PC4PM: A Tool for Privacy/Confidentiality Preservation in Process Mining

Majid Rafiei and Alexander Schnitzler

Technical Report
Issue: This report provides a user manual for PC4PM.
Date: June 11, 2021

RWTH Aachen University
Chair of Process and Data Science (PADS)
Aachen, Germany

Abstract

Process mining techniques such as process discovery, conformance checking, and performance analysis provide insights into actual processes by analyzing event data that are widely available in different types of information systems. These data are very valuable, but often contain sensitive information, and process analysts need to balance privacy/confidentiality and utility. Privacy issues in process mining are recently receiving more attention from researchers and new techniques are being introduced. The PADS group of RWTH Aachen University, as one of the pioneer groups of process mining, provides a range of privacy and confidentiality preservation techniques. In this technical report, we provide a comprehensive explanation of the techniques developed by the PADS group. The report describes the main concepts of process mining and privacy-related activities. It provides a guide for users through the web-based application which integrates all the techniques designed and developed. The provided privacy/-confidentiality preservation techniques have been peer-reviewed and published as scientific papers.

1 Introduction

Process mining provides fact-based insights into actual business processes using event data, which are often stored in the form of event logs. The three main types of process mining are *process discovery*, *conformance checking*, and *process enhancement* [1]. An event log is a collection of events, and each event is described by its attributes. The main attributes required for process mining are *case id*, *activity*, *timestamp*, and *resource*. Some of the event attributes may refer to individuals, e.g., in the health-care context, the *case id* attribute may refer to the patients whose data are stored, and the *resource* attribute may refer to the employees who perform activities for the patients, e.g., nurses. Privacy issues in process mining are highