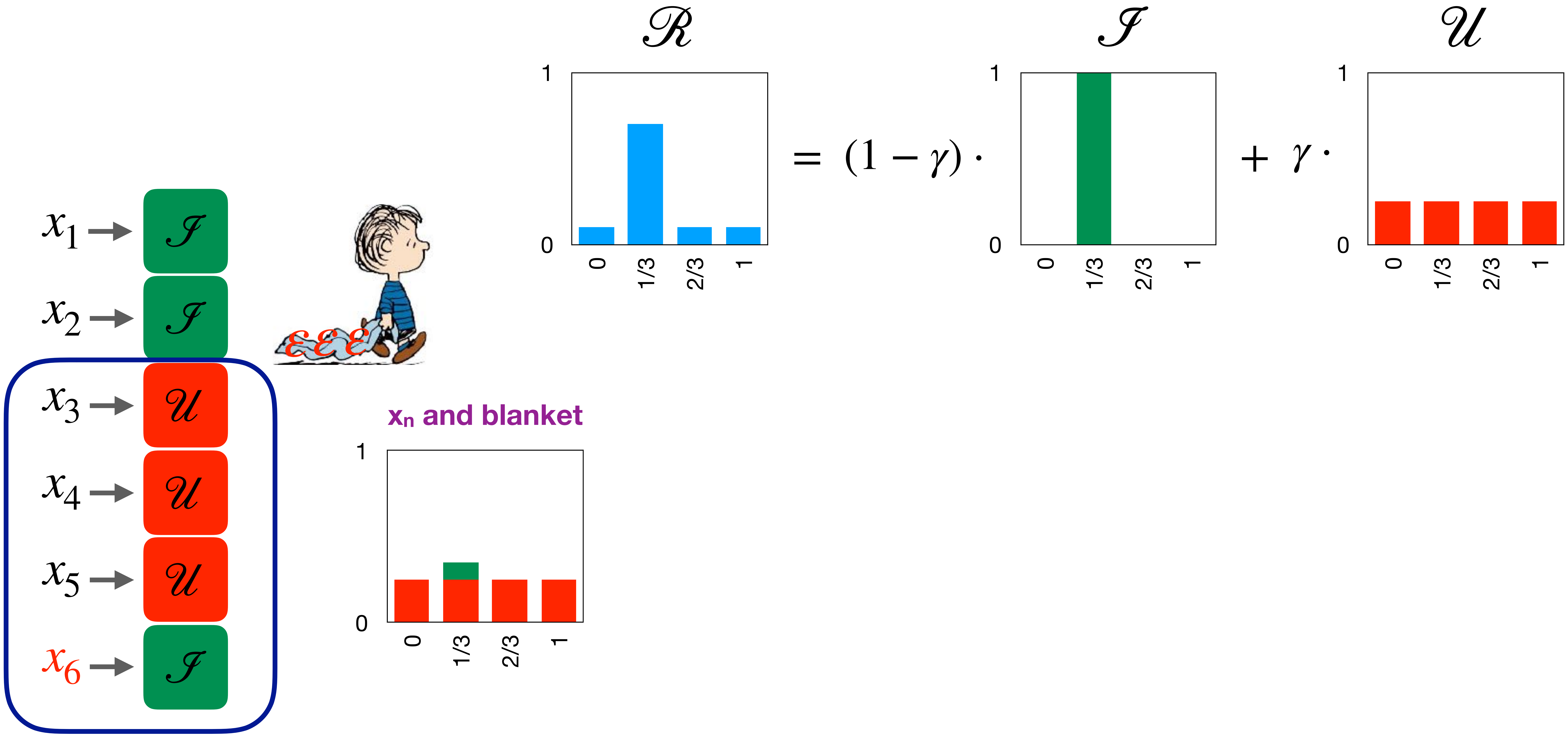




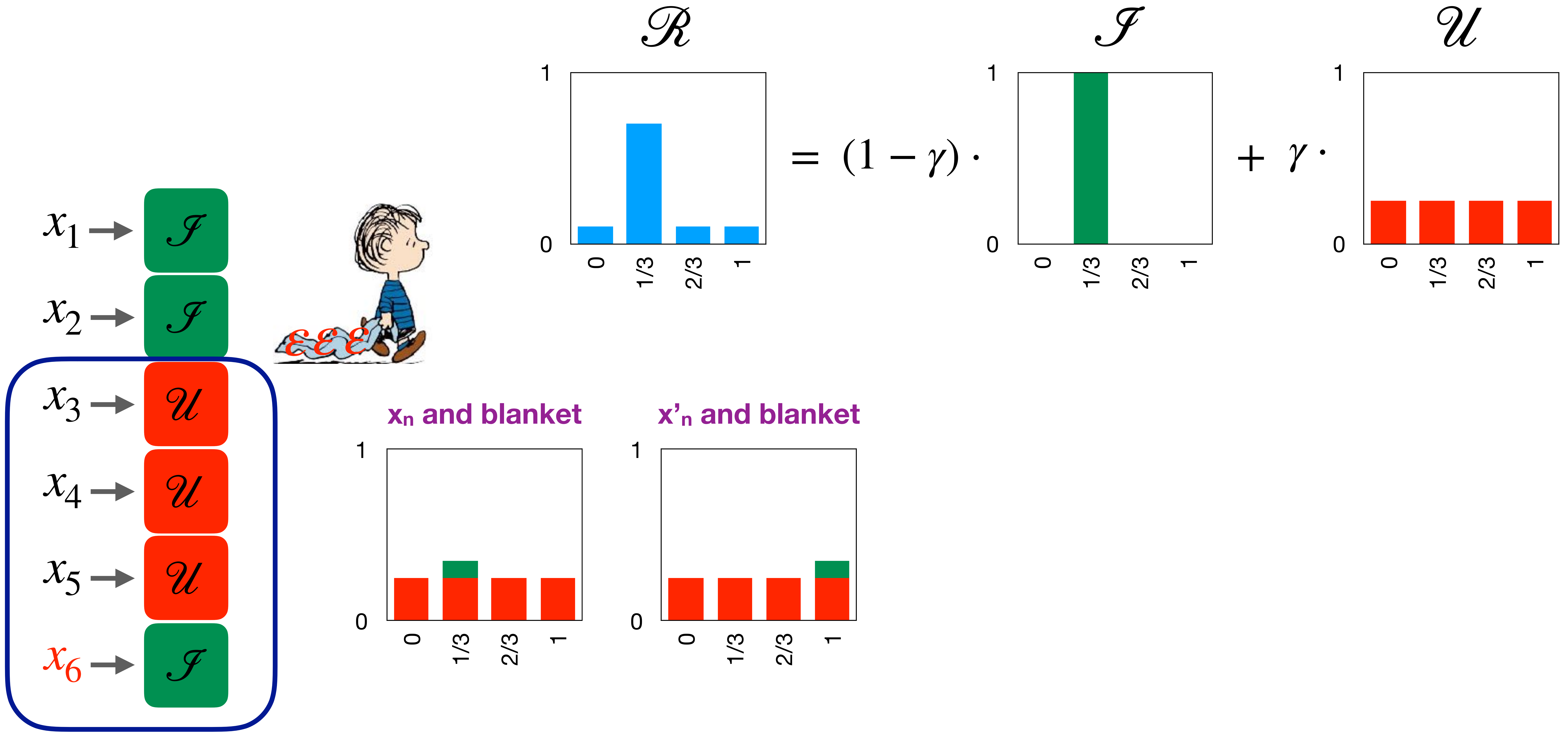
Privacy Analysis



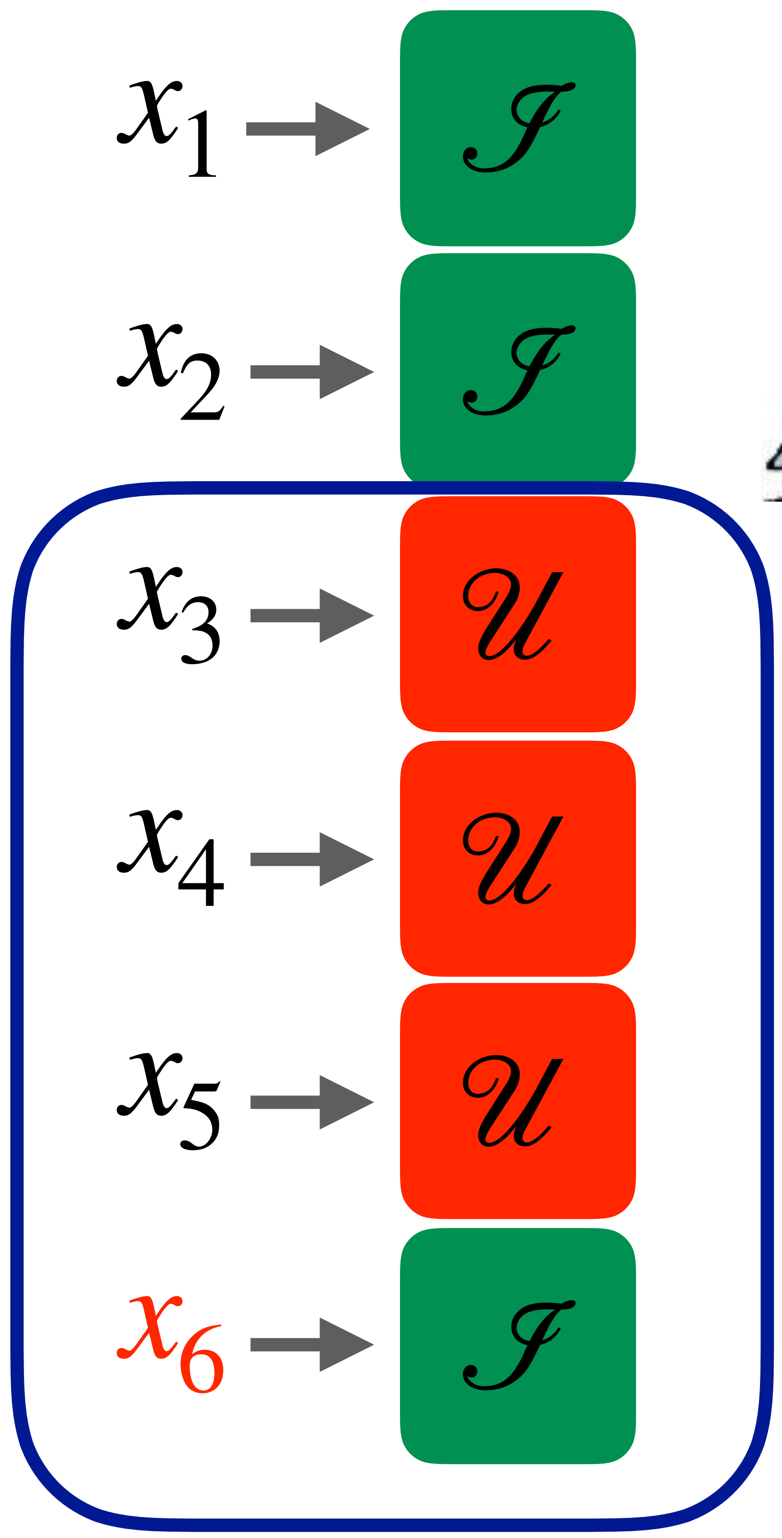
Privacy Analysis

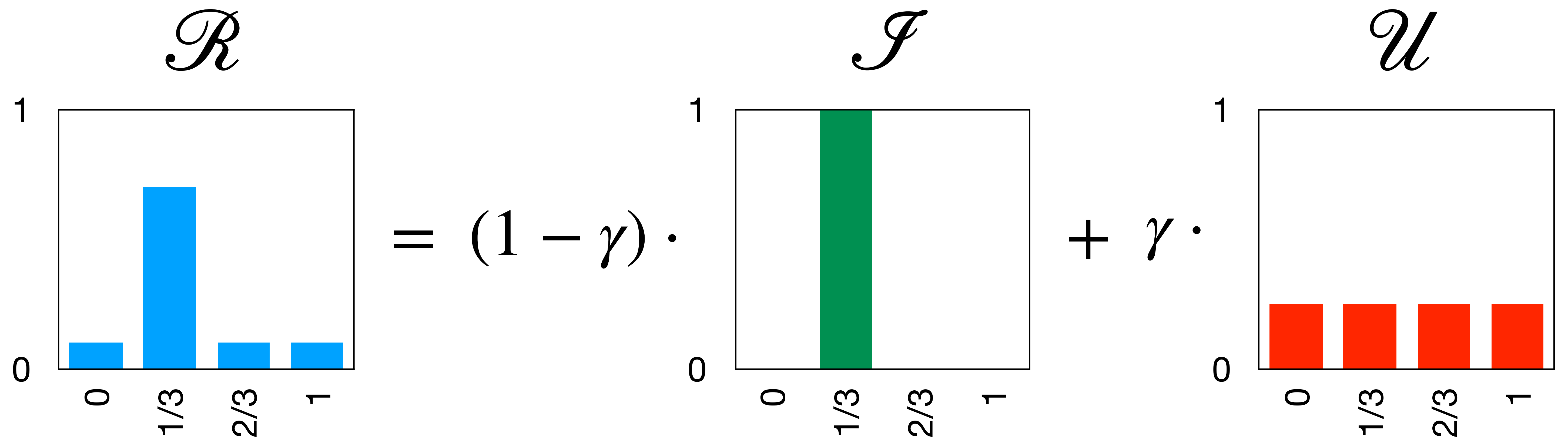


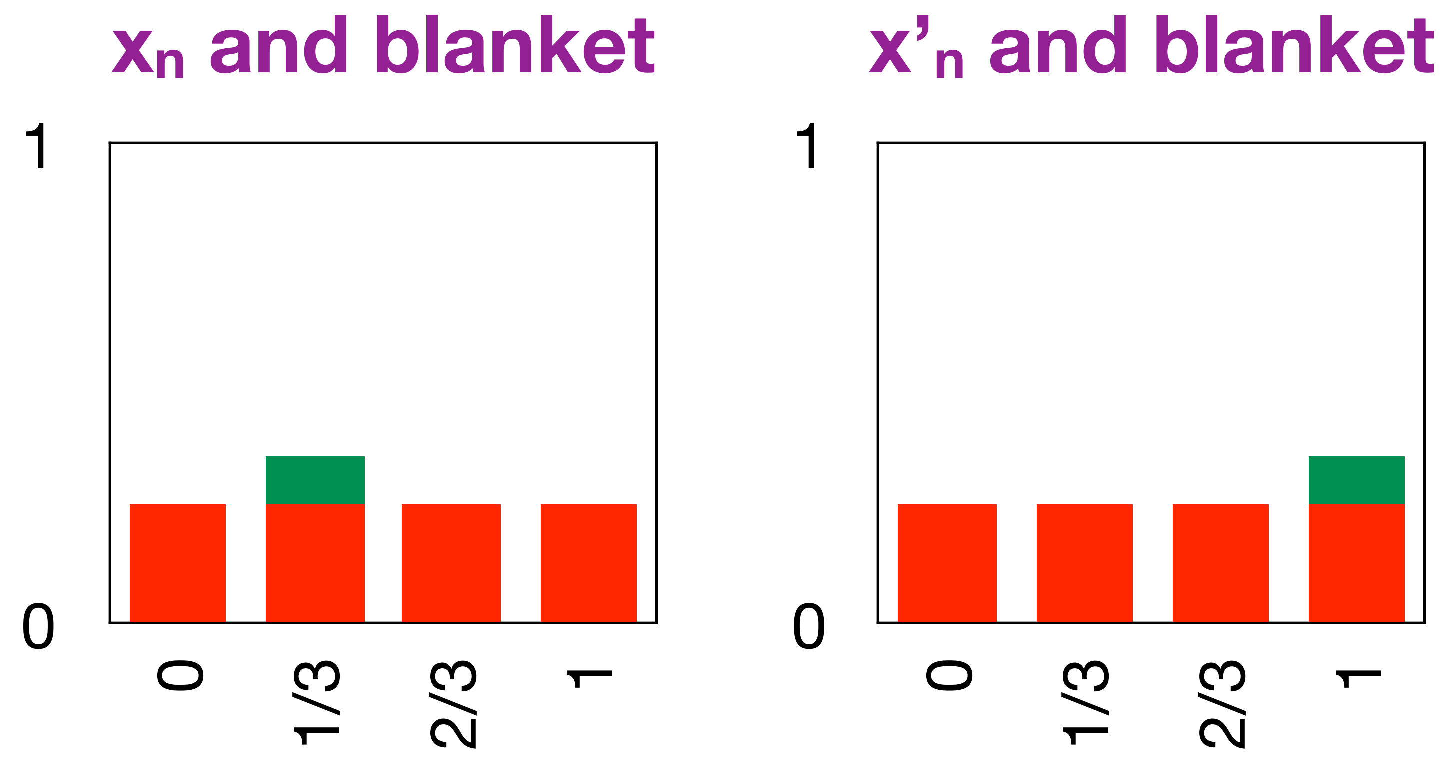
Privacy Analysis



Privacy Analysis







$$\mathbb{P}\left[\frac{Bin\left(n-1,\frac{\gamma}{k+1}\right)+1}{Bin\left(n-1,\frac{\gamma}{k+1}\right)} \geq e^{\varepsilon}\right] \leq \delta$$

$$\gamma = O\left(\frac{k\log(1/\delta)}{n\varepsilon^2}\right)$$

Lower Bound

