

Dependable and Secure Privacy Trading in Federated Learning via MFG-RegretNet

Kangkang Sun, *Member, IEEE*, Jianhua Li, *Senior Member, IEEE*, Xiuzhen Chen, *Member, IEEE*,
Weizhi Meng, *Senior Member, IEEE*, Minyi Guo, *Fellow, IEEE*

Abstract—Federated learning (FL) keeps raw data local, but repeated model-update sharing still creates cumulative privacy exposure and weak incentives for privacy-sensitive clients. This paper develops a dependable and secure privacy-trading framework in which each client’s differential-privacy budget is treated as a tradable commodity. We formalize this interaction as a Privacy Auction Game (PAG) and derive an analytical Privacy-Aware Auction Mechanism (PAC) with exact dominant-strategy incentive compatibility, individual rationality, and budget feasibility. For large client populations, we introduce MFG-RegretNet, a differentiable auction solver that combines mean-field game approximation with regret-minimizing mechanism learning. It clears a round with one forward pass whose cost grows linearly in N , and under secure computation its number of communication rounds does not grow with N ; the finite- N analytical mechanism converges to its closed-form mean-field limit at rate $\mathcal{O}_P(N^{-1/2})$. For MFG-RegretNet, which targets approximate rather than exact incentive compatibility, we prove uniform regret generalization bounds, a held-out regret certificate, and an $\mathcal{O}(1/N)$ bound on manipulation through the mean-field statistic. We further attach Gaussian RDP and Laplace composition accountants to track long-horizon privacy exposure and evaluate coordinated bid manipulation under 972 attack configurations. Experiments on MNIST, CIFAR-10, Bank, and NSL-KDD show that MFG-RegretNet achieves low empirical regret with a certified held-out bound and no IR violations, high auction efficiency, near-linear scalability up to $N=500$ clients, and competitive downstream FL utility under heterogeneous privacy valuations.

Index Terms—Federated Learning, Differential Privacy, Privacy Auction, Collusion-Resistant Mechanism Design, Mean-Field Game, Dependable Distributed Computing

I. INTRODUCTION

FEDERATED Learning (FL) [1], [2], [3] has emerged as the predominant paradigm for privacy-aware distributed machine learning, keeping raw data on local devices while sharing only model gradients with a central server. Yet a fundamental tension persists: *service providers require high-quality data to build accurate models, while users are increasingly reluctant to disclose sensitive information without adequate compensation or privacy guarantees*. Beyond accuracy, large-scale FL deployments also require *dependability*: the incentive layer must remain computationally scalable, robust

to strategic manipulation, and consistent with long-horizon privacy accounting when the same clients participate over many rounds. They also require *security*: privacy valuations and budget reports are sensitive market signals whose leakage or manipulation can affect both payments and privacy exposure [4]. Accordingly, the privacy-incentive layer must be evaluated as part of a secure distributed system: it should specify what information is visible to the broker, server, clients, and external observers; how bid confidentiality and gradient confidentiality can be protected; how strategic and coordinated manipulation are handled; and how cumulative privacy loss is accounted for across repeated FL rounds. This tension is exacerbated by two structural deficiencies that undermine practical FL adoption at scale.

First, FL does not eliminate privacy risk. Model gradients themselves constitute a serious privacy vulnerability: gradient inversion attacks [5] can reconstruct original training samples from uploaded gradients with high fidelity. Differentially private FL (DP-FL) [6] mitigates this risk by injecting calibrated noise into gradients before upload, but this introduces a new dilemma: different clients have heterogeneous privacy *valuations*: some are willing to upload relatively clean gradients (ϵ large) for higher compensation, while others require strong protection (ϵ small) and therefore contribute noisier, less useful gradients. A one-size-fits-all privacy policy fails to respect this heterogeneity, resulting in either privacy harm or avoidable accuracy loss.

Second, FL lacks effective incentive mechanisms for privacy-sensitive participants. In the absence of monetary compensation, rational users have little motivation to expend local computation, communication bandwidth, and crucially privacy budget. Most existing Federated Learning Market (FLM) approaches trade raw datasets, trained model weights, or FL training capacity [7], [8], [9]. Within FL, only a few works price privacy itself, notably the privacy-price contracts of Lin et al. [10] and the grouping-competition privacy allocation of Wu et al. [11]. Following this line, we treat the differential privacy budget ϵ_i , rather than data, models, or computation, as the *primary object of trade*, with users compensated in proportion to the privacy they sacrifice, and ask how such a market can be made truthful, budget-feasible, and scalable to large client populations.

Three Key Technical Challenges. Realizing this vision raises three fundamental challenges that prior work has not jointly addressed.

Challenge 1: Privacy Market Design. How should the privacy trading market be formally structured? A rigorous market

Kangkang Sun, Jianhua Li, Xiuzhen Chen and Minyi Guo are with the Shanghai Key Laboratory of Integrated Administration Technologies for Information Security, School of Computer Science, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: szpsunkk@sjtu.edu.cn; lijh888@sjtu.edu.cn; xzchen@sjtu.edu.cn; guo-my@cs.sjtu.edu.cn).

Weizhi Meng is with the School of Computing and Communications, Lancaster University, Lancaster LA1 4YW, United Kingdom, Email: weizhi.meng@ieee.org.

Corresponding author: Jianhua Li.

design must specify: (i) how users express their heterogeneous privacy valuations; (ii) how a data buyer (e.g., an FL server) specifies its accuracy requirements and budget; and (iii) how a semi-trusted broker/auctioneer allocates privacy budgets and determines payments. Critically, the mechanism must satisfy *Dominant-Strategy Incentive Compatibility* (DSIC), so that truthful reporting of privacy valuations is each user’s optimal strategy, and *Individual Rationality* (IR), so that participation is always at least as good as abstaining. Without these guarantees, rational users will strategically misreport valuations, corrupting both market efficiency and FL model quality.

Challenge 2: Scalability of Auction Equilibrium. Even with a well-defined mechanism, computing the Nash Equilibrium (NE) of a privacy auction game among N participants is intrinsically difficult. Each user’s optimal bid depends on the strategies of all other $N-1$ participants, so the NE problem has dimension $\mathcal{O}(N^2)$ and is PPAD-hard in general [12]. For the large-scale FL settings of interest (N in the hundreds to thousands), exact NE computation is computationally intractable. *How can we approximate the auction equilibrium efficiently without losing theoretical guarantees?*

Challenge 3: Near-Optimal Mechanism Design under Unknown Valuation Distributions. A mean-field view summarizes the population, but does *not* specify how to find a revenue-maximizing allocation and payment rule. Classical approaches (e.g., Myerson’s optimal auction theory [13]) yield closed-form solutions only under restrictive distributional assumptions. In FL, privacy valuations are heterogeneous, correlated across rounds, and drawn from an *unknown* distribution. *How can a mechanism be learned that is approximately revenue-optimal with low empirical regret, without assuming any closed-form prior over valuations?*

The three challenges above are addressed through a unified framework combining a *privacy trading market*, a *mean-field game (MFG) approximation*, and a *deep learning-based automatic auction solver*, termed **MFG-RegretNet**. Inspired by the foundational privacy auction work [14] and recent progress in differentiable mechanism design [15], this framework makes the following specific contributions to dependable and secure FL deployment.

- **Privacy Trading Market for Secure FL.** A formal privacy trading market for FL is established, in which each FL client treats its personalized differential privacy budget ϵ_i as a tradable commodity. Clients bid for compensation commensurate with their privacy cost $c(v_i, \epsilon_i) = v_i \cdot \epsilon_i$, and the FL server purchases privacy budgets subject to a total expenditure constraint B . Unlike the privacy–price contracts of Lin et al. [10], which are designed for a known distribution of privacy-preference types without a hard procurement budget, the market is organized as a budget-constrained reverse auction in which clients report both their privacy valuation and their privacy budget. The market model is paired with an explicit security and threat model that separates incentive compatibility from bid confidentiality, broker trust, client collusion, and gradient leakage.
- **Privacy Auction Game (PAG) with Theoretical Guarantees.** The interaction among clients and the server

is modeled as a *Privacy Auction Game* (PAG), and an analytical PAC procurement mechanism is designed to satisfy DSIC, IR, and Budget Feasibility (BF); its release does not depend on the declared privacy budget, so budget reports cannot be manipulated. It is proven (Theorem 1) that under these properties, truthful reporting constitutes a Bayesian-Nash Equilibrium of the PAC-based PAG, and Proposition 1 characterizes, for any mechanism, when truthful reporting of both the valuation and the privacy budget is optimal.

- **Mean-Field Game Approximation for Dependability.** We model the population of FL clients as a *mean-field flow* $\mu_t \in \mathcal{P}(\mathcal{V} \times \mathcal{E})$ and reduce the multi-agent Nash problem to a coupled Hamilton-Jacobi-Bellman / Fokker-Planck system, in which each client responds to the population law instead of to the individual bids of all other clients (Theorems 2 and 3). For the analytical mechanism, the mean-field limit is computed in closed form, and the deployed finite- N mechanism provably converges to it at rate $\mathcal{O}_P(N^{-1/2})$ (Theorem 4), as confirmed by simulations up to $N=10^4$.
- **MFG-RegretNet: Deep Learning Auction Solver.** MFG-RegretNet is proposed as a differentiable mechanism that addresses Challenge 3 by parameterizing allocation and payment rules as neural networks conditioned on the mean-field statistic b_{MFG} ; one forward pass clears a round, at a cost linear in N . Beyond using b_{MFG} as a context feature, an optional *MFG alignment regularizer* can align learned payments with the mean-field payment functional, so that MFG equilibrium structure enters the training gradient. Approximate DSIC is enforced by minimizing ex-post regret via augmented Lagrangian optimization, without any closed-form prior over privacy valuations. These incentive properties are approximate rather than exact PAC-style guarantees, and we bound them formally: a uniform regret generalization bound, a held-out regret certificate, an $\mathcal{O}(1/N)$ bound on manipulation through the mean-field statistic, and an exact characterization of truthfulness over two-dimensional reports (Sec. VII-C, Proposition 1).
- **Security- and Dependability-Oriented Validation.** A validation suite is provided, spanning calibrated CIFAR-10 No-DP baselines, Gaussian RDP and Laplace cumulative accounting over 200 rounds, coordinated collusion under six attack modes, and privacy-sensitive Bank/NSL-KDD FL tasks. These experiments directly address downstream utility, cumulative privacy exposure, collusion robustness, and scalability.

The remainder of this paper is organized as follows. Section II and Section III present related work and background on differential privacy and federated learning, respectively. Section IV and Section V present the privacy trading market and auction game, and the PAG-FL framework, respectively. Section VI and Section VII present the MFG approximation and MFG-RegretNet, respectively. Section VIII presents experimental results, and Section IX concludes the paper.

TABLE I
COMPARISON OF REPRESENTATIVE RELATED WORKS ACROSS KEY
DESIGN DIMENSIONS

Work	FL	Priv.	Auc.	MFG	DL	DSIC
Ghosh & Roth [14]	N/A	✓	✓	N/A	N/A	✓
Fleischer & Lyu [16]	N/A	✓	✓	N/A	N/A	<i>Approx.</i>
Yang et al. [17]	✓	N/A	✓	N/A	N/A	✓
Zheng et al. [8]	✓	N/A	✓	N/A	✓	<i>Approx.</i>
Sun et al. [18]	✓	N/A	N/A	✓	N/A	N/A
Hu et al. [19]	✓	N/A	N/A	✓	N/A	N/A
Dütting et al. [15]	N/A	N/A	✓	N/A	✓	<i>Approx.</i>
Lin et al. [10]	✓	✓	N/A	N/A	N/A	✓ [†]
Wu et al. [11]	✓	✓	✓	N/A	N/A	✓
MFG-RegretNet (Ours)	✓	✓	✓	✓	✓	<i>Approx.</i>

✓ = supported; N/A = not supported; *Approx.* = approximately satisfied via augmented Lagrangian training.

Priv. = privacy as commodity; Auc. = auction mechanism; MFG = MFG scalability; DL = deep learning; [†] incentive compatibility of contract self-selection.

II. RELATED WORK

This study lies at the intersection of four research streams, summarized in Table I.

Privacy-preserving FL. Gradient inversion attacks [5], [20], [21] demonstrate that an honest-but-curious server can reconstruct raw samples from shared gradients. The dominant defense is Differential Privacy: DP-SGD and its federated variants [6], [22] inject calibrated Gaussian noise into gradients before upload, offering (ϵ, δ) -DP at the cost of model utility. Secure aggregation [23], [2] uses cryptographic protocols so that the server observes only the aggregate. These approaches treat ϵ as a fixed system parameter rather than as a user-controlled commodity tradable for compensation.

FL incentive mechanisms. Incentive design for FL has been studied via Stackelberg games [24], [25], MFG-based reputation auctions [18], matching games [26], and DRL-based pricing [27]. Auction-based FL (AFL) [28] has emerged as a prominent sub-field, with CSRA [17] providing robust reverse auctions for DP-FL with truthfulness and budget-feasibility guarantees. Most AFL works treat privacy as a cost constraint and trade data, models, or compute capacity rather than the privacy budget itself; the exceptions are discussed next.

Privacy auctions and data markets. Ghosh and Roth [14] laid the theoretical foundation for privacy-as-commodity by formulating a reverse auction that elicits individual privacy valuations. Fleischer and Lyu [16] extended this to correlated privacy costs; Ghosh et al. [29] addressed unverifiable valuations. More recently, Hu et al. [19] proposed MFG-based data pricing for federated markets. Two FL works are closest to ours. Lin et al. [10] bound the effect of heterogeneous DP noise on FL convergence and derive optimal privacy-price contracts for discrete and continuous privacy-preference types, so that clients truthfully reveal their privacy attitudes over multi-round training. Wu et al. [11] allocate privacy budgets in FL through a grouping competition and treat privacy as a truthfully reported resource. Our framework differs from both in three respects: (i) PAC is a budget-feasible reverse

auction that is truthful in dominant strategies and whose release does not depend on the declared budget, whereas the contracts of [10] are optimized for a known type distribution without a hard procurement budget; (ii) MFG-RegretNet learns allocation and payment rules without a closed-form valuation prior; and (iii) the learned mechanism clears a round with one forward pass, so its cost is linear in N and its number of rounds under secure computation is constant. Most other work in this line studies static, single-round, centralized settings.

