GEORGETOWN UNIVERSITY

2024 Privacy and Public Policy Conference

OPTIMAL RESOLUTION OF A DATA SHARING TRILEMMA:
STATISTICAL POWER, SAMPLE COMPLEXITY, AND PRIVACY BUDGET

Yuxin Liu
Pitt

M. Amin Rahimian
Pitt

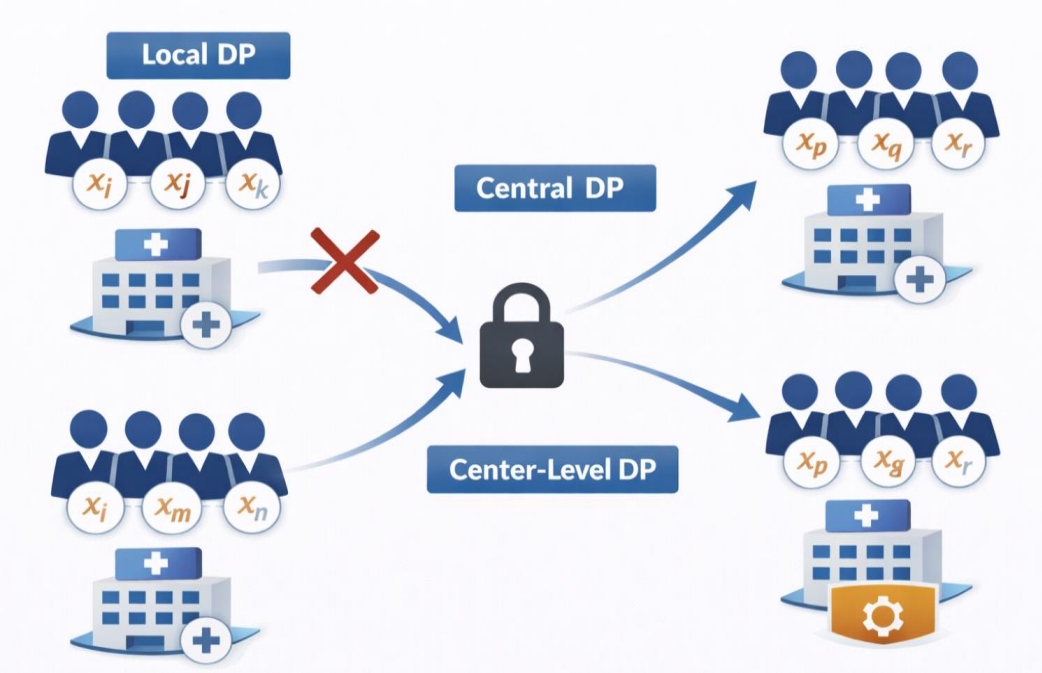
Marios Papachristou
ASU

Motivation

- In many real-world collaborations, *institutions*—not individuals—decide whether to share data.

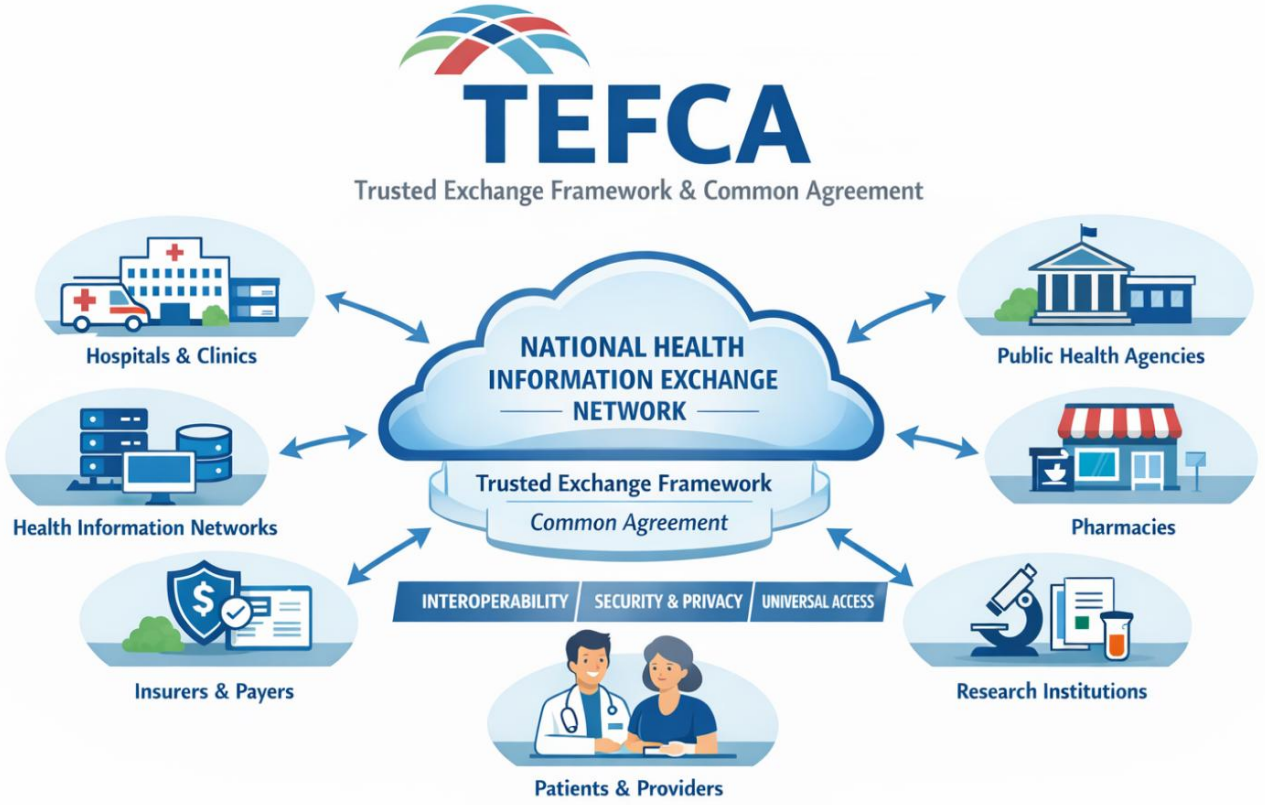
- Existing privacy models focus on two extremes:
  - **Local DP**: protects each individual record
  - **Central DP**: assumes a fully trusted aggregator