- **Theorem:** Any (ε,δ)-DP one-message shuffled protocol for real summation with n inputs in [0,1] and $\delta < 0.5$ must have

$$\text{MSE} = \Omega\left(n^{1/3}\min\left\{e^{-\varepsilon}, \frac{1}{2} - \delta\right\}\right)$$

- **Proof Sketch**

  1. Reduction to i.i.d. case where aggregation is summation and randomizer maps to [0,1] (apply optimal Bayesian denoising)
  2. Take inputs to be uniform on partition of [0,1] in $n^{1/3}$ equally spaced points
  3. Prove two lower bounds on MSE, interpolate them, and couple them through privacy

Amplification by Shuffling

- **Theorem:** Shuffling n copies of any ε_0-LDP randomizer with blanket parameter γ gives (ε,δ)-DP with

$$\frac{\gamma(e^{\varepsilon}+1)^2(e^{\varepsilon_0}-e^{-\varepsilon_0})^2}{4n(e^{\varepsilon}-1)}\cdot\exp\left(-0.86n\left(\gamma\wedge\frac{(e^{\varepsilon}-1)^2}{\gamma(e^{\varepsilon}+1)^2(e^{\varepsilon_0}-e^{-\varepsilon_0})^2}\right)\right)\leq\delta$$

Amplification by Shuffling

- **Theorem:** Shuffling n copies of any ε_0-LDP randomizer with blanket parameter γ gives (ε,δ)-DP with

