Practical Fully-Decentralized Secure Aggregation for Personal Data Management Systems

Julien Mirval, Luc Bouganim, Iulian Sandu Popa

Practical Fully-Decentralized Secure Aggregation for Personal Data Management Systems

Julien Mirval
julien.mirval@cozycloud.cc
Cozy Cloud
Inria-Saclay
UVSQ, Université Paris-Saclay
France

Luc Bouganim
luc.bouganim@inria.fr
Inria-Saclay
UVSQ, Université Paris-Saclay
France

Iulian Sandu-Popa
iulian.sandu-popa@uvsq.fr
UVSQ, Université Paris-Saclay
Inria-Saclay
France

Personal Data Management Systems (PDMS) are flourishing, boosted by legal and technical means like smart disclosure, data portability and data altruism. A PDMS allows its owner to easily collect, store and manage data, directly generated by her devices, or resulting from her interactions with companies or administrations. PDMSs unlock innovative usages by crossing multiple data sources from one or many users, thus requiring aggregation primitives. Indeed, aggregation primitives are essential to compute statistics on user data, but are also a fundamental building block for machine learning algorithms. This paper proposes a protocol allowing for secure aggregation in a massively distributed PDMS environment, which adapts to selective participation and PDMSs characteristics, and is reliable with respect to failures, with no compromise on accuracy. Preliminary experiments show the effectiveness of our protocol which can adapt to several contexts with varying PDMSs characteristics in terms of communication speed or CPU resources and can adjust the aggregation strategy to the estimated selective participation.

The new privacy-protection regulations (e.g., GDPR) and smart disclosure initiatives in the last decade have boosted the development and adoption of Personal Data Management Systems (PDMSs) [1]. A PDMS (e.g., Cozy Cloud, Nextcloud, Solid) is a data platform allowing users to easily collect, store and manage into a single place data directly generated by user devices (e.g., quantified-self data, smart home data, photos) and data resulting from user interactions (e.g., social interaction data, health, bank, telecom). Users can then leverage the power of their PDMS to benefit from their personal data for their own good and in the interest of the community [2].

Consequently, the PDMS paradigm leads to an important shift in the personal data ecosystem since data becomes massively distributed, at the user-side. It also holds the promise of unlocking innovative usages. An individual can now cross her data from different data silos, e.g., health records and physical activity data. Moreover, individuals can cross data within large communities of users, e.g., to compute statistics for epidemiological studies or to train a machine learning model (ML) for recommender systems or automatic classification of user data. However, these exciting perspectives should not eclipse the security issues –user data must be kept private– and the right for any PDMS user to consent, or not, in participating in each computation.

Aggregation primitives (e.g., sum or average) are obviously essential to compute basic statistics on user data but are also a fundamental building block for machine learning algorithms. Thus, to enable such new usages, we need scalable, privacy-preserving protocols implementing data aggregation primitives with selective (i.e., consenting) participants. Ideally, the proposed protocol should provide an accurate result that fully takes advantage of high-quality data available in PDMSs. Efficiency (i.e., protocol latency and total load of the system) is of prime importance and the protocol should adapt to several contexts: the PDMSs could be limited by their communication speed or by their computation power. Finally, given the scale of such decentralized aggregation, such protocols must also be robust to node failures. To summarize, our goal is to propose an aggregation protocol for basic aggregate functions that fulfills the following properties:

- *fully decentralized and highly scalable*, with the number of participants.
- *privacy-preserving*, i.e., it protects the confidentiality of user data.
- *accurate*, i.e., it does not require a trade-off between accuracy and privacy.
- *adaptable*, i.e, it can adapt to a large spectrum of computation selectivity values (reflecting the subset of contributor nodes) and system configurations (network and cryptographic latency).
- *reliable*, i.e., it handles node failures or voluntary disconnections.

REFERENCES

[1] Nicolas Anciaux, Philippe Bonnet, Luc Bouganim, Benjamin Nguyen, Philippe Pucheral, Iulian Sandu Popa, and Guillaume Scerri. 2019. Personal data management systems: The security and functionality standpoint. *Information Systems* 80 (2019), 13–35.

[2] EU Commission. 25 October 2020. Proposal for a Regulation on European data governance (Data Governance Act), COM/2020/767. [eur-lex].