PAPAYA: A Platform for Privacy Preserving Data Analytics

by Eleonora Ciceri (MediaClinics Italia), Marco Mosconi (MediaClinics Italia), Melek Önen (EURECOM) and Orhan Ermis (EURECOM)

The PAPAYA project is developing a dedicated platform to address privacy concerns when data analytics tasks are performed by untrusted data processors. This platform regrouping will allow stakeholders to ensure their clients' privacy and comply with the General Data Protection Regulation (GDPR) [L1] while extracting valuable and meaningful information from the analysed data. PAPAYA targets two digital health use cases, namely arrhythmia detection and stress detection, whereby patients' data are protected through dedicated privacy enhancing technologies.

Recent advances in information technology make it easy for businesses and other organisations to collect large amounts of data and use data analytics techniques to derive valuable information and improve predictions. The information obtained, however, is usually sensitive, and may endanger the privacy of data subjects. Whilst the General Data Protection Regulation (GDPR) necessitates a technological means to protect privacy, it is vital that this is achieved in a way that still allows healthcare stakeholders to extract meaningful information and make good predictions (e.g., about diseases). The PAPAYA project aims to provide solutions that minimise privacy risks while increasing trust in third-party data processors and the utility of the underlying analytics.

The newly developed PAPAYA platform will integrate several privacy-preserving data analytics modules, ensuring compliance with the GDPR. The project considers different settings involving various actors (single/multiple data sources, queriers) and ensuring different privacy levels. The project will facilitate user experience for data subjects while providing transparency and control measures.

The PAPAYA project focuses on three main data analytics techniques, namely, neural networks (training and classification), clustering, and basic statistics (counting) and aims at developing their privacy - preserving variants while optimising the resulting performance overhead and assuring an acceptable utility/accuracy. More specifically, privacy-preserving neural networks (inspired by the architecture of neurons in human brains) learn prediction models about a certain characteristic/capability using some test datasets and further apply this model over new data to make accurate predictions while keeping the input data confidential. On the other hand, privacy-preserving clustering algorithms allow data owners to group similar (but confidential) data objects in clusters. Finally, privacy-preserving counting primitives enable parties to encrypt one or several datasets related to individuals and further count the number of individuals in the set. The main cryptographic tools that will be used to design these new solutions are homomorphic encryption, secure multiparty computation, differential privacy and functional encryption.

Privacy-preserving neural networks for two digital health use cases

The PAPAYA project defines two digital health use cases, namely privacy-preserving arrhythmia detection and privacy-preserving stress detection. While both use cases rely on neural networks, the former (arrhythmia detection) only considers the classification phase and the latter (stress detection) involves multiple data sources, such as hospitals, that collaboratively train a stress detection neural network model. Both use cases and the underlying PAPAYA solutions are summarised below.

Privacy-preserving arrhythmia detection

This use case targets scenarios whereby patients need to perform cardiac parameters analyses with the goal of verifying the presence/absence of arrhythmia. The patient wears a device that collects his/her ECG data for a fixed amount of time (e.g., 24 hours). Once the patient returns the device to the pharmacy, the ECG data are protected and submitted to the PAPAYA platform, as illustrated in Figure 1. The data are then analysed to predict whether the patient suffers from arrhythmia.

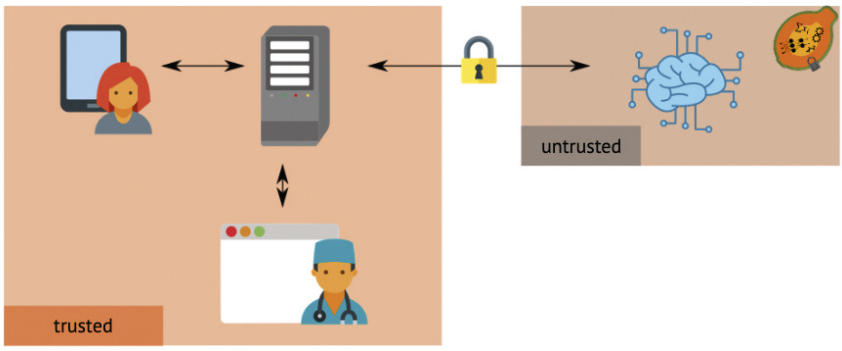

Figure 1: A patient's ECG data are collected by a pharmacist, sent to a trusted cloud, protected and submitted to the PAPAYA platform to predict arrhythmia. The detected arrhythmia are used by a cardiologist to redact the patient's report.