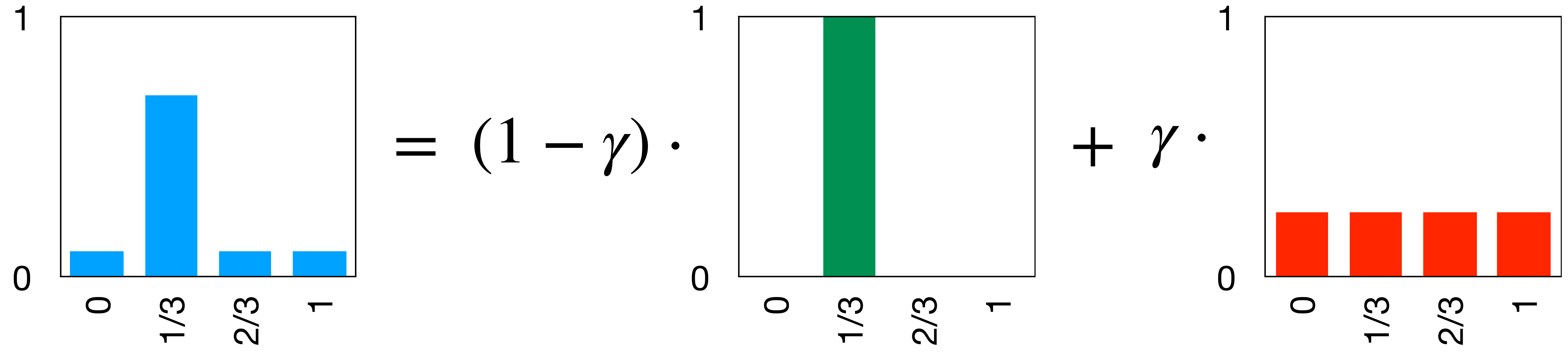
$$\frac{\gamma(e^{\varepsilon}+1)^2(e^{\varepsilon_0}-e^{-\varepsilon_0})^2}{4n(e^{\varepsilon}-1)} \cdot \exp\left(-0.86n\left(\gamma \wedge \frac{(e^{\varepsilon}-1)^2}{\gamma(e^{\varepsilon}+1)^2(e^{\varepsilon_0}-e^{-\varepsilon_0})^2}\right)\right) \leq \delta$$

- **Corollary:** Shuffling n copies of an ε_0-LDP randomizer gives (ε,δ)-DP with

$$\varepsilon = O\left((\varepsilon_0 \wedge 1)e^{\varepsilon_0}\sqrt{\log(1/\delta)/n}\right) \qquad \varepsilon_0 \leq \log(n/\log(1/\delta))/2$$

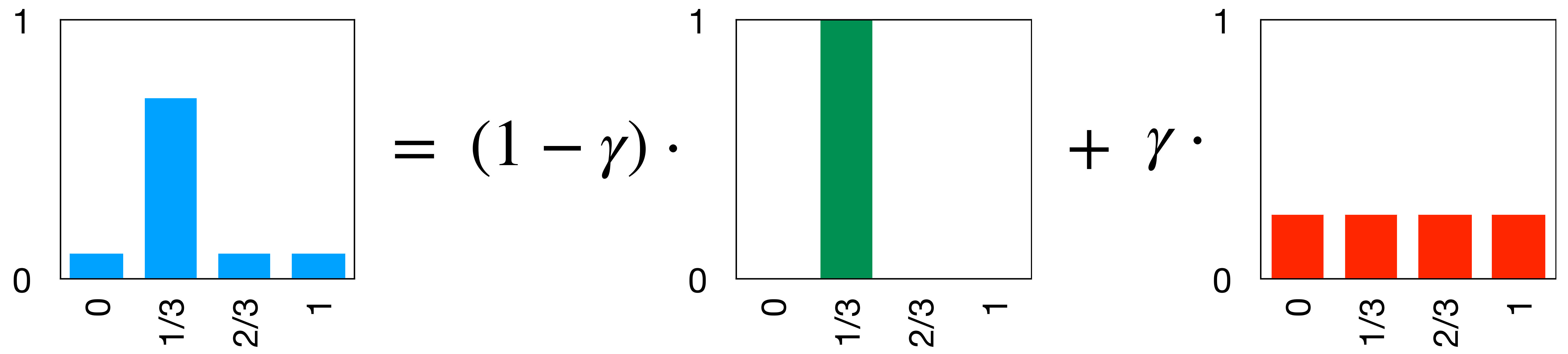
Blanket of a Local Randomizer

Randomized Response



Blanket of a Local Randomizer

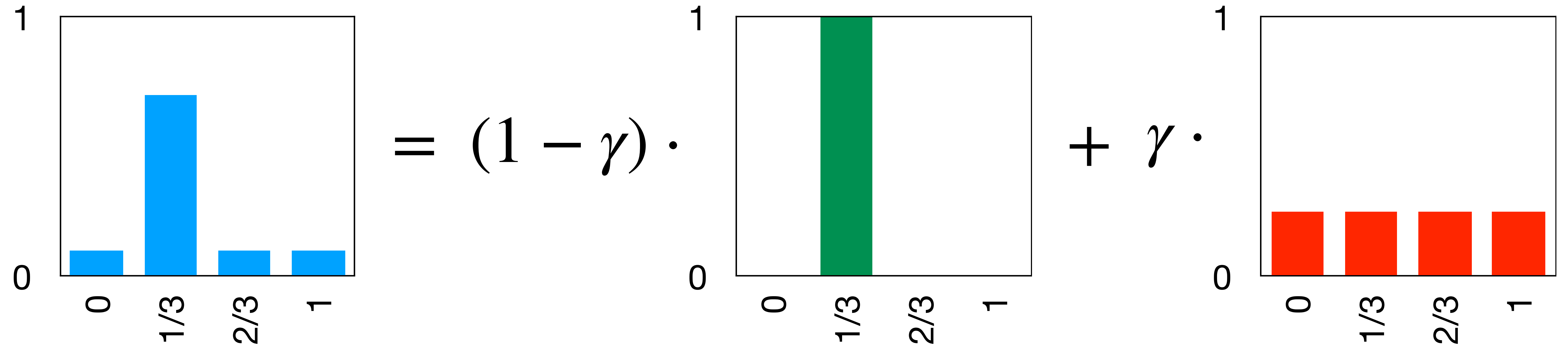
Randomized Response



- **Theorem (Blanket Decomposition):** Every ε_0-LDP randomizer admits a (unique maximal) mixture decomposition where one of the components is independent of the input