Differentiable mechanism design. Myerson [13] gives revenue-optimal auctions under restrictive distributional assumptions inapplicable to heterogeneous FL clients. RegretNet [15] parameterizes allocation and payment rules as neural networks trained to minimize ex-post regret, circumventing distributional priors; DM-RegretNet [8] applied this to FL model trading. MFG-RegretNet builds on RegretNet: it adds the mean-field statistic to the input, a budget projection that makes budget feasibility structural, and a regret search over misreports of the valuation, the privacy budget, and the data size, and it trades privacy budgets as the commodity.

Analytical DP-FL selection and pricing. Recent analytical approaches address heterogeneous- ϵ selection and compensation directly. Privacy-aware client selection [30] and heterogeneous DP aggregation [31] optimize selection probabilities under non-uniform noise variance, but provide no economic mechanism or incentive-compatibility guarantees. Jointly-optimal selection-and-allocation mechanisms (JSAM-style, e.g., [32]) derive structural results (threshold selection, $\epsilon_k^* \propto p_k^{2/3} v_k^{-1/3}$) under Bayesian priors, but do not scale to unknown distributions or large N . Our key distinctions are: (i) *distribution-free* learned mechanism via regret minimization; (ii) *MFG-aware regret* that accounts for population-level strategic effects; and (iii) a certified bound on the regret of the deployed network.

III. PRELIMINARIES

A. Differential Privacy

We adopt *Personalized Local Differential Privacy* (PLDP) [33], [34], which generalizes standard LDP [35] to allow each client to hold a distinct privacy budget.

Definition 1 (PLDP [33]) For client $i \in \mathcal{N}$ with $\epsilon_i > 0$ and $\delta_i \geq 0$, a randomized mechanism $\mathcal{M}^i$ is (ϵ_i, δ_i) -PLDP if for all adjacent inputs $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^i$ and all measurable sets $S \subseteq \mathcal{Y}^i$:

$$\Pr[\mathcal{M}^i(\mathbf{x}) \in S] \leq e^{\epsilon_i} \Pr[\mathcal{M}^i(\mathbf{x}') \in S] + \delta_i. \quad (1)$$

Smaller ϵ_i implies stronger privacy at the cost of lower data utility.

In practice, private gradients are perturbed via the Gaussian Mechanism [35]: for $\epsilon_i < 1$, noise scale $\sigma_i \geq \Delta f \sqrt{2 \ln(1.25/\delta_i)}/\epsilon_i$ makes the perturbed gradient (ϵ_i, δ_i) -DP, where Δf is the ℓ_2 -sensitivity of the gradient function; for larger budgets, σ_i is calibrated numerically from the Rényi-DP curve of the mechanism [36], as in our experiments. Either way, the allocated budget ϵ_i^{out} determines the noise injected into each client's update (see Algorithm 2).

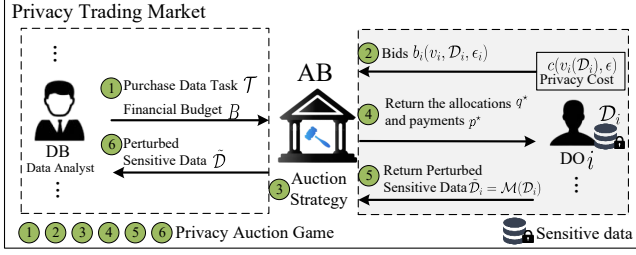


Fig. 1. Privacy Trading Market: the six-step auction process among DOs, the Auction Broker (AB), and DBs.

B. Federated Learning

In standard FL, a server S coordinates N clients, each holding a private dataset $\mathcal{D}_i$. Client i minimizes its local loss $\mathcal{L}_i(\mathbf{w}, \mathcal{D}_i)$ to obtain local model $\mathbf{w}_i$; the server aggregates:

$$\mathbf{w}_G = \sum_{i=1}^N \alpha_i \mathbf{w}_i, \quad \sum_{i=1}^N \alpha_i = 1, \quad (2)$$

targeting $\mathbf{w}_G^* = \arg \min_{\mathbf{w}} \sum_i \alpha_i \mathcal{L}_i(\mathbf{w}_i)$. Under DP-FL, each client perturbs its gradient with Gaussian noise calibrated to ϵ_i^{out} before upload, so the server can only observe the noisy aggregate. The ℓ_2 -sensitivity of client i 's gradient is:

$$\Delta f_{\mathbf{w}_i} = \max_{\mathcal{D}_i, \mathcal{D}'_i} \|\nabla \mathcal{L}_i(\mathbf{w}, \mathcal{D}_i) - \nabla \mathcal{L}_i(\mathbf{w}, \mathcal{D}'_i)\|_2, \quad (3)$$

where $\mathcal{D}'_i$ differs from $\mathcal{D}_i$ in exactly one sample. The allocated budget ϵ_i^{out} thus determines the noise scale σ_i and directly affects gradient quality and downstream model accuracy, making it the natural economic object to trade in the privacy market.

IV. PRIVACY TRADING MARKET AND PRIVACY AUCTION GAME

In this section, we introduce the *privacy trading market*, where privacy as a commodity trades among data owners (DOs) and data buyers (DBs). We then design a *privacy auction game* to incentivize participants to share data.

A. Privacy Trading Market

The privacy trading market comprises three roles: *Data Owners* (DOs), an *Auction Broker* (AB), and *Data Buyers* (DBs). As illustrated in Fig. 1, the auction proceeds in six steps. DOs $\mathcal{N} = \{1, \dots, n\}$ hold local sensitive datasets $\mathcal{D} = \{\mathcal{D}_1, \dots, \mathcal{D}_n\}$. We assume that DOs are *single-minded*: a DO will provide raw sensitive data only if the privacy preservation it requests is satisfied [37]. DO i protects her sensitive data through a mechanism $\mathcal{M}_{\epsilon_i}$ satisfying (ϵ_i, δ_i) -PLDP. DBs purchase sensitive data from DOs to complete analysis or model training tasks. The AB is a semi-trusted auctioneer that performs multiple rounds of auctions. Since a DO may misreport its privacy valuation $v(\mathcal{D})$, the mechanism must incentivize truthful reporting.

We model the AB as a semi-trusted auctioneer: it is expected to execute the published auction rule faithfully, but the bids it observes are privacy-sensitive and must be protected by protocol design in practical deployments.

Assumption 1 All market participants agree to use privacy as a common evaluation metric, making the auction a fair game. Each DO knows the true value of its own data; v represents the DO's valuation of its data privacy during the auction. DOs do not collude with each other during the auction process.

Assumption 1 is the precondition for the formal incentive analysis of PAC and MFG-RegretNet. The threat model below characterizes a broader practical attack surface, including risks that are not covered by the exact DSIC/BF proof.

B. Security and Threat Model

Our mechanism operates at the auction layer and is orthogonal to the secure aggregation layer. The AB observes bid tuples such as (v_i, ϵ_i, d_i) , the winner set, and payments. It does not need raw data or individual plaintext gradients; these can remain protected by local DP [10] and secure aggregation protocols [23]. The FL server observes the aggregate model update and payment outcome, but not losing bids or individual valuations if the auction is executed inside a protected broker implementation; Supplementary Table S14 lists what each party sees.

We consider three security risks. First, an honest-but-curious AB may try to infer clients' privacy preferences from bids. DSIC addresses strategic misreporting, not bid confidentiality; preventing bid leakage requires a hardware-protected or cryptographic implementation, i.e., a trusted execution environment (TEE) or secure multi-party computation (SMPC). A practical TEE implementation proceeds as follows: the AB publishes an enclave measurement for the auction code; clients verify remote attestation and encrypt bids to the enclave; the enclave computes allocation and payments, then signs a commitment to the winner set and payment vector. An SMPC implementation can instead split bid processing across several non-colluding computation parties, using secure sorting and payment computation for PAC or an MPC-friendly neural forward pass for MFG-RegretNet, revealing only the final allocation and payment outputs; RQ2 measures a secure SMPC round and the channel cost of the TEE path. A malicious broker or FL server cannot weaken the DP guarantee, since each client adds its noise before upload, and in the TEE path the enclave attests the published network and signs the outcome, so each client can check that it is paid the signed amount and that $\sum_i \bar{p}_i \leq B$. Second, clients may collude to manipulate bids or inflate payments. Our formal analysis assumes independent clients (Assumption 1); collusion-resistant auction design is outside the present theoretical guarantee. We therefore evaluate coordinated overbid, underbid, privacy-budget manipulation, and joint attacks empirically in RQ1. Third, gradient inversion is handled at the FL layer by DP noise and can be further mitigated by secure aggregation; it is not solved by incentive compatibility alone.

In deployment, the auction runs as a protected privacy-budget service: clients submit encrypted bids to a TEE- or SMPC-backed AB, which releases only the winners, payments, and budgets; the FL server runs DP-FL with secure aggregation; and an accountant records each client's cumulative exposure. PAC gives exact incentive guarantees, whereas MFG-

RegretNet trades exactness for a learned rule with certified regret, structural budget feasibility, and a number of secure-computation rounds that does not grow with N .

C. Privacy Auction Game (PAG)

To incentivize privacy trading, we design a privacy auction game (Fig. 1). Each DO i has a valuation function $v_i(\mathcal{D}_i; b_i)$, where bidding profile b_i is given at privacy budget (ϵ_j, δ_j) -PLDP. DOs report their valuations (possibly untruthfully) and the AB decides an allocation and payment $p_1, \dots, p_n$.

For DO i , the cost of having data used at privacy level ϵ is $c(v_i, \epsilon)$, linear in both v_i and ϵ : $c(v_i, \epsilon) \leq c(v_j, \epsilon)$ if and only if $v_i \leq v_j$. Specifically: $c_i(\epsilon) = \epsilon \cdot v_i$ [16].

The utility of DO i is:

$$u_i = p_i - c(v_i, \epsilon_i). \quad (4)$$

Summing client utilities defines *participant social welfare* (client-side economic surplus): monetary compensation net of privacy disutility. In one auction round,

$$\text{SW} = \sum_{i=1}^n u_i = \sum_{i=1}^n (p_i - v_i \epsilon_i^{\text{out}}), \quad (5)$$

where ϵ_i^{out} is the privacy level actually applied to client i after allocation (in FL, the noise scale of the uploaded gradient). Under IR, $p_i \geq c(v_i, \epsilon_i^{\text{out}})$ and each summand in Eq. (5) is non-negative. Payments are transfers from the server's budget B , so SW does not include the server's utility from model quality; the latter is evaluated via downstream accuracy (RQ4, Sec. VIII). In experiments we evaluate Eq. (5) using each client's *true* v_i when forming $c(\cdot)$, so that welfare aligns with realized privacy cost and is comparable across mechanisms. Across T_{FL} federated rounds we report the time average $\overline{\text{SW}} = \frac{1}{T_{\text{FL}}} \sum_{t=1}^{T_{\text{FL}}} \text{SW}^{(t)}$, where $(p_i^{(t)}, \epsilon_i^{(t), \text{out}})$ denote round- t outcomes. We write SW to distinguish welfare from the type distribution denoted $\mathcal{W}$ in Theorem 1.

Define bidder i 's type as $\theta_i = (v_i, \epsilon_i) \in \Theta_i$, where v_i is the privacy valuation and ϵ_i is the declared privacy-budget requirement. Let $\Theta = \Theta_1 \times \dots \times \Theta_n$ and θ_{-i} denote the joint type profile excluding agent i .

D. Privacy Auction Equilibrium

All bidders aim to maximize $u_i(v_i, b)$. To obtain a *Nash Equilibrium*, we design a *privacy procurement mechanism*.

Definition 2 (Privacy Procurement Mechanism) A privacy procurement mechanism $\mathcal{M} = (q, p)$ consists of:

- an allocation function $q : \Theta \rightarrow \{0, 1\}^n$;
- a payment function $p : \Theta \rightarrow \mathbb{R}^n$.

The mechanism $\mathcal{M}$ must satisfy the following properties:

- **Two-Dimensional Dominant-Strategy Incentive Compatibility (DSIC):** For each bidder $i \in \mathcal{N}$ and $\forall \theta_{-i} \in \Theta_{-i}$, truthful reporting of both valuation and privacy-budget requirement, $\theta_i = (v_i, \epsilon_i)$, is a dominant strategy regardless of others' reports:

$$u_i(\theta_i, \theta_{-i}) \geq u_i(\hat{\theta}_i, \theta_{-i}), \quad \forall \hat{\theta}_i \in \Theta_i. \quad (6)$$

- **Individual Rationality (IR):** For each bidder $i \in \mathcal{N}$, participation yields non-negative utility:

$$p_i(v) \geq c(v_i, \epsilon_i). \quad (7)$$

- **Budget Feasibility (BF):**

$$\sum_{i=1}^n p_i(v_i) \leq B. \quad (8)$$

Under DSIC and IR, the expected revenue given joint distribution $\mathcal{W}$ of (v, ϵ, δ) is:

$$\mathcal{R} := \mathbb{E}_{(v, \epsilon) \sim \mathcal{W}} \left[\sum_{i=1}^n p_i(v, \epsilon) \right]. \quad (9)$$

The optimal privacy auction problem is:

$$\mathcal{P}_1 : \max_{(q, p)} \mathcal{R} \quad \text{s.t.: DSIC, IR, BF.} \quad (10)$$

Definition 3 (Privacy Auction Equilibrium) A strategy profile $s = (s_1, \dots, s_n)$, where $s_i : \Theta_i \rightarrow \Theta_i$ maps a type to a report, is a *Bayesian-Nash Equilibrium (BNE)* if, for all $i \in \mathcal{N}$, $\theta_i \in \Theta_i$, and $\theta'_i \in \Theta_i$:

$$\mathbb{E}_{\theta_{-i}}[u_i(s_i(\theta_i), s_{-i}(\theta_{-i}); \theta_i)] \geq \mathbb{E}_{\theta_{-i}}[u_i(\theta'_i, s_{-i}(\theta_{-i}); \theta_i)]. \quad (11)$$

Theorem 1 (Privacy Auction Game Equilibrium) In a PAG where privacy valuations v_i of DOs are drawn from joint distribution $\mathcal{W}$, the cost function is linear $c(v_i, \epsilon_i) = \epsilon_i \cdot v_i$, and the mechanism $\mathcal{M} = (q, p)$ satisfies two-dimensional DSIC, IR, and BF, there exists a Bayesian-Nash Equilibrium in which all rational bidders truthfully report both their privacy valuations v_i and their privacy budgets ϵ_i : for all i and $\theta_i, \theta'_i \in \Theta_i$,

$$\mathbb{E}_{\theta_{-i}}[u_i(\theta_i, \theta_{-i}; \theta_i)] \geq \mathbb{E}_{\theta_{-i}}[u_i(\theta'_i, \theta_{-i}; \theta_i)]. \quad (12)$$

Proofs of all formal results are in Supplementary Sec. S5.

