12-1-2019

How to Encourage Sharing Sensitive Data for Large-Scale Research through Blockchain and Smart Contract

Peter Golubtsov
Lomonosov Moscow State University, pgolubtsov@hse.ru

How to Encourage Sharing Sensitive Data for Large-Scale Research through Blockchain and Smart Contract

Peter Golubtsov

Lomonosov Moscow State University, Moscow, Russia;

National Research University Higher School of Economics, Moscow, Russia;

pgolubtsov@hse.ru

Currently, a huge amount of data is being generated routinely, which opens up opportunities for large-scale medical research, in particular, for revealing hidden patterns that are not detected in small or fragmented data sets. Such data includes standard Electronic Medical Records (EMRs) collected during physical examinations and tests; Personal Health Records (PHRs) generated by wearable devices; DNA sequencing results, etc. As a rule, these data are stored in the institutions on the basis of which they were collected. Often, arrays of such data are used in statistical analysis to identify hidden patterns. However, since the data sets at the disposal of individual research groups are relatively small, the results of such studies often have low statistical reliability. To increase the level of reliability, research groups can combine the data at their disposal, or even sell them, however, the required "anonymization" of the data can lead to a serious loss of information. In addition, such data transfers may be of concern to patients, as it could potentially violate their confidentiality. All these factors dramatically reduce the availability of such data for research. Moreover, customers who are actually data sources are not interested in the availability of their confidential data.

It seems that considerations of the secrecy of personal data makes it impossible for them to be used on a large scale in multiple research projects, because as soon as the data becomes available to the members of the research group, there is the possibility of leakage.

In this research, we propose a model for handling sensitive personal (in particular, medical) data, which is designed to eliminate the contradictions noted and to increase the availability of personal data for research while maintaining a high level of security in the analysis of these data. This model relies on blockchain, smart contract and Trusted Execution Environment (TEE) technologies to ensure that no sensitive information can be revealed during the whole processing. Besides, every instance of private personal data should be sufficiently "diluted" and cannot be extracted from the "average" research results. To verify that the program code conforms the dilution condition, it must be open and accessible for free study.

We also discuss ways to encourage data donors to share their personal data for large-scale research, providing them not only monetary compensation, but also generating encrypted individual health reports. The ability to repeatedly sell their data and receive personal reports and recommendations will motivate patients to invest in their own medical research. More complete data will not only be much higher evaluated, but also provide more detailed information about their own health.

The study also considers the possibilities of scaling and parallelizing data processing both at the stage of their aggregation and at the stage of preparing personal reports and identifying anomalies. It is important to emphasize that all calculations should be carried out in trusted computing environments. Data and intermediate results should be transmitted to the computing agents in encrypted form and can only be decrypted by the keys of the corresponding agents inside isolated enclaves.

Keywords: big data; health data; blockchain; smart contract; trusted execution environment; dilution of personal data; data ownership; privacy protection; data sharing; parallel processing.