Harnessing the Power of Federated Learning in Federated Contextual Bandits

Chengshuai Shi
University of Virginia
cs7ync@virginia.edu

Ruida Zhou
University of California, Los Angeles
ruida@g.ucla.edu

Kun Yang
University of Virginia
ky9tc@virginia.edu

Cong Shen
University of Virginia
cong@virginia.edu

Abstract

Federated contextual bandits (FCB), as a pivotal instance of combining federated learning (FL) and sequential decision-making, have received growing interest in recent years. However, existing FCB designs often adopt FL protocols tailored for specific settings, deviating from the canonical FL framework (e.g., the celebrated FedAvg design). Such disconnections not only prohibit these designs from flexibly leveraging canonical FL algorithmic approaches but also set considerable barriers for FCB to incorporate growing studies on FL attributes such as robustness and privacy. To promote a closer relationship between FL and FCB, we propose a novel FCB design, FedIGW, which can flexibly incorporate both existing and future FL protocols and thus is capable of harnessing the full spectrum of FL advances.

1 Introduction

Federated learning (FL), since proposed in [1, 2], has received great attention due to its attractive features in handling distributed multi-agent machine learning [3, 4]. With the popularity of FL, many attempts have been made to generalize it beyond the original focus of supervised learning to other learning paradigms [5–8]. Among these attempts, the study of federated decision-making is one particularly promising direction [9, 10]. As one pivotal instance of federated decision-making, federated contextual bandits (FCB) [11–20] have gained significant interest recently, which has found broad practical applications (e.g., in cognitive radio and recommender systems).

However, the existing FCB designs [11–20] mostly adopt tailored FL protocols for their specific settings, which often deviate from the canonical FL framework [1, 2] (see Table 1 for a brief summary). In particular, most of them perform *one-shot aggregation* of *compressed local data* per epoch (e.g., combining local covariance matrices). Such choices are rare (and even undesirable) in canonical FL designs, where agents typically communicate and aggregate their *model parameters* (e.g., gradients) for *multiple rounds*. Due to such disconnections, these designs cannot

Table 1: A brief summary of FCB designs; a comprehensive illustration can be found in Appendix A

Ref.	Setting	FL	CB
[11]	Tabular	Mean Averaging	AE
[11, 12]	Linear	Linear Regression	AE
[11, 13–16]	Linear	Ridge Regression	UCB
[17]	Gen. Lin.	Distributed AGD	UCB
[18, 19]	Kernel	Nyström approx.	UCB
[20]	Neural	NTK approx.	UCB
FedIGW	Realizable	Flexible	IGW

AE: arm elimination; UCB: upper confidence bound; Gen. Linear: generalized linear model; AGD: accelerated gradient descent; NTK: neural tangent kernel; IGW: inverse gap weighting

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Multi-Agent Security Workshop.

effectively leverage advances in FL studies, ranging from basic algorithmic approaches to appealing additional guarantees (e.g., privacy, robustness, and beyond).

Motivated by this disconnection from FCB to FL, this work proposes a novel FCB design, termed FedIGW, where inverse gap weighting [21], a regression-based CB strategy, is adopted for contextual bandits (CB) while flexible FL routines can be incorporated. With the flexible choice of FL, FedIGW not only accommodates a wide range of basic FL protocols but also allows for great extensibility (e.g., to privacy and robustness) as both existing and future advances from FL can be effectively leveraged.

2 Problem Formulation

Agents. A total of M agents simultaneously participate in a contextual bandit (CB) system for T time steps. At each time step t, for each agent m, the environment samples a context $x_{m,t} \in \mathcal{X}_m$ and a context-dependent reward vector $r_{m,t} \in [0,1]^{\mathcal{A}_m}$ according to a fixed but unknown distribution $\mathcal{D}_m$, where $\mathcal{A}_m$ is the action set of agent m with size $|\mathcal{A}_m| = K_m$. Then, the agent m observes the context $x_{m,t}$, selects an action $a_{m,t}$ from $\mathcal{A}_{m,t}$, and then receives the associated reward $r_{m,t}(a_{m,t})$ as in the standard CB [22]. The expected reward of playing action a_m facing context x_m is denoted as $\mu_m(x_m, a_m) := \mathbb{E}[r_{m,t}(a_m)|x_{m,t} = x_m]$. The agents gradually learn their optimal policies, denoted as $\pi^*_m(x_m) := \arg\max_{a_m \in \mathcal{A}_m} \mu_m(x_m, a_m)$ for agent m with context x_m.

Federation. In federated learning, it is commonly considered that there exists a central server in the system, and the agents can share information with the server, which can then broadcast aggregated information back to the agents. The later discussions in this work also follow this scenario, while we note that the proposed FedIGW design can be effectively extended to handle general (instead of star-shaped) communication graphs such as in [23].

Realizable Rewards. Moreover, to capture the common interests motivating collaboration among agents, we initiate this work by considering the agents' expected reward functions are globally shared and are within a function class $\mathcal{F}$, to which the agents have access. This assumption, rigorously stated in the following, is often referred to as the *realizability* assumption.

Assumption 2.1 (Realizability)**.** *There exists f^* in $\mathcal{F}$ such that $f^*(x_m, a_m) = \mu_m(x_m, a_m)$ for all $m \in [M]$, $x_m \in \mathcal{X}_m$ and $a_m \in \mathcal{A}_m$.*

This assumption is a natural extension from its commonly adopted single-agent version [21, 24–26] to a federated one, and it incorporates many previously studied FCB scenarios as special cases. For example, the federated linear bandits [12–16] are with a linear function class $\mathcal{F}$.

Algorithm 1 FedIGW (Agent m)

