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

Yuxin Liu[1], *M. Amin Rahimian*[1], *Marios Papachristou*[2]

[1]University of Pittsburgh, [2]Arizona State University

Motivation. A recurring challenge in privacy-aware data collaboration is that sharing data enhances collective accuracy, yet this introduces privacy risks. Institutions in healthcare, finance, and education frequently face this *dilemma*: acting alone yields limited statistical power, while sharing data requires privacy protection that may degrade accuracy. This raises a central question at the heart of our study: *When does privacy-preserving data sharing outperform non-sharing?*

To illustrate the tension, we begin with a simple example of hypothesis testing. Consider n agents, each receiving a binary private signal that is correct with probability $p > 1/2$. Acting alone, each agent performs the most powerful level-α test and achieves some statistical power, denoted β_{ind}. Alternatively, the agents may share privatized signals using the randomized response (RR) mechanism, which flips each signal with probability depending on a privacy budget ε. A central aggregator then tests the hypothesis using the sum of these noisy reports, yielding collective power $\beta_{\text{RR}}(\varepsilon)$. Sharing becomes attractive when collective inference is more powerful than individual inference, i.e.

$$\beta_{\text{RR}}(\varepsilon) \;\geq\; \beta_{\text{ind}}.$$

This condition defines a *critical privacy budget* ε^* which corresponds to the minimum privacy level at which sharing noisy data is worth it. Figure 1 shows this relationship. For a fixed signal accuracy p, increasing the number of agents n reduces the required privacy budget, as aggregation compensates for noise. Conversely, when the private signals themselves are very accurate (high p), individual inference is already strong, and the critical privacy budget increases. Across typical parameter ranges, we find that ε^* remains modest ($\varepsilon^* < 1$), suggesting that privacy-preserving collaboration can be incentive-compatible even under relatively strict privacy guarantees.

This toy example highlights the core trade-off that motivates our study: balancing *statistical power*, *sample complexity*, and *privacy budget*. It also reveals why understanding participation incentives is essential: even with privacy guarantees in place, agents will only share if collective inference surpasses what they can do alone.

Methodology. We study statistical hypothesis testing as a fundamental setting in which agents decide whether to share information while protecting their privacy. Across various privacy regimes, we examine how statistical power, sample complexity, and privacy budget interact. Our goal is to identify conditions under which collective inference with privacy dominates individual inference without sharing. We consider three canonical levels of privacy protection.

First, under a *local privacy model*, each individual perturbs their own data prior to sharing, for example, using randomized response. This model captures decentralized environments where no trusted intermediary exists, and individuals require strong deniability guarantees for their personal data.

Second, under a center-level privacy model, institutions (e.g., clinical centers) are willing to share summary statistics but do not trust an external aggregator with their exact local data. Accordingly, each center independently applies

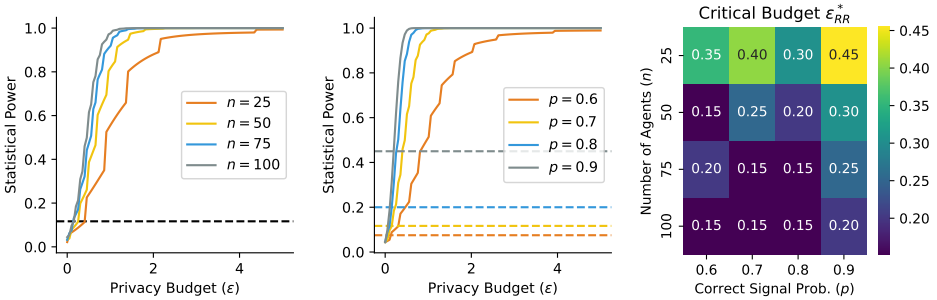

Figure 1. Critical privacy thresholds for hypothesis testing. Each panel compares individual statistical power (dotted line) with collective power under randomized response. **Left:** Increasing sample size n lowers the critical privacy budget ε^*. **Middle:** More accurate private signals (higher p) increase ε^*, since individuals already perform well. **Right:** Critical thresholds across (n, p); most lie below $\varepsilon = 1$.

a differentially private mechanism to its own local summary before transmission. The aggregator receives only privatized statistics and conducts inference based on these already-noised quantities. Privacy protection is therefore enforced separately at each center prior to aggregation. Compared to the local model, privacy operates at the institutional rather than individual level; compared to the central model described below, noise is injected before pooling rather than after aggregation.

Third, under a central privacy model, centers transmit their exact local summaries to a trusted aggregator. Differential privacy noise is added only once to the pooled statistic before public release. This model reflects settings in which internal collaboration and centralized trust are feasible, but public disclosure requires privacy guarantees. Because noise is introduced after aggregation rather than independently at each center, this approach avoids the compounding of noise across institutions and typically yields greater statistical efficiency. In this regime, we apply the Laplace mechanism on the aggregated statistics

By comparing statistical power under these mechanisms with the baseline of acting alone, we derive the *critical privacy budget* ε^*, namely the minimum budget above which agents strictly prefer to participate. Our analysis reveals how ε^* depends on group size, signal accuracy, and the type of mechanism. The results show that even under tight privacy constraints, collaborative inference can be incentive-compatible when mechanisms are well-designed.

Conclusion. This work develops an incentive-theoretic framework for the privacy–power–sample complexity trilemma, showing how different differential privacy regimes (local, center-level, and central) shape data-sharing incentives and statistical efficiency. We show that strong privacy and accurate inference can coexist when mechanisms align with the underlying trust and data access structure. By quantifying trade-offs across regimes, the framework informs data governance and policy design in settings such as multicenter clinical trials, federated inference, and epidemiological surveillance, offering guidance for building privacy-aware systems that balance protection, accuracy, and participation.