Remark 1 (Scope of Incentive Guarantees) Theorem 1 is formally established for the PAC mechanism (Algorithm 1), where DSIC is defined over the two-dimensional type (v_i, ϵ_i) . MFG-RegretNet satisfies only *approximate* DSIC via augmented-Lagrangian regret minimization; exact DSIC for arbitrary neural payment rules requires monotonicity conditions not verifiable for deep networks [15], which Proposition 1 states for the present setting. All exact DSIC/BNE claims apply to PAC; the approximate DSIC of MFG-RegretNet is bounded formally in Sec. VII-C and measured in RQ1 (Section VIII).

PAC selects the k lowest-valuation clients as winners (monotone allocation: adding a client can only increase k , not decrease it), assigns each winner $\epsilon_i^{\text{out}} = 1/(n - k)$ (equal-share DP budget), and pays each winner $p_i = \min(B/k, c(v_{k+1}, 1/(n - k)))$, i.e., a threshold payment equal to the critical bid at which the marginal client would be selected. The three properties follow directly: (i) *DSIC*: reporting $v'_i > v_i$ cannot increase p_i (threshold payment is determined by v_{k+1} , not v_i); misreporting $v'_i < v_i$ either leaves p_i unchanged (if i remains a winner) or yields $p_i = 0$ (if i drops

Algorithm 1: PAC Mechanism

Input: Privacy valuations v , datasets $\mathcal{D}$, budget B

Output: Perturbed data $\mathcal{D}'$, payments p

```

1 All participants submit privacy valuation  $v$  to AB;
2 Sort valuations:  $v_1 \leq v_2 \leq \dots \leq v_n$ ;
3 Let  $k$  be the largest integer such that  $c(v_k, \frac{1}{n-k}) \leq \frac{B}{k}$ ;
4 for  $i = 1$  to  $n$  do
5   Compute perturbed data:  $\mathcal{D}'_i \leftarrow \mathcal{D}_i + \mathcal{N}(0, \Delta_f^2 \sigma^2)$ ;
6   if  $i \leq k$  then
7      $p_i \leftarrow \min(\frac{B}{k}, c(v_{k+1}, \frac{1}{n-k}))$ ;
8   else
9      $p_i \leftarrow 0$ ;
10  end
11 end
12 return  $\mathcal{D}'$ ,  $p$ ;
```

out), so $u_i(v_i) \geq u_i(v'_i)$. (ii) *IR*: $p_i = \min(B/k, c(v_{k+1}, \epsilon)) \geq c(v_k, \epsilon) \geq c(v_i, \epsilon)$ for $i \leq k$ (since $v_i \leq v_k \leq v_{k+1}$). (iii) *BF*: $\sum_i p_i \leq k \cdot B/k = B$. The choice $\epsilon = 1/(n-k)$ is a conservative uniform allocation rule: all selected clients use the same DP budget, which simplifies the threshold payment and avoids heterogeneous noise scales in aggregation. Since this release does not depend on the declared budgets, PAC applies to winners whose budget is at least $1/(n-k)$. The winner index k is chosen only when $c(v_k, \epsilon) \leq B/k$, so the rule is compatible with the per-client affordability condition. More general non-uniform or learned privacy-budget rules are possible, but would require a separate monotonicity and critical-payment analysis to retain exact DSIC. *Complexity*: $\mathcal{O}(N \log N)$ (sorting).

The two-dimensional statement holds for any mechanism under a simple condition. Fix the other reports and the reported data size, and let $e(\hat{v}, \hat{\epsilon})$ and $p(\hat{v}, \hat{\epsilon})$ be client i 's release and payment. A declared cap commits the client to accept any release up to it, so a single-minded client never declares more than its cap ϵ , but it may declare less; likewise, a client cannot report more data than it holds.

Proposition 1 (Truthfulness with One-Sided Caps) *Let $U(v, \hat{\epsilon}) = \int_v^{\hat{\epsilon}} e(s, \hat{\epsilon}) ds + h(\hat{\epsilon})$. Truthful reporting is optimal for every type $(v, \epsilon) \in [0, \bar{v}] \times [0, \bar{\epsilon}]$ over all reports with $\hat{\epsilon} \leq \epsilon$ if and only if, for some function h , (i) $e(\cdot, \hat{\epsilon})$ is non-increasing and $p(\hat{v}, \hat{\epsilon}) = \hat{v} e(\hat{v}, \hat{\epsilon}) + U(\hat{v}, \hat{\epsilon})$ for every $\hat{\epsilon}$, and (ii) $U(v, \cdot)$ is non-decreasing for every v .*

Condition (i) is Myerson's characterization [13] at each cap, and (ii) rules out profitable cap understatement (proof in Supplementary Sec. S5.B). PAC satisfies (ii) trivially, with equality, because its release $1/(n-k)$ ignores the declared cap. A rule whose payment ignores the cap while releasing $e = z\hat{\epsilon}$ violates (ii): the client gains $vz(\epsilon - \epsilon_{\text{floor}})$ by declaring the smallest admissible cap ϵ_{floor} .

V. PRIVACY-AWARE AUCTION FEDERATED LEARNING MECHANISM

We combine the privacy auction game with FL to incentivize users to share data for three reasons. First, users are concerned about privacy risks from local model leakage. Second, without

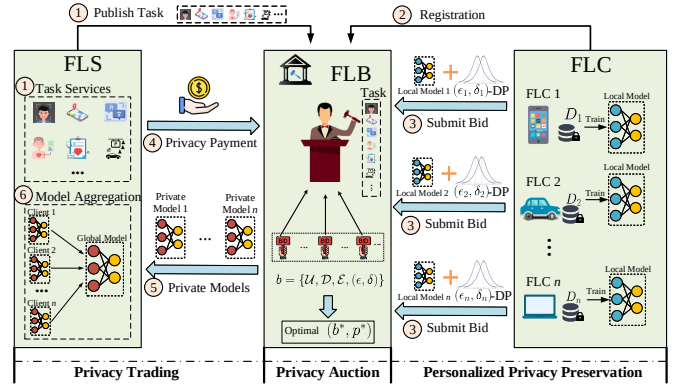


Fig. 2. The PAG-FL framework showing six operational steps between the FL Server (FLS), FL Broker (FLB), and FL Clients (FLC).

revenue, users have limited incentive to participate in FL training. Third, a unified privacy policy ignores user heterogeneity, reducing model accuracy. We propose the *Privacy Auction Game-based FL* (PAG-FL) framework (Fig. 2).

A. PAG-FL Framework

PAG-FL involves three entities:

- 1) *FL Server (FLS)*. FLS publishes model training tasks (e.g., image classification, text translation) and purchases client gradients via the FL architecture.
- 2) *FL Broker (FLB)*. FLB is the auction institution for user data privacy. It recruits FLCs eligible for a given task and runs the privacy auction to determine allocation and payment.
- 3) *FL Clients (FLC)*. FLCs are users or organizations that provide raw training data. Each FLC trains locally, applies a DP mechanism, and uploads a private gradient to FLB for auction.

The operational steps are:

- 1) *Publish Task*. FLS publishes a learning task to FLB, which recruits eligible FLCs.
- 2) *Registration*. FLCs register with FLB and declare initial privacy requirements (ϵ, δ) .
- 3) *Submit Bid*. Each FLC i submits a bid $b_i = (v_i, \epsilon_i, d_i)$, where v_i is a valuation, ϵ_i is the privacy budget, and $d_i = |\mathcal{D}_i|$ is the size of its local dataset.
- 4) *Privacy Payment*. FLB runs the auction (PAC mechanism, Algorithm 1) and FLS pays selected FLCs.
- 5) *Private Model*. Each selected FLC applies its personalized DP mechanism to the local gradient.
- 6) *Model Aggregation*. FLS aggregates received private gradients and broadcasts the updated global model.

B. FL Privacy Auction Model

1) *Valuation Function of FLC*: The valuation v_i of FLC i reflects the economic value that client i places on its privacy. It is determined by three factors: (i) the *sensitivity* of the local dataset $\mathcal{D}_i$, which is higher for medical or financial records than for generic browsing data; (ii) the *data quality*, quantified by the local model accuracy acc_i and dataset size $|\mathcal{D}_i|$; and

(iii) the *non-IID degree* of the data distribution, measured by the divergence between the local and global data distribution. A concrete parameterization is:

$$v_i = \alpha_v \cdot \frac{\text{acc}_i}{\epsilon_i} + \beta_v \cdot |\mathcal{D}_i| + \gamma_v \cdot \Delta_{\text{noniid},i}, \quad (13)$$

where $\alpha_v, \beta_v, \gamma_v > 0$ are weighting coefficients and $\Delta_{\text{noniid},i}$ quantifies the non-IID divergence (e.g. total variation between local and global class frequencies). This formulation captures the key observation that higher-quality and more privacy-sensitive data should command higher valuations. In the theoretical analysis (Sections VI through VII), v_i is treated as drawn from a generic distribution $\mathcal{W}$, so the results hold independently of the specific form of Eq. (13). In experiments (Section VIII), valuations are drawn directly from the scenarios of Supplementary Table S10: $v_i \sim \text{Uniform}[0, 1]$, independent of ϵ_i , in the uniform and bimodal scenarios, and $v_i \sim \text{LogNormal}(\mu=0, \sigma=0.5)$ in the realistic scenario, reflecting heterogeneous real-world privacy sensitivities.

The cost $c_i(\epsilon_i^{\text{out}}) = \epsilon_i^{\text{out}} \cdot v_i$ depends on the *allocated* ϵ_i^{out} (mechanism output), not the declared type ϵ_i ; the privacy and cost trade-off is genuine.

Obtaining valuations in practice. The mechanism asks client i only for its own per-unit privacy price v_i ; it needs neither the valuations of other clients nor the distribution $\mathcal{W}$. Because truthful reporting is a dominant strategy under PAC (Sec. IV-D), a client that does not know how others bid gains nothing by shading its report, so its best response is its own subjective estimate; for MFG-RegretNet, the gain from misreporting is kept small and measured by ex-post regret (RQ1). A deployment can set v_i in three ways. (i) Organizational clients such as hospitals or banks derive v_i from their data-governance policy, mapping the sensitivity class of $\mathcal{D}_i$ and the associated contractual or regulatory liability to a price per unit of ϵ . (ii) Individual users select one entry from a short menu of privacy levels with posted compensation in the client application, a stated-preference format used in empirical studies of privacy valuation [38], and the selected entry is converted into v_i . (iii) Otherwise, the client software computes a default from Eq. (13) using its own $|\mathcal{D}_i|$, acc_i , and $\Delta_{\text{noniid},i}$. Since bids are re-submitted in every round (Algorithm 2), a client can revise v_i when its context changes, for example by raising it as its remaining privacy allowance (Sec. VIII-A1) shrinks. The surrounding protocol, i.e., encrypted bid submission to a protected broker, release of only the winners, payments, and budgets, DP-FL with secure aggregation, and per-client privacy accounting, is described in Sec. IV-B.

2) *Bidding Information:* In FL, global model accuracy depends on training dataset size, non-IID data distribution, and model privacy level [39], [40]. FLC i 's bid is:

$$b_i = (v_i, \epsilon_i, d_i), \quad (14)$$

where ϵ_i and $d_i = |\mathcal{D}_i|$ indicate the privacy budget and the data size; the local accuracy acc_i enters only through the valuation, Eq. (13).

3) *FL Aggregation Mechanism:* Under privacy preservation, the aggregated global gradient error is:

$$\mathcal{L}(\mathbf{w}; \alpha, \epsilon) = \|\tilde{\mathbf{w}}_G - \mathbf{w}_G^*\|_\ell = \left\| \sum_{i=1}^N \alpha_i \mathcal{M}_{\epsilon_i}(\mathbf{w}_i) - \mathbf{w}_G^* \right\|_\ell. \quad (15)$$

The FL privacy auction objective is:

$$\mathcal{P}_2 : \min_{\mathbf{w}, \alpha, \epsilon} \mathbb{E}_{(\mathbf{b}, B)} [\mathcal{L}(\mathbf{w}; \alpha, \epsilon)] \quad \text{s.t.: DSIC, IR, BF.} \quad (16)$$

4) *Economic Meaning of the Traded Budget:* Following the privacy-auction literature [14], [16], the unit of trade is the DP parameter: in each round, client i sells the right to release its local update at privacy level ϵ_i^{out} and is compensated for the resulting privacy cost $v_i \epsilon_i^{\text{out}}$. For the server, the same number measures the quality of what it buys. If the local update is released with the classical Gaussian calibration of Sec. III, the per-coordinate noise variance is $\sigma_i^2 = 2 \ln(1.25/\delta) \Delta f^2 / (\epsilon_i^{\text{out}})^2$, so a larger budget yields a less perturbed and thus more useful update. Under the budget-proportional weights $\alpha_i = \epsilon_i^{\text{out}} / \sum_{i' \in \mathcal{I}} \epsilon_{i'}^{\text{out}}$ of Algorithm 2 and independent zero-mean noise, the squared error of Eq. (15) with $\ell = 2$ decomposes as

$$\mathbb{E} \|\tilde{\mathbf{w}}_G - \mathbf{w}_G^*\|_2^2 = \|\mathbf{w}_G - \mathbf{w}_G^*\|_2^2 + \frac{C_{\text{DP}} |\mathcal{I}|}{(\sum_{i \in \mathcal{I}} \epsilon_i^{\text{out}})^2}, \quad (17)$$

where $\mathcal{I}$ is the winner set, $\mathbf{w}_G = \sum_{i \in \mathcal{I}} \alpha_i \mathbf{w}_i$ is the noise-free aggregate, $C_{\text{DP}} = 2 d_{\mathbf{w}} \Delta f^2 \ln(1.25/\delta)$ with $d_{\mathbf{w}}$ the number of model parameters, and $\epsilon_i^{\text{out}} \geq \epsilon_{\min}$. The first term depends on data quality and heterogeneity, which d_i and acc_i describe; the second term is the DP noise that the server pays to reduce. A smaller budget thus gives a client stronger privacy at the price of a noisier contribution, and the payment rule sets the exchange rate between the two: the noise term decreases with the square of the total purchased budget, while each client's privacy cost grows linearly in the budget it sells. With DP-SGD (RQ4), the noise enters during local training, and Eq. (17) describes the same trade-off qualitatively. Across rounds, the purchased budgets compose into each client's end-to-end guarantee as given in Eq. (42) of Sec. VIII-A1.