Input: epoch number $l = 1$, reward function $\widehat{f}^l_m(\cdot,\cdot) = 0$, local dataset $\mathcal{S}^l_m = \emptyset$

for time step $t = 1, 2, \cdots$ **do**

  observe context $x_{m,t}$ ▷ *CB: IGW*

  compute $\widehat{a}^*_m = \arg\max_{a_m \in \mathcal{A}_m} \widehat{f}^l(a_m, x_{m,t})$ and set action selection distribution as

$$p^l_m(a_m|x_{m,t}) \leftarrow \begin{cases} 1/\left(K_m + \gamma^l\left(\widehat{f}^l(\widehat{a}^*_m, x_{m,t}) - \widehat{f}^l(a_m, x_{m,t})\right)\right) & \text{if } a_m \neq \widehat{a}^*_m \\ 1 - \sum_{a'_m \neq \widehat{a}^*_m} p^l_m(a'_m|x_{m,t}) & \text{if } a_m = \widehat{a}^*_m \end{cases}$$

  select action $a_{m,t} \sim p^l_m(\cdot|x_{m,t})$; observe reward $r_{m,t}(a_{m,t})$

  update the local dataset, $\mathcal{S}^l_m \leftarrow \mathcal{S}^l_m \cup \{(x_{m,t}, a_{m,t}, r_{m,t}(a_{m,t}))\}$

  if $t = \tau^l$ **then** ▷ *FL*

    perform FL $\widehat{f}^{l+1} \leftarrow \texttt{FLroutine}_m(\mathcal{S}^l_m)$

    update dataset $\mathcal{S}^{l+1}_m \leftarrow \emptyset$; update epoch $l \leftarrow l + 1$

  end if

end for

3 Algorithm Design

In this section, we present a novel FCB design, termed FedIGW. In particular, FedIGW proceeds in epochs, as illustrated in Algorithm 1, which are separated at time slots $\tau^1, \tau^2, \cdots$ w.r.t. the global

time step t, i.e., the l-th epoch starts from $t = \tau^{l-1} + 1$ and ends at $t = \tau^l$, and the overall number of epochs is denoted as $l(T)$. In each epoch l, we describe the FL and CB components, respectively, as follows, while emphasizing on how they are compatible yet decoupled, which thus enables the incorporation of flexible FL choices.

CB: Inverse Gap Weighting (IGW). For CB, we use the method of inverse gap weighting [27], which has received growing interest in the single-agent setting recently [21, 28–30] but has not been fully investigated in the federated setting. At any time step in epoch l, when encountering the context x_m, agent m first estimates the optimal arm by $\widehat{a}^*_m = \arg\max_{a_m \in \mathcal{A}_m} \widehat{f}^l(x_m, a_m)$ from an estimated function $\widehat{f}^l$ (provided by the to-be-discussed FL). Then, she randomly selects her action a_m according to the following distribution, which is inversely proportional to each action's estimated reward gap from the estimated optimal action $\widehat{a}^*_m$:

$$p^l_m(a_m|x_m) \leftarrow \begin{cases} 1/\left(K_m + \gamma^l \left(\widehat{f}^l(\widehat{a}^*_m, x_m) - \widehat{f}^l(a_m, x_m)\right)\right) & \text{if } a_m \neq \widehat{a}^*_m \\ 1 - \sum_{a'_m \neq \widehat{a}^*_m} p^l_m(a'_m|x_m) & \text{if } a_m = \widehat{a}^*_m \end{cases},$$

where γ^l is the learning rate in epoch l that controls the exploration-exploitation tradeoff.

FL: Flexible Designs. By IGW, all agents perform stochastic arm sampling, and thus each agent m collects a set of data samples $\mathcal{S}^l_m := \{(x_{m,t}, a_{m,t}, r_{m,t} : t \in [\tau^{l-1} + 1, \tau^l])\}$ in epoch l. In order to enhance the CB interactions with IGW in the subsequent epoch $l + 1$, an improved estimate $\widehat{f}^{l+1}$ based on all agents' data is desired. This objective aligns precisely with the aim of standard FL, which aggregates local models for better global estimates [1, 2].

With this match, the agents can perform a standard FL protocol (e.g., FedAvg [1] or SCAFFOLD [31]) with the server. To highlight the flexibility and generality, we denote the adopted FL protocol as `FLroutine`$(\cdot)$ with datasets $\mathcal{S}^l_{[M]} := \{\mathcal{S}^l_m : m \in [M]\}$. `FLroutine`$(\mathcal{S}^l_{[M]})$ targets at solving the following standard FL problem:

$$\min_{f \in \mathcal{F}} \widehat{\mathcal{L}}(f; \mathcal{S}^l_{[M]}) := \sum\nolimits_{m \in [M]} (n_m/n) \cdot \widehat{\mathcal{L}}_m(f; \mathcal{S}^l_m), \tag{1}$$

where $n_m := |\mathcal{S}^l_m|$ is the number of samples in dataset $\mathcal{S}^l_m$, $n := \sum_{m \in [M]} n_m$ is the total number of samples, and $\widehat{\mathcal{L}}_m(f; \mathcal{S}^l_m) := (1/n_m) \cdot \sum_{i \in [n_m]} \ell_m(f(x^i_m, a^i_m); r^i_m)$ is the empirical local loss of agent m with $\ell_m(\cdot;\cdot) : \mathbb{R}^2 \to \mathbb{R}$ as the loss function and (x^i_m, a^i_m, r^i_m) as the i-th sample in $\mathcal{S}^l_m$. The output function of this FL process is then used as the estimated reward function $\widehat{f}^{l+1}$ for IGW sampling in the next epoch $l + 1$.

It is worth particularly emphasizing that there is no restriction on the FL protocol in FedIGW as long as it follows the canonical FL framework formulated in Eq. (1), which is a goal commonly adopted in FL studies. Such flexibility is remarkable as it enables FedIGW to incorporate any existing or future FL protocols just by plugging a new FL component into it.

4 Flexible Extensions

Besides flexibly incorporating canonical FL protocols, another major advantage offered by the decoupled FL choices is to bring appropriate appendages from FL that directly benefit FCB, as illustrated in Fig. 2. In the following, we discuss how to leverage techniques of robustness and privacy from FL in FedIGW, while presenting intriguing avenues for future exploration.

Robustness. One important direction in FCB studies is to improve robustness against malicious attacks. Some advances have been achieved in attaining this desirable property, e.g., robust aggregation schemes are studied in [32–34]. However, these designs are still tailored to specific settings and require performing careful construction on previous basic FCB designs.

With the FL component as a largely decoupled component in the design of FedIGW, it is more convenient to achieve robustness as suitable techniques from FL studies can be directly applied with only minor modifications. The key is that as long as such FL protocols can provide an estimated function (which is a common goal of FL), they can be adopted in FedIGW to achieve additional robustness guarantees in FCB. For example, [35–39] studied how to handle malicious agents, who can deviate arbitrarily from the FL protocol and tamper with their own updates, during learning.

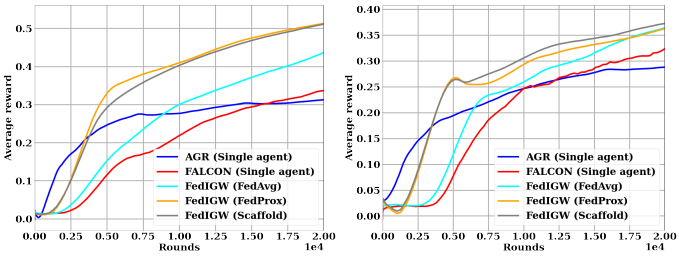

Figure 1: Experiments with Bibtex (left) and Delicious (right).

