Economy-motivated Federated Crowdsourcing

Xiaoqian Jiang

Abstract

...

Introduction

With the prosperity of Artificial Intelligence, crowdsourcing is widely used to solve the imbalance problems with supervised-learning data sets(e.g., barriers to expertise, regional restriction (Ye et al., 2018; Sigurdsson et al., 2016; Amgad et al., 2022)). It is necessary for a platform to make it convenient for workers differing in professional knowledge or regions to participate in Crowdtasks.MTurk, Zooniverse, Datatang, and Baidu-crowdsourcing are several typical crowdsourcing systems. However, studies showed that information workers uploaded may cause their privacy disclosure, especially about their physical traits, personality bias, traces of life, and so on (Xia and McKernan, 2020). Due to people's growing awareness of privacy protection, privacy issues in the field of crowdsourcing require extensive research by scholars. State-of-the-art privacy mechanisms being proposed all come at the expense of the accuracy of crowdsourcing data(e.g.cloaking (Pournajaf et al., 2014; Ren et al., 2022) or inaccuracy(e.g. obfuscation like local differential privacy (Wang et al., 2018; Wei et al., 2019)))(Wang et al., 2020a). Unfortunately, these mechanisms sacrifice the quality of crowdsourcing, because they have to blur the corresponding original information.

For the above problem, Federated learning(FL), a new distributed learning framework is put forward (McMahan et al., 2017). FL allows multiple clients to collaborate on training shared models by iteratively aggregating model updates without exposing the raw data (Wang et al., 2020b; Gao et al., 2022). In traditional crowdsourcing projects, workers are organized to perform tasks as required, and then data is aggregated and processed (Yu et al., 2020; Tu et al., 2020; Wu et al., 2021a; Zhang et al., 2022). However, Centralized crowdsourcing organizations can easily reveal workers' privacy. Fortunately, we can introduce the framework of FL to develop mobile crowdsourcing. Mobile crowdsourcing under the FL framework allows workers to collect and process data locally, without the need to upload raw data directly. Federal crowdsourcing is a great protection for worker privacy (Li et al., 2020a; Ciftler et al., 2020; Zhang et al., 2021a).

FL can effectively alleviate the problem of privacy leakage in crowdsourcing. However, in FL, each client maintains its local data, forming the situation of the data island. Therefore, we must consider how to incentivize large numbers of clients to contribute data to model training to break data constraints in the form of islands (Zhan et al., 2021a). Without adequate incentives, clients are reluctant to volunteer their data, computing, and communications resources to participate in FL. Furthermore, FL, while allowing clients to train models locally and update them to remote servers without exposing the raw data, does not protect against inference attacks (Lyu et al., 2020; Suri et al., 2022). Such potential privacy concerns make clients less willing to participate in FL or crowdsourcing (Mothukuri et al., 2021). Unless there's enough compensation that they're willing to take those risks and contribute their passion and resources. Moreover, FL is a form of distributed machine learning, with each client performing tasks independently on their own devices (Liu et al., 2022). In other words, clients have the right to determine their own participation strategy(e.g., Participation time and frequency and learning accuracy (Li et al., 2020b)). To sum up, the incentive mechanism can improve the performance of the model by encouraging clients to choose the best participation strategy, which is an essential link in both FL and crowdsourcing.

Currently, it is mainly about the incentive mechanism of FL, which is designed around the driving factors of clients' contribution, reputation, and resource allocation (Zhan et al., 2020a, 2021b; Li et al., 2023). Incentives in FL are designed to motivate clients to contribute their own local resources (e.g., data, device resources, bandwidth) through iterative aggregation model updates to collaboratively train shared models. However, it is not appropriate to use these incentive mechanisms for FL in a federated crowdsourcing project. The reasons are as follows:(1) Federated crowdsourcing requires workers to annotate training samples on their own equipment, so the incentive mechanism of federated crowdsourcing also needs to motivate workers to contribute their enthusiasm and interest to the projects. (2) Since the equipment resources of crowdsourcing workers are highly heterogeneous (e.g., computing resources, communication resources), and the knowledge reserves of crowdsourcing workers are also uneven, time control should be considered. (3) Federated crowdsourcing needs to both assess the quality of submitted data to prevent malicious workers from submitting low-quality data for quick rewards, and respond to client delays in updating the model due to emergencies(e.g., Workers themselves, client equipment failure, network quality). (4) In federal crowdsourcing, platforms need to recruit and retain high-quality workers, workers(clients) need to be paid fairly on time, and task publishers(servers) need to pay as little as possible to maximize their own benefits. The incentive mechanism of federated crowdsourcing must meet the requirements of crowdsourcing platforms, crowdsourcing workers, and task publishers at the same time.