VI. MEAN-FIELD GAME APPROXIMATION FOR THE PRIVACY AUCTION

In Section IV, we formulated the PAG as the N -player optimization problem $\mathcal{P}_1$ (Eq. (10)). Although Theorem 1 guarantees the existence of a Bayesian-Nash Equilibrium under the PAC mechanism (Algorithm 1), *computing* that equilibrium becomes computationally intractable as the number of FL clients N grows. This section introduces a Mean-Field Game (MFG) approximation in which each client responds to the population law rather than to the individual bids of the other $N - 1$ clients. At the exact Mean-Field Equilibrium (MFE), the induced strategy profile is an $\mathcal{O}(N^{-1/2} \log N)$ -Nash equilibrium of the N -player PAG (Theorem 3); this bound concerns the MFE itself, and Remark 2 states how the deployed MFG-RegretNet relates to it. For the analytical rule of Algorithm 1, the mean-field solution is computed in closed form and the deployed finite- N mechanism converges

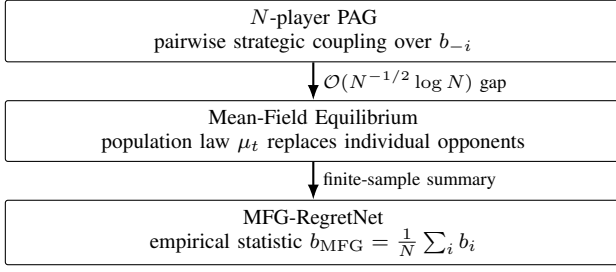


Fig. 3. Approximation hierarchy from the original N -player privacy auction to the deployable MFG-RegretNet statistic. The gap of Theorem 3 refers to the upper arrow. For the lower arrow, Sec. VII-C bounds the regret of the deployed network directly and shows that the mean-field statistic contributes at most $\mathcal{O}(1/N)$ to it.

to it at rate $\mathcal{O}_P(N^{-1/2})$ (Theorem 4). We model FL rounds by a continuous time index $t \in [0, T]$ so that the HJB-FP formalism describes the *law* of valuations and privacy types; the auction *within* each round remains the static PAG of Sec. IV, embedded in the representative-agent utility Eq. (19).

The intuition is simple. In the original N -player auction, each client's best response depends on the bids of all other $N - 1$ clients. The MFG approximation replaces these pairwise dependencies with a population distribution μ_t . MFG-RegretNet then gives its networks the empirical mean bid $b_{\text{MFG}} = \frac{1}{N} \sum_i b_i$ as a finite-population summary of that distribution. The MFE approximates the true N -player equilibrium; the network is not trained to solve the MFE, and Sec. VII-C bounds its incentive properties directly.

A. Scalability and MFG State Space

For a mechanism without dominant strategies, the NE of the N -player PAG requires solving N coupled optimizations, which is PPAD-hard in general [12]; best-response iteration repeats an $\mathcal{O}(N \log N)$ evaluation of a sorting-based mechanism in each of $\mathcal{O}(N)$ iterations. (PAC needs none of this, since truthful reporting is dominant.) The MFG framework [41], [42] replaces explicit agent-to-agent coupling with a representative agent interacting with the aggregate population law μ_t (Corollary 1).

Each DO's state at round t is $\mathbf{s}_i(t) = (v_i(t), \epsilon_i(t)) \in \mathcal{S} \triangleq \mathcal{V} \times \mathcal{E}$, evolving via the controlled SDE:

$$d\mathbf{s}_i(t) = f(\mathbf{s}_i(t), b_i(t), \mu_t) dt + \sigma dW_i(t), \quad (18)$$

where $b_i(t) \in \mathcal{B}$ is the bid, f is the drift, $\sigma > 0$ is a noise coefficient, $\{W_i(t)\}$ are independent Brownian motions, and the state is reflected at the boundary of $\mathcal{S}$ (R6). The empirical distribution $\mu_t^N = \frac{1}{N} \sum_i \delta_{\mathbf{s}_i(t)}$ converges weakly to the mean-field flow μ_t under standard Lipschitz regularity conditions on f (Assumption 2).

Assumption 2 (Regularity Conditions) (R1) Lipschitz

drift: $\|f(s, b, \mu) - f(s', b, \nu)\| \leq L_f(\|s - s'\| + W_1(\mu, \nu)).$

(R2) Compact controls: $\mathcal{B} \subset \mathbb{R}^m$ compact and convex.

(R3) Bounded costs: $c(v, \epsilon) = v \cdot \epsilon$ bounded and continuous, and $\tilde{p}(b; \mu)$ continuous in b .

(R4) Finite-moment initial law: $\mu_0 = \mathcal{W} \in \mathcal{P}_2(\mathcal{S})$.

(R5) Exogenous types: bids do not affect the type dynamics, $f(s, b, \mu) = f(s, \mu)$, and in each round the clients' types are i.i.d. draws from μ_t .

(R6) Compact state space: $\mathcal{S}$ is compact and convex, and the state is reflected at its boundary.

B. Hamilton-Jacobi-Bellman and Fokker-Planck Equations

Full derivations are provided in Supplementary Material S3. In the mean-field regime, each representative agent maximizes cumulative *mean-field utility*

$$u^{\text{MFG}}(s, b; \mu_t) = \tilde{p}(b; \mu_t) - c(v, \epsilon(b)), \quad (19)$$

where $\tilde{p}(b; \mu_t)$ is the mean-field payment, the payment for bid b against opponents whose types have law μ_t (Assumption 3). The *mean-field bid* summarizes the competitive environment:

$$b_{\text{MFG}}(t) = \mathbb{E}_{\mu_t}[b] = \int_{\mathcal{B}} b(s; \mu_t) d\mu_t(s). \quad (20)$$

The optimal bid induced by the HJB Hamiltonian is

$$b^*(s, t; \mu) = \arg \max_{b \in \mathcal{B}} [\tilde{p}(b; \mu_t) - c(v, \epsilon(b)) + (\nabla_s \Phi)^\top f(s, b, \mu_t)]. \quad (21)$$

Under (R5) the last term does not depend on b , so b^* maximizes the round's mean-field utility Eq. (19), consistent with the static auction within each round. The value function Φ and the population density μ_t are jointly characterized by the coupled HJB-Fokker-Planck system (Eqs. (S2) and (S3) in the Supplementary Material); their fixed point defines the Mean-Field Equilibrium below.

C. Mean-Field Equilibrium: Definition and Existence

Definition 4 (Mean-Field Equilibrium (MFE)) A pair (Φ^*, μ^*) is a MFE if:

- (i) *Optimality:* Given μ^* , Φ^* solves the HJB equation, Eq. (S2), and b^* is the maximizer of Eq. (21);
- (ii) *Consistency:* μ^* solves the FP equation, Eq. (S3), under $b^*(\cdot; \mu^*)$, with $\mu_0^* = \mathcal{W}$.

Theorem 2 (Existence of MFE) Under Assumption 2, the privacy auction MFG admits at least one Mean-Field Equilibrium (Φ^*, μ^*) .

D. MFG Approximation Guarantee

Assumption 3 (Anonymous Lipschitz Payment) The

mechanism is anonymous: client i 's N -player payment is $\tilde{p}(b_i; \mu_{-i}^N)$, where μ_{-i}^N is the empirical law of the other clients' types, and $\tilde{p}(b; \mu)$ is L_u -Lipschitz in μ w.r.t. W_1 :

$$|\tilde{p}(b; \mu) - \tilde{p}(b; \nu)| \leq L_u W_1(\mu, \nu), \quad \forall b \in \mathcal{B}, \mu, \nu \in \mathcal{P}(\mathcal{S}). \quad (22)$$

This is an assumption on a smooth payment rule. The threshold rule of Algorithm 1 decides through an order statistic and is analyzed separately in Theorem 4 with the Kolmogorov distance.

Definition 5 (ϵ -Nash Equilibrium) A strategy profile $\mathbf{b}^N = (b_1^N, \dots, b_N^N)$ is an ϵ -Nash Equilibrium if, for all $i \in \mathcal{N}$ and $\hat{b}_i \in \mathcal{B}$:

$$\mathbb{E}[u_i(\hat{b}_i, b_{-i}^N) - u_i(b_i^N, b_{-i}^N)] \leq \epsilon. \quad (23)$$

Theorem 3 (MFG Approximation of Nash Equilibrium)

Let Assumptions 2 and 3 hold, and let (Φ^*, μ^*) be a MFE. Define the MFE-induced strategy profile:

$$b_i^{N,*} = b^*(s_i, t; \mu^*), \quad i = 1, \dots, N. \quad (24)$$

Then $\mathbf{b}^{N,*} = (b_1^{N,*}, \dots, b_N^{N,*})$ is an ε_N -Nash Equilibrium (Definition 5) of the PAG, where:

$$\varepsilon_N = \frac{2L_u C \log N}{\sqrt{N-1}} = \mathcal{O}\left(\frac{\log N}{\sqrt{N}}\right), \quad (25)$$

and $C > 0$ depends only on the diameter of the compact state space $\mathcal{S} \subset \mathbb{R}^2$.

In words, as the number of clients grows, replacing explicit client-to-client strategic coupling with a mean-field population summary becomes increasingly accurate at rate $\log N / \sqrt{N}$. The logarithm comes from the two-dimensional state (v, ϵ) : the empirical measure of N i.i.d. points in $\mathbb{R}^2$ approaches its law in W_1 at rate $N^{-1/2} \log N$ [43], and in general this rate cannot be improved to $N^{-1/2}$ [44]. The theorem is stated for the exact MFE profile $\mathbf{b}^{N,*}$; it does not by itself bound the regret of the learned mechanism of Sec. VII (Remark 2).

Corollary 1 (Computational Complexity Reduction)

Once (Φ^*, μ^*) is computed at a cost independent of N (via the coupled HJB-FP system, Eqs. (S2) and (S3)), evaluating $b^*(s_i; \mu^*)$ for each agent costs $\mathcal{O}(1)$, so the MFE profile of N clients costs $\mathcal{O}(N)$, against $\mathcal{O}(N^2 \log N)$ for best-response iteration with a mechanism without dominant strategies.

E. Mean-Field Limit of the Analytical Mechanism

For the analytical rule of Algorithm 1, the mean-field solution is available in closed form. Because the rule decides through an order statistic, we measure the population error with the Kolmogorov distance $\|F_N - F\|_\infty$ between the empirical and population CDFs of the reported valuations.

Theorem 4 (Mean-Field Limit of PAC) Let $v_1, \dots, v_N$ be i.i.d. with CDF F on $[0, \bar{v}]$ and density $f \geq f_{\min} > 0$, let the budget $B > 0$ be fixed, and let $\kappa^* \in (0, 1)$ be the unique root of

$$\Psi(\kappa) = \kappa(F^{-1}(\kappa) + B) - B. \quad (26)$$

Let k_N be the number of winners of Algorithm 1 and $\hat{q}_N = k_N/N$. If $N \geq \max\{(1 - \kappa^*)^{-1}, 1 + \bar{v}/B\}$, then

$$|\hat{q}_N - \kappa^*| \leq \frac{\|F_N - F\|_\infty}{B f_{\min}} + \frac{1}{N}, \quad B - \frac{B + \bar{v}}{N - k_N} \leq \sum_i p_i \leq B. \quad (27)$$

Since $\Pr(\|F_N - F\|_\infty > s) \leq 2e^{-2Ns^2}$ [45], $\hat{q}_N \rightarrow \kappa^*$ at rate $\mathcal{O}_P(N^{-1/2})$; the per-winner release and payment satisfy $N\epsilon_{\text{win}}^{\text{out}} \rightarrow (1 - \kappa^*)^{-1}$ and $Np_{\text{win}} \rightarrow B/\kappa^*$ at the same rate; and the budget is exhausted at rate $\mathcal{O}(1/N)$.

Proof idea. Algorithm 1 accepts k if and only if $\Psi_N(k/N) \leq 0$, where Ψ_N is Ψ with F replaced by F_N . A density bounded below turns $\|F_N - F\|_\infty$ into a uniform perturbation of Ψ , whose slope is at least B ; at κ^* the two arguments of the PAC payment coincide. The full proof is in Supplementary Sec. S5.E.

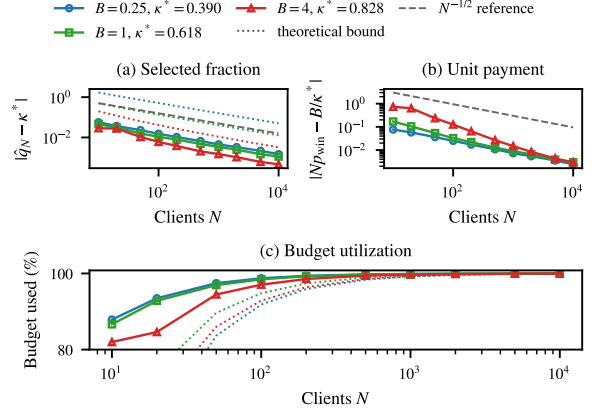


Fig. 4. Finite- N PAC against its mean-field limit (Theorem 4), $v \sim \mathcal{U}(0, 1)$, 1024 markets per point, every simulated N plotted. (a) Selected fraction and (b) normalized per-winner payment; error bars are one standard error of the mean, and dotted lines in (a) are the bound $\sqrt{\pi/(2N)}/B + 1/N$ implied by Eq. (27). (c) Budget used; dotted lines are the deterministic lower bound of Eq. (27), averaged over markets.

Fig. 4 checks the theorem on the PAC implementation of Sec. VIII with $v \sim \mathcal{U}(0, 1)$, three budgets, and 1024 markets for each $N \in [10, 10^4]$. Both inequalities of Eq. (27) hold in all 30,720 markets, and the fitted log-log slopes (-0.52 to -0.63 for the selected fraction, -0.51 to -0.85 for the unit payment, and -0.96 to -1.00 for the unused budget) match the predicted $N^{-1/2}$ and N^{-1} rates.