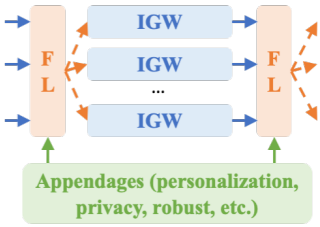

Figure 2: Flexible FL appendages

The commonly adopted scheme is to invoke certain robust estimators (e.g., median and trimmed mean). Under suitable assumptions, existing approaches have shown that as long as the proportion of malicious agents does not exceed a threshold (typically, $1/2$), the estimators calculated by federation can still converge within certain amounts of error due to the malicious agents.

Privacy. Moreover, many mechanisms have also been studied in FL [40–42], to guarantee differential privacy (DP), where the most common approach is to insert noises of suitable scales [40, 43, 44]. Similar to the study of robustness, while there have been some attempts to insert tailored noises into FCB designs for privacy guarantees [13, 45, 46], the proposed FedIGW algorithm can effortlessly leverage the existing fruitful investigations on FL with privacy guarantees [40–42]. The key is still that all favorable properties during learning the estimate $\widehat{f}$ (which is used in IGW interactions) are naturally inherited by FedIGW.

Other Possibilities. There have been many studies on personalization [47, 48], fairness guarantees [49, 50], client selections [51, 52], and practical communication designs [53–55] in FL among many other directions, which are all conceivably applicable in FedIGW.

5 Performance Evaluation

Following the single-agent study of IGW [21], theoretical analyses can be performed on FedIGW to demonstrate its efficiency, where the convergence results of FL algorithms play a critical role. Due to the space limitation, the detailed theoretical results are deferred to the full version of this work.

In this section, we report the empirical performances of FedIGW on two real-world datasets: Bibtex [56] and Delicious [57]. The reported Fig. 1 compares the averaged rewards collected by FedIGW using different FL choices, including FedAvg [1], SCAFFOLD [31], and FedProx [3], and $M = 10$ agents with two single-agent designs, where FALCON [21] can be viewed as the single-agent version of FedIGW and AGR [58] is an alternative strong single-agent CB baseline. This is the first time, to the best of our knowledge, FedAvg is practically integrated with FCB experiments, let alone other FL choices. It can be observed that on both datasets, FedIGW achieves better performance than the single-agent baselines with more rewards collected by each agent on average, which validates its effectiveness in leveraging agents' collaborations. Also, it can be observed that using the more developed SCAFFOLD and FedProx provides improved performance compared with the basic FedAvg, demonstrating FedIGW's capability of harnessing advances in FL protocols. Additional experimental details are discussed in Appendix B with more results provided, including error bars.

6 Conclusions

In this work, we studied the problem of federated contextual bandits (FCB) and recognized that existing FCB designs are largely disconnected from canonical FL studies in their adopted FL protocols, which hinders the integration of crucial FL advancements. To bridge this gap, we introduced a novel design, FedIGW, capable of accommodating a wide range of FL protocols, provided they address a standard FL problem. Moreover, we explored how advancements in FL can seamlessly bestow additional desirable attributes upon FedIGW. Specifically, we delved into the incorporation of robustness and privacy, presenting intriguing opportunities for future research. Empirical validations on real-world datasets underscored its practicality and flexibility. It would be valuable to pursue further exploration of alternative CB algorithms within FCB, e.g., [25, 59, 60], and investigate whether the FedIGW design can be extended to more general federated RL [9, 10].

Acknowledgments and Disclosure of Funding

The work of CSs and KY was supported in part by the US National Science Foundation (NSF) under awards 2029978, 2143559, 2002902, Virginia Commonwealth Cyber Initiative, and the Bloomberg Data Science Ph.D. Fellowship.

References

[1] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[2] Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

[3] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020.

[4] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[5] Fengda Zhang, Kun Kuang, Zhaoyang You, Tao Shen, Jun Xiao, Yin Zhang, Chao Wu, Yueting Zhuang, and Xiaolin Li. Federated unsupervised representation learning. *arXiv preprint arXiv:2010.08982*, 2020.

[6] Bram van Berlo, Aaqib Saeed, and Tanir Ozcelebi. Towards federated unsupervised representation learning. In *Proceedings of the third ACM international workshop on edge systems, analytics and networking*, pages 31–36, 2020.

[7] Weiming Zhuang, Yonggang Wen, and Shuai Zhang. Divergence-aware federated self-supervised learning. *arXiv preprint arXiv:2204.04385*, 2022.

[8] Ekdeep Lubana, Chi Ian Tang, Fahim Kawsar, Robert Dick, and Akhil Mathur. Orchestra: Unsupervised federated learning via globally consistent clustering. In *International Conference on Machine Learning*, pages 14461–14484. PMLR, 2022.

[9] Abhimanyu Dubey and Alex Pentland. Provably efficient cooperative multi-agent reinforcement learning with function approximation. *arXiv preprint arXiv:2103.04972*, 2021.

[10] Yifei Min, Jiafan He, Tianhao Wang, and Quanquan Gu. Multi-agent reinforcement learning: Asynchronous communication and linear function approximation. *arXiv preprint arXiv:2305.06446*, 2023.

[11] Yuanhao Wang, Jiachen Hu, Xiaoyu Chen, and Liwei Wang. Distributed bandit learning: Near-optimal regret with efficient communication. *arXiv preprint arXiv:1904.06309*, 2019.

[12] Ruiquan Huang, Weiqiang Wu, Jing Yang, and Cong Shen. Federated linear contextual bandits. *Advances in neural information processing systems*, 34:27057–27068, 2021.

[13] Abhimanyu Dubey and Alex Pentland. Differentially-private federated linear bandits. *Advances in Neural Information Processing Systems*, 33:6003–6014, 2020.

[14] Chuanhao Li and Hongning Wang. Asynchronous upper confidence bound algorithms for federated linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 6529–6553. PMLR, 2022.

[15] Jiafan He, Tianhao Wang, Yifei Min, and Quanquan Gu. A simple and provably efficient algorithm for asynchronous federated contextual linear bandits. *Advances in neural information processing systems*, 2022.

[16] Sanae Amani, Tor Lattimore, András György, and Lin F Yang. Distributed contextual linear bandits with minimax optimal communication cost. *arXiv preprint arXiv:2205.13170*, 2022.

[17] Chuanhao Li and Hongning Wang. Communication efficient federated learning for generalized linear bandits. *Advances in Neural Information Processing Systems*, 2022.

