Workshop on Privacy Challenges in Public and Private Organizations

Alessandra Bagnato, Paulo Silva, Ala Sarah Alaqra, Orhan Ermis

Workshop on Privacy Challenges in Public and Private Organizations

Alessandra Bagnato[1], Paulo Silva[2], Ala Sarah Alaqra[3], and Orhan Ermis[4]

[1] SOFTEAM R&D Department, Guyancourt, France
alessandra.bagnato@softeam.fr
[2] CISUC, Department of Informatics Engineering, University of Coimbra, Portugal
pmgsilva@dei.uc.pt
[3] Karlstad University, Sweden
alaa.alaqra@kau.se
[4] EURECOM, France
orhan.ermis@eurecom.fr

Abstract. Recent developments in information technology such as the Internet of Things and the cloud computing paradigm enable public and private organisations to collect large amounts of data to employ various data analytic techniques for extracting important information that helps improve their businesses. Unfortunately, these benefits come with a high cost in terms of privacy exposures given the high sensitivity of the data that are usually processed at powerful third-party servers. Given the ever-increasing of data breaches, the serious damage they cause, and the need for compliance to the European General Data Protection Regulation (GDPR), these organisations look for secure and privacy-preserving data handling practices. During the workshop, we aimed at presenting an approach to the problem of user data protection and control, currently being developed in the scope of the PoSeID-on and PAPAYA H2020 European projects.

Keywords: Privacy-enhancing Dashboard · Privacy-preserving Data Analytics · GDPR

1 Introduction

Several European projects address the topical area of privacy, data protection, and digital identities. PAPAYA [1], and PoSeID-on [2], described next, are exploring ways of cooperation to enhance privacy assurances to end-users.

The goal of the PAPAYA project is to devise and develop a platform of privacy-preserving modules that protects the privacy of users on an end-to-end basis without sacrificing data analytics functionalities. The PAPAYA platform will integrate several privacy-preserving data analytics modules each of them dedicated to specific analytics operations and specific settings (e.g., single data owner, multiple data owners). The platform aims to be usable in the sense that it also includes proper transparency and control mechanisms through a dashboard.

The PoSeID-on solution is based on innovative technologies such as blockchain, cloud and smart contracts. These technologies provide targeted benefits for end-users, enabling them to manage personal data and data access authorizations easily, securely and independently. This helps both public and private entities identify new business opportunities, be compliant with the GDPR while processing personal data, as well as undergo a substantial ICT-driven transformation, which will ensure higher security of end-users' data. PoSeID-on also impacts society as a whole, as it leads to increased trust in the digital single market, in addition to supporting fundamental rights in the digital society. The project will be evaluated through four different pilot studies (in Italy, France, Spain and Malta) that will test its functionalities in public, private and mixed contexts. Initially, pilots will involve a basic set of users to be enlarged during the evaluation months. The pilots (described in Deliverable 2.1 [5]) will run in a controlled environment to simulate real-life services and conditions.

2 Motivation and Objectives

Recent advances in information technology such as the Internet of Things and/or the cloud computing paradigm enable public and private organisations to collect large amounts of data and use advanced techniques to infer valuable insights and improve their businesses. Unfortunately, these benefits come with a high cost in terms of privacy exposures given the high sensitivity of the data that are usually analysed/processed at powerful third-party servers. Given the ever-increasing of data breaches, the serious damage they cause, and the need for compliance to the European General Data Protection Regulation (GDPR), these organisations look for secure and privacy-preserving data handling practices.

Both projects decided it was worth to get feedback on the work done so far. Not only bilaterally, but also from all the workshop attendees. During the workshop, we presented our approach to the problem of user data protection and control that is currently being developed in the scope of the PoSeID-on H2020 European project. The presented solution complies with EU's General Data Protection Regulation and explores the use of Blockchain technology to provide data transactions protection and accountability, as well as full control of users over their data.