F. MFG-Reduced Privacy Auction Problem

Theorem 3 and Corollary 1 motivate replacing $\mathcal{P}_1$ with the tractable MFG-reduced problem:

$$\mathcal{P}_3 : \max_{q, p, \mu^*} \mathbb{E}_{s \sim \mu_0^*} \left[\sum_{i=1}^N p_i(b_i^{N,*}; b_{\text{MFG}}) \right] \quad (28)$$

subject to FP, HJB, penalized DSIC/IR/BF violations.

$\mathcal{P}_3$ has three advantages over $\mathcal{P}_1$: (a) *reduced dimensionality*: opponent interaction is summarized by the low-dimensional statistic $b_{\text{MFG}} \in \mathbb{R}^m$; (b) *equilibrium foundation*: the exact MFE of $\mathcal{P}_3$ induces an $\mathcal{O}(N^{-1/2} \log N)$ -Nash profile of $\mathcal{P}_1$ by Theorem 3; (c) *neural-network solvability*: gradient-based optimization is applicable.

Remark 2 (Approximation Hierarchy) Theorem 3 guarantees that the MFE strategy $b^*(\cdot; \mu^*)$ is an ε_N -Nash Equilibrium given the exact mean-field flow μ^* (solved from the coupled HJB-FP system). MFG-RegretNet does not solve the HJB-FP system: its networks read the reports together with the empirical mean bid $b_{\text{MFG}} = \frac{1}{N} \sum_i b_i$, a finite-particle estimate of $\mathbb{E}_{\mu_t}[b]$, and are trained to minimize regret (Sec. VII-B). The theoretical $\mathcal{O}(N^{-1/2} \log N)$ guarantee therefore applies to the exact MFE solution of $\mathcal{P}_3$ and is not inherited by the trained network, whose incentive properties are instead bounded directly in Sec. VII-C and measured in Sec. VIII-B1.

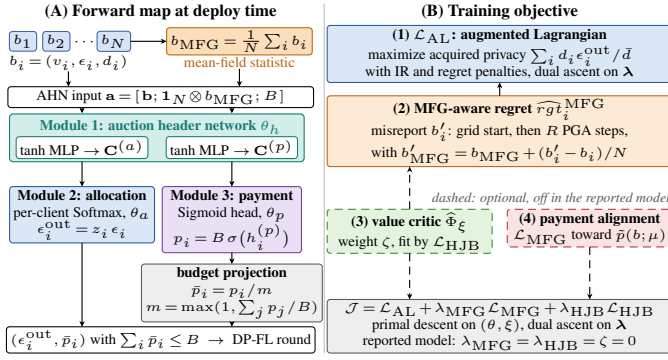


Fig. 5. MFG-RegretNet: (A) forward map (Sec. VII-A); (B) training objective (Sec. VII-B).

VII. MFG-REGRETNET: DEEP LEARNING AUCTION SOLVER

Remark 2 identifies two layers: (i) the MFE ε_N approximates the N -player Nash Equilibrium (Theorem 3), with $\varepsilon_N = \mathcal{O}(N^{-1/2} \log N)$; and (ii) computing the exact MFE requires solving the coupled HJB-FP system (Eqs. (S2) and (S3)), which is costly when $\mathcal{W}$ is unknown and must be run iteratively. MFG-RegretNet does not solve this system. It *parameterizes* (q, p) as neural networks that read the reports together with the empirical mean bid b_{MFG} , computed in $\mathcal{O}(Nm)$, and penalizes empirical DSIC/IR violations through differentiable augmented Lagrangian terms, while a projection makes budget feasibility structural. It needs no closed-form distributional assumptions, and the optional alignment regularizer (Sec. VII-B) can tie it to the MFG equilibrium structure.

We propose **MFG-RegretNet**. Fig. 5 spans two parts: (A) architecture and forward map (Sec. VII-A) with MFG statistic b_{MFG} , AHN, allocation, payment, and budget projection; (B) training objective (Sec. VII-B): augmented Lagrangian on MFG-aware regret, plus the optional value critic and payment alignment (Sec. VII-B3).

A. Architecture and Forward Mechanism

Given $\mathbf{b}$, the broker sets $b_{\text{MFG}} = \frac{1}{N} \sum_i b_i$. Three modules apply in sequence; b_{MFG} augments the input Eq. (29), unlike vanilla RegretNet [15].

1) *Module 1: Auction Header Network (AHN)*: Given bid profile $\mathbf{b} = (b_1, \dots, b_N)$, the statistic b_{MFG} , and budget B , the AHN encodes the augmented input

$$\mathbf{a} = [b_1; \dots; b_N; \mathbf{1}_N \otimes b_{\text{MFG}}; B], \quad (29)$$

with b_{MFG} repeated once per client and every field rescaled to $[0, 1]$, into the context embedding

$$\mathbf{C} = (\mathbf{C}^{(a)}, \mathbf{C}^{(p)}) = \text{AHN}_{\theta_h}(\mathbf{a}) \in \mathbb{R}^H \times \mathbb{R}^H, \quad (30)$$

where θ_h are the AHN parameters. The AHN consists of two tanh multilayer perceptrons of width H that read the same input: $\mathbf{C}^{(a)}$ feeds the Allocation Network and $\mathbf{C}^{(p)}$ the Payment Network, so the two heads do not share hidden features.

2) *Module 2: Allocation Network*: The Allocation Network maps $\mathbf{C}^{(a)}$ to a logit $h_{ij}^{(a)}$ for each client i and item $j \in [m]$ together with an abstention logit $h_{i0}^{(a)}$, and normalizes them per client with a Softmax, giving the allocation matrix $\mathbf{Z} \in [0, 1]^{N \times m}$:

$$z_{ij} = \frac{\exp(h_{ij}^{(a)})}{\exp(h_{i0}^{(a)}) + \sum_{j'=1}^m \exp(h_{ij'}^{(a)})}, \quad i \in \mathcal{N}, j \in [m], \quad (31)$$

ensuring $z_{ij} \geq 0$ and $\sum_j z_{ij} \leq 1$ for every client (allocation feasibility: the broker never buys more than a client offers). The privacy released by client i is

$$\epsilon_i^{\text{out}} = \sum_{j=1}^m z_{ij} \epsilon_{ij}; \quad (32)$$

with a single privacy item ($m=1$, as in our experiments), $z_i = \sigma(h_{i1}^{(a)} - h_{i0}^{(a)})$ and $\epsilon_i^{\text{out}} = z_i \epsilon_i$.

3) *Module 3: Payment Network*: The Payment Network maps $\mathbf{C}^{(p)}$ to one logit $h_i^{(p)}$ per client and scales its Sigmoid by the budget:

$$p_i = B \sigma(h_i^{(p)}) = \frac{B}{1 + \exp(-h_i^{(p)})} \in (0, B). \quad (33)$$

The projection $\bar{p}_i = p_i / \max(1, \frac{1}{B} \sum_{i'=1}^N p_{i'})$ then enforces Eq. (8) for every report profile, and the outputs $(z_{ij}, \bar{p}_i)$ fix ϵ_i^{out} for DP-FL.

Payments depend on the reports only through the network, so no bid field enters p_i mechanically; any gain from misreporting v_i , ϵ_i , or d_i is what the **ex-post regret penalty** of Eq. (39) removes during training. The rule is not DSIC by construction (cf. Remark 1), so Sec. VII-C bounds its regret formally and RQ1 measures and certifies it on held-out profiles.

Algorithm 2 summarizes one FL training round at deploy time: FLCs report bids (Sec. V), FLB forms the mean-field statistic Eq. (20), runs the MFG-RegretNet forward map (AHN, allocation, payment), and FLS aggregates DP private gradients weighted by effective privacy budgets.

B. Training Objective: Regret, Value Baseline, and MFG Alignment

The training objective combines (1) $\mathcal{L}_{\text{AL}}$, the augmented Lagrangian on the acquired privacy, IR hinge, and *regret penalties*, with (2) the MFG-aware regret of Eq. (38); two optional terms, (3) a value critic and (4) a payment alignment, complete the objective in Sec. VII-B3.

1) *Augmented-Lagrangian backbone*: Let $\mathcal{L}_{\text{AL}}(\theta, \lambda)$ denote the RegretNet-style Lagrangian. The projection of Module 3 lets the network spend up to B , so the revenue objective of $\mathcal{P}_3$ saturates at $\sum_i \bar{p}_i = B$, and the buyer-side term left is the acquired privacy, which drives the aggregation error of $\mathcal{P}_2$. $\mathcal{L}_{\text{AL}}$ thus combines $-\mathbb{E}[\sum_i d_i \epsilon_i^{\text{out}} / \bar{d}]$ ($\bar{d}$ the mean data size) with the IR violations $[c(v_i, \epsilon_i^{\text{out}}) - \bar{p}_i]^+$ and regrets $\widehat{r}_{gt_i}$ (Eq. (38)), each weighted by a multiplier and divided by $\bar{p}_i$. Multipliers λ_i follow dual ascent with a step ρ that grows across epochs.

Algorithm 2: MFG-Federated Privacy Auction: Online FL Round

Input: Trained $\theta^* = (\theta_h^*, \theta_a^*, \theta_p^*)$; global model $\mathbf{w}_G^{(t-1)}$; bids $\{b_i^{(t)}\}_{i=1}^N$ (Sec. V-B); budget B ; DP parameter δ ; smallest budget $\epsilon_{\min} > 0$ that the local DP mechanism can calibrate

Output: Updated $\mathbf{w}_G^{(t)}$; payments $\{\bar{p}_i^{(t)}\}$

```

1  $b_{\text{MFG}}^{(t)} \leftarrow \frac{1}{N} \sum_{i=1}^N b_i^{(t)}$ ;
2  $\mathbf{C}^{(t)} \leftarrow \text{AHN}_{\theta_h^*}([b_1^{(t)}; \dots; b_N^{(t)}; \mathbf{1}_N \otimes b_{\text{MFG}}^{(t)}; B])$ ;
3  $\mathbf{Z}^{(t)} \leftarrow \text{AllocNet}_{\theta_a^*}(\mathbf{C}^{(t)})$ ;
4  $(p_i^{(t)})_{i=1}^N \leftarrow \text{PayNet}_{\theta_p^*}(\mathbf{C}^{(t)})$  for  $i = 1$  to  $N$  do
5    $\bar{p}_i^{(t)} \leftarrow p_i^{(t)} / \max(1, \frac{1}{B} \sum_{i'=1}^N p_{i'}^{(t)})$ ;
6    $\epsilon_i^{\text{out}} \leftarrow \sum_{j=1}^m z_{ij}^{(t)} \epsilon_{ij}^{(t)}$ ;
7 end
8  $\mathcal{I}^{(t)} \leftarrow \{i : \epsilon_i^{\text{out}} \geq \epsilon_{\min}\}$  (the others do not upload);
9 for  $i \in \mathcal{I}^{(t)}$  do
10   From  $\mathbf{w}_G^{(t-1)}$ , train on  $\mathcal{D}_i$  with a local DP mechanism calibrated to  $(\epsilon_i^{\text{out}}, \delta)$ , e.g., DP-SGD, giving  $\tilde{\mathbf{w}}_i^{(t)}$ ;
11 end
12  $\alpha_i^{(t)} \leftarrow \epsilon_i^{\text{out}} / \sum_{i' \in \mathcal{I}^{(t)}} \epsilon_{i'}^{\text{out}}$  for  $i \in \mathcal{I}^{(t)}$ ;
13  $\mathbf{w}_G^{(t)} \leftarrow \sum_{i \in \mathcal{I}^{(t)}} \alpha_i^{(t)} \tilde{\mathbf{w}}_i^{(t)}$ ;
14 return  $\mathbf{w}_G^{(t)}$ ,  $\{\bar{p}_i^{(t)}\}$ ;
```

2) *MFG-aware ex-post regret*: DSIC is measured by ex-post regret:

$$\text{rgt}_i(\theta; \mathbf{b}) = \max_{b'_i \in \mathcal{B}_i} u_i(b'_i, b_{-i}; \theta) - u_i(b_i, b_{-i}; \theta). \quad (34)$$

Definition 6 (ϵ -Approximate DSIC) The mechanism $\mathcal{M}_\theta$ is ϵ -approximately DSIC in expected ex-post regret if:

$$\mathbb{E}_{\mathbf{b} \sim \mathcal{W}}[\text{rgt}_i(\theta; \mathbf{b})] \leq \epsilon, \quad \forall i \in \mathcal{N}. \quad (35)$$

MFG-Aware Regret Computation. In MFG-RegretNet, when agent i misreports, the mean-field input b_{MFG} updates self-consistently:

$$b_{\text{MFG}}' = b_{\text{MFG}} + \frac{b'_i - b_i}{N}, \quad (36)$$

reflecting the agent's impact on the population-level statistic. We approximate the regret via R steps of projected gradient ascent (PGA):

$$b_i^{(r+1)} = \Pi_{\mathcal{B}_i} \left(b_i^{(r)} + \eta_r \nabla_{b'_i} u_i(b'_i, b_{-i}; \theta) \right), \quad r = 0, \dots, R-1, \quad (37)$$

initialized at the best point of a grid over $\mathcal{B}_i$ in $(v'_i, \epsilon'_i, d'_i)$ that includes $\epsilon'_i=0$ (withholding).

This yields the MFG-aware regret estimator:

$$\begin{aligned} \widehat{\text{rgt}}_i^{\text{MFG}}(\theta) &= \frac{1}{L} \sum_{l=1}^L \left[u_i(b_i^{(R,l)}, b_{-i}^{(l)}; b_{\text{MFG}}^{(R,l)}, \theta) - u_i(b_i^{(l)}, b_{-i}^{(l)}; b_{\text{MFG}}^{(l)}, \theta) \right]^+, \end{aligned} \quad (38)$$

which accounts for both the agent's individual incentive to deviate and the resulting shift in the mean-field statistic.

3) *Optional terms and full objective*: Two optional terms tie training more closely to the MFG structure (Supplementary Sec. S4). A value critic $\hat{\Phi}_\xi(b, \nu)$, fit by minimizing an HJB residual $\mathcal{L}_{\text{HJB}}(\xi)$, turns the regret of Eq. (38) into an advantage form that is mixed in with weight $\zeta \in [0, 1]$; and an

alignment loss $\mathcal{L}_{\text{MFG}}$ pulls the learned payments toward the mean-field payment $\tilde{p}(b; \mu_t)$ of Sec. VI-B. The full objective

$$\mathcal{J}(\theta, \xi) = \mathcal{L}_{\text{AL}}(\theta, \boldsymbol{\lambda}) + \lambda_{\text{MFG}} \mathcal{L}_{\text{MFG}}(\theta) + \lambda_{\text{HJB}} \mathcal{L}_{\text{HJB}}(\xi) \quad (39)$$

is optimized by alternating primal descent on (θ, ξ) and dual ascent on $\boldsymbol{\lambda}$; the reported model sets $\lambda_{\text{MFG}} = \lambda_{\text{HJB}} = \zeta = 0$.