[18] Chuanhao Li, Huazheng Wang, Mengdi Wang, and Hongning Wang. Communication efficient distributed learning for kernelized contextual bandits. *Advances in Neural Information Processing Systems*, 2022.

[19] Chuanhao Li, Huazheng Wang, Mengdi Wang, and Hongning Wang. Learning kernelized contextual bandits in a distributed and asynchronous environment. *The Eleventh International Conference on Learning Representations*, 2023.

[20] Zhongxiang Dai, Yao Shu, Arun Verma, Flint Xiaofeng Fan, Bryan Kian Hsiang Low, and Patrick Jaillet. Federated neural bandit. *The Eleventh International Conference on Learning Representations*, 2023.

[21] David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 47 (3):1904–1931, 2022.

[22] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[23] Lie He, An Bian, and Martin Jaggi. Cola: Decentralized linear learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[24] Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert Schapire. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, pages 19–26. PMLR, 2012.

[25] Yunbei Xu and Assaf Zeevi. Upper counterfactual confidence bounds: a new optimism principle for contextual bandits. *arXiv preprint arXiv:2007.07876*, 2020.

[26] Rajat Sen, Alexander Rakhlin, Lexing Ying, Rahul Kidambi, Dean Foster, Daniel N Hill, and Inderjit S Dhillon. Top-k extreme contextual bandits with arm hierarchy. In *International Conference on Machine Learning*, pages 9422–9433. PMLR, 2021.

[27] Naoki Abe and Philip M Long. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, pages 3–11. Citeseer, 1999.

[28] Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

[29] Sanath Kumar Krishnamurthy, Vitor Hadad, and Susan Athey. Adapting to misspecification in contextual bandits with offline regression oracles. In *International Conference on Machine Learning*, pages 5805–5814. PMLR, 2021.

[30] Avishek Ghosh, Abishek Sankararaman, and Kannan Ramchandran. Model selection for generic contextual bandits. *arXiv preprint arXiv:2107.03455*, 2021.

[31] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

[32] Ilker Demirel, Yigit Yildirim, and Cem Tekin. Federated multi-armed bandits under byzantine attacks. *arXiv preprint arXiv:2205.04134*, 2022.

[33] Ali Jadbabaie, Haochuan Li, Jian Qian, and Yi Tian. Byzantine-robust federated linear bandits. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 5206–5213. IEEE, 2022.

[34] Aritra Mitra, Arman Adibi, George J Pappas, and Hamed Hassani. Collaborative linear bandits with adversarial agents: Near-optimal regret bounds. *Advances in neural information processing systems*, 2022.

[35] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pages 5650–5659. PMLR, 2018.

[36] Krishna Pillutla, Sham M Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022.

[37] Shuhao Fu, Chulin Xie, Bo Li, and Qifeng Chen. Attack-resistant federated learning with residual-based reweighting. *arXiv preprint arXiv:1912.11464*, 2019.

[38] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pages 6357–6368. PMLR, 2021.

[39] Banghua Zhu, Lun Wang, Qi Pang, Shuai Wang, Jiantao Jiao, Dawn Song, and Michael I Jordan. Byzantine-robust federated learning with optimal statistical rates. In *International Conference on Artificial Intelligence and Statistics*, pages 3151–3178. PMLR, 2023.

[40] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020.

[41] Xuefei Yin, Yanming Zhu, and Jiankun Hu. A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. *ACM Computing Surveys (CSUR)*, 54(6):1–36, 2021.

[42] Ziyao Liu, Jiale Guo, Wenzhuo Yang, Jiani Fan, Kwok-Yan Lam, and Jun Zhao. Privacy-preserving aggregation in federated learning: A survey. *IEEE Transactions on Big Data*, 2022.

[43] Antonious Girgis, Deepesh Data, Suhas Diggavi, Peter Kairouz, and Ananda Theertha Suresh. Shuffled model of differential privacy in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2521–2529. PMLR, 2021.

[44] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Hang Su, Bo Zhang, and H Vincent Poor. User-level privacy-preserving federated learning: Analysis and performance optimization. *IEEE Transactions on Mobile Computing*, 21(9):3388–3401, 2021.

[45] Xingyu Zhou and Sayak Ray Chowdhury. On differentially private federated linear contextual bandits. *arXiv preprint arXiv:2302.13945*, 2023.

[46] Tan Li and Linqi Song. Privacy-preserving communication-efficient federated multi-armed bandits. *IEEE Journal on Selected Areas in Communications*, 40(3):773–787, 2022.

[47] Filip Hanzely, Boxin Zhao, and Mladen Kolar. Personalized federated learning: A unified framework and universal optimization techniques. *arXiv preprint arXiv:2102.09743*, 2021.

[48] Alekh Agarwal, John Langford, and Chen-Yu Wei. Federated residual learning. *arXiv preprint arXiv:2003.12880*, 2020.

[49] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR, 2019.

[50] Wei Du, Depeng Xu, Xintao Wu, and Hanghang Tong. Fairness-aware agnostic federated learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 181–189. SIAM, 2021.

[51] Ravikumar Balakrishnan, Tian Li, Tianyi Zhou, Nageen Himayat, Virginia Smith, and Jeff Bilmes. Diverse client selection for federated learning via submodular maximization. In *International Conference on Learning Representations*, 2022.

[52] Yann Fraboni, Richard Vidal, Laetitia Kameni, and Marco Lorenzi. Clustered sampling: Low-variance and improved representativity for clients selection in federated learning. In *International Conference on Machine Learning*, pages 3407–3416. PMLR, 2021.

[53] Mingzhe Chen, Deniz Gündüz, Kaibin Huang, Walid Saad, Mehdi Bennis, Aneta Vulgarakis Feljan, and H Vincent Poor. Distributed learning in wireless networks: Recent progress and future challenges. *IEEE Journal on Selected Areas in Communications*, 39(12):3579–3605, 2021.

[54] Xizixiang Wei and Cong Shen. Federated learning over noisy channels: Convergence analysis and design examples. *IEEE Transactions on Cognitive Communications and Networking*, 8(2): 1253–1268, 2022.

[55] Sihui Zheng, Cong Shen, and Xiang Chen. Design and analysis of uplink and downlink communications for federated learning. *IEEE Journal on Selected Areas in Communications*, 39(7):2150–2167, 2020.

[56] Ioannis Katakis, Grigorios Tsoumakas, and Ioannis Vlahavas. Multilabel text classification for automated tag suggestion. *ECML PKDD discovery challenge*, 75:2008, 2008.

[57] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Effective and efficient multilabel classification in domains with large number of labels. In *Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD'08)*, volume 21, pages 53–59, 2008.