- In practice, institutions are often willing to share *summary statistics*,
  but must protect the privacy of their *internal datasets*.



A Real-World Example: TEFCA

- **TEFCA** establishes a national framework for health data exchange in the U.S.
- Under TEFCA, **institutions—not individuals—control data sharing**.
- Data are shared primarily as **aggregated or summary statistics**, under strict governance.
- Institutions must protect the privacy of their **internal datasets** while participating.



From Practice to Modeling — Three Privacy Mechanisms

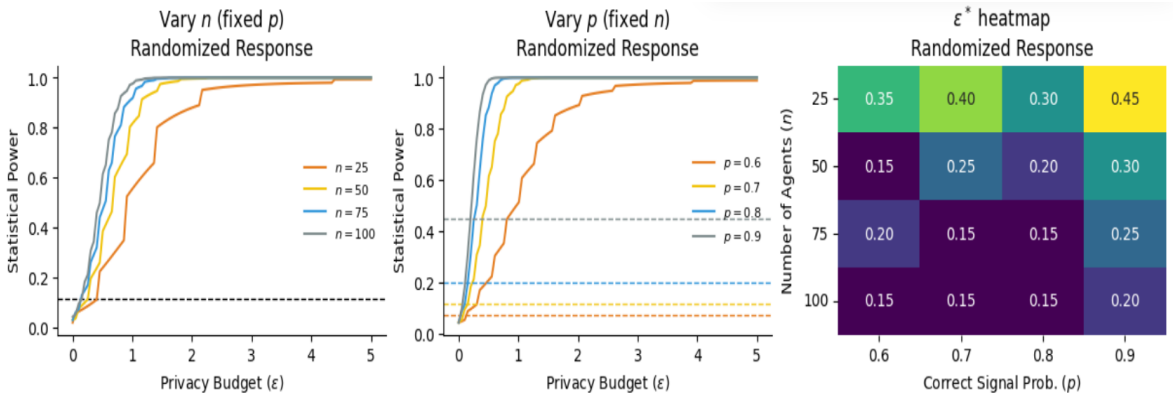
- **Local Privacy (Individual-Level DP)**
  Each individual perturbs their own data before sharing.
  → Strong individual protection, but often low statistical power.

- **Center-Level Privacy (Institution-Level DP)**
  Each institution shares privatized **summary statistics**, protecting its internal dataset.
  → Matches real-world collaborations such as TEFCA.

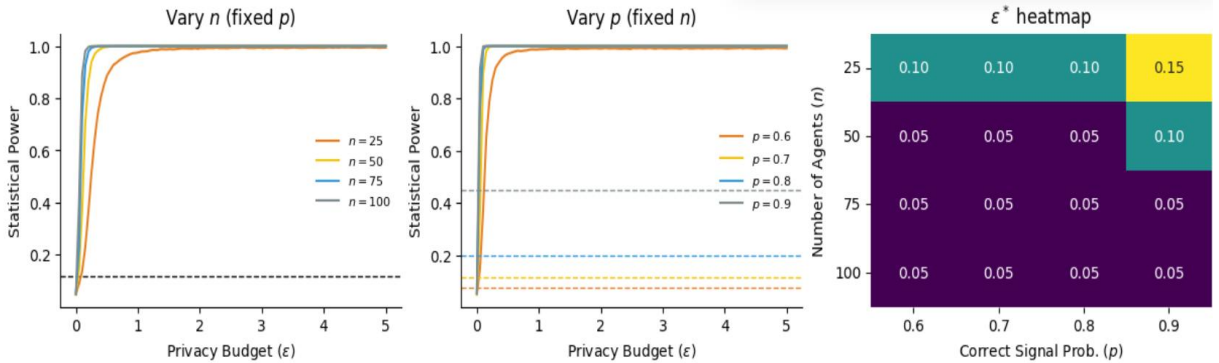
- **Central Privacy (Aggregator-Level DP)**
  Institutions share exact summaries with a trusted aggregator, which adds noise only at release.
  → High utility, but requires strong trust assumptions.