The project aims to develop a privacy-preserving classification solution whereby the neural network model is executed over confidential data. These solutions use advanced cryptographic schemes such as homomorphic encryption [1] or secure multiparty computation [2]. The main challenge in using such tools is the complexity of the neural network in terms of size and the underlying operations. Therefore,

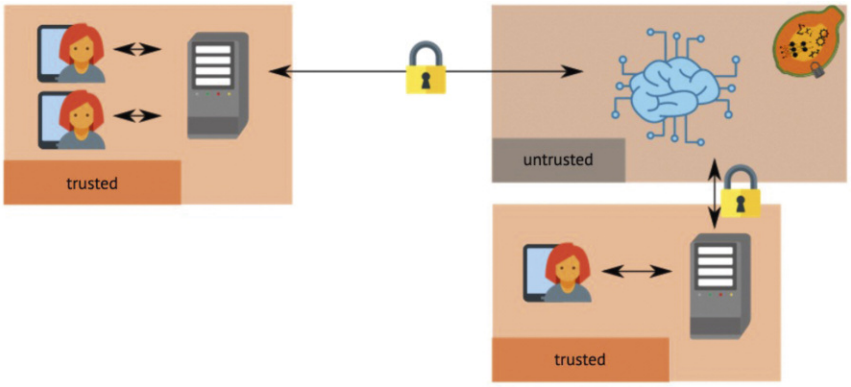

Figure 2: Health-related parameters are collected from workers, aggregated locally and outsourced to train a collaborative model, which can be later used to perform real-time detection of stress and anxiety conditions.

PAPAYA will use these advanced cryptographic tools once the original neural network is modified in order to make it compatible with the actual cryptographic tool (for example, complex operations are approximated to low degree polynomials). This modified neural network will still maintain a good level of accuracy.

Privacy-preserving stress management

This use case targets workers who suffer from stress. It would be very helpful to have an automatic solution that would help anxious and stressed people to recognise symptoms at their onset and suggest mitigation strategies to help the person take preventative action and keep stress levels in check. To this end, sensitive health data from IoT sensors are collected by multiple sources and used to train a collaborative model via the PAPAYA platform as shown in Figure 2, with the goal of automatically detecting stress conditions in workers.

As a potential solution for this use case, we are studying the problem of privacy-preserving collaborative training based on differential privacy [3] involving many data owners who need to jointly construct a neural network model. Differential privacy prevents participants' individual datasets from being leaked, but allows the joint model to be computed.

This project is a joint work of the PAPAYA project consortium. The PAPAYA project is funded by the H2020 Framework of the European Commission under grant agreement no. 786767. In this project, six renowned research institutions and industrial players with balanced expertise in all technical aspects of both applied cryptography, privacy and machine learning are working together to address the challenges of the project: EURECOM (project coordinator), Atos Spain, IBM Research Israel, Karlstad University Sweden, MediaClinics Italia and Orange France.

Link:

[L1] https://kwz.me/hyK

References:

[1] R. Rivest, L. Adleman and M. L. Dertouzos: "On data banks and privacy homomorphisms," Foundations of secure computation, pp. 169--180, 1978.

[2] A. Chi-Chih Yao: "Protocols for secure computations" (extended abstract), in 23rd Annual Symposium on Foundations of Computer Science, Chicago, Illinois, USA, 1982, p.160–164, n.a., 1982. IEEE Computer Society.

[3] C. Dwork: "Differential privacy", in Proc. of the 33rd International Conference on Automata, Languages and Programming - Volume Part II, ICALP'06, pages 1–12, Springer, 2006.

Please contact:

Orhan Ermis
EURECOM, France
orhan.ermis@eurecom.fr

Resilient Network Services for Critical mHealth Applications over 5G Mobile Network Technologies

by Emmanouil G. Spanakis and Vangelis Sakkalis (FORTH-ICS)

DAPHNE is aiming to develop a resilient networking service for critical related applications, as a novel approach for next generation mHealth information exchange. Our goal is to provide in-transit persistent information storage, allowing the uninterruptible provision of crucial services. Our system will overcome network instabilities, capacity efficiency problems, incompatibilities, or even absence of end-to-end homogeneous connectivity, with an emphasis on future networks and services (i.e. 5G). We aim to provide a set of tools for the appropriate management of communication networks during their design time and avoid the "build it first, manage later" paradigm.

Future mHealth informatics rely on innovative technologies and systems for transparent and continuous collection of evidence-based medical information at anytime, anywhere, regardless of coverage and availability of communication means. Such an emerging critical infrastructure is influenced by factors such as biomedical and clinical incentives, advances in mobile telecommunications, information technology developments, and the socioeconomic environment. This cross dependency has led to concerns about reliability and resilience of current network deployments, hence it is imperative that communication networks be designed to adequately respond to failures, especially in cloud, mobile and Internet Of Things (IoT) / Web Of Things (WoT) environments that have traditional boundaries.