The recent technology developments enable millions of people to collect and share data on a massive scale. Such data allow deriving relevant information about people through advanced analytics such as statistical analysis or machine learning. The analytical findings can help companies improve their customer services, or hospitals identify patterns based on patients' historical data and come up with early treatments.

Unfortunately, this new data collection paradigm raises serious privacy concerns mainly because of the high sensitivity of the collected data.
Moreover, end-users are usually even not aware of the collection of their personal information. Therefore, there is a strong need for protecting the systems collecting and processing/mining this large amount of data and keeping these data confidential against unauthorized parties.

While data encryption techniques would help ensure data privacy, traditional symmetric encryption mechanisms would unfortunately fail in addressing the need for allowing the search or some other advanced analytics such as machine learning over the collected data. Despite some very promising recent advancements in fully-homomorphic encryption which enables any arbitrary operations over encrypted data, this technology is not ready yet to be applied as a general purpose solution to the problem of privacy preserving data analytics.

One of the first data analytics operations researchers focused on was the computation of the sum over encrypted data, which is usually called privacy-preserving data aggregation. Privacy preserving data aggregation consists of the collection of data by an Aggregator from several users and computing the aggregate sum without having access to the individual data originating from each user. This has led to an active area of research whereby proposed solutions mainly rely on partially homomorphic encryption such as Elgamal or symmetrically HE and ensure aggregator obliviousness either using a trusted key dealer or later without any trusted infrastructure; in these solutions, each data source receives a partial encryption key and the aggregator who receives individually encrypted data, holds the decryption key only and hence can only decrypt the aggregate.
Recently, researchers focus on more complex operations than the basic sum. Machine learning algorithms are more and more used in the context of data analytics since they solve many problems faced nowadays by businesses and institutions such as prediction (where from a set of training data, the machine estimates possible values for some output
data) or classification (which is the process of labelling some data based on previously trained data whose labels are known). Recently, neural networks (NN), have been considered as the one of the most useful machine learning tools. The main challenge in designing privacy-preserving neural network is the combination of linear and non-linear operations over the data that have to be protected. While early solutions leave the server with the execution of the linear operations over encrypted data and require the data owner to perform locally the non-linear operations over intermediary decrypted data, current solutions apply approximation of the non-linear functions (the activation functions, like sigmoid or hyperbolic tangent) with low-degree polynomial over which techniques such as HE can be performed. This causes a non-negligible defect on the accuracy of the classification.

Research plan

The goal of the PhD is to design and evaluate customized privacy preserving and security primitives that will on the one hand protect the confidentiality of the data and on the other hand enable data centers to perform these data mining or machine learning techniques over the encrypted data. To this end, the successful candidate will first study the privacy and security challenges associated with Big Data

applications leveraging different data mining techniques such as statistical data analysis and/or machine learning. The PhD candidate will further investigate privacy preserving variants of some specific data mining techniques while leveraging more practical homomorphic encryption solutions that are not fully homomorphic but can support some operations (such as addition only). These operations will be tailored to improve the efficiency of the underlying primitives while not sacrificing their accuracy. Another possible approach is to use secure multiparty computation (SMC) which can be used for the protection of data both at the collection and processing phases. Parties involved in the SMC protocol (such as different hospitals) will be able to perform collaborative machine learning over the entire dataset and retrieve the desired results without revealing any information on their respective datasets (patients' health records, eg.). Moreover, the PhD candidate will also focus on the multi-user case where data is coming from distinct independent sources and during the analysis of the data these sources only discover the output of the data mining algorithm and do not learn any information about each other's' data.