Orchestrating Privacy Enhancing Technologies and Services with BPM Tools

The WITDOM Data Protection Orchestrator

Nicolás Notario
Atos
Spain
nicolas.notario@atos.es

Eleonora Ciceri
Fondazione Centro San Raffaele
Italy
ciceri.eleonora@hsr.it

Alberto Crespo
Atos
Spain
alberto.crespo@atos.es

Eduardo González Real
Atos
Spain
eduardo.gonzalezreal@atos.net

Ilio Catallo
Fondazione Centro San Raffaele
Italy
catallo.ilio@hsr.it

Sauro Vicini
Fondazione Centro San Raffaele
Italy
vicini.sauro@hsr.it

ABSTRACT

Privacy is a highly complex subject, especially when it comes to balancing data subjects' expectations, requirements and needs with i) the objectives of service providers and data controllers, and ii) the variety of legal obligations that dictate protection rights of data subjects and responsibilities of data controllers. This requires to provide technical solutions capable of matching different and adequate levels of privacy, while still attending to data subjects' preferences and business objectives. The Data Protection Orchestrator (DPO) developed in the context of the WITDOM project[1] meets this challenge by interacting with different Protection Enhancing Technologies or Services following a set of pre-defined protection processes, so as to support automated management trade-offs between privacy, performance and utility. By leveraging Business Process Management standards, the DPO is capable of making data protection processes and practices (such as automated anonymization or management of data subject's consent) integral to other business core services, as intended with the data protection by design and by default approach in the EU's GDPR. The DPO capabilities will be explained in the context of two complementary scenarios: the eHealth scenarios, where the DPO will be used for protecting genomic data and the financial scenario where the DPO will be responsible for protecting the transaction history and personal attributes of the bank's customers.

CCS CONCEPTS

•Security and privacy →; Pseudonymity anonymity and untraceability; Information flow control; Privacy protections; •Software and its engineering →Orchestration languages; Integration frameworks; Software as a service orchestration system; Software design tradeoffs;

KEYWORDS

Data Protection, Business Process Management, Privacy, Anonymization, Trust, Privacy Enhancing Technologies, Privacy Enhancing Services

ACM Reference format:
Nicolás Notario, Eleonora Ciceri, Alberto Crespo, Eduardo González Real, Ilio Catallo, and Sauro Vicini. 2017. Orchestrating Privacy Enhancing Technologies and Services with BPM Tools. In *Proceedings of ARES '17, Reggio Calabria, Italy, August 29-September 01, 2017,* 7 pages.
DOI: 10.1145/3098954.3104057

1 INTRODUCTION

Following the General Data Protection Regulation (GDPR)[4], to be enforced in May 2018, data controllers interacting with European citizens' personal data will have to comply with a set of more demanding privacy requirements, and will be subjected to higher penalties if they fail to fulfil them. For instance, breaching requirements related to either international transfers or basic processing principles may lead to penalties up to either 20 million EUR or the 4% of the total worldwide annual turnover of the preceding financial year, whichever is higher. It is thus vital to integrate data protection by design and by default principles in data controllers business processes, so that they can easily demonstrate their compliance with the regulation. In order to minimize the impact of such integration on data controllers daily operations, it is essential to embed privacy practices via tools and methods which can fit in their already established business processes.

Data controllers have already at hand many existing Privacy Enhancing Technologies (PETs) that can support the provisioning of some privacy properties in their services. However, and among many other challenges, these technologies implement several types of APIs and use different underlying components related to big data, data streaming, etc. Furthermore, while the integration of individual PETs (which are able to provide specific but limited privacy

[1] http://witdom.eu/

ARES '17, Reggio Calabria, Italy

DOI: 10.1145/3098954.3104057

properties) is supposed to lead to a more complete privacy framework, it often results in a deterioration of the achievable privacy levels, due to incompatibilities and requirements inconsistencies.

The Data Protection Orchestrator (DPO) developed in the context of the WITDOM project solves these issues by easing the integration of heterogeneous PETs, so as to successfully compose their privacy-preserving capabilities while taking into account performance requirements. The combination of different PETs, which is achieved via the application of Business Process Management Notation (BPMN)[1], still requires the intervention of privacy experts, but thanks to the Orchestrator, integration efforts can be greatly reduced.