The goal of this workshop was to identify and present the privacy challenges related to data analysis by public and private organisations and to encompass research advances in the privacy-enhancing technologies that will enable privacy-preserving data management and GDPR compliance. Moreover, the workshop served as a discussion environment for those familiar with cryptographic tools, and discuss possible concerns and risks when it comes to applying such tools in different areas when data is critical and sensitive. We intended to understand and shed light on the mental models, trust factors, and the possible risks and

[5] https://www.poseidon-h2020.eu/documents/d2-1-use-case-analysis-and-user-scenarios/

concerns when it comes to data analysis on the protected data and how these discussions might be used to foster collaboration among potential privacy-enhancing technology outputs of PoSeID-on and PAPAYA projects.

3 Workshop Format

The allocated time for the workshop was two hours. The objective was to first introduce both projects and then focus on the specific objectives and challenges. It was possible to present and discuss the previously agreed topics. The workshop was organized as follows:

- **Opening (5 min.):**
  The workshop program presentation (Orhan Ermis).
- **PoSeID-on Project Presentation (20 min.):**
  The presentation introduced the PoSeID-on H2020 EU project Project [6] (Alessandra Bagnato).
- **PoSeID-on Project Dashboard Demo (10 min.):**
  Offered an insight about the dashboard and its functionalities (Paulo Silva).
- **Papaya Presentation (25 min.):**
  The presentation introduced the Papaya Project [7] H2020 EU project (Orhan Ermis).
- **The End User requirements in PAPAYA Presentation (10 min.):**
  Highlighted human aspects and results from end-user studies (Ala Sarah Alaqra).
- **Questionnaire Session on Human Factors (10 min.):**
  Participants have provided feedback on PAPAYA e-Health use cases (Ala Sarah Alaqra).
- **Discussion Session (25 min.):**
  Questions and debate regarding the use cases and respective approaches. Each speaker had a set of comments/questions to spark the discussion. Nevertheless, the discussion was naturally flowing in interventions from the audience.
- **PAPAYA-PoSeID-on collaboration (15 min.):**
  A discussion with the participants from both projects to describe the needs of the dashboard and the needs of the analysis. More specifically, answering the questions "What are PAPAYA's needs with respect to the PoSeID-on's Dashboard?" and "What are PoSeID-on's needs with respect to Papaya's analytics?" was the objective of the discussion.

4 Lessons Learned from the PoSeID-on Project Perspective

In this session of the workshop, the PoSeID-on general objectives, the PoSeID-on architecture and the privacy challenges in one PoSeID-on Public and one PoSeID-on Private Organization were described.

[6] https://www.poseidon-h2020.eu/
[7] https://www.papaya-project.eu

4.1 Use Cases

The first presented PoSeID-on public organization use case and the respective challenge was related to the General Administration, Personnel and Services Department (DAG) of the Italian Ministry of Economy and Finance (MEF). The MEF is in charge of the management of payroll functions for approximately 2.1 million Italian public sector employees. Such service is provided through a unique payroll function, NoiPA, which annually manages more than €51B in payments. NoiPA is a portal created to manage administrative and economic data of central and peripheral Public Administration employees. Therefore, NoiPA has a big experience in personal data management and it could be very close to PoSeID-on project because this platform aims to collect the users' given authorizations (for sharing personal information) that are stored in the platform itself.

The second presented PoSeID-on Private sector organization use case was related to Softeam, a private French software vendor with about 1000 employees. Softeam develops a software called e-Citiz – a platform for Business Process Management for both e-government and companies, and which has been on the market since 2004. Softeam has a big experience in personal data management due to several business projects and some research projects. With the e-Citiz platform, Softeam proposes the SVE ("Saisine par Voie Electronique" which means Seizure by Electronic Way), an eService product allowing users to apply for a claim or any sort of demand to the company. The SVE pilot privacy challenge will imply the customization of the SVE product to integrate PoSeID-on solution to provide the users of the current SVE services with a single platform for personal data management, as well as to support SVE to be compliant with the GDPR.