Blanket of a Local Randomizer

Randomized Response



- **Theorem (Blanket Decomposition):** Every ε_0-LDP randomizer admits a (unique maximal) mixture decomposition where one of the components is independent of the input

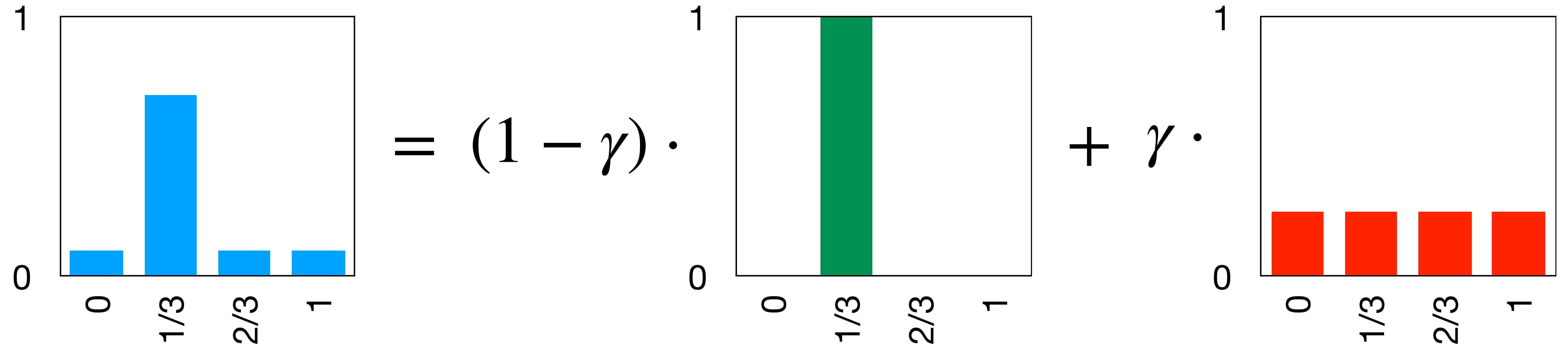
$$\mathcal{R}(x) = (1-\gamma)\mathcal{R}'(x) + \gamma\omega$$

$$\mathcal{R} : \mathbb{X} \to \mathbb{Y}$$
$$\mathcal{R}' : \mathbb{X} \to \mathbb{Y}$$
$$\omega \in Dist(\mathbb{Y})$$

$$e^{-\varepsilon_0} \leq \gamma \leq 1$$

Blanket of a Local Randomizer

Randomized Response



- **Theorem (Blanket Decomposition):** Every ε_0-LDP randomizer admits a (unique maximal) mixture decomposition where one of the components is independent of the input