2 WITDOM

The WITDOM (Empowering Privacy and Security in Non-trusted Environments) project is a Research and Innovation Action funded by the European Commission's Horizon 2020 programme and the Swiss State Secretariat for Education, Research and Innovation. WITDOM's main objective is to technically enable the processing of large amounts of personal data in untrusted environments with minimal impact on privacy risks for data subjects. For such, it relies in improving and applying a number of PETs and will be demonstrated in two use-case scenarios where privacy and confidentiality constraints are a true barrier to using outsourced architectures such as public clouds. The first use-case is a health scenario that involves outsourcing genetic data processing. The second scenario deals with outsourced financial analyses based on the management of both customers' data, to enable financial services over private and public cloud environments.

3 BUSINESS PROCESS MODELLING FOR PRIVACY

The popularity of Business Process Modelling keeps increasing at a solid pace every year. The latest " The state of Business Process Management" report [7] shows how BPMN has become the most popular process standard: 67% of organizations are interested in its adoption. The same report proves that organizations developing operational applications use BPM products mainly for: a) defining the organization's business processes; b) monitoring existing processes; and c) fully automatizing processes.

The application of BPMN techniques for privacy purposes is not a completely novel approach. For instance, [10] and [9] propose a BPMN privacy-aware extension for modelling privacy requirements. Also, [6] discusses the application of a consent management extension of BPMN. However, the Data Protection Orchestrator brings some innovations to the usage of BPMN towards the provision of privacy-enabled services, listed in the following:

(1) Application of standard BPMN 2.0 for the definition of data protection requirements;
(2) Orchestration of heterogeneous PETs and other services;
(3) Support for complex key management schemes considering different scopes for keys (e.g., request, user, service, application and platform keys);
(4) Integration of privacy metrics resulting from data protection as process variables, so as to use them to take business decisions (e.g., publishing protected data);
(5) Interaction with data subjects for implementing data protection principles such as the rights for objection or the concept of dynamic consent [8].

By using BPMN to integrate privacy or data protection perspectives into business process, the following objectives are met:

(1) Facilitation of the dialogue between different stakeholders (i.e., service providers, customers and data subjects);
(2) Combination of different PETs to maximize the achievable level of data protection;
(3) Provisioning of a mechanism to automatically balance services privacy, performance and utility, via the evaluation of relevant Key Performance Indicators (KPIs);
(4) Coverage of key GDPR requirements like accountability, privacy by design or informed consent.

4 THE DATA PROTECTION ORCHESTRATOR

The Data Protection Orchestrator was developed to solve the issue of orchestrating several PETs in a protection process that combines end-user preferences with trade-offs related to privacy, performance and utility. While several PETs allow achieving the functional objectives of a customer (e.g., a genomic analysis), there are other non-functional objectives in terms of performance, cost and accuracy that vary from one use case to another. In the following, we list the requirements that drove the creation and implementation of the concept of the DPO. This list of requirements was the result of a elicitation requirement process in the context of WITDOM project, with the collaboration of several experts from the legal domain, from the different functional domains involved in the scenarios (eHealth and financial) but also with privacy engineers system architects.

(1) The DPO should not be restricted to be used with a subset of PETs, it should be interoperable with online services following standards such as REST.
(2) The DPO must allow secured services' privacy experts to specify their own protection processes using standard configuration mechanisms (e.g., XML, BPMN) according to the specific needs of their customers. The specification may establish some minimal privacy thresholds according to user preferences or sensitivity of the data.
(3) The DPO must allow end-users and operators of end-user applications to specify some security preferences that may affect the execution of a given protection process. I.e. a user could be interested in a faster execution in exchange of less accuracy.
(4) The DPO must allow the possibility for interaction between the data subject and the protection process. This interaction will allow, e.g., to gather informed consent before the processing or outsourcing of the data. This means that the DPO must expose an API that will allow to integrate user interfaces for data subjects.
(5) Besides the protection configurations that are directly developed to address the protection of data before its outsourcing or processing, the DPO must allow the creation of additional protection configurations, that can be used to take care of transversal issues. A clear example of this would be to have a specific protection configuration that