The public and private challenges were discussed as well as the solutions proposed by the Papaya and PoSeID-on projects. The participants were particularly interested in the public aspects and challenges and potential help that the project results could give to European citizens. There were also questions regarding the trust issues that the PoSeID-on platform should provide for its users (for example the fact that it could be guaranteed by governmental organizations).

4.2 Dashboard and Data Analysis Modules

The team then presented the PoSeID-on Web Privacy Enhanced Dashboard as a web application giving Data Subjects access to the PoSeID-on functionalities. Access to the Web Dashboard is managed by national systems compliant with eIDAS (e.g., SPID, @firma, FRANCEconnect). Such systems guarantee users secured access to the digital services of Public Administrations. These "electronic Identities" are released by Identity Providers, accredited bodies that release the credentials (User ID and Password), after verifying the user's identity. The session was very positive and interactive. The dashboard mock-ups were well accepted and had positive feedback. The participants wondered about the dashboard compliance with W3C Standards, which is one of the aspects that PoSeID-on project takes into consideration.

The architecture of PoSeID-on was also presented and the audience was particularly interested in the PoSeID-on analytics capabilities, in particular, the Risk Management Module (RMM) and Personal Data Analyser (PDA). These components will be used to evaluate and manage a risk score as well as to monitor all personal data flow and usage in addition to related warnings generated, to detect and prevent anomalies and misbehaved transactions (data flow and usage).

The RMM leverages Apache Spark Streaming and associated machine learning library MLlib to build a data analysis pipeline in order to perform anomaly detection on incoming logs from system components. All data is stored in CassandraDB to guarantee throughput of writes and the partition of distributed data according to the deployment each RMM instance and Cassandra node.

The PDA explores the potentialities of Natural Language Processing (NLP), more precisely Named Entity Recognition (NER) to analyse personal information. It comprises an ensemble learning mechanism with the best-performing machine learning algorithms (e.g., Random Forest, Conditional Random Fields, Convolutional Neural Networks), NLP tools (e.g., Stanford CoreNLP or SpaCy) and regular expressions to provide an accurate analysis of data. Moreover, since it does not store any data and is open source, it is trustworthy and transparent.

With the Blockchain, Dashboard, RMM and PDA, the user could be aware of data privacy exposure and have control over his/her data. The audience particularly appreciated that PoSeID-on Users can be advised on which service they could eventually disable in case of anomalies or high exposure of their data to privacy risks. Moreover, the participants were also very interested in the way PoSeID-on Dashboard handles such amount of information.

One aspect that came out from the workshop discussion was that how much it is necessary to write a statement which clearly details and clarifies the provided transparency and add it to PoSeID-on Dashboard web site. It was also noticeable the interest in better understanding how we handle the data flowing through PoSeID-on. More specifically, over how the data is managed, kept, analysed by our modules. Therefore, this valuable input is going to be considered during the next stages of development as well as dissemination and presentation of the project.

5 Lessons Learned from the PAPAYA Project Perspective

In this session of the workshop, general objectives, use cases and underlying privacy-enhancing technologies were briefly explained. Moreover, the architecture and the planned dashboards for the players of PAPAYA were also explained. Another important benefit for the PAPAYA project is that we have the chance to discuss human aspects, particularly end-users, with privacy experts.

Workshop participants on human aspects The workshop also included a discussion regarding potential human aspects and privacy concerns for the privacy-enhancing technologies presented in the workshop. It was followed up

by a small questionnaire, where we inquired about participants' perspectives of data privacy, key challenges and concerns, and factors to be considered in the use-case scenario presented. We chose to collect their opinions in a questionnaire format to allow some freedom of expressing their sharing concerns and opinions without revealing such information to fellow participants, especially in a privacy-aware group of experts, i.e., participants of the summer school.