C. Formal Regret and Truthfulness Guarantees

We now bound the incentive properties of the deployed network itself. Client i has type $t_i = (v_i, \epsilon_i, d_i)$ with $v_i \in [0, \bar{v}]$, $\epsilon_i \in [0, \bar{\epsilon}]$, $d_i \in [1, \bar{d}]$, and may report any $b'_i \in \mathcal{B}_i(t_i) = [0, \bar{v}] \times [0, \epsilon_i] \times [1, d_i]$, since caps and data sizes cannot be overstated (Proposition 1). Under parameters θ it receives the projected payment $\bar{p}_{\theta, i}$ and releases $z_{\theta, i} \hat{\epsilon}_i$, so $u_i = \bar{p}_{\theta, i} - v_i z_{\theta, i} \hat{\epsilon}_i \in [-\bar{v} \bar{\epsilon}, B]$; let $\bar{U} = B + \bar{v} \bar{\epsilon}$. The budget projection of Algorithm 2 makes budget feasibility structural: for every θ and every report profile, $0 \leq \bar{p}_{\theta, i} \leq B$ and $\sum_i \bar{p}_{\theta, i} \leq B$. Let $\text{rgt}_i(\theta; \mathbf{t}) \in [0, \bar{U}]$ be Eq. (34) maximized over $\mathcal{B}_i(t_i)$ at the truthful profile $\mathbf{t}$, and $r_i(\theta) = \mathbb{E}_{\mathbf{t} \sim \mathcal{W}}[\text{rgt}_i]$. Because the network recomputes b_{MFG} from the deviating profile, rgt_i is the MFG-aware regret of Eq. (38) with an exact inner maximization. Proofs are in Supplementary Sec. S5.F.

Proposition 2 (From Regret to Truthfulness) If $r_i(\theta) \leq \epsilon$, then (a) for every $\tau > 0$, $\Pr[\text{rgt}_i \geq \tau] \leq \epsilon/\tau$; (b) the interim gain of any misreport against truthful opponents is at most $\mathbb{E}[\text{rgt}_i | t_i]$, so if all clients have $r_i \leq \epsilon$, truthful reporting is an ex-ante ϵ -Bayes-Nash equilibrium; (c) misreporting the valuation, the cap, or both gains at most rgt_i pointwise; and (d) if types are drawn afresh in each round, independently of past outcomes, then, since the network keeps no state across rounds, a strategy that adapts to all past outcomes gains at most $\sum_{\tau \leq T} \text{rgt}_i(\theta; \mathbf{t}^{(\tau)})$ over T rounds against truthful opponents, i.e., $T r_i(\theta)$ in expectation.

Theorem 5 (Uniform Regret Generalization) Let Θ_W contain the networks of Sec. VII-A with n_h hidden tanh layers, d_θ parameters, per-layer induced norms $\|w_\ell\|_1 \leq W$, inputs bounded by X , a per-client Softmax allocation, and a Sigmoid payment head followed by the budget projection of Algorithm 2. For L i.i.d. profiles from $\mathcal{W}$, with probability at least $1 - \delta_c$, for all $\theta \in \Theta_W$,

$$\frac{1}{N} \sum_i r_i(\theta) \leq \hat{r}_L(\theta) + c_1 \bar{U} \sqrt{\frac{\mathcal{C}_\theta}{L}} + c_2 \bar{U} \sqrt{\frac{\log(1/\delta_c)}{L}}, \quad (40)$$

where $\hat{r}_L$ is the average of rgt_i over clients and profiles, $\mathcal{C}_\theta = (n_h + 1)d_\theta \log(LWX d_\theta)$, and c_1, c_2 are absolute constants.

Theorem 5 adapts Dütting et al. [15]: b_{MFG} enters through a parameter-free linear lift of the reports, and the projection is a 2-Lipschitz radial retraction onto an ℓ_1 -ball, so neither changes the form of the bound.

Corollary 2 (Held-Out Regret Certificate) Let $\hat{\theta}$ be independent of L' i.i.d. test profiles $\mathbf{t}^{(1)}, \dots, \mathbf{t}^{(L')}$, and let $\rho_i^{(l)} \in [\text{rgt}_i(\hat{\theta}; \mathbf{t}^{(l)}), \bar{U}]$ be certified per-profile bounds with mean $\bar{r}_i$ and sample variance V_i . With probability at least $1 - \delta_c$, for

all i ,

$$r_i(\hat{\theta}) \leq \bar{r}_i + \sqrt{\frac{2V_i\ell}{L'}} + \frac{7\bar{U}\ell}{3(L'-1)}, \quad \ell = \ln \frac{2N}{\delta_c}. \quad (41)$$

Eq. (41) is the empirical Bernstein inequality [46] with a union bound over clients (for the client average, $\ell = \ln(2/\delta_c)$). It has no capacity term, and since PGA only lower-bounds the inner maximum, $\rho_i^{(l)}$ must come from a certified search, e.g., the linear-relaxation bounds [47] with branch-and-bound used in RQ1 (Supplementary Sec. S1.I).

Proposition 3 (The Mean-Field Channel Is $\mathcal{O}(1/N)$)

Suppose client i 's utility is L_m -Lipschitz in the statistic input of the network (ℓ_1 norm), and D bounds the ℓ_1 -diameter of $\mathcal{B}_i(t_i)$. Holding the statistic at its truthful value changes rgt_i by at most $L_m D/N$. If client types are i.i.d. with coordinate standard deviations σ_k , holding it at the population mean changes $\mathbb{E}[\text{rgt}_i]$ by at most $L_m D/N + 2L_m \sum_k \sigma_k / \sqrt{N}$.

Hence, as long as L_m does not grow with N , the statistic cannot be used for manipulation beyond $\mathcal{O}(1/N)$, and its fluctuation costs $\mathcal{O}(N^{-1/2})$, a rate free of the report dimension (the constant grows with $\sum_k \sigma_k$), because the network reads the population through a mean rather than through the empirical measure.

Large-population behavior. Unlike the statistic channel, the cap rent of Proposition 1 is per client and does not shrink with N , so the payment rule must price it through the regret penalty; and since d_θ grows linearly in N , keeping the capacity term of Eq. (40) fixed requires training profiles linear in N .

VIII. PERFORMANCE EVALUATION

We evaluate MFG-RegretNet through four research questions (RQs):

- 1) *RQ1 (Incentive Security)*: Does MFG-RegretNet keep ex-post regret and IR violations near zero under a binding budget, can its regret be certified, and does this incentive stability persist under coordinated bid manipulation?
- 2) *RQ2 (Dependable Scalability)*: How do latency and memory scale with the client population size N , in plaintext and with the bids protected by secure computation?
- 3) *RQ3 (Privacy-Accounted Auction Efficiency)*: Does MFG-RegretNet achieve higher budget utilization η_{rev} and participant social welfare $\overline{SW}$ while keeping long-horizon privacy exposure interpretable?
- 4) *RQ4 (End-to-End FL Utility)*: After full federated training under a common budget, how does the privacy each mechanism buys translate into model utility on image and privacy-sensitive tabular tasks?

A. Experimental Setup

Data and models. We use MNIST [48] with a CNN of two 5×5 convolutional and two fully connected layers and CIFAR-10 [49] with a lightweight CNN of three convolutional blocks, partitioned among clients by a Dirichlet distribution with $\alpha \in \{0.1, 0.5\}$ (severe and mild non-IID). The CIFAR-10 model is a controlled FL benchmark rather than a state-of-the-art classifier, so the No-DP row of Table IV is the upper bound of

this lightweight setup. The privacy-sensitive Bank marketing and NSL-KDD tasks use a one-hidden-layer MLP.

Mechanisms. RQ1 and RQ3 use $N=10$ clients and one privacy item ($m=1$), as in the RegretNet benchmarks [15]; RQ2 varies $N \in \{10, \dots, 500\}$, and RQ4 uses $N=10$. Profiles draw $v_i \sim \text{Uniform}[0, 1]$, $\epsilon_i \sim \text{Uniform}[0.1, 5]$, and $d_i \sim \text{Uniform}\{10, \dots, 500\}$. Every learned mechanism is trained with five seeds and evaluated on 10,000 held-out profiles (mean $\pm$ std over seeds); the reported MFG-RegretNet uses neither the alignment term nor the critic ($\lambda_{\text{MFG}} = \lambda_{\text{HJB}} = \zeta = 0$). Supplementary Sec. S1.G gives the training hyperparameters, the valuation scenarios (Table S10), and the FL settings (Table S11). The code is available at the public code repository¹.

Baselines. PAC (Algorithm 1), VCG [50] in reverse-procurement form, RegretNet [15] without b_{MFG} , DM-RegretNet [8] (at epoch 150, before its training diverges), MFG-Pricing [19], CSRA [17], and, in RQ4, Lin et al. [10]; Supplementary Sec. S1.G describes each.

1) *Privacy Composition Across Rounds*: Per-round budgets compose across the T_{FL} rounds: basic composition [51] gives $\epsilon_{\text{total}} \leq T\epsilon_i^{\text{out}}$, which Rényi and concentrated DP [36], [52] tighten: the RDP costs $\rho_i^{(t)}(\alpha)$ of a client's releases add up to $\rho_i^{\text{tot}}(\alpha)$ over the rounds in which it participates and convert to an $(\epsilon_i^{\text{tot}}, \delta')$ -DP guarantee (Supplementary Proposition S1). For a round- t Gaussian output perturbation with the classical calibration of Sec. III to $\epsilon_i^{(t), \text{out}}$, which is $(\alpha, \alpha \Delta f^2 / (2(\sigma_i^{(t)})^2))$ -RDP [36], this gives the multi-round guarantee directly in the budgets allocated by the auction:

$$\epsilon_i^{\text{tot}} \leq \Gamma_i + 2\sqrt{\Gamma_i \log(1/\delta')}, \quad \Gamma_i = \frac{\sum_{t \in \mathcal{H}_i} (\epsilon_i^{(t), \text{out}})^2}{4 \ln(1.25/\delta)}, \quad (42)$$

where $\mathcal{H}_i$ is the set of rounds in which client i is selected, δ is the per-round calibration parameter, and δ' is the end-to-end target. Unselected rounds add nothing, so the guarantee grows with the sum of squared allocated budgets: for a constant budget, its leading term grows as $\sqrt{|\mathcal{H}_i|}$ while $\Gamma_i \leq \log(1/\delta')$, rather than linearly; under Laplace noise, basic composition gives $\epsilon_i^{\text{tot}} \leq \sum_{t \in \mathcal{H}_i} \epsilon_i^{(t), \text{out}}$. For DP-SGD, as in RQ4, the same composition is applied to the RDP curve of the subsampled Gaussian mechanism instead. Because noise is added before upload, the guarantee holds against the FL server and every party that observes the updates. It is record-level: it protects each training example of a client, not whether the client participates, and the bids and payments are protected by the broker implementation of Sec. IV-B. The accountant runs beside the per-round auction, with the same settings for every mechanism: after each round it updates $\rho_i^{\text{tot}}(\alpha)$, and before each release it excludes a client whose release would take it above its cap E_i^{tot} , so no cap is exceeded; pricing the remaining privacy inside the auction is left as an extension (Sec. VIII-C).

B. Experimental Results

1) *RQ1: Incentive Compatibility*: We measure approximate DSIC by the ex-post regret of Eq. (34) and IR by the share

¹<https://github.com/szpsunkk/MFG-RegretNet>

TABLE II
RQ1 AT THE BINDING BUDGET $B=25$ (10,000 HELD-OUT PROFILES):
MEAN REGRET (GRID SEARCH PLUS PGA OVER $\mathcal{B}_i(t_i)$), SHARE OF
CLIENTS WITH NEGATIVE UTILITY, AND BUDGET USE; MEAN $\pm$ STD OVER
FIVE TRAINING SEEDS.

Mechanism	Regret (10^{-3})	IR viol.	$\mathcal{R}/B$
PAC / VCG	0	0	0.327
CSRA	261	1.5%	0.734
MFG-Pricing	22.0	6.7%	1.000
DM-RegretNet	62.8 $\pm$ 16.4	7.6%	1.000
MFG-RegretNet (Ours)	2.43$\pm$0.15	0	1.000

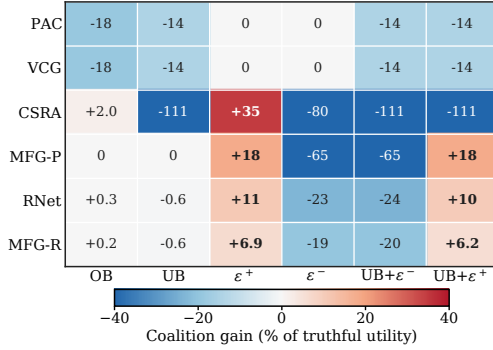


Fig. 6. RQ1: coalition gain under extreme coordinated attacks, in % of the coalition’s truthful utility under the true valuations (pooled over three seeds and three collusion ratios); positive values mean that the coalition profits. OB/UB denote overbidding/underbidding and ϵ^+/ϵ^- cap inflation/deflation; ϵ^+ lies outside the one-sided report set. MFG-P, RNet, and MFG-R abbreviate MFG-Pricing, RegretNet, and MFG-RegretNet.

of clients with negative utility, for all mechanisms on the same 10,000 held-out profiles, with utility $u_i = \bar{p}_i - v_i \epsilon_i^{\text{out}}$ under the true v_i and the report set $\mathcal{B}_i(t_i)$ of Sec. VII-C. Each client’s best deviation is found by a $21 \times 21 \times 5$ grid over $\mathcal{B}_i(t_i)$ plus 25 PGA steps from the best grid point, keeping the better of the two, which finds $3.3 \times$ more regret on MFG-RegretNet than 25 PGA steps from the truthful report; analytical baselines are searched on a $21 \times 21 \times 3$ grid. Table II uses the binding budget $B=25$ (then B/N is below the largest cost $\bar{v}\bar{\epsilon}=5$); we also evaluate the default $B=50$. We observe that: 1) MFG-RegretNet’s regret is $(2.43 \pm 0.15) \times 10^{-3}$, about 0.1% of B/N ($(3.56 \pm 0.50) \times 10^{-3}$, 0.07%, at $B=50$), with zero IR violations and full budget use at both budgets. 2) Its regret is $9 \times$ lower than MFG-Pricing’s, $26 \times$ lower than DM-RegretNet’s, and two orders of magnitude lower than CSRA’s, and it leaves no client with negative utility, whereas MFG-Pricing does so for 6.7%, DM-RegretNet for 7.6%, and CSRA for 1.5%. 3) PAC and VCG are exactly truthful but spend only 33% of the budget (16% at $B=50$). DM-RegretNet, like MFG-Pricing, buys every client’s whole cap. RegretNet (not shown) has regret $(1.75 \pm 0.12) \times 10^{-3}$ and no IR violations but buys 10.2 units of privacy per profile against 13.6 for MFG-RegretNet (21.0 against 23.3 at $B=50$).

