RANDOMIZED REQUANTIZATION WITH LOCAL DIFFERENTIAL PRIVACY

Sijie Xiong[1], Anand Sarwate[1], Narayan Mandayam[1]

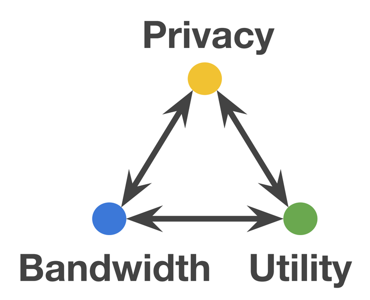
[1] Rutgers, The State University of New Jersey

Why interesting?

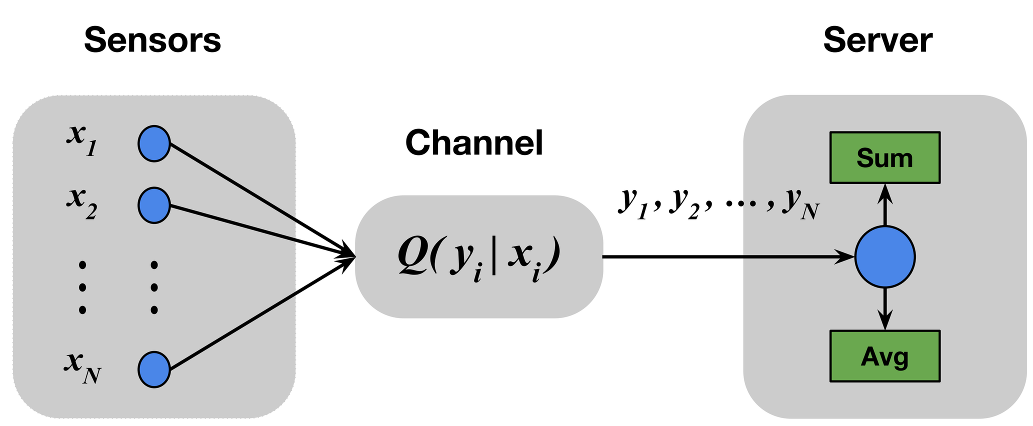
Goals for future sensor networks such as IoT:

- limit resource consumption
- protect private information
- maintain data fidelity

What are the tradeoffs between these criteria?



The system model



Each sensor $i = 1, 2, \ldots, N$

- measures a r.v. $X_i \sim P$ where the distribution P is unknown but in a set $\mathcal{P}$ of distributions on an alphabet $\mathcal{X} \subset \mathbb{R}$.
- transmits private version $Y_i \in \mathcal{Y}$, where $|\mathcal{Y}| \leq |\mathcal{X}|$.

Randomized requantization: map $X_i \to Y_i$ using *channel* $Q(y|x)$.

Server goal: estimate a linear combination of X_i's.

Performance metrics

Local differential privacy [Duchi et al. '13] **:**

The adversary's likelihood of guessing that the input sample was x over x' doesn't increase more than e^{ϵ} after observing the released value y:

$$\frac{P(X=x)}{P(X=x')} \leq \frac{P(X=x|Y=y)}{P(X=x'|Y=y)} \cdot e^{\epsilon}$$

$$\frac{Q(y|x)}{Q(y|x')} \leq e^{\epsilon} \qquad \text{(by Bayes's rule)}$$

Compression ratio:

Bit Rate $\propto \log_2 |\mathcal{X}|$

Cmp. Ratio $\rho = \dfrac{\log_2 |\mathcal{Y}|}{\log_2 |\mathcal{X}|}$

Utility (mse):

$$\delta = \mathbb{E}_{P \times Q}[d(X,Y)] = \sum_{i=1}^{N} \sum_{j=1}^{\hat{N}} P(x_i) Q(y_j|x_i)(x_i - y_j)^2$$

Goal: find privacy-utility tradeoff and optimal Q

The set of ε-locally differentially private channels and the set of channels yielding expected distortion no greater than δ are defined by

$$\mathbb{Q}_{\text{LDP}}(\epsilon) = \left\{ Q(y|x) : \log \frac{Q(y|x)}{Q(y|x')} \leq \epsilon, \quad \forall (x, x', y) \in \mathcal{X} \times \mathcal{X} \times \mathcal{Y} \right\}$$

$$\mathbb{Q}_{\text{MSE}}(\delta) = \left\{ Q(y|x) : \max_{P \in \mathcal{P}} \mathbb{E}_{P \times Q}(d(X,Y)) \leq \delta \right\}$$