To solve the above problems well, we put forward Economy-motivated Federated Crowdsourcing (eFedCrowd) which inspires mobile data owners in federated crowdsourcing projects to actively contribute their passion and resources to training the client model and even optimizing the server model. eFedCrowd modeled the above issue as a two-stage Stackelberg game (Li and Sethi, 2017) scenario for analysis and discussion. In the second stage, eFedCrowd equitably distributes rewards based on the local accuracy of the client training model. At the same time, the corresponding costs paid by workers to complete federated crowdsourcing tasks are considered, mainly computing costs and communication costs. In the first phase, eFedCrowd maximizes the task publisher's net utility, which is the total utility gained from aggregating the model on the server, and deducts the total cost of motivating the client to complete model training and updating. And finally, we deduce the Nash equilibrium in the Stackelberg game. Figure 1 shows the working flow diagram of the eFedCrowd.The main contributions of this paper are as follows:

(1) eFedCrowd only considers the local accuracy of model training when awarding rewards to workers involved in crowdsourcing tasks, which reduces the complexity of the federated crowdsourcing system.

(2) eFedCrowd hands the control of time limit and data freshness to the task publisher, which applies to both instant-time and non-instant-time crowdsourcing projects, and improves the generalization ability of the federated crowdsourcing system.

(3) eFedCrowd allocates rewards according to contributions, and the rewards workers get are only related to the model accuracy level received by the server, which maintains the fairness of the crowdsourcing market.

(4) eFedCrowd sets a time threshold to preliminarily screen the quality of crowdsourcing workers, and excludes the possibility that malicious workers sacrifice accuracy for shorter training time and thus get rich rewards, thus enhancing the robustness of federated crowdsourcing system.

(5) eFedCrowd rules in this paper are fair and simple, with good interpretability, which is conducive to the long-term retention of high-quality crowdsourcing workers, and reflects the responsibility of the federated crowdsourcing market.

Related Work

The research in this paper is divided into two aspects: one is privacy protection in crowdsourcing, and the other is incentive mechanism in FL.

Since crowdsourcing needs to collect data from workers, it is inevitable that there will be some crowdsourcing tasks involving workers' sensitive information, so workers involved in the crowdsourcing tasks will face the risk of privacy disclosure (Wu et al., 2019; Zhang et al., 2020). Differential privacy (Dwork, 2006) has been widely favored in the research field of Internet privacy protection since it was proposed. Nevertheless, differential privacy works by injecting different levels of noise into the model, undoubtedly at the cost of model accuracy (Bagdasaryan et al., 2019). In addition, there are also techniques to protect the privacy of crowdsourcing workers that introduce various cryptographic algorithms. For example, Shu et al. (2018) proposed a privacy-preserving task recommendation scheme for crowdsourcing, which exploits polynomial functions to express multiple keywords of task requirements and worker interests, and then designs a key derivation method based on matrix decomposition. Joshi et al. (2020) used SALT cryptography in the proposed solution to ensure privacy. Zhang et al. (2019) proposed a privacy-preserving traffic monitoring scheme through both adopting a homomorphic Paillier cryptosystem and super-increasing sequence. However, These encryption algorithms are complex, expensive, and cannot resist inference attacks (Lin et al., 2020; Wang et al., 2019).

To alleviate the above defects, FL provides a secure way to work together so that participants can share and leverage data without exposing their privacy. Mean teacher semisupervised FL (Zhang et al., 2021b) trains a deep neural network ensemble under a novel semisupervised FL framework, achieving highly accurate and privacy-protected crowdsourcing. Li et al. (2020c) proposed a crowdsourcing framework named CrowdSFL, which combines blockchain with FL to help users implement crowdsourcing with less overhead and higher security. Zhao et al. (2021) proposed a privacy-preserving mobile crowdsensing (MCS) system, which integrates FL into MCS and allows participants to locally process sensing data via FL. Nevertheless, These approaches to privacy protection in crowdsourcing by introducing the FL framework are all based on the ideal condition that participants are fully willing to make any contribution.