[58] David Cortes. Adapting multi-armed bandits policies to contextual bandits scenarios. *arXiv preprint arXiv:1811.04383*, 2018.

[59] Dylan J Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *arXiv preprint arXiv:2010.03104*, 2020.

[60] Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Conference on Learning Theory*, pages 4300–4354. PMLR, 2021.

[61] Eshcar Hillel, Zohar S Karnin, Tomer Koren, Ronny Lempel, and Oren Somekh. Distributed exploration in multi-armed bandits. *Advances in Neural Information Processing Systems*, 26, 2013.

[62] Balazs Szorenyi, Róbert Busa-Fekete, István Hegedus, Róbert Ormándi, Márk Jelasity, and Balázs Kégl. Gossip-based distributed stochastic bandit algorithms. In *International conference on machine learning*, pages 19–27. PMLR, 2013.

[63] Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. On distributed cooperative decision-making in multiarmed bandits. In *2016 European Control Conference (ECC)*, pages 243–248. IEEE, 2016.

[64] David Martínez-Rubio, Varun Kanade, and Patrick Rebeschini. Decentralized cooperative stochastic bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

[65] Chengshuai Shi and Cong Shen. Federated multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9603–9611, 2021.

[66] Chengshuai Shi, Cong Shen, and Jing Yang. Federated multi-armed bandits with personalization. In *International Conference on Artificial Intelligence and Statistics*, pages 2917–2925. PMLR, 2021.

[67] Clémence Réda, Sattar Vakili, and Emilie Kaufmann. Near-optimal collaborative learning in bandits. In *NeurIPS 2022-36th Conference on Neural Information Processing System*, 2022.

[68] Zhaowei Zhu, Jingxuan Zhu, Ji Liu, and Yang Liu. Federated bandit: A gossiping approach. In *Abstract Proceedings of the 2021 ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems*, pages 3–4, 2021.

[69] Zhirui Chen, PN Karthik, Vincent YF Tan, and Yeow Meng Chee. Federated best arm identification with heterogeneous clients. *arXiv preprint arXiv:2210.07780*, 2022.

[70] Sudeep Salgia and Qing Zhao. Distributed linear bandits under communication constraints. *arXiv preprint arXiv:2211.02212*, 2022.

[71] Jiabin Lin and Shana Moothedath. Federated stochastic bandit learning with unobserved context. *arXiv preprint arXiv:2303.17043*, 2023.

[72] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

[73] Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with ucb-based exploration. In *International Conference on Machine Learning*, pages 11492–11502. PMLR, 2020.

[74] Baihe Huang, Xiaoxiao Li, Zhao Song, and Xin Yang. Fl-ntk: A neural tangent kernel-based framework for federated learning analysis. In *International Conference on Machine Learning*, pages 4423–4434. PMLR, 2021.

[75] Alekh Agarwal, H Brendan McMahan, and Zheng Xu. An empirical evaluation of federated contextual bandit algorithms. *arXiv preprint arXiv:2303.10218*, 2023.

[76] Deepayan Chakrabarti, Ravi Kumar, Filip Radlinski, and Eli Upfal. Mortal multi-armed bandits. *Advances in neural information processing systems*, 21, 2008.

A Related Works

We provide a more detailed review of federated multi-armed bandits (FMAB) and federated contextual bandits (FCB) in the following.

• **Tabular.** There have been many studies on cooperative designs in multi-armed bandits (i.e., the tabular setting), e.g., [61–64], focusing on different learning targets and different communication schemes (e.g., through a communication graph or with some randomly selected peers). Notably, in [11], communication-efficient designs are proposed via periodically aggregating local estimates and performing arm elimination globally. We here also discuss another line of works on FMAB [65–69]. In their considered setting, the global rewards are (weighted) averages of local observations; however the former is not directly observable. With maximizing global rewards as the learning target, the agents need to collaboratively perform explorations and aggregate local information. Especially, [65–69] all commonly employ UCB-based exploration schemes while the adopted FL routine is to average local sample means as globe ones and to construct global confidence bounds.

• **Linear.** The most commonly studied FCB setting is federated linear bandits. There have been many investigations in this direction. Especially, different environments have been tackled in different works, e.g., the finite-armed fixed-context setting [11, 12], the finite-armed stochastic-context setting [16], the infinite-armed fixed-context setting [70], and the infinite-armed adversarial-context setting [11, 13–15]. Furthermore, many other settings (e.g., unobserved context [71]) and additional properties (e.g., privacy [13, 45], robustness [33]) have been investigated. As summarized in the main paper, these works mainly select arm elimination (AE) or LinUCB [72] as their CB designs, which require both model estimates and confidence bounds. Thus, in their designed communication schemes, compressed local data (e.g., aggregated local rewards and covariance matrices) are often directly shared to solve a global ridge regression and to construct tighter confidence bounds. Compared with these studies, FedIGW can effectively solve the finite-armed stochastic-context setting without sharing any raw or compressed local data but only communicate processed model parameters (e.g., gradients).

• **Generalized Linear and Kernelized.** As extensions of the linear reward functions, [17] considers the generalized-linear class, and [18, 19] study the kernelized one. The adopted basic techniques are similar to the aforementioned ones in federated linear bandits, while efforts are focused on fine-tuning communications (e.g., via Nyström approximation [18, 19]). It is worth noting that [17] invokes the distributed accelerated gradient descent algorithm to solve their considered distributed optimization with a generalized linear function class, which can be viewed as a preliminary attempt to involve FL or distributed optimization designs in FCB. However, the motivation there is the lack of a closed-form solution as in the linear case, while [17] additionally needs to share the local covariance matrices to construct better confidence bounds. The FedIGW proposed in this work, instead, can leverage flexible FL designs.

• **Neural.** A recent work [20] extends the advances on single-agent neural bandits [73] to the federated setting, where the neural tangent kernel (NTK) analyses are incorporated. With NTK to "linearize" the considered over-parameterized neural network, [20] still largely follows the designs in the aforementioned federated linear bandits while some additional attempts have been made, e.g., an extra one-round averaging of model parameters besides aggregating NTK. This work, instead, takes a step further to fully leverage FL designs which often perform multiple (instead of one) rounds of model aggregations that are often necessary to guarantee convergence. Moreover, the optimization and generalization errors of a FedAvg variant with overparameterized neural networks are provided in [74], which is conceivably compatible with FedIGW for the corresponding analyses. Moreover, as shown by the additional experimental results in Appendix B.4 FedIGW empirically outperforms FN-UCB [20] on different tasks and is more computationally efficient.