Given $\mathcal{P}, \rho, \delta$, the optimal ϵ becomes

$$\epsilon^*(\mathcal{P}, \rho, \delta) = \{ \mathbb{Q}_{\text{LDP}} \cup \mathbb{Q}_{\text{MSE}} \neq \emptyset \}$$

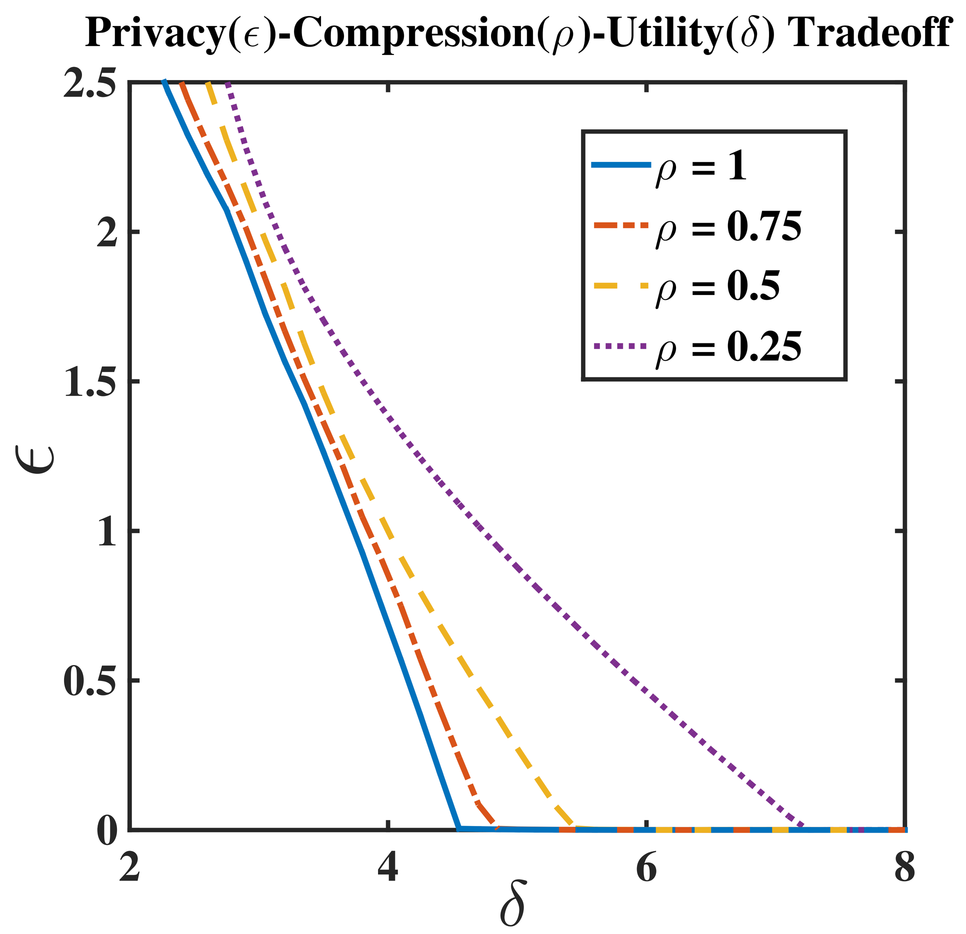
$$\Downarrow$$

$$\begin{aligned} \text{minimize} \quad & e^{\varepsilon} \\ \text{s.t.} \quad & \max_{P \in \mathcal{P}} \mathbb{E}_{P \times Q}[d(X,Y)] \leq \delta, \\ & 0 \preceq Q \preceq 1, \\ & Q \cdot \hat{\mathbf{1}}_{|\mathcal{Y}|} = \mathbf{1}_{|\mathcal{X}|}. \end{aligned}$$

Theorem

The above optimization problem is a constrained quasi-convex optimization problem, and can be solved by bisection method.