$$\mathscr{R}(x) = (1-\gamma)\mathscr{R}'(x) + \gamma\omega$$

$$\mathscr{R} : \mathbb{X} \to \mathbb{Y}$$

$$\mathscr{R}' : \mathbb{X} \to \mathbb{Y}$$

$$\omega \in Dist(\mathbb{Y})$$

$$e^{-\varepsilon_0} \leq \gamma \leq 1$$

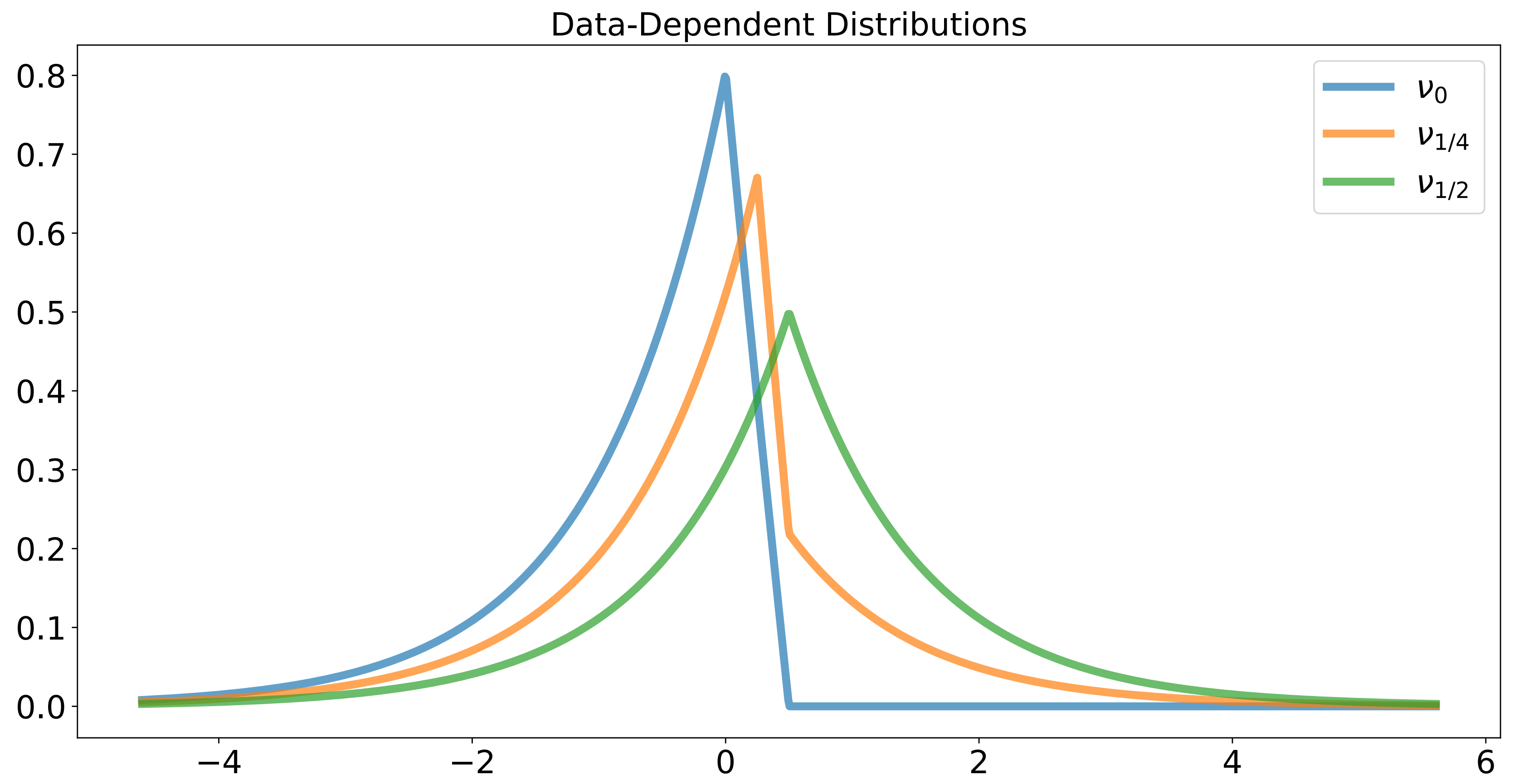
Blanket Construction

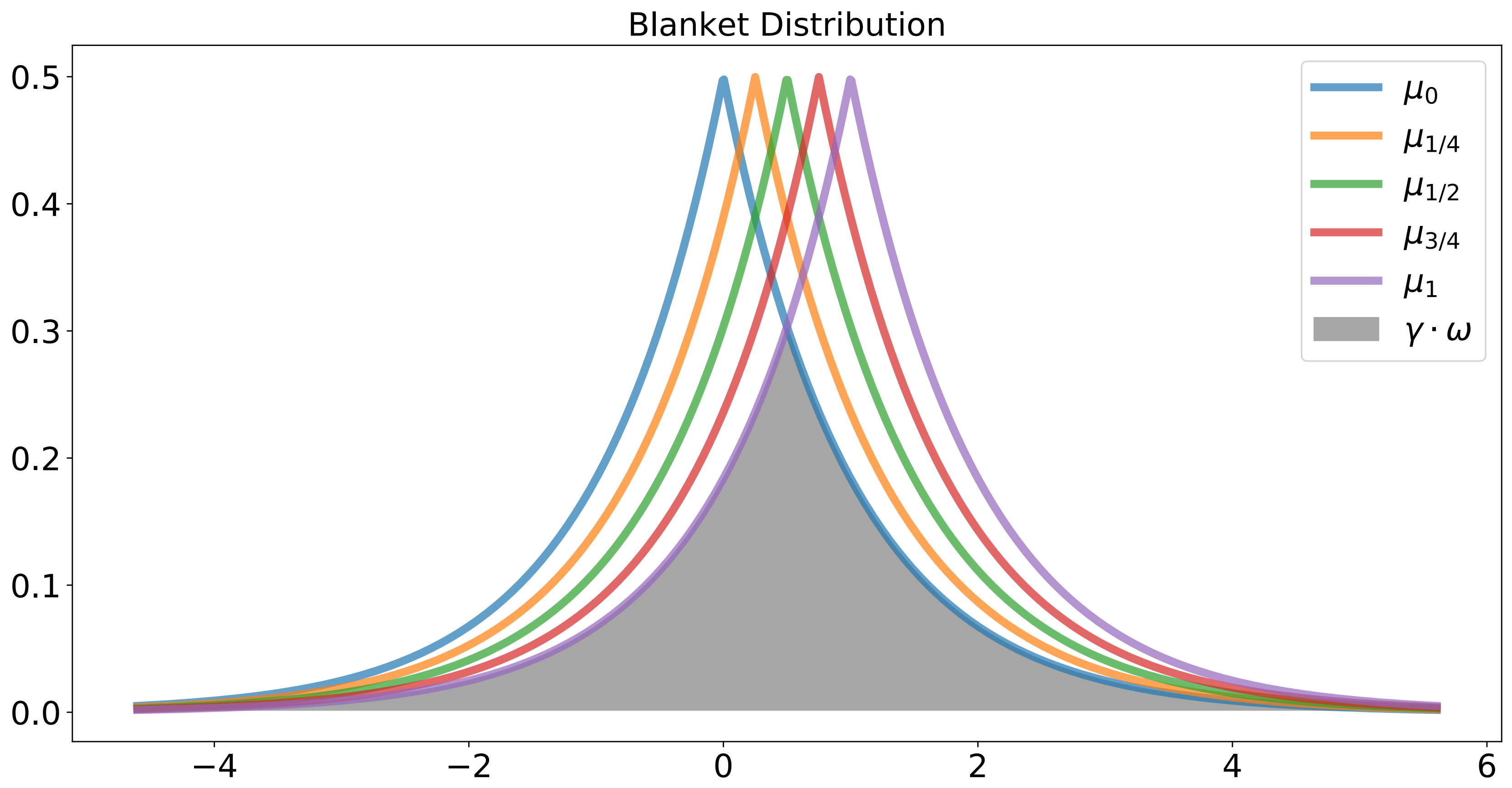
$$\gamma = \int_{\mathbb{Y}} \min_{x \in \mathbb{X}} p_{\mathscr{R}(x)}(y)dy$$

$$p_\omega(y) = \frac{\min_{x\in\mathbb{X}} p_{\mathscr{R}(x)}(y)}{\gamma}$$

Example Blanket Decompositions

ε_0-LDP RR on [k]	$\gamma = \dfrac{k}{e^{\varepsilon_0} + k - 1}$
ε_0-LDP Laplace on [0,1]	$\gamma = e^{-\frac{\varepsilon_0}{2}}$
σ^2 Gaussian on [0,1]	$\gamma = 2\mathbb{P}[N(0,\sigma^2) \leq -1/2]$





Amplification: Proof Idea

- General idea
  - Couple who samples from the blanket in both executions
  - Reveal the identity of who samples from the blanket (joint convexity)
  - Remove the data from the users in 1…n-1 who sampled from R’ (post-processing)

- Define privacy amplification random variable

$$\mathbb{E}[L] = 1 - e^{\varepsilon} < 0$$

$$Y \sim \omega \qquad L = L^{\mathcal{R}}_{x,x'} = \frac{p_{\mathcal{R}(x)}(Y) - e^{\varepsilon} p_{\mathcal{R}(x')}(Y)}{p_{\omega}(Y)}$$

$$\gamma(e^{-\varepsilon_0} - e^{\varepsilon+\varepsilon_0}) \leq L \leq \gamma(e^{\varepsilon_0} - e^{\varepsilon-\varepsilon_0})$$

- Reduce to bounding expectation, apply concentration for bounded r.v.’s

$$\sup_{E} \left(\mathbb{P}[\mathcal{S} \circ \mathcal{R}^n(\overrightarrow{x}) \in E] - e^{\varepsilon} \mathbb{P}[\mathcal{S} \circ \mathcal{R}^n(\overrightarrow{x}') \in E] \right) \leq \frac{1}{\gamma n} \mathbb{E}\left[\sum_{i=1}^{Bin(n,\gamma)} L_i \right]_+$$