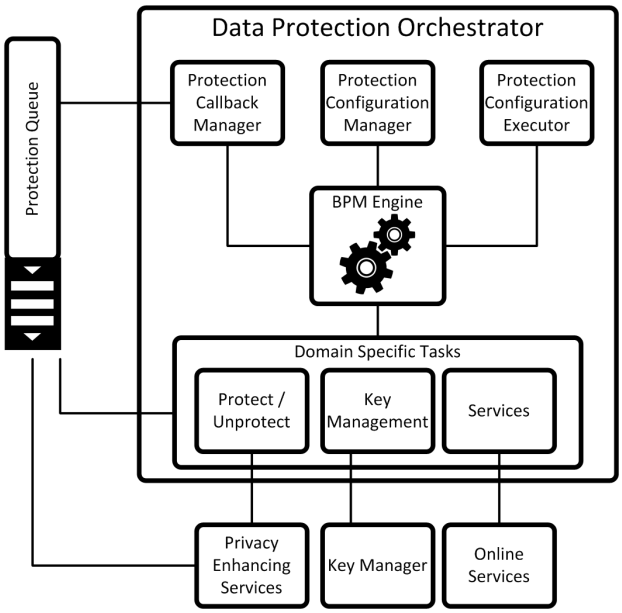

Figure 1: Architecture of the DPO

can automatically invoke a mechanism for the secure deletion of data after the retention period.

4.1 Architecture

Figure 1 presents a high level architecture of the DPO.

At the core of the component, a **BPM Engine** is responsible for interpreting business processes in BPM notation, isolating processes from the rest and persisting them in a database whenever the lapse between two steps in a process is expected to take some time. Besides keeping a state of each process, the engine is also in charge of triggering next actions when requested by external inputs (e.g., a user interaction or the activation of a time-based trigger), and branching the execution of the process. Applications using the Data Protection Orchestrator do not interact directly with the BPM Engine, but make use of a set of higher-level services:

(1) **Protection Callback Manager**. This service is responsible for i) handling callbacks from other services or components, and ii) listening to the Protection Queue so as to notify the BPM Engine about the occurrence of events (e.g., a protection service finished its execution or a data subject consented to process her data).
(2) **Protection Configuration Manager**. This service deals with the management of protection configurations. A protection configuration is a file that contains logic and visual information about a business process in BPMN 2.0 notation. The visual representation of the protection business process can be used to track process status or to align with other stakeholders.
(3) **Protection Configuration Executor**. This service is responsible for starting and ensuring the completion of protection business processes, notifying relevant stakeholders when appropriate.
(4) **Protection Queue**. This component enables a message queue paradigm to ensure reliable asynchronous communication between the DPO and external components/services. This aspect is crucial, due to the nature of the protection algorithms, which are not generally expected to produce a result instantly, but rather to take some time (even minutes) to complete the protection process.

4.2 Integration with PETs and Services

The interface of the DPO towards external components and services is based on **Domain Specific Tasks** (DSTs). The level of specificity of a DST may vary from case to case. This means that one could define either a generic DST (e.g., a REST DST that can invoke any REST interface), or a very specific DST (e.g., a DST for the integration with an API of a content management system).

From a development point of view, any developer with a minimal knowledge of the DPO or its framework can easily create a new DST with its own custom logic by implementing a plugin following a common DST interface. The DPO includes a mechanism that enables the seamless deployment of these additional DSTs without the need to rebuild the DPO.

4.2.1 Implementation of generic DSTs. The current implementation of the DPO incorporates DSTs used to deal with generic requirements. An example is the key management DST that connects to Barbican [2], which is an open source solution designed for the secure storage, provisioning and management of secrets such as passwords, encryption keys and X.509 Certificates. This DST fulfils crucial management need from the aforementioned privacy enhancing services: it retrieves system-wide keys under some access control mechanism and unprotects user-keys using some passphrase.

4.2.2 Implementation of specific WITDOM DSTs. The current implementation of the DPO includes DSTs for invoking the privacy enhancing services developed in the WITDOM project, in a context in which computation is distributed between trusted and untrusted environments:

(1) **Data Masking (DM)**: introduces a model for dynamic data masking which desensitizes data using a cryptographically proven secure method (maintaining reference integrity) and allows key updates without the need of desensitizing all data from scratch.
(2) **End-to-end encryption (E2EE)**: based on Crypton[12] implementation, it encrypts data in a trusted environment and outsources it to the untrusted environment. Properties like integrity, write serializability, read-freshness are also provided but with the need of a trusted third party.
(3) **Integrity & Consistency Verification (ICV)**: includes protocols that guarantee the consistency notion of fork-linearizability [11] by adding condensed data about the causal evolution of the users' views into their interaction with the remote service. It does not require a trusted third party.

