Oblivious Federated Analytics for Mobile Devices

Donghyun Sohn, Northwestern University, donghyun.sohn@u.northwestern.edu

As the volume of data generated by mobile devices surges, the need for efficient data analysis while preserving privacy is increasingly urgent. With federated analytics [4], edge devices perform distributed data analysis without participants uploading their records to a trusted third party. It presents a compelling solution to this quandary. For example, Google's "Now Playing" feature on mobile enables their servers to refine their song recommendations by securely aggregating over the listening habits of individual users [6]. Here, the user's phone locally recognizes and records the songs they are hearing. It periodically encrypts their listening habits, and sends them to the a centralized server for analysis with those of other users. Although federated analytics offers efficient privacy-preserving data analysis, it inherently permits centralized servers to learn these results. The Fundamental Law of Information Recovery [1] indicates that revealing these raw statistics leaves users vulnerable to reconstruction attacks. We posit that this personalization is possible with better security by pushing this computation into mobile devices.

Secure multi-party computation (MPC) enables users to securely aggregate the union of their data without relying on a trusted third party. It guarantees, within the semi-honest model, that participants only learn what is revealed by the final aggregated result without gaining additional knowledge about each other's data. To advance federated analytics under MPC, we draw inspiration from SMCQL [2], refining our approach to maximize pre-computation on local devices prior to entering MPC. Devices start by aggregating their data locally. Users then secret-share their data, and these secret shares serve as input to a query's MPC phase. This cryptographic protocol readies the data for evaluation under MPC — it is analogous to each site having a copy of the unioned input data but none having enough information to decrypt it. Within this MPC framework, devices execute cryptographic protocols collaboratively, computing the final aggregate without revealing their own aggregated data to others. Once this process is complete, our framework shares the initial aggregate results with participating devices, subsequently leveraging these shared results and their local data to personalize recommendations. Our methodology combines MPC with post-processing for local personalization, setting it apart from OLAP querying using MPC like SMCQL.

Furthermore, we are tailoring MPC protocols to a continuous querying environment, where the same queries are periodically executed on a data stream instead of persistent tables alone. Recognizing the limited computational capacity of edge devices [3], our approach optimizes partial aggregation through adaptive sampling and incremental aggregation techniques. This allows for selective querying based on data relevance and likelihood of change, substantially reducing local computational overhead. Despite potential hints from data selection patterns, information leakage in this context is acceptable, as these steps are prior to secret sharing. Our approach pioneers the integration of MPC into continuous query processing over data streams, a realm not explored by prior approaches such as SMPAI [5], which introduced MPC to federated learning but did not support continuous querying environment.

By integrating MPC and local data aggregation within federated analytics for edge devices in a continuous querying environment, we will eliminate the need to reveal the results of federated analytics to a centralized server. This effort opens pathways for secure and efficient federated analytics across edge devices. Furthermore, leveraging aggregated data improves models for personalization without the use of a trusted third party.

[1] John Abowd, Lorenzo Alvisi, Cynthia Dwork, et al. Privacy-preserving data analysis for the federal statistical agencies. *arXiv preprint arXiv:1701.00752*, 2017.

[2] Johes Bater, Gregory Elliott, Craig Eggen, Satyender Goel, Abel Kho, and Jennie Rogers. SMCQL: Secure Querying for Federated Databases. *VLDB*, 10(6):673–684, 2016.

[3] Keyan Cao, Yefan Liu, Gongjie Meng, and Qimeng Sun. An overview on edge computing research. *IEEE Access*, 8:85714–85728, May 2020.

[4] Ahmed Ramzy Elkordy, Youssef H Ezzeldin, and Shuguang Han. Federated analytics: A survey. *IEEE Transactions on Signal and Information Processing over Networks*, 2023.

[5] Vaikkunth Mugunthan, Antigoni Polychroniadou, David Byrd, and Tucker Hybinette Balch. Smpai: Secure multi-party computation for federated learning. In *Proceedings of the NeurIPS 2019 Workshop on Robust AI in Financial Services*. jpmorgan.com, 2019.

[6] Daniel Ramage and Stefano Mazzocchi. Federated analytics: Collaborative data science without data collection. https://blog.research.google/2020/05/federated-analytics-collaborative-data.html, 2020.