Getting the Bound

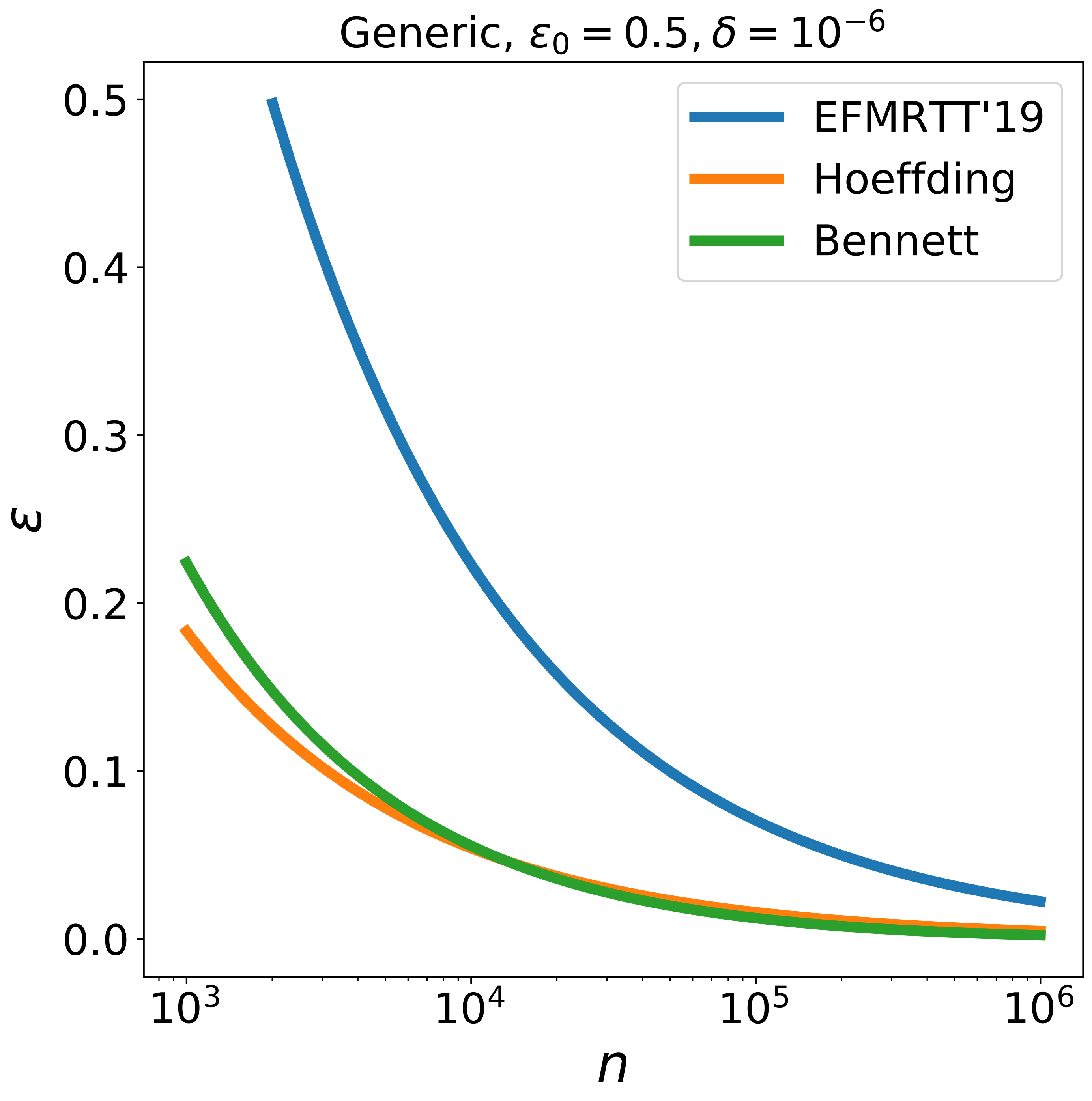
- Applying Hoeffding's inequality we get

$$\frac{1}{\gamma n}\mathbb{E}\left[\sum_{i=1}^{Bin(n,\gamma)} L_i\right]_+ \leq \frac{\gamma(e^\varepsilon+1)^2(e^{\varepsilon_0}-e^{-\varepsilon_0})^2}{4n(e^\varepsilon-1)} \cdot \exp\left(-0.86n\left(\gamma \wedge \frac{(e^\varepsilon-1)^2}{\gamma(e^\varepsilon+1)^2(e^{\varepsilon_0}-e^{-\varepsilon_0})^2}\right)\right)$$

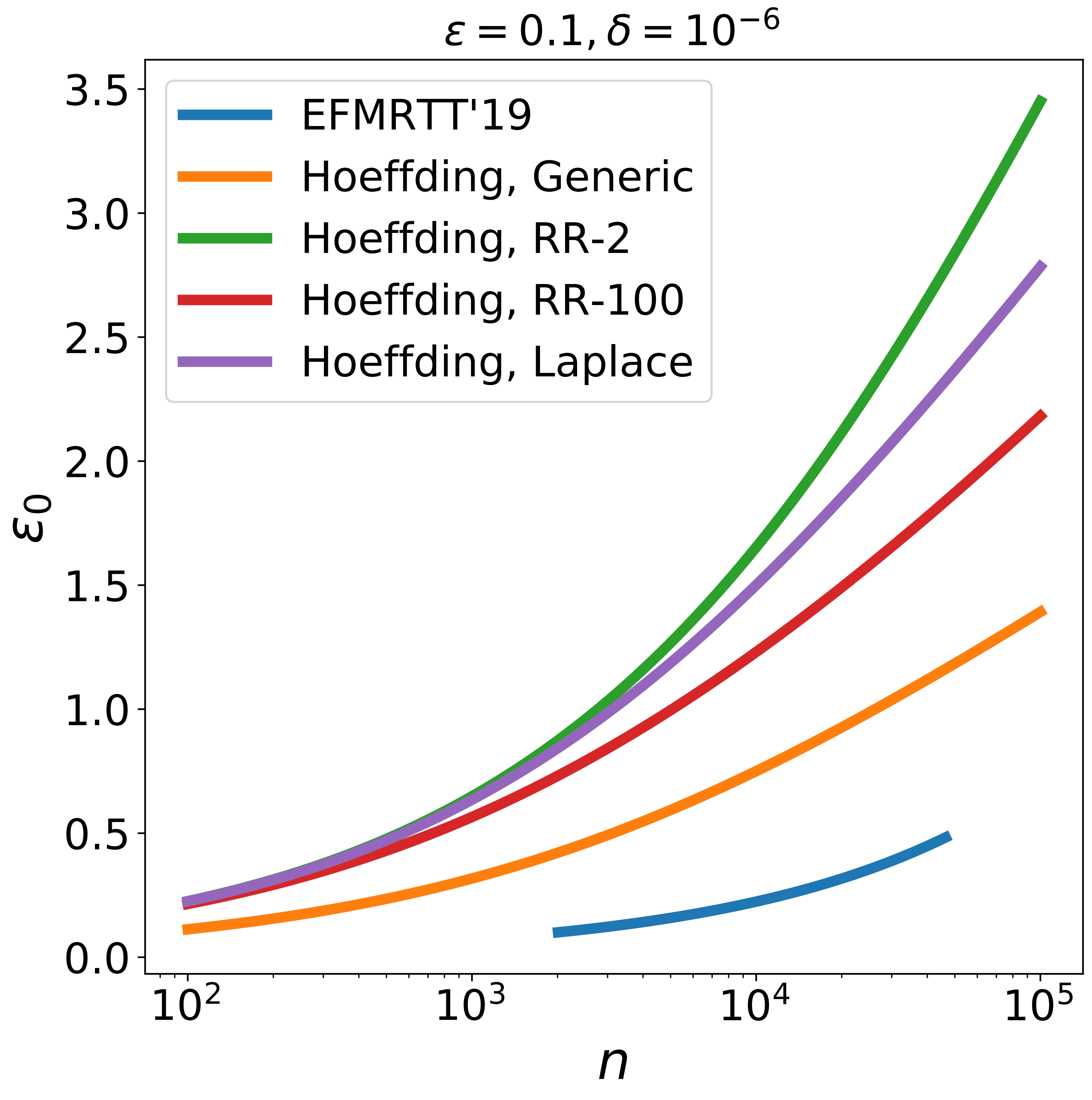
- Refinements:
  - Use mechanism-specific bounds on L and γ
  - Alternative concentration bounds, eg. Bennett's inequality