B Experiment Details and Additional Results

This section first provides a comprehensive description of the experimental settings and procedures. **The codes and detailed instructions have been uploaded in the supplementary materials so as to execute the experiments and reproduce the results.**

Additional results are also provided to deepen the understanding of the impact of adopting different FL protocols in FedIGW and the effect of involving different numbers of agents. Moreover, a performance comparison between FedIGW and the state-of-the-art FCB design, FN-UCB [20], is

reported. These results reveal that our proposed FedIGW not only outperforms the single-agent designs but also supersedes the strong FCB baselines.

B.1 Experiment Settings

In the following, we report a comprehensive description of the experimental details adopted in the simulations.

Datasets. Our experiments employ two distinct real-world multi-label classification datasets, Bibtex [56] and Delicious [57], which are also used in other practical CB investigations such as [58]. The aim of CB is considered to be recommending one of the correct labels at any given time. Especially, in the experiments, at each time step, a context is randomly sampled from the dataset while the true labels are concealed from the agents. The agents then determine which label to select (i.e., pull one arm) with their CB algorithms; thus the number of arms is the number of possible labels in each dataset. Upon pulling one arm, a reward of 1 is granted if the pulled arm corresponds to one of the true labels, while a reward of 0 is granted otherwise. Details of each task are listed in Table 2, from which we can observe that these scenarios are challenging given their high-dimensional contexts and large numbers of arms.

Table 2: The context dimension and number of arms in Bibtex and Delicious

Task	Context dimension	Number of arms
Bibtex	1835	159
Delicious	500	983

Environments. In the experiments, the environments sample contexts for all agents from the same set of datasets described above. For simplicity, the system is also designed as a synchronous one, i.e., $t_m(t) = t, \forall m \in [M]$.

FedIGW. As described in Section 5, for both tasks, two-layer multi-layer perceptrons (MLPs) with a hidden layer having a constant 256 width are used to approximate the reward functions in FedIGW. Moreover, multiple standard FL protocols including FedAvg [1], SCAFFOLD [31] and FedProx [3] are adopted as the FL component in FedIGW. During each FL process, the local batch size, the number of communications, and the local learning rate are specified in Table 3. Moreover, the epoch length is designed to be growing exponentially, while culminating at an upper limit of 4096 to maintain timely updates.

Additionally, for practical conveniences, we set the parameter γ as a constant hyper-parameter and perform some preliminary manual selections with the final adopted values reported in Table 3. We believe this approach is more practically appealing as it does not need to scale γ consistently; a similar choice of using constant γ's is also adopted in [75].

Table 3: Hyperparameter choices for FedIGW in Bibtex and Delicious

Task	Learning Rate	Batch Size	Communications	Parameter γ
Bibtex	0.1	64	100	7000
Delicious	0.2	64	100	7000

Single-agent baseline: AGR. The adaptive greedy (AGR) algorithm [76] is selected as one of the single-agent baselines due to its strong empirical performance on Bibtex and Delicious reported in [58]. The algorithmic details can be found in [58], and we also leveraged the code provided in [58] to build this baseline.

Single-agent baseline: FALCON. The other single-agent baseline, FALCON, is proposed in [21], which is essentially the single-agent version of FedIGW. We still adopt the same algorithmic configurations as FedIGW (i.e., epoch length, parameter γ, local batch size, and local learning rate) except that the MLP is optimized locally instead of in a federation, i.e., there are no communications.

Performance evaluation. All the reported results except Fig. 8 are averaged from 10 independent runs, whose error bars are further provided. In Fig. 8, the unpractical long running time of the baseline algorithm prohibits us from performing repeated experiments as discussed in the following description of computing resources.

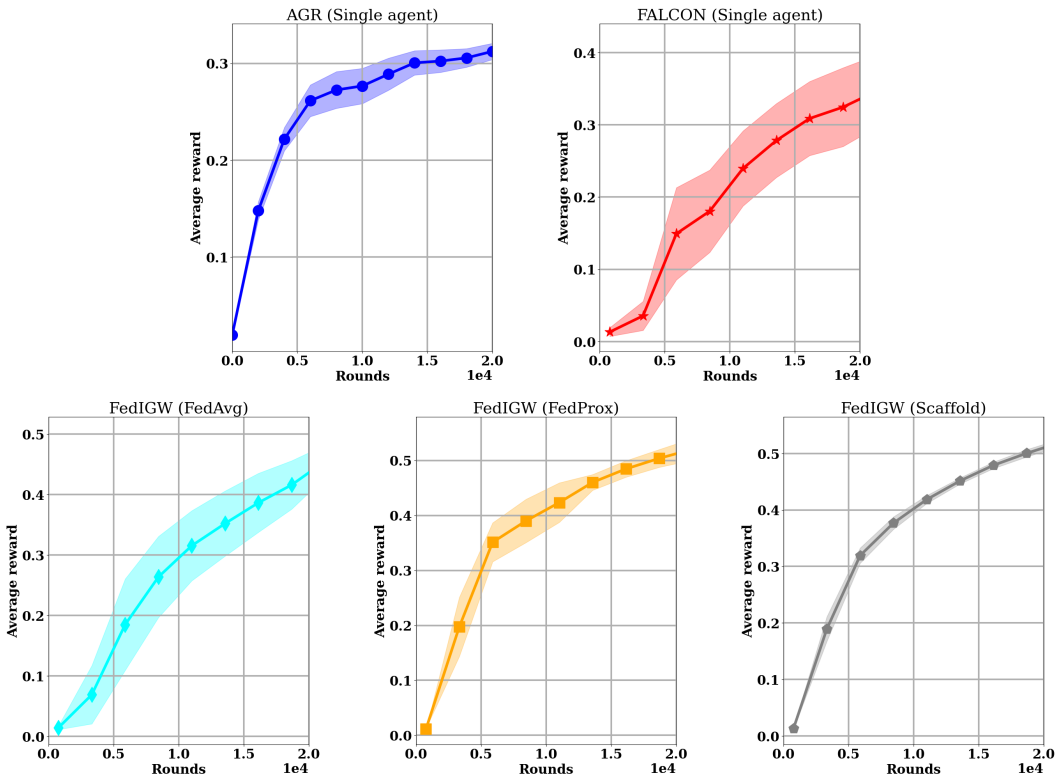

Figure 3: Averaged rewards and error bars on Bibtex from two single-agent baselines and FedIGW using FedAvg, FedProx, and SCAFFOLD as its FL protocol. The continuous curves represent the empirical average values, and the shadowed areas are the standard deviations.