An incentive mechanism is necessary to ensure the quality and efficiency of FL. Participating in FL consumes computing resources on clients, occupies network bandwidth on clients, and even shortens the battery life of client devices. Clients are not willing to make sacrifices to participate in FL without any return. Accordingly, there is a growing body of research on the incentive mechanism of FL. Zhan et al. (2020b) designed a deep reinforcement learning-based incentive mechanism to determine the optimal pricing strategy for the parameter server and the optimal training strategies for edge nodes. Le et al. (2021) formulated the incentive mechanism between the base station and mobile users as an auction game and further proposed the primal-dual greedy auction mechanism to decide winners in the auction and maximize social welfare. Zhang et al. (2021c) proposed an incentive mechanism of FL based on reputation and reverse auction theory, which selects and rewards participants by combining the reputation and bids of the participants under a limited budget. Wu et al. (2021b) modelled each data owner's contribution and the three categories of computing, communication, and privacy costs based on a multi-dimensional contract approach. Pandey et al. (2019) introduced the crowdsourcing framework into FL and developed a two-stage Starkelberg game to analyze and solve the interests maximization of the client and central server respectively. Exploiting the non-trivial dependence of the training loss on clients' hidden efforts and private local models, Zhao et al. (2023) devised Labeling and Computation Effort and local Model Elicitation mechanisms which incentivize strategic clients to make truthful efforts as desired by the server in local data labeling and local model computation.

Unfortunately, none of these incentive Mechanisms for FL are designed to work in a federated crowdsourcing program that needs to collect data samples manually. Moreover, they all ignore the impact of time on the effectiveness of federated crowdsourcing and fail to respond to the instability of participants and networks. The eFedCrowd proposed in this paper sets a time threshold to assign the data freshness level and task completion time to the determination of task publication. Furthermore, the contribution is measured against the accuracy of the client's local training model, and the rewards are distributed fairly in an economical manner, so as to motivate workers to complete tasks efficiently and with high quality.

References

Cheng Ye, Joseph Coco, Anna Epishova, Chen Hajaj, Henry Bogardus, Laurie Novak, Joshua Denny, Yevgeniy Vorobeychik, Thomas Lasko, Bradley Malin, et al. A crowdsourcing framework for medical data sets. *AMIA Summits on Translational Science Proceedings*, 2018:273, 2018.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016.

Mohamed Amgad, Lamees A Atteya, Hagar Hussein, Kareem Hosny Mohammed, Ehab Hafiz, Maha AT Elsebaie, Ahmed M Alhusseiny, Mohamed Atef AlMoslemany, Abdelmagid M Elmatboly, Philip A Pappalardo, et al. Nucls: A scalable crowdsourcing approach and dataset for nucleus classification and segmentation in breast cancer. *GigaScience*, 11, 2022.

Huichuan Xia and Brian McKernan. Privacy in crowdsourcing: a review of the threats and challenges. *Computer Supported Cooperative Work (CSCW)*, 29: 263–301, 2020.

Layla Pournajaf, Li Xiong, Vaidy Sunderam, and Slawomir Goryczka. Spatial task assignment for crowd sensing with cloaked locations. In *2014 IEEE 15th International Conference on Mobile Data Management*, volume 1, pages 73–82. IEEE, 2014.

Yanbing Ren, Xinghua Li, Yinbin Miao, Bin Luo, Jian Weng, Kim-Kwang Raymond Choo, and Robert H Deng. Towards privacy-preserving spatial distribution crowdsensing: A game theoretic approach. *IEEE Transactions on Information Forensics and Security*, 17:804–818, 2022.

Leye Wang, Gehua Qin, Dingqi Yang, Xiao Han, and Xiaojuan Ma. Geographic differential privacy for mobile crowd coverage maximization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Jianhao Wei, Yaping Lin, Xin Yao, and Jin Zhang. Differential privacy-based location protection in spatial crowdsourcing. *IEEE Transactions on Services Computing*, 15(1):45–58, 2019.