[2] https://wiki.openstack.org/wiki/Barbican

(4) **Secure Signal Processing (SSP)** [? , Pedrouzo-UlloaT16] protects data using different techniques at the trusted domain so that the subsequent computation can be performed efficiently in the untrusted domain. The SSP component will rely on Polynomially-approximated general functions through applied Somewhat Homomorphic Encryption (SHE), binary SHE for string comparisons and approximate search, combination of SHE and interactive protocols for linear regression.
(5) **Secure Computation (SC)**: protects data using homomorphic encryption (HE) techniques, which enable further processing in the untrusted domain without unprotecting the data. Hardware-based accelerators are used in order to address typical performance limitations of HE.
(6) **Anonimization**: protects data applying statistical principles [13] that support the reduction of the accuracy of the data (i.e. generalization of data or introducing some nose using some statistical model) minimizing the risk of re-identification and minimizing the impact on the data utility.

These services receive through their API information of the location and destination of the data that need protection, as well as information regarding the protection/un-protection algorithms and related parameters (including cryptographic keys).

5 PRACTICAL APPLICATION IN THE E-HEALTH DOMAIN

A *genome* is the genetic material of an individual. The genetic instructions it contains are used in the growth, development, functioning and reproduction of individuals, and define one's phenotype (that is, one's observable characteristics or traits). Genomes are composed of a sequence of nucleotides, and naturally divide themselves in *chromosomes*, which in turn contain *genes*, i.e., regions of DNA that encode specific functions. Genes can acquire mutations in their sequence of nucleotides, leading to different variants in the population, and consequently to different phenotypic traits. Over the years, scientists have published several archives of genomic variants, i.e., collections of alterations that occur in specific positions in the genome. Every variant is decorated with a set of *genomic annotations*, that state which are its semantics (e.g., specifying whether it is associated with an increased probability of developing a pathology) and its biological structure.

Over the last years, many improvements in genome sequencing technologies have been introduced. Such technologies allow determining with high speed the precise sequence of nucleotides within individuals' genomes. This, added to an increased quality of genomic annotations, allows doctors and biologists to recognize patients' pathological variants in short time, so as to devise timely appropriate therapies. Nevertheless, this exponential growth of interest in the genomic field has also brought to the production of an unprecedented mass of genomic data, leaving genomics laboratories to cope with a huge amount of needed effort.

Thus, genomic analysis could greatly benefit from the usage of a cloud-based infrastructure, since shared storage and processing resources enable ubiquitous, on-demand access to a set of services and machines that can be rapidly provisioned and released with minimal effort. Unfortunately, cloud environments are typically untrustworthy with such data, since they are generally located outside laboratories premises, and thus expose sensitive data to the risk of attacks and unwanted disclosures. This obviously generates problems: i) genomic data is highly sensitive, for it provides a full description of an individual's traits, including her health status; ii) since genetic material is passed through generations, disclosing it to unauthorized people would invade family members' privacy too. Consequently, it is necessary to build a secure environment in which genomic data can processed and stored while ensuring patients and their relatives' privacy.

5.1 Genome analysis pipeline

In this section we present the pipeline used to analyse a DNA sequence so as to extract its genomic variants and related annotations. The pipeline is composed of three steps: *alignment*, *variant calling* and *variant annotation*. In the following, we discuss each step in detail.

5.1.1 Alignment. A *reference genome* is a digital genome assembled using genetic material of several individuals, with the objective of having a representative example of human genome. When looking for a patient's genomic variants, one has to compare her DNA with the reference genome, to list all the positions in which the two DNA strands (and in case their phenotypes) differ. However, depending on the set of pathologies doctors want to investigate on, laboratory technicians extract only some *reads*, i.e., small genome portions that are known to be related to such pathologies. Thus, to compare the extracted DNA sequence with a reference genome $\mathcal{R}$, one has to *align* each read r_i with $\mathcal{R}$, i.e., find the exact area in $\mathcal{R}$ where r_i is positioned. The outcome of the alignment phase is a list of tuples $\langle r_i, C, p\rangle$, where C is the chromosome of $\mathcal{R}$ in which r_i was found, and p is the position on chromosome C where r_i starts.