Computing Resources. The computational requirement is relatively low for testing the two single-agent baselines (AGR and FALCON) and our proposed FedIGW. Specifically, we use a dual Nvidia-RTX 3090 workstation with an overall 20 GB RAM, which is more than needed as only 2 GB RAM is occupied during the experiments of AGR, FALCON, and FedIGW using the aforementioned MLP having a hidden layer with width 256. However, as illustrated later in Section B.4, when testing FN-UCB [20], the overall 20 GB RAM is not sufficient for running it when the MLP hidden layer has a width larger than 10; thus, we down-scaled the width to be 5 for smooth testing.

B.2 Additional Results: Varying FL Choices

We here provide additional details of Fig. 1 in Figs. 3 and 4, especially with error bars, and numerically verify the flexibility of FedIGW with different FL protocols, including FedAvg [1], SCAFFOLD [31] and FedProx [3]. As observed in Section 5, using the further optimized SCAFFOLD and FedProx provides better performances compared with using the basic FedAvg, which credits to that FedIGW can seamlessly leverage algorithmic advances in FL protocols. Moreover, it can be observed that the performance obtained by using SCAFFOLD as the FL choice in FedIGW is also particularly stable, which demonstrates its superiority in the application of sequential decision-making.

B.3 Additional Results: Varying Numbers of Agents

Fig. 5 further reports the averaged rewards of FedIGW with $M = 10, 20, 30, 50$ involved agents and their associated error bars in the Bibtex dataset, while Fig. 6 reports the same set of results in the Delicious dataset. Moreover, Fig. 7 compares the averaged rewards of FedIGW with varying numbers of agents. In these results, the FL protocol in FedIGW is selected to be FedAvg.

From Figs. 5 and 6, we can observe that FedIGW is capable of collecting more rewards than the two single-agent baselines, which demonstrates its efficiency. Moreover, Fig. 7 elucidates that the final performance of FedIGW is positively correlated with an increasing number of clients, which verifies

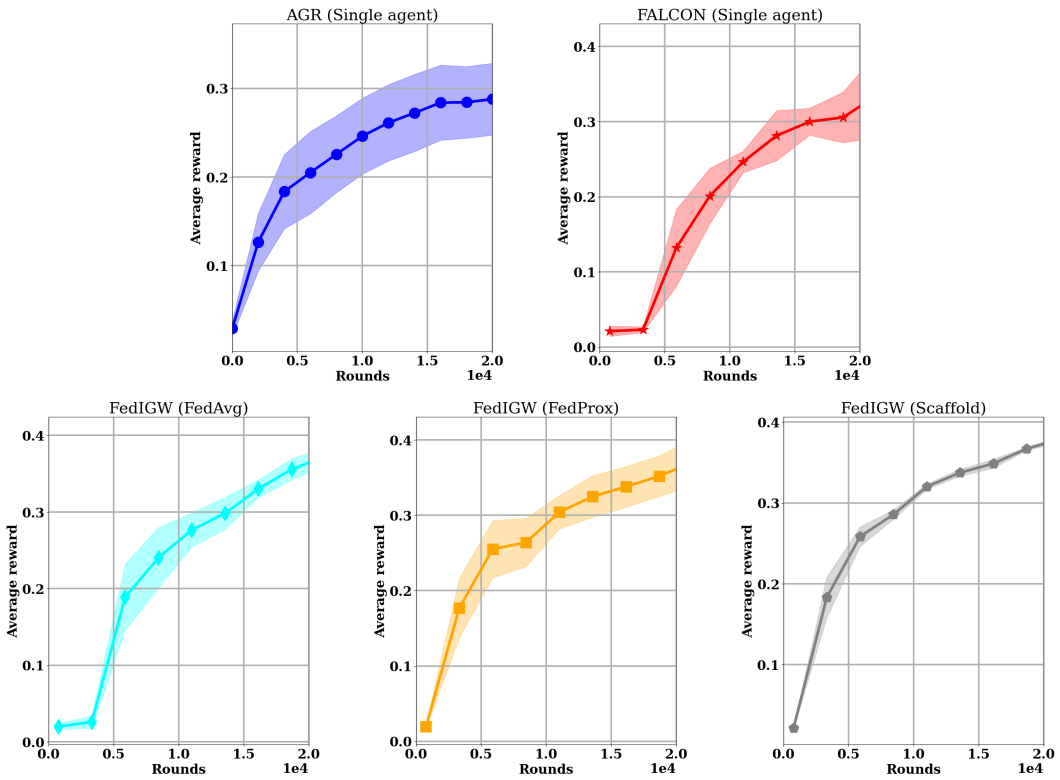

Figure 4: Averaged rewards and error bars on Delicious from two single-agent baselines and FedIGW using FedAvg, FedProx, and SCAFFOLD as its FL protocol. The continuous curves represent the empirical average values, and the shadowed areas are the standard deviations.

the benefits of learning in a larger federation. Furthermore, the variance (i.e., error bar) in Figs. 5 and 6 begins at a relatively small value and subsequently expands due to the initially intense explorations. Eventually, when the algorithm gradually approaches convergence, the variance begins to reduce, reflecting the stabilization of the learning process.

B.4 Additional Results: Comparison with Federated Neural Bandits

To further verify the performance of FedIGW, additional comparisons are conducted with a state-of-the-art FCB baseline, specifically, the federated neural-upper confidence bound (FN-UCB) design proposed in [20]. FN-UCB is capable of leveraging neural networks to approximate rewards and [20] has reported superior performance compared to many other designs, which makes it a strong FCB baseline. When conducting the experiments, we first notice that FN-UCB necessitates multiple matrix inversions over the entire set of neural network parameters and such operations lead to substantial memory consumption when handling the high-dimensional context and numerous arms in both Bibtex and Delicious (which evidence the difficulty of these two employed datasets). To accommodate our already powerful computing resources (dual Nvidia-RTX 3090 and 20 GB RAM), we had to use a small-size MLP in both FALCON and FN-UCB for smooth testing and fair comparison. Especially, the width of the MLP hidden layer is down-scaled from the originally adopted 256 (in Fig. 1) to only 5, as our 20 GB RAM cannot support FN-UCB using an MLP with its hidden layer wider than 10. Also, due to the inefficiency of FN-UCB in our testing scenario, the error bars are omitted in Fig. 8. We believe this experimental observation demonstrates the computational efficiency of FedIGW over FN-UCB.