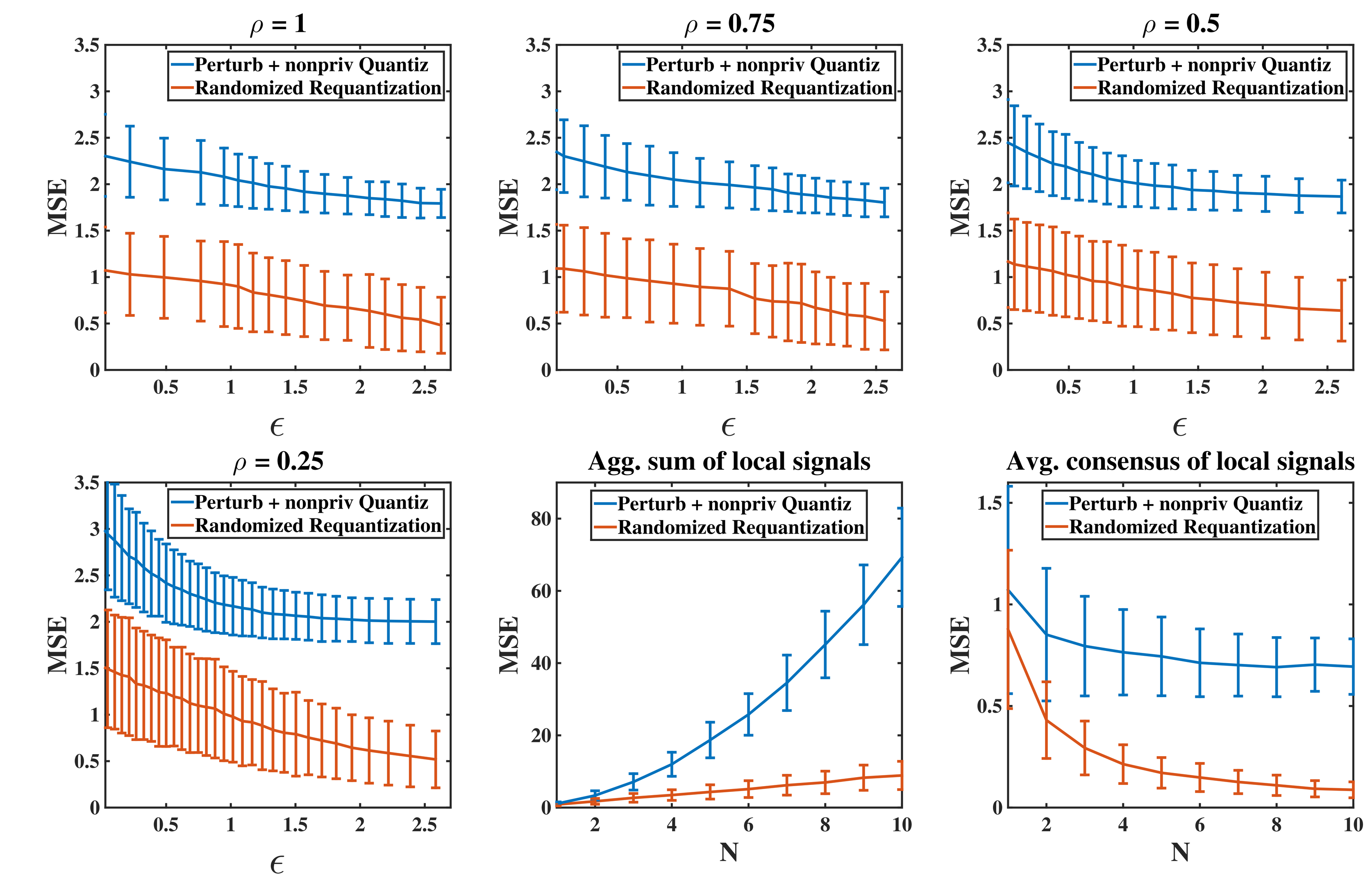
Solving the optimization problem



Minimum achievable privacy level ϵ^* given (δ, ρ) value pairs, finding (ϵ, δ, ρ)-tradeoff.

- for fixed ρ, standard $\delta \uparrow \leftrightarrow \epsilon \downarrow$ tradeoff
- across cmp. ratios, achievable ϵ quite small under small δ
- can halve bit rate without sacrificing privacy

Validation on synthetic data



Compare *randomized requantization* (RR) with perturbation method in the sparse Fourier transform domain

- RR works better, more consisitent
- RR adds in much smaller noise
- RR scales better with network size

Ongoing work and further directions

- Optimizing over reconstruction $\mathcal{Y}$ (c.f. Lloyd-Max).
- Use privacy allocation to apportion resources in networks:
  - individuals have different privacy budget $\epsilon_1, \epsilon_2, \ldots, \epsilon_N$
  - multiple servers trying to access the same data
  - gateway has to manage constraints and demands