5.1.2 Variant Calling. During the variant calling phase (performed with custom scripts by a bioinformatician), each nucleotide in a read r_i is compared with the related nucleotide in the reference genome $\mathcal{R}$. Their differences are called *variants*, and are expressed as tuples $\langle C, p, \mathtt{r}, \mathtt{a}\rangle$, where C and p are respectively the chromosome and position where the variant was found, $\mathtt{r}$ is the nucleotide found in $\mathcal{R}$ and $\mathtt{a}$ is the alternative nucleotide found in the patient's DNA.

5.1.3 Variant annotation. During the variant annotation phase, each patient's variant is decorated with a set of annotations found in public archives of genomic variants. Some of these annotations will suggest biologists and doctors a possible correlation of such variant with specific pathologies.

5.2 Protection Scenarios

In the following, we present the details about four protection scenarios that would greatly benefit from an automatic, on-cloud genome analysis pipeline. For each of the protection scenarios, we propose a protection process orchestrated by the DPO, with reference to the involved PETs. Variant calling is excluded from these automatized scenarios due to the high level of customization of the calling pipeline, which requires the bioinformatician to build tailored scripts for each test and run them in the laboratory environment.

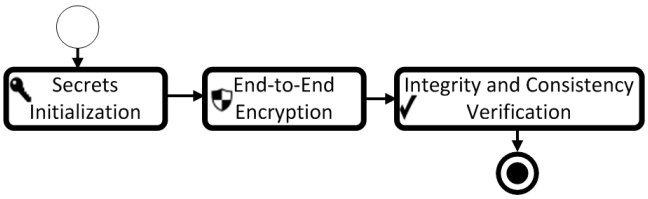

Figure 2: Backup protection configuration

5.2.1 Backup Scenario. Genomic data is intrinsically complex to be stored: i) an individual's genome, when sequenced, may take up to 300*GB*; ii) patients' genomic variant annotations (and related annotations) change over time and for diagnostic purposes it is essential to store every update. Thus, due to the number and size of produced files, one would like to back them up in a cloud environment with large storage capabilities. This would avoid unnecessary re-sequencing and re-alignments when a new genomic analysis on already sequenced reads is requested, and keep an archive of patients' reads and variant annotations to track their changes over time.

The protection configuration that would enable this processing in an untrusted environment has to ensure the integrity and confidentiality of the genomic information. To achieve this result, the DPO orchestrates the invocation of the E2EE and the ICV services, as shown in Figure 2. While the E2EE service is responsible for creating an encrypted version of the genomic information in the cloud, the ICV ensures the protected backup is never corrupted. Whenever the backup is retrieved from the untrusted environment, the same pipeline is executed in reverse-order: ICV ensures data integrity before data is decrypted in the trusted environment.

5.2.2 Alignment. Reference genome and patients' DNA usually contain large amount of genomic data, so aligning them requires a large computational effort. Cloud computing can cover such a necessity by allowing to instantiate new resources when needed. However, while reference genomes are provided as public resources (and thus not subjected to privacy concerns), patients' DNA has to be safely delivered to an untrusted environment, where large computational resources are available, but attackers could take advantage of vulnerabilities to access this sensitive data.

In order to enable the secure alignment of genomic data in the cloud, the data must be cleared of all personal identifiers (via DM) and then protected (via SSP) in such a way that the alignment can still be performed. Figure 3 shows these steps in BPM notation.

5.2.3 Variant Annotation. Genomic research allows us to discover new variants and annotations every day. Such findings continuously enrich public sources of genomic variants, which become larger and larger as time passes. The Single Nucleotide Polymorphism database (dbSNP[3]), for instance, currently contains more than 300 million variants (along with their annotations), making it one of the most used annotation sources. However, working with huge variant archives makes variant annotation difficult to be performed in a laboratory environment, where scarce IT resources are in place. Again, cloud computing can help by offering large computational resources on demand, although at the cost of introducing protection mechanisms to safely process patients' sensitive data in untrusted environments.