Moreover, the statistical performance of FedIGW and FN-UCB is presented in Fig. 8. As evident from the reported results, the substantial reduction in the neural network size has affected the performance of FedIGW compared with Figs. 5 and 6. In particular, FedIGW achieves only about 30% of the previously reported performance on Bibtex and 50% on Delicious. Nevertheless, despite this diminished performance, our proposed FedIGW still outperforms FN-UCB significantly, surpassing it by nearly 70% on both tasks as shown in Fig. 8. These comparisons further validate the advantages

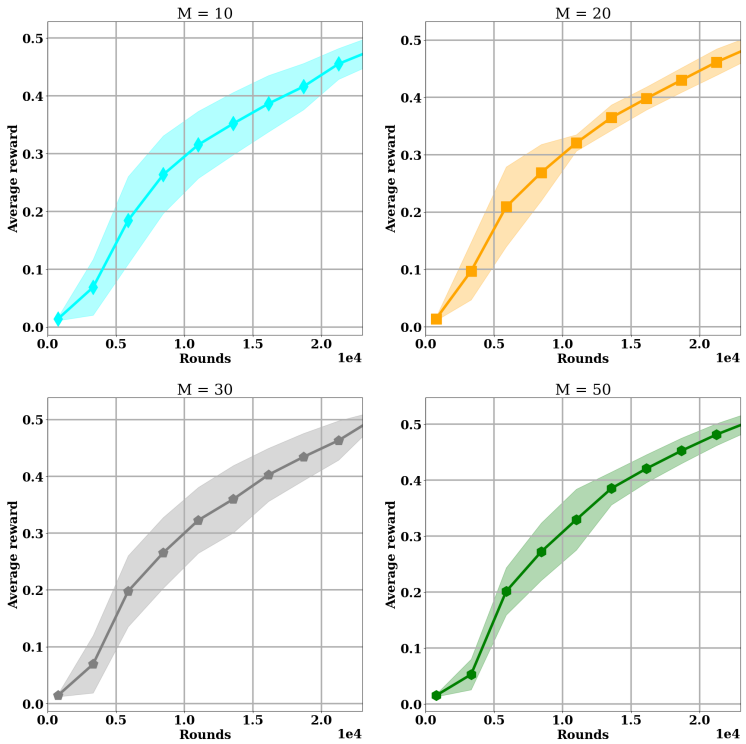

Figure 5: Averaged rewards and error bars on Bibtex from FedIGW (using FedAvg) with varying numbers of involved agents, i.e., $M = 10, 20, 30, 50$. The continuous curves represent the empirical average values, and the shadowed areas are the standard deviations.

of FedIGW, demonstrating its relatively easy implementation and superior performance compared to existing FCB designs.

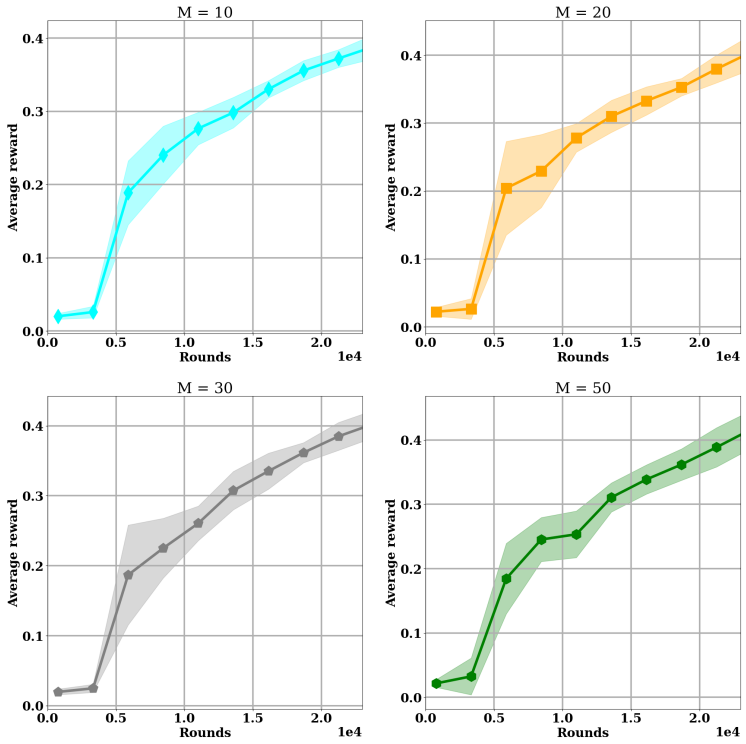

Figure 6: Averaged rewards and error bars on Delicious from FedIGW (using FedAvg) with varying numbers of involved agents, i.e., $M = 10, 20, 30, 50$. The continuous curves represent the empirical average values, and the shadowed areas are the standard deviations.

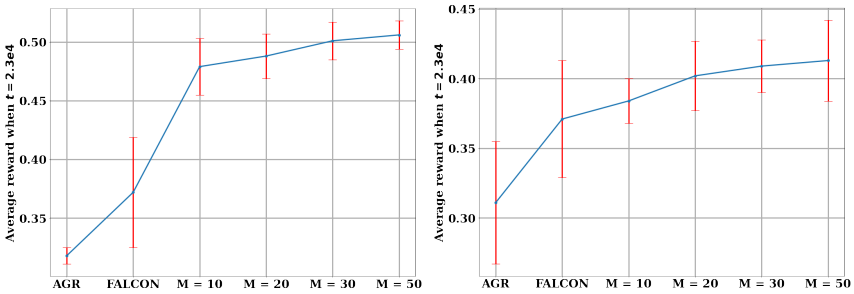

Figure 7: Averaged rewards and error bars on Bibtex (left) and Delicious (right) at time step 2.3×10^4 from two single-agent baselines and FedIGW (using FedAvg) with varying numbers of involved agents, i.e., $M = 10, 20, 30, 50$.

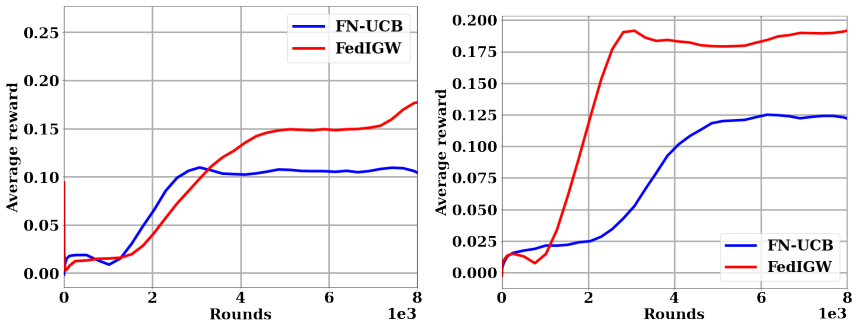

Figure 8: Comparisons between FedIGW and FN-UCB on Bibtex (left) and Delicious (right) with 10 involved agents.