Although we presented both of the eHealth use-cases of PAPAYA project [3] by **MediaClinics Italia** (project partner in the PAPAYA project), namely *UC1: Privacy preserving Arrhythmia Detection* and *UC2: Privacy preserving Stress Management* [7], we focused on UC1 during the presentation. The three different human actors in the use case scenario are the patient, pharmacist, and cardiologist. The eHealth use-case involves electrocardiogram (ECG) medical data that is to be collected from the patient via a wearable device and uploaded to the medical healthcare platform by the pharmacist. Data is then encrypted and send to the PAPAYA platform, where data is analyzed and sent back to the platform. Finally, the cardiologist receives the results of the analysis alongside raw data and other medical records to perform the diagnosis and send a report back to the patient.

In total, there were 11 participants in the workshop. We had optional demographic questions included in the questionnaire and the following are the responses for work/field, ages and genders for those who chose to answer those fields. There were only 5 who stated their field of work to be in research, academia, information technology, and computer science. There were 9 responses for age, of which 2 were in the age range of 21-30, 5 in the age range of 31-40, 1 in the age range of 41-50, and 1 preferred not to answer. There were 8 responses for genders, 2 were females, 4 males, and 2 preferred not to answer.

Workshop feedback on human aspects In human centered design approaches, end-users are taken into account early and throughout the development process of technologies. Following such approaches yield quality results that target human aspects throughout the development process by highlighting concerns and requirements which suit end-user's mental models[4,5]. In this workshop, we had the opportunity to engage privacy experts in our projects discussions. Below is a summary of the human factors part discussion of the PAPAYA.

Overview. Despite the limited number of participants, we had gathered varied inputs from participants, some included some UC2 points in their feedback sheets and were left-out from the summary. Main points raised by participants are summarized into their perspective/mental models on privacy, key challenges and concerns, communication requirements found in the following paragraphs.

Perspectives/mental models. In order to understand the mental model of participants, we inquired about their sharing behaviours and their opinions about privacy. We asked participants whether they are active on social media, personal (e.g., Facebook) and professional (e.g., Researchgate), so that we get an indication about their data sharing behaviour. Similarly in previous work, where it

was shown that different stakeholders have different values for privacy that affect their privacy sharing behaviours [6]. 10 out of 11 participants indicated their responses, 6 indicated yes/occasionally on personal social media where 4 indicated no, whereas 7 indicated yes/occasionally on professional social media and 3 no. When asked whether privacy is not always the most important incentive on 5-point Likert scale from strongly disagree to strongly agree. 3 participants disagreed, 2 were neutral, 5 agreed, and 1 strongly agreed. A comment related to that question highlighted that privacy is the most important incentive just in research environment.

Key challenges. Several challenges and concerns were highlighted by participants, they specified the need to address challenges with technologies, actors involved and trust factors for both use-case scenarios. It was indicated that technical challenges to PAPAYA for implementing and guaranteeing data accuracy should be addressed, as well as confidentiality, data security and privacy, and scope of data use and limitations. Additionally, considerations for ethical, legal and social factors pose trust challenges, such as the safety and the need for guarantees for patient's safety in the first use case scenario.

Communication requirements. Communication with different actors and users was shown to be of significance by some participants; explanations of why and how data is being used, processed, stored, and who is involved. Also communicating procedure for safety, privacy and security, as in the case of data privacy breaches and how it is handled.

6 Conclusion

During the workshop, within the scope of PAPAYA and PoSeID-on project, we presented potential solutions to overcome privacy challenges related to data management and processing for public and private organizations. Additionally, we had the chance to discuss human aspects, particularly end-users, with privacy experts. Inputs from these experts was a helpful feedback for developing the privacy enhancing technologies of the use-cases particularly for the next stages of the project. In terms of PoSeID-on project, the most important inputs raised during discussions were "how the platform should provide trust to its users" and "how PoSeID-on project handles data flow, management and processing to achieve transparency". That is for sure, by considering these inputs, the consortium will have the chance to easily achieve the dissemination goals of the project. In terms of PAPAYA project, the most significant learned lessons was that we had the opportunity to get inputs on our questionnaire from privacy experts. These inputs are not only important for the dissemination activities of PAPAYA project but also plays an important role in the Data Subject Toolbox, which is mainly designed and will be used at the end of the project to explain privacy preserving data analytics techniques to data subjects.