Certified regret. The search yields lower bounds; at $B=50$ we also certify upper bounds by linear relaxation [47] with branch-and-bound over $\mathcal{B}_i(t_i)$ (tolerance 0.05). On all 10,000 profiles of a pre-specified seed the certified bound averages 0.051 (median 0.050), 99.9% of client–profile pairs close,

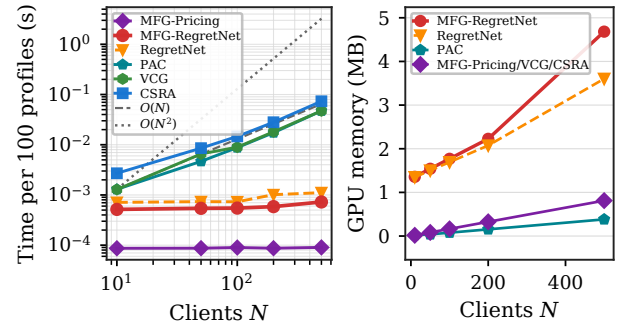


Fig. 7. RQ2: (a) time to clear a batch of 100 auction profiles: one forward pass for the networks, whereas our reference implementations of PAC, VCG, and CSRA process the profiles one at a time; (b) incremental GPU memory of the auction step. MFG-Pricing, VCG, and CSRA have identical memory traces and are merged.

and no lower bound exceeds its certified bound (average 0.051 on 1,000 profiles of each other seed). Corollary 2 then certifies, with probability at least 0.95, an expected regret of at most 0.129 for every client, 2.6% of B/N (0.098, 2.0%, averaged over clients). The certificate concerns the floating-point network and the expected regret over profiles, not every single profile.

Adaptive attackers. The search above knows the network and the whole profile. We also let one client learn its misreport online from its own outcomes: over 20,000 fresh markets at $B=25$, it bins its type into eight contexts, tries each of 70 misreports of valuation, cap, and data size ten times per context, and then commits to the empirically best one. Against MFG-RegretNet the committed policy earns less than truthful reporting (-0.004 to -0.006 per round over the five seeds), and even the best context-dependent misreport in hindsight gains at most 0.0016, below 0.1% of the client’s truthful utility; the same learner gains 0.15 per round against CSRA (8.7%) and 0.067 against DM-RegretNet (7.0%; Supplementary Table S6). Proposition 2(d) extends the per-round bounds to any adaptive strategy over T rounds of such fresh markets.

Coordinated attacks. We further evaluate 972 bid-manipulation configurations (six mechanisms, three seeds, three collusion ratios, six attack modes, three magnitudes) by the coalition’s gain, the change in its members’ utility under their true valuations; Fig. 6 reports the extreme magnitude and Supplementary Tables S4 and S5 the gains and the changes of the winner set. Within their caps, coalitions of up to half of the clients gain at most 0.5% of their truthful utility under MFG-RegretNet over all magnitudes and ratios, against 0.9% under RegretNet, 7.6% under CSRA, and 11.8% under PAC/VCG. Overstating the caps, which the one-sided report set excludes because the release then exceeds them, gains 6.9% at the extreme magnitude, the least among the mechanisms whose release follows the caps (RegretNet 11.0%, MFG-Pricing 18.1%, CSRA 34.9%).

2) *RQ2: Scalability Analysis:* Fig. 7 reports the time to clear a batch of 100 auction profiles (a) and the incremental GPU memory (b), excluding local training, for $N \in \{10, 50, 100, 200, 500\}$ on the same hardware. We observe that: 1) MFG-RegretNet clears the batch in one forward pass

TABLE III

RQ3: BUDGET USE, LARGEST IR VIOLATION RATE, AND WELFARE $\overline{SW}$ OVER THE BUDGET SWEEP (10,000 ROUNDS PER BUDGET; MFG-REGRETNET STD AT MOST 0.11 OVER FIVE SEEDS).

Method	$\mathcal{R}/B$	IR viol.	$\overline{SW}$ at budget B			
			25	50	75	100
PAC / VCG	0.08–0.33	0	4.08	4.08	4.08	4.08
CSRA	0.51–0.73	2.6%	17.14	29.56	39.56	47.50
MFG-Pricing	1.00	6.7%	12.23	37.23	62.23	87.23
DM-RegretNet	1.00	7.6%	12.23	37.23	62.23	87.23
MFG-RegretNet (Ours)	1.00	0	18.21	38.34	62.39	87.28

in about 5×10^{-4} – 7×10^{-4} s from $N=10$ to $N=500$, so its throughput stays nearly flat in N ; for a single auction, every mechanism, including MFG-RegretNet (about 0.2 ms on the GPU), needs under 2 ms up to $N=500$, which is negligible next to local training. 2) MFG-RegretNet and RegretNet stay below 5 MB of incremental GPU memory at $N=500$, against sub-MB to about 1 MB for the analytical baselines. 3) *Cost of protecting the bids.* We run one round under two-party secure computation with CrypTen [53] (semi-honest, online phase, loopback; secret bids, public weights; Supplementary Table S13). A secure MFG-RegretNet round takes 0.22–0.27 s for $N=10$ –500 with a constant 280 communication rounds and 0.62–1.38 MB, matching the plaintext outputs within 0.003 (allocation) and 0.14% of B (payment), whereas the oblivious sort alone that PAC, VCG, and CSRA need takes 0.09–0.39 s (91–406 rounds, 0.04–5.35 MB), exceeding the whole MFG-RegretNet round from $N=100$ on. A TEE broker adds 0.03–0.46 ms for bid decryption (AES-GCM) and outcome signing, plus at most about 14,000 cycles per enclave transition [54] (in-enclave execution not measured; our CPU lacks SGX).

3) *RQ3: Auction Efficiency:* We measure (i) *auction revenue* $\mathcal{R}$ (total payment per round), (ii) *budget utilization* $\eta_{\text{rev}} = \mathcal{R}/B$ and the budget-feasibility rate $\text{BF} = \frac{1}{T} \sum_t \mathbf{1}[\mathcal{R}^{(t)} \leq B]$, and (iii) *participant social welfare* $\overline{SW}$ (client surplus after privacy costs, Eq. (5)). For each budget $B \in \{25, 50, 75, 100\}$ (multipliers 0.5–2 of the default), every mechanism runs one auction round on each of the 10,000 held-out bid profiles of RQ1 with truthful reports, and the privacy costs use the clients’ *true* valuations v_i . Every mechanism attains $\text{BF}=1$ (up to float32 rounding); for MFG-RegretNet this follows from the projection of Module 3. Table III summarizes the results, from which we observe that: 1) MFG-RegretNet spends the whole budget at every B with zero IR violations, whereas PAC/VCG use 8–33% and CSRA 51–73%. 2) It yields higher welfare than MFG-Pricing, DM-RegretNet, CSRA, and PAC/VCG at every budget and in each training seed; the gain over MFG-Pricing is largest when the budget binds (18.21 vs. 12.23, +48.8%, at $B=25$, where MFG-Pricing leaves 6.7% of the clients with negative utility), and over CSRA it grows from 6.3% to 83.7%. 3) MFG-RegretNet buys selectively when the budget binds (13.62 units at $B=25$) and 99.7% of the offered privacy at $B=100$. RegretNet, not listed, has slightly higher welfare (19.9 at $B=25$) because it buys less privacy (10.2 units against 13.6): with the budget spent, $\overline{SW} = B - \sum_i v_i \epsilon_i^{\text{out}}$, so welfare also rises when less privacy is bought.

Auction efficiency must also be read against the privacy con-

sumed over repeated rounds. Supplementary Figs. S1 and S2 and Tables S2 and S3 report the cumulative ϵ over 200 rounds under Laplace composition and Gaussian RDP, with the pre-release check of Sec. VIII-A1 at $E_i^{\text{tot}}=60$: the mechanisms that spend the whole budget (MFG-Pricing, RegretNet, and MFG-RegretNet) release the most per round, and their clients reach $\epsilon=10$ after 4.5–5.2 rounds on average under Laplace composition (9.5–13.4 under RDP); PAC/VCG and CSRA release less, and no client exceeds its cap. Strict long-horizon caps therefore call for an intertemporal privacy state.

4) *RQ4: Final Global Model Accuracy:* Finally, we evaluate how the privacy that each mechanism buys under the same budget translates into downstream FL accuracy. For every mechanism we run the PAG-FL loop (Algorithm 2) for $T_{\text{FL}}=100$ rounds with $N=10$ clients and $B=50$ per round (Supplementary Table S11); within a seed, all mechanisms share the data partition, the bid draws, the initial model, and the noise streams. On MNIST and CIFAR-10, each selected client runs record-level DP-SGD whose noise is calibrated so that its participation is $(\epsilon_i^{\text{out}}, \delta)$ -DP, and its participations are composed by RDP accounting (Sec. VIII-A1). Lin et al. [10] enters through the menu adaptation of Supplementary Sec. S1.H, Uniform-DP gives every client the mean ϵ_i^{out} of MFG-RegretNet without an auction, and No-DP trains without noise. We report the final accuracy $\mathcal{A}_{\text{final}}$, the held-out test accuracy averaged over the last five rounds and over three seeds. Table IV lists $\mathcal{A}_{\text{final}}$ together with the privacy each mechanism releases and its payment, and Supplementary Figs. S5 and S6 show the convergence curves. We observe that: 1) All mechanisms converge within 100 rounds. On MNIST, MFG-RegretNet reaches 94.7% ($\alpha=0.5$) and 89.9% ($\alpha=0.1$), above Lin et al. (+0.40 and +1.43 points), PAC/VCG (+0.80, +1.14), and CSRA (+2.59, +7.23) in every seed, and within 0.2 points of RegretNet, DM-RegretNet, and MFG-Pricing, which buy 2.1–2.6 units of privacy per client and round; a separate five-seed paired comparison with Lin et al. (Supplementary Table S12) finds it more accurate on average in all four settings and in every seed in three. 2) On MNIST, accuracy follows the privacy bought. MFG-RegretNet spends the budget and releases $\bar{\epsilon}=2.34$ per client and round, against 1.44 for Lin et al., 0.90 for PAC/VCG, and 0.52 for CSRA; after 100 rounds the median client’s cumulative ϵ is 14.2–14.8, against 9.2–10.0, 5.6–6.3, and 4.0–4.6. The mechanisms therefore sit at different points of the privacy–utility trade-off, which the buyer sets through B . 3) On CIFAR-10 the lightweight CNN is noise-limited: No-DP reaches 65.6% ($\alpha=0.5$) and 58.5% ($\alpha=0.1$), whereas all auction mechanisms, including MFG-RegretNet (34.7%, 27.3%), lie within 1.4 points of each other, so additional privacy buys no measurable accuracy at this model capacity; with ResNet-18, No-DP training reaches $85.28\% \pm 0.09\%$ (Supplementary Fig. S3 and Table S1).

On the privacy-sensitive tabular tasks (Supplementary Sec. S1.E), MFG-RegretNet has the highest AUC among the DP mechanisms on Bank (0.703, against 0.47–0.59), whereas on the imbalanced five-class NSL-KDD all-winner allocation is a weak point (L3 in Sec. VIII-C).

TABLE IV

RQ4: FINAL TEST ACCURACY $\mathcal{A}_{\text{final}}$ (% , MEAN $\pm$ STD OVER THREE SEEDS) AFTER 100 ROUNDS OF DP-SGD FL WITH $N=10$ AND $B=50$, WITH THE MEAN ϵ RELEASED PER CLIENT AND ROUND ($\bar{\epsilon}$), THE MEDIAN CLIENT’S CUMULATIVE ϵ AFTER 100 ROUNDS (RDP, $\delta=10^{-5}$; RANGE OVER THE FOUR SETTINGS), AND THE PAYMENT PER ROUND. VCG SELECTS THE SAME CLIENTS AS PAC IN EVERY ROUND; UNIFORM-DP GIVES EVERY CLIENT MFG-REGRETNET’S MEAN ϵ WITHOUT AN AUCTION.