Result

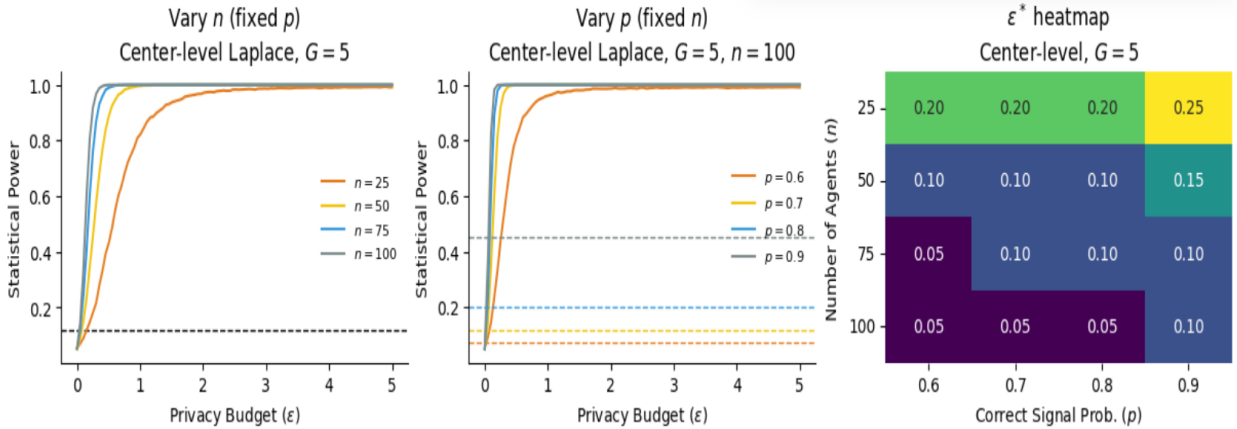
- **Local Privacy Mechanism**



- **Central Privacy Mechanism**

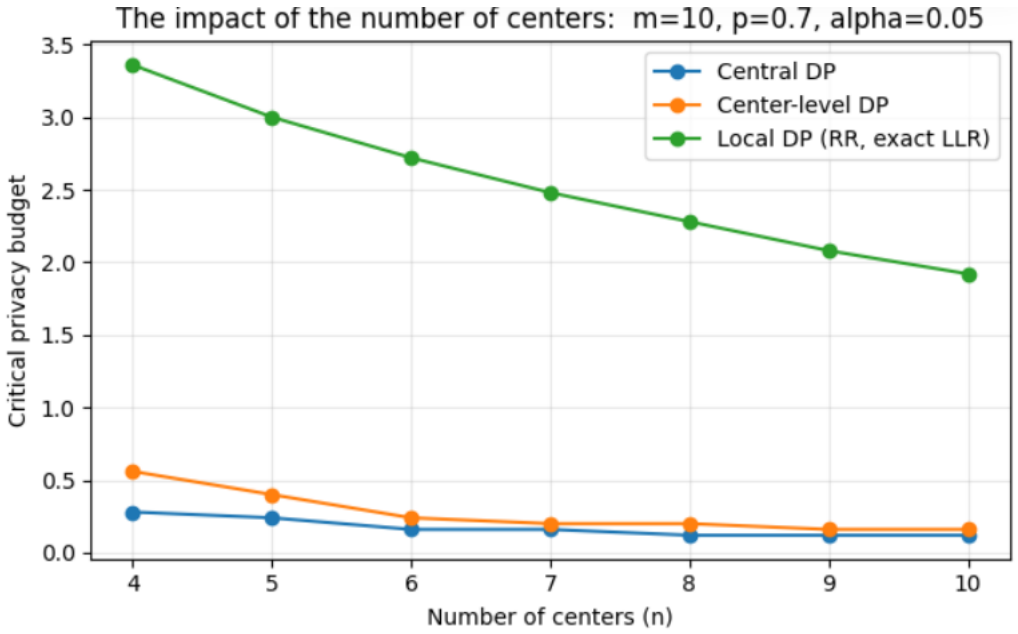


- **Center-Level Privacy Mechanism**



Result

- **The number of Centers**



- **The number of Datapoints in Center 1**