Similarly to the "Alignment" case, enabling variant annotation in a cloud environment requires the genomic information to be protected in such a way that further processing can be performed over the protected data. In the context of WITDOM, there are two services that match this requirement: the SC and the SSP. The branch selection (performed according to the preferences of hospital and bioinformaticians) codifies the trade-offs between different business objectives: the SC service may be more accurate but much slower than SSP, and thus more indicated, for instance, to confirm diagnoses with high accuracy. Figure 4 shows these steps in BPMN notation.

5.2.4 Variant reannotation. Since new annotations are discovered continuously, genomic test results need to be kept updated over time: new information that at the time of the test was not available may change diagnoses and give more insights about patients' health status. This is where *variant reannotation* comes into play. By periodically reannotating all patients' genomic variants, we enrich them with new information. Biologists can be thus alerted whenever there is a relevant change in the semantics of variants. This obviously needs the informed consent to be modified, so that patients can choose between a continuously updated genomic analysis and a one-off analysis.

Based on the previous protection scenarios and the BPM notion of subprocesses, timers, events, messages and user tasks, rather complex privacy enhanced business processes can be built. Figure 5 shows how after an initial alignment, simultaneously, the results can be backed up and annotated. At a later stage (e.g., after one year), the system automatically sends a message to patients requesting their consent for a reannotation process taking into account the

[3]https://www.ncbi.nlm.nih.gov/projects/SNP/

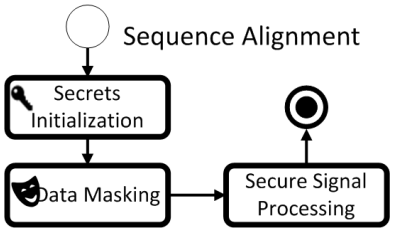

Figure 3: Sequence Alignment Protection Configuration

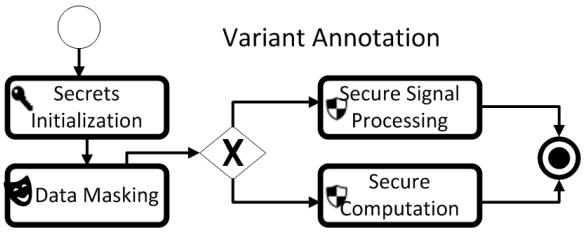

Figure 4: Variant Annotation Protection Configuration

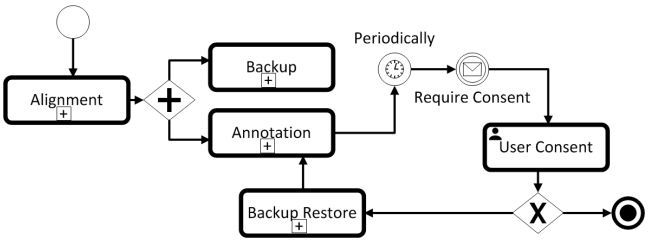

Figure 5: Variant Reannotation Protection Configuration

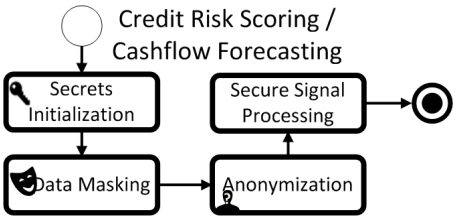

Figure 6: Cashflow Forecasting and Credit Risk Scoring Protection Configurations

latest genomic discoveries. Once the user provides consent, the system restores the backup and performs a reannotation process.

6 PRACTICAL APPLICATION IN THE FINANCIAL DOMAIN

In the financial domain the usage of mathematical functions over large amounts of data is one of the basic tools that enables daily operations. For instance, regulations such as Basil II encourage financial institutions to build mathematical models over data of customers and operations to estimate their Value at Risk, awarding benefits for those institutions with lower values. Also, more than two decades ago neural networks were proposed [5] as an effective tool to detect fraudulent transactions. Both cases do not need any customers' personal data, since they only need financial data (e.g., history of credit card transactions) to build predictive models.