7 Organizers

Alessandra Bagnato is a research scientist and the Head of the Research Unit within the Softeam R&D Department. She holds a Ph.D. degree in Computer Science from TELECOM SudParis and Université Evry Val d'Essonne, France and a MSc in Computer Science from the University of Genoa, Italy. She has served on the technical program committee of several international events, such as ECMFA, MISE @ ICSE, AMMoRe @ MODELS, VVIoT @ ICST. She was also co-organizer of SEC-MDA 2009 and 2010 at ECMFA, DeCPS at AdaEurope since 2015, and of the MeGSuS workshops at the ESEIW since 2015. Her main research interests include software engineering in the context of big data, cyber-physical systems design, security and data privacy.

Paulo Silva is a PhD student enrolled in the Doctoral Program in Information Science and Technology of the University of Coimbra where he works as a researcher for the Center for Informatics and Systems of the University of Coimbra (CISUC). He holds an MSc. in Communications, Services and Infrastructures. His main research interests are data privacy protection and security services for Cloud Computing.

Ala Sarah Alaqra is a PhD candidate within the Privacy and Security research group (PriSec) at the department of Computer Science in Karlstad University. Related background includes an interdisciplinary MSc. from Umeå University in Human Computer Interaction (HCI). Current research is focused on human aspects regarding privacy trade-offs, trust, and adoption of privacy enhancing technologies within the scope of PAPAYA.

Orhan Ermis is a postdoctoral researcher of the Digital Security Department at EURECOM. He received his PhD degree from Department of Computer Engineering at Boğaziçi University, 2017, where he was a post-doctoral researcher and a part-time faculty between 2017 and 2018. His current research interests are topics related to applied cryptography such as privacy preserving technologies, network security and security protocols.

Acknowledgement

This work was partly supported by the PAPAYA project and PoSeID-on project funded by the European Union's Horizon 2020 Research and Innovation Programme, under Grant Agreement no. 786767 and no. 786713, respectively.

References

1. Papaya Project - Platform for privacy preserving data analytics., `www.papaya-project.eu`. Last accessed 17 October 2019

2. PoSeID-on Project - Protection and control of secured information by means of a privacy enhanced dashboard., www.poseidon-h2020.eu. Last accessed 17 October 2019
3. Ciceri, Eleonora, Marco Mosconi, Melek Önen, and Orhan Ermis. 2019. "PAPAYA: A Platform for Privacy Preserving Data Analytics.", https://ercim-news.ercim.eu/en118/special/papaya-a-platform-for-privacy-preserving-data-analytics. Last accessed 22 October 2019
4. Abras, Chadia, Diane Maloney-Krichmar, and Jenny Preece. 2004. "User-Centered Design." Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications 37(4): 445–56.
5. Anderson, Nancy S., Donald A. Norman, and Stephen W. Draper. 1988. "User Centered System Design: New Perspectives on Human-Computer Interaction." The American Journal of Psychology 101(1): 148.
6. Alaqra, Ala Sarah, and Erik Wästlund. 2019. "Reciprocities or Incentives? Understanding Privacy Intrusion Perspectives and Sharing Behaviors." In International Conference on Human-Computer Interaction, Springer, 355–70.
7. Ciceri, Eleonora and Galliani, Stefano and Mosconi, Marco and Azraoui, Monir and Canard, Sébastien, D2.1: Use Cases and Requirements, PAPAYA Deliverable D2.1, 2019.