Leye Wang, Han Yu, and Xiao Han. Federated crowdsensing: framework and challenges. *arXiv preprint arXiv:2011.03208*, 2020a.

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.

Leye Wang, Han Yu, and Xiao Han. Federated crowdsensing: framework and challenges. *arXiv preprint arXiv:2011.03208*, 2020b.

Dashan Gao, Xin Yao, and Qiang Yang. A survey on heterogeneous federated learning. *arXiv preprint arXiv:2210.04505*, 2022.

Guoxian Yu, Jinzheng Tu, Jun Wang, Carlotta Domeniconi, and Xiangliang Zhang. Active multilabel crowd consensus. *IEEE Transactions on Neural Networks and Learning Systems*, 32(4):1448–1459, 2020.

Jinzheng Tu, Guoxian Yu, Carlotta Domeniconi, Jun Wang, Guoqiang Xiao, and Maozu Guo. Multi-label crowd consensus via joint matrix factorization. *Knowledge and Information Systems*, 62:1341–1369, 2020.

Ming Wu, Qianmu Li, Muhammad Bilal, Xiaolong Xu, Jing Zhang, and Jun Hou. Multi-label active learning from crowds for secure iiot. *Ad Hoc Networks*, 121:102594, 2021a.

Jing Zhang, Ming Wu, Cangqi Zhou, and Victor S Sheng. Active crowdsourcing for multilabel annotation. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Ziyuan Li, Jian Liu, Jialu Hao, Huimei Wang, and Ming Xian. Crowdsfl: A secure crowd computing framework based on blockchain and federated learning. *Electronics*, 9(5):773, 2020a.

Bekir Sait Ciftler, Abdullatif Albaseer, Noureddine Lasla, and Mohamed Abdallah. Federated learning for rss fingerprint-based localization: A privacy-preserving crowdsourcing method. In *2020 International Wireless Communications and Mobile Computing (IWCMC)*, pages 2112–2117. IEEE, 2020.

Chen Zhang, Yu Guo, Xiaohua Jia, Cong Wang, and Hongwei Du. Enabling proxy-free privacy-preserving and federated crowdsourcing by using blockchain. *IEEE Internet of Things Journal*, 8(8):6624–6636, 2021a.

Yufeng Zhan, Peng Li, Song Guo, and Zhihao Qu. Incentive mechanism design for federated learning: Challenges and opportunities. *IEEE Network*, 35(4): 310–317, 2021a.

Lingjuan Lyu, Han Yu, Jun Zhao, and Qiang Yang. Threats to federated learning. *Federated Learning: Privacy and Incentive*, pages 3–16, 2020.

Anshuman Suri, Pallika Kanani, Virendra J Marathe, and Daniel W Peterson. Subject membership inference attacks in federated learning. *arXiv preprint arXiv:2206.03317*, 2022.

Viraaji Mothukuri, Reza M Parizi, Seyedamin Pouriyeh, Yan Huang, Ali Dehghantanha, and Gautam Srivastava. A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115:619–640, 2021.

Ji Liu, Jizhou Huang, Yang Zhou, Xuhong Li, Shilei Ji, Haoyi Xiong, and Dejing Dou. From distributed machine learning to federated learning: A survey. *Knowledge and Information Systems*, 64(4):885–917, 2022.

Li Li, Yuxi Fan, Mike Tse, and Kuo-Yi Lin. A review of applications in federated learning. *Computers & Industrial Engineering*, 149:106854, 2020b.

Yufeng Zhan, Peng Li, Zhihao Qu, Deze Zeng, and Song Guo. A learning-based incentive mechanism for federated learning. *IEEE Internet of Things Journal*, 7(7):6360–6368, 2020a.

Yufeng Zhan, Jie Zhang, Zicong Hong, Leijie Wu, Peng Li, and Song Guo. A survey of incentive mechanism design for federated learning. *IEEE Transactions on Emerging Topics in Computing*, 10(2):1035–1044, 2021b.

Beibei Li, Yaxin Shi, Qinglei Kong, Qingyun Du, and Rongxing Lu. Incentive-based federated learning for digital twin driven industrial mobile crowdsensing. *IEEE Internet of Things Journal*, 2023.