Similarly to the eHealth domain, the amount of produced data is very large, and thus financial institutions could greatly benefit from the usage of a cloud-based infrastructure to perform their calculations over the bank's endogamic data (e.g., credit card transactions, customers' financial history). However, data protection and sector-specific regulations still hamper the adoption of such technologies. This means that most banks are constrained to use their in-house computational power, which leads to large upfront costs and non-scalable solutions and increases the risk of interfering with banks' daily operation. The WITDOM protection scenarios that are described in the following demonstrate how it is possible to overcome such issues: the right combination of privacy properties allows a bank to process financial data in an untrusted environment.

6.1 Protection Scenarios

In this section we present the details of the three protection scenarios for the financial domain.

6.1.1 Credit Risk Scoring. Customers' Credit Risk in the banks is managed with two main purposes. The first purpose is to understand the Value at Risk of the money lent to customers, which requires to estimate their probability that they be unable to meet its debt obligations and the loaned amount. The second purpose, which is the one discussed in the context of the WITDOM project, is to apply credit risk scoring to financial operations and use its result in the context of a business decision (e.g., to approve or deny an operation). In Europe, the legal framework stipulates that, in order to guarantee customers' fundamental right, this model has to be transparent (i.e., explained to the regulator) and bound to financially explainable variable. Banks typically apply linear regression algorithms over past data to build a strong model that will consistently predict future operations.

Figure 6 shows the protection configuration, which is very similar to the one used for the e-Health domain's Alignment use case. Personal identifiers associated with credit card transactions, financial positions, incomes and expenses have to be masked via DM (guaranteeing referential integrity). However, masked financial data still contain a large amount of personal information that can lead to customer re-identification if correlated with other observations. That is why an anonymization step is introduced before finalizing the protection pipeline. Finally, SSP introduces an additional level of privacy, which allows computations to be securely performed in an untrusted environment.

6.1.2 Cashflow forecasting. A new generation of financial services is aimed to tailor the bank value proposition to each customer's needs. One option would be to estimate future cash flows, based on customers' past behaviour and current financial situation, in order to propose financially adequate actuations (e.g., ask for a low-interest temporary loan to avoid cash shortages, or open a pension plan with existing savings). These predictions can be performed using autoregressive integrated moving average (ARIMA [3]) models built upon customers' position, income and expense time series.

The protection scenario is the same used for Credit Risk Scoring, as described in Figure 6.

6.1.3 Credit Card Fraud Detection. In 2013, the total value of fraudulent transactions conducted using cards issued within SEPA and acquired worldwide amounted to 1.44 billion EUR [2]. When required to detect fraud, banks are not constrained by any regulation on the technology or model to use, and may rely on "black box" algorithms (e.g., neural networks) to build models that estimate the probability of each transaction to be fraudulent or not.

One of the main differences between the scenarios in this domain and the ones pertaining to the eHealth scenario is that in the financial one the required accuracy is lower: transactions could be aggregated, deleted or randomly modified (increasing the level of privacy of outsourced data) and this would not modify significantly the results. That is why the anonymization component, which implies the loss of information is only used in this domain (and not in the eHealth one).

The protection process proposed for this scenario (see Figure 7) is similar to the variant annotation one: bank operators, according

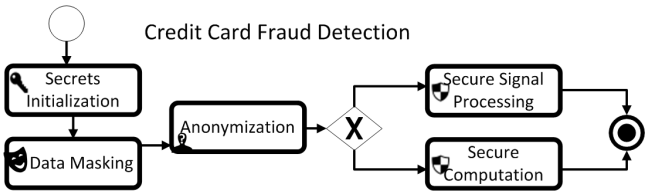

Figure 7: Credit Card Fraud Detection Protection Configuration

to their preferences, can choose between SSP or SC to protect data prior to processing.

7 CONCLUSIONS

When data protection and privacy is approached from a BPM perspective, we can combine several PETs in order to achieve higher levels of privacy, so as to provide a tool for customers that enables secure computation on untrusted environments with high accuracy and performance. Benefits of BPM are highly appreciated and understood in organizations and can be leveraged as a ground for discussions among different stakeholders: the proposed tool allows a company to explicitly include privacy aspects in its business processes, which are now mandatory according to the GDPR. E.g. **Accountability**: the protection configuration itself, where the list of processings, data protection mechanisms, privacy metrics and thresholds are identified, can be used as an accountability measure towards data protection supervisory authorities . **Informed Consent** Organizations may bind data processes to the consent of the user collected in collaboration with the DPO. The consent collection step may inform the user of real privacy risks based on the privacy metrics KPIs obtained during previous steps within protection process. While **PbD** is a much broader concept that cannot be solved by one technical component, the DPO supports privacy by design and by default by orchestrating PETs that minimize the amount of personal data outsourced to untrusted environments.