Numerical Comparison

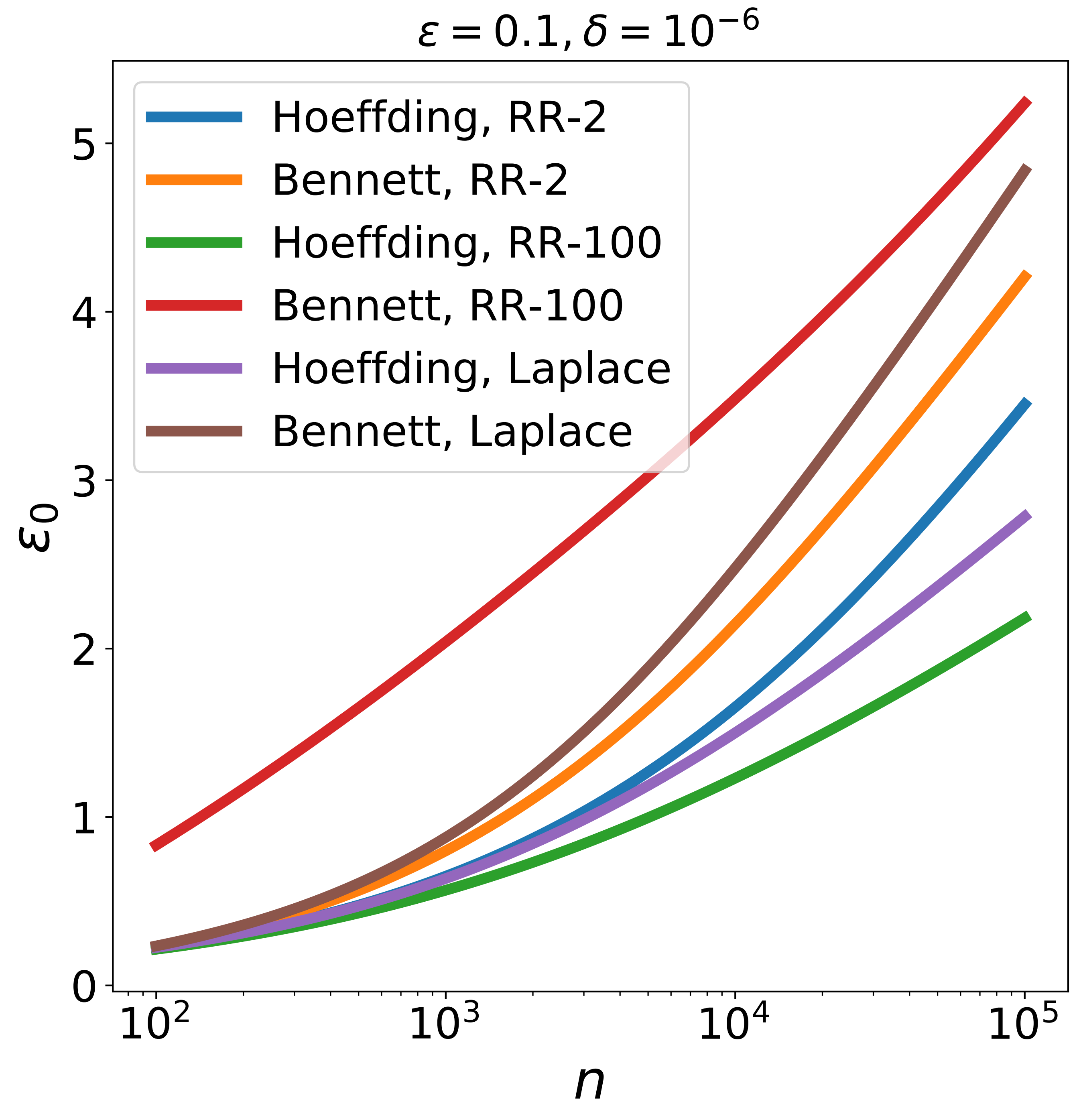
$\varepsilon(n, \varepsilon_0)$



$\varepsilon_0(n, \varepsilon)$



$\varepsilon_0(n, \varepsilon)$



Conclusion

- Matching upper and lower bounds for one-message, one-randomizer real summation in the shuffle model
  - Error $\Theta(n^{1/6})$ and communication O(log n)
  - First tight shuffle-native lower bound
- General and flexible privacy amplification bounds for randomize-then-shuffle one-randomizer protocols in the shuffle model
  - Simple analysis via privacy blanket, without subsampling