Method	MNIST		CIFAR-10		Privacy and payment		
	$\alpha=0.5$	$\alpha=0.1$	$\alpha=0.5$	$\alpha=0.1$	$\bar{\epsilon}$ per round	Cumulative ϵ	Payment
PAC / VCG [50]	93.85 $\pm$ 0.65	88.71 $\pm$ 1.81	35.58 $\pm$ 2.19	27.80 $\pm$ 2.18	0.90	5.6–6.3	8.2
CSRA [17]	92.06 $\pm$ 0.65	82.62 $\pm$ 2.24	35.74 $\pm$ 0.92	27.51 $\pm$ 1.54	0.52	4.0–4.6	31.8
MFG-Pricing [19]	94.71 $\pm$ 0.77	89.96 $\pm$ 1.04	34.69 $\pm$ 2.84	27.27 $\pm$ 1.28	2.55	15.5–16.3	50.0
RegretNet [15]	94.60 $\pm$ 0.81	89.69 $\pm$ 1.18	34.69 $\pm$ 2.84	27.19 $\pm$ 1.37	2.10	12.7–13.4	50.0
DM-RegretNet [8]	94.71 $\pm$ 0.77	89.96 $\pm$ 1.04	34.70 $\pm$ 2.84	27.33 $\pm$ 1.34	2.55	15.5–16.3	50.0
Lin et al. [10]	94.25 $\pm$ 0.90	88.43 $\pm$ 2.22	34.53 $\pm$ 3.11	26.41 $\pm$ 0.99	1.44	9.2–10.0	12.7
Uniform-DP (no auction)	94.77 $\pm$ 0.65	90.47 $\pm$ 1.05	35.60 $\pm$ 2.64	28.77 $\pm$ 1.93	2.34	11.8–13.1	0.0
No-DP FL (no noise)	99.22 $\pm$ 0.05	98.97 $\pm$ 0.06	65.64 $\pm$ 0.87	58.49 $\pm$ 3.16	–	–	–
MFG-RegretNet (Ours)	94.65 $\pm$ 0.80	89.86 $\pm$ 1.10	34.68 $\pm$ 2.82	27.28 $\pm$ 1.39	2.34	14.2–14.8	50.0

C. Discussion of Results and Limitations

Experiments validate three points: low regret and coordinated manipulation robustness (RQ1), near-linear runtime and memory up to $N=500$ with secure rounds under 0.3 s (RQ2), and strong auction efficiency with explicit cumulative privacy accounting (RQ3). Downstream accuracy follows the privacy each mechanism buys, with MFG-RegretNet level with or above the auction baselines on MNIST and the lightweight CIFAR-10 model noise-limited, and privacy-sensitive Bank/NSL-KDD results clarify where all-winner privacy allocation helps or fails (RQ4). *Scale and heterogeneity.* PAC follows its mean-field limit up to $N=10^4$ (Fig. 4), and RQ4 covers non-IID and tabular data. The networks trained on uniform valuations also run, without retraining, the three bimodal and realistic settings of Supplementary Table S10, with results in Supplementary Table S7 (5,000 fresh profiles, RQ1 search; realistic valuations in units of their 99th percentile, top 1% capped, B rescaled alike). In each, MFG-RegretNet keeps regret at 0.09–0.11% of B/N with no IR violations and full budget use, whereas DM-RegretNet’s regret is 0.11–0.25% of B/N , CSRA’s is 2.7–4.8% with 2.3–13.9% of the clients below zero utility, and PAC/VCG spend 8–20% of the budget. Five limitations remain. *(L1) Dynamic privacy:* The RDP accountant of Sec. VIII-A1 and Supplementary Tables S2 and S3 track cumulative privacy exposure, but the current PAG still treats each FL round as an independent auction primitive; a fully dynamic PAG should include the remaining privacy budget as a mechanism state and price future budget scarcity explicitly. *(L2) Trust and strategic security:* the formal mechanism assumes a semi-trusted AB. RQ2 measures a semi-honest two-party SMPC broker and the channel cost of a TEE broker; malicious-broker robustness and collusion resistance beyond the tested attack modes require audit mechanisms or formal coalition-proof extensions. *(L3) Adaptive allocation on difficult FL tasks:* the privacy-sensitive NSL-KDD results show that all-winner allocation can be suboptimal under highly imbalanced multi-class data; adaptive winner-set sizing or task-aware client filtering remains future work. *(L4) Fairness across privacy sensitivities:* high-sensitivity clients may be under-served under skewed valuation distributions. Service floors, compensation floors, or group-regularized regret can

mitigate this issue without changing the core privacy-trading model. *(L5) Simulation:* all results are single-machine simulations with synthetic valuations; a multi-node deployment with elicited valuations remains future work.

IX. CONCLUSION

This paper develops a privacy-trading framework for federated learning incentive design that treats the personalized privacy budget ϵ_i as a *tradable commodity* rather than a fixed system constraint. We formalized the Privacy Auction Game with DSIC, IR, and budget feasibility guarantees for the analytical PAC mechanism (Theorem 1) and a characterization of truthful two-dimensional reporting (Proposition 1), and used a mean-field approximation in which each client responds to the population law; the analytical mechanism converges to its closed-form mean-field limit at rate $\mathcal{O}_P(N^{-1/2})$ (Theorem 4). Built on this foundation, MFG-RegretNet learns allocation and payment rules without closed-form valuation priors via augmented Lagrangian optimization conditioned on the mean-field statistic. Experiments on MNIST and CIFAR-10 show that MFG-RegretNet achieves low empirical ex-post regret, certified on held-out profiles, and strong auction efficiency while remaining competitive in downstream FL model accuracy compared with VCG, PAC, RegretNet, and MFG-Pricing baselines. These learned incentive properties are approximate rather than exact PAC-style DSIC guarantees, and Sec. VII-C bounds them through uniform generalization, a held-out regret certificate, and an $\mathcal{O}(1/N)$ bound on manipulation through the mean-field statistic.

REFERENCES

- [1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. PMLR, Apr. 2017, pp. 1273–1282.
- [2] J. Wu and W. Zhang, “On the security of verifiable and oblivious secure aggregation for privacy-preserving federated learning,” *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 5, pp. 4324–4326, 2024.
- [3] K. Gu, W. Zhao, J. Tan, Z. Yang, X. Li, and W. Jia, “Two-dimensional privacy-preserving federated learning scheme against poisoning attacks,” *IEEE Transactions on Dependable and Secure Computing*, 2026.

- [4] J. Wu, F. Luo, T. Sun, H. Wang, and W. Zhang, "Privacy-preserving federated learning scheme with mitigating model poisoning attacks: Vulnerabilities and countermeasures," *IEEE Transactions on Dependable and Secure Computing*, 2025.
- [5] H. Yang, M. Ge, D. Xue, K. Xiang, H. Li, and R. Lu, "Gradient leakage attacks in federated learning: Research frontiers, taxonomy and future directions," *IEEE Network*, 2023.
- [6] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," *arXiv preprint arXiv:1712.07557*, 2017.
- [7] Q. Li, Z. Liu, Q. Li, and K. Xu, "martFL: Enabling Utility-Driven Data Marketplace with a Robust and Verifiable Federated Learning Architecture," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*. ACM, Nov. 2023, pp. 1496–1510.
- [8] S. Zheng, Y. Cao, M. Yoshikawa, H. Li, and Q. Yan, "FI-market: Trading private models in federated learning," in *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 2022, pp. 1525–1534.
- [9] T. Mai, H. Yao, J. Li, C. Sang, S. Hu, and S. Guo, "Automatic Double-Auction Mechanism for Federated Learning Service Market in Internet of Things," *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 5, pp. 3123–3135, Sept. 2022.
- [10] X. Lin, J. Wu, J. Li, C. Sang, S. Hu, and M. J. Deen, "Heterogeneous differential-private federated learning: Trading privacy for utility truthfully," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 6, pp. 5113–5129, 2023.
- [11] X. Wu, L. Zhang, H. Zhou, J. Tao, Y. Gu, and S. Shen, "Grouping competition-based privacy allocation mechanism for federated learning," *Knowledge-Based Systems*, vol. 332, p. 114906, 2026.
- [12] K. Daskalakis, *The complexity of Nash equilibria*. University of California, Berkeley, 2008.
- [13] R. B. Myerson, "Optimal auction design," *Mathematics of operations research*, vol. 6, no. 1, pp. 58–73, 1981.
- [14] A. Ghosh and A. Roth, "Selling privacy at auction," in *Proceedings of the 12th ACM conference on Electronic commerce*, 2011, pp. 199–208.
- [15] P. Dütting, Z. Feng, H. Narasimhan, D. C. Parkes, and S. S. Ravindranath, "Optimal Auctions through Deep Learning: Advances in Differentiable Economics," *Journal of the ACM*, vol. 71, no. 1, pp. 1–53, Feb. 2024.
- [16] L. K. Fleischer and Y.-H. Lyu, "Approximately optimal auctions for selling privacy when costs are correlated with data," in *Proceedings of the 13th ACM conference on electronic commerce*, 2012, pp. 568–585.
- [17] Y. Yang, M. Hu, Y. Zhou, X. Liu, and D. Wu, "CSRA: Robust Incentive Mechanism Design for Differentially Private Federated Learning," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 892–906, 2024.
- [18] K. Sun, J. Wu, and J. Li, "Reputation-aware incentive mechanism of federated learning: A mean field game approach," in *2024 9th IEEE International Conference on Smart Cloud (SmartCloud)*. IEEE, 2024, pp. 48–53.
- [19] J. Hu-Bolz, M. Reed, K. Zhang, Z. Liu, and J. Hu, "Federated data acquisition market: Architecture and a mean-field based data pricing strategy," *High-Confidence Computing*, vol. 5, no. 1, p. 100232, 2025.
- [20] S. Jiang, Y. Liu, Z. Wang, Y. Li, H. Chi, and X. Du, "Hybrid makes better: Hybrid differential privacy medical image classification based on federated learning," *IEEE Transactions on Dependable and Secure Computing*, 2025.
- [21] C. Zhang, J. Weng, J. Weng, Y. Zhong, J.-N. Liu, and C. Deng, "Robust and secure federated learning with verifiable differential privacy," *IEEE Transactions on Dependable and Secure Computing*, 2025.
- [22] Y. Zhou, X. Liu, Y. Fu, D. Wu, J. H. Wang, and S. Yu, "Optimizing the numbers of queries and replies in convex federated learning with differential privacy," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 6, pp. 4823–4837, 2023.
- [23] R. Xu, B. Li, C. Li, J. B. Joshi, S. Ma, and J. Li, "Tapfed: Threshold secure aggregation for privacy-preserving federated learning," *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 5, pp. 4309–4323, 2024.
- [24] Z. Yi, Y. Jiao, W. Dai, G. Li, H. Wang, and Y. Xu, "A Stackelberg Incentive Mechanism for Wireless Federated Learning With Differential Privacy," *IEEE WIRELESS COMMUNICATIONS LETTERS*, vol. 11, no. 9, 2022.
- [25] L. Xie, Z. Su, Y. Wang, N. Chen, Y. Liu, R. Wang, X. Liu, D. Liu, and H. Zhang, "A privacy-preserving incentive scheme for uav-aided federated learning: A contract method with prospect theory," *IEEE Transactions on Dependable and Secure Computing*, 2025.
- [26] M. B. Singh, H. Singh, and A. Pratap, "Stable Matching based Revenue Maximization for Federated Learning in UAV-Assisted WBANs," *IEEE Transactions on Services Computing*, pp. 1–12, 2024.
- [27] Y. Zhan, P. Li, Z. Qu, D. Zeng, and S. Guo, "A learning-based incentive mechanism for federated learning," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6360–6368, 2020.
- [28] X. Tang, H. Yu, and X. Li, "A reinforcement learning-based bidding strategy for data consumers in auction-based federated learning," *Advances in Neural Information Processing Systems*, vol. 38, pp. 82 666–82 684, 2026.
- [29] A. Ghosh, K. Ligett, A. Roth, and G. Schoenebeck, "Buying private data without verification," in *Proceedings of the fifteenth ACM conference on Economics and computation*, 2014, pp. 931–948.
- [30] T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Privacy-aware federated learning with heterogeneous data," in *International Conference on Learning Representations*, 2022.
- [31] M. Noble, A. Bellet, and A. Dieuleveut, "Differentially private federated learning on heterogeneous data," in *International conference on artificial intelligence and statistics*. PMLR, 2022, pp. 10 110–10 145.
- [32] N. Ding, Y. Guo, N. Chakraborty, and X. Ying, "Differentially private and incentive compatible recommendation," in *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, 2021, pp. 382–391.
- [33] Z. Jorgensen, T. Yu, and G. Cormode, "Conservative or liberal? personalized differential privacy," in *2015 IEEE 31st international conference on data engineering*. IEEE, 2015, pp. 1023–1034.
- [34] Y. Hu, Y. Wang, J. Lou, W. Liang, R. Wu, W. Wang, X. Li, J. Liu, and Z. Qin, "Privacy risks of federated knowledge graph embedding: New membership inference attacks and personalized differential privacy defense," *IEEE Transactions on Dependable and Secure Computing*, vol. 22, no. 3, pp. 2788–2805, 2024.
- [35] C. Dwork, A. Roth, et al., "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.
- [36] I. Mironov, "Rényi differential privacy," in *Proc. IEEE 30th Computer Security Foundations Symposium (CSF)*, 2017, pp. 263–275.
- [37] M. Zhang, F. Beltran, and J. Liu, "Selling Data at an Auction under Privacy Constraints," in *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*. PMLR, Aug. 2020, pp. 669–678.
- [38] A. Acquisti, L. K. John, and G. Loewenstein, "What is privacy worth?" *The Journal of Legal Studies*, vol. 42, no. 2, pp. 249–274, 2013.
- [39] W. Cheng, Y. Zou, J. Xu, and W. Liu, "Dynamic games for social model training service market via federated learning approach," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 1, pp. 64–75, 2021.
- [40] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.
- [41] J.-M. Lasry and P.-L. Lions, "Mean field games," *Japanese Journal of Mathematics*, vol. 2, no. 1, pp. 229–260, 2007.
- [42] M. Huang, R. P. Malhamé, and P. E. Caines, "Large population stochastic dynamic games," *Communications in Information and Systems*, vol. 6, no. 3, pp. 221–252, 2006.
- [43] N. Fournier and A. Guillin, "On the rate of convergence in wasserstein distance of the empirical measure," *Probability Theory and Related Fields*, vol. 162, no. 3, pp. 707–738, 2015.
- [44] M. Ajtai, J. Komlós, and G. Tusnády, "On optimal matchings," *Combinatorica*, vol. 4, no. 4, pp. 259–264, 1984.
- [45] P. Massart, "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality," *The Annals of Probability*, vol. 18, no. 3, pp. 1269–1283, 1990.
- [46] A. Maurer and M. Pontil, "Empirical Bernstein bounds and sample variance penalization," in *Proc. 22nd Annu. Conf. Learn. Theory (COLT)*, 2009.
- [47] H. Zhang, T.-W. Weng, P.-Y. Chen, C.-J. Hsieh, and L. Daniel, "Efficient neural network robustness certification with general activation functions," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 31, 2018.
- [48] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [49] A. Krizhevsky, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009, technical Report.
- [50] W. Vickrey, "Counterspeculation, auctions, and competitive sealed bids," *The Journal of Finance*, vol. 16, no. 1, pp. 8–37, 1961.
- [51] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy," in *International conference on machine learning*. PMLR, 2015, pp. 1376–1385.

- [52] M. Bun and T. Steinke, “Concentrated differential privacy: Simplifications, extensions, and lower bounds,” in *Theory of cryptography conference*. Springer, 2016, pp. 635–658.
- [53] B. Knott, S. Venkataraman, A. Hannun, S. Sengupta, M. Ibrahim, and L. van der Maaten, “CrypTen: Secure multi-party computation meets machine learning,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021.
- [54] P. Yuhala, M. Paper, T. Zerbib, P. Felber, V. Schiavoni, and A. Tchana, “SGX switchless calls made configless,” in *Proc. 53rd Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. (DSN)*, 2023.