The protection scenarios from the eHealth and financial domain are currently under development and will be validated by WITDOM project consortium.

There are a set of additional functionalities that can be codified in protection configurations handled by the DPO and that are not explicitly displayed (for simplicity) in any of the previously presented protection configurations:

(1) Considering results of individual steps in the protection pipeline before choosing next steps (e.g., if the privacy level achieved is not satisfactory then stop the processing or choose a different execution branch);
(2) Re-execute protection services with different parameters in order to achieve desired privacy levels;
(3) Introducing time-based events to deal with data protection requirements such as data retention periods.

REFERENCES

[1] *Business Process Model and Notation (BPMN).* Technical Report. Object Management Group, Inc. (OMG). http://www.omg.org/spec/BPMN/2.0/PDF/
[2] European Central Bank. 2015. *Fourth report on card fraud.* Technical Report. https://www.ecb.europa.eu/pub/pdf/other/4th_card_fraud_report.en.pdf
[3] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. 2015. *Time series analysis: forecasting and control.* John Wiley & Sons.
[4] 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union* L119/59 (4 May 2016). http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC
[5] Brian Patrick Green and Jae Hwa Choi. 1997. Assessing the risk of management fraud through neural network technology. *Auditing* 16, 1 (1997), 14.
[6] Nils Gruschka and Meiko Jensen. 2014. Aligning User Consent Management and Service Process Modeling. In *44. Jahrestagung der Gesellschaft für Informatik, Informatik 2014, Big Data - Komplexität meistern, 22.-26. September 2014 in Stuttgart, Deutschland.* 527–538. http://subs.emis.de/LNI/Proceedings/Proceedings232/article214.html
[7] Paul Harmon. 2016. *The State of the BPM Market.* Technical Report. BPTrends. http://www.bptrends.com/bpt/wp-content/uploads/2015-BPT-Survey-Report.pdf
[8] Jane Kaye, Edgar A Whitley, David Lund, Michael Morrison, Harriet Teare, and Karen Melham. 2014. Dynamic consent: a patient interface for twenty-first century research networks. *European Journal of Human Genetics* 23, 2 (may 2014), 141–146. DOI: http://dx.doi.org/10.1038/ejhg.2014.71
[9] Wadha Labda, Nikolay Mehandjiev, and Pedro Sampaio. 2013. *Privacy-Aware Business Processes Modeling Notation (PrvBPMN) in the Context of Distributed Mobile Applications.* Springer International Publishing, Cham, 120–134. DOI: http://dx.doi.org/10.1007/978-3-319-03737-0_13
[10] Wadha Labda, Nikolay Mehandjiev, and Pedro Sampaio. 2014. Modeling of Privacy-aware Business Processes in BPMN to Protect Personal Data. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing (SAC '14).* ACM, New York, NY, USA, 1399–1405. DOI: http://dx.doi.org/10.1145/2554850.2555014
[11] David Mazières and Dennis Shasha. 2002. Building Secure File Systems out of Byzantine Storage. In *Proceedings of the Twenty-first Annual Symposium on Principles of Distributed Computing (PODC '02).* ACM, New York, NY, USA, 108–117. DOI: http://dx.doi.org/10.1145/571825.571840
[12] SpiderOak. 2015. Crypton. https://github.com/SpiderOak/crypton. (2015).
[13] Latanya Sweeney. 2002. K-anonymity: A Model for Protecting Privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10, 5 (Oct. 2002), 557–570. DOI: http://dx.doi.org/10.1142/S0218488502001648

ACKNOWLEDGMENTS

The work described in this document has been conducted within the project WITDOM, started in January 2015. This project has received funding from the European Union's Horizon 2020 research and innovation programme (H2020-ICT-2014-1) under grant agreement No. 64437. This work was supported in part by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract No. 15.0098. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the European Commission or the Swiss Government.