Tao Li and Suresh P Sethi. A review of dynamic stackelberg game models. *Discrete & Continuous Dynamical Systems-B*, 22(1):125, 2017.

Yiming Wu, Shaohua Tang, Bowen Zhao, and Zhiniang Peng. Bptm: Blockchain-based privacy-preserving task matching in crowdsourcing. *IEEE access*, 7:45605–45617, 2019.

Junwei Zhang, Fan Yang, Zhuo Ma, Zhuzhu Wang, Ximeng Liu, and Jianfeng Ma. A decentralized location privacy-preserving spatial crowdsourcing for internet of vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 22(4):2299–2313, 2020.

Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pages 1–12. Springer, 2006.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32, 2019.

Jiangang Shu, Xiaohua Jia, Kan Yang, and Hua Wang. Privacy-preserving task recommendation services for crowdsourcing. *IEEE Transactions on Services Computing*, 14(1):235–247, 2018.

Shailja Joshi, Hemraj Saini, and Geetanjali Rathee. Salt cryptography for privacy in mobile crowdsourcing. *International Journal of Information Technology*, 12:585–591, 2020.

Chuan Zhang, Liehuang Zhu, Chang Xu, Xiaojiang Du, and Mohsen Guizani. A privacy-preserving traffic monitoring scheme via vehicular crowdsourcing. *Sensors*, 19(6):1274, 2019.

Chao Lin, Debiao He, Sherali Zeadally, Neeraj Kumar, and Kim-Kwang Raymond Choo. Secbcs: a secure and privacy-preserving blockchain-based crowdsourcing system. *Science China Information Sciences*, 63:1–14, 2020.

Zhibo Wang, Jingxin Li, Jiahui Hu, Ju Ren, Zhetao Li, and Yanjun Li. Towards privacy-preserving incentive for mobile crowdsensing under an untrusted platform. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 2053–2061. IEEE, 2019.

Chenhan Zhang, Yuanshao Zhu, Christos Markos, Shui Yu, and JQ James. Toward crowdsourced transportation mode identification: A semisupervised federated learning approach. *IEEE Internet of Things Journal*, 9(14):11868–11882, 2021b.

Ziyuan Li, Jian Liu, Jialu Hao, Huimei Wang, and Ming Xian. Crowdsfl: A secure crowd computing framework based on blockchain and federated learning. *Electronics*, 9(5):773, 2020c.

Bowen Zhao, Ximeng Liu, and Wei-neng Chen. When crowdsensing meets federated learning: Privacy-preserving mobile crowdsensing system. *arXiv preprint arXiv:2102.10109*, 2021.

Yufeng Zhan, Peng Li, Zhihao Qu, Deze Zeng, and Song Guo. A learning-based incentive mechanism for federated learning. *IEEE Internet of Things Journal*, 7(7):6360–6368, 2020b.

Tra Huong Thi Le, Nguyen H Tran, Yan Kyaw Tun, Minh NH Nguyen, Shashi Raj Pandey, Zhu Han, and Choong Seon Hong. An incentive mechanism for federated learning in wireless cellular networks: An auction approach. *IEEE Transactions on Wireless Communications*, 20(8):4874–4887, 2021.

Jingwen Zhang, Yuezhou Wu, and Rong Pan. Incentive mechanism for horizontal federated learning based on reputation and reverse auction. In *Proceedings of the Web Conference 2021*, pages 947–956, 2021c.

Maoqiang Wu, Dongdong Ye, Jiahao Ding, Yuanxiong Guo, Rong Yu, and Miao Pan. Incentivizing differentially private federated learning: A multidimensional contract approach. *IEEE Internet of Things Journal*, 8(13):10639–10651, 2021b. doi: 10.1109/JIOT.2021.3050163.

Shashi Raj Pandey, Nguyen H Tran, Mehdi Bennis, Yan Kyaw Tun, Zhu Han, and Choong Seon Hong. Incentivize to build: A crowdsourcing framework for federated learning. In *2019 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2019.

Yuxi Zhao, Xiaowen Gong, and Shiwen Mao. Truthful incentive mechanism for federated learning with crowdsourced data labeling. *arXiv preprint arXiv:2302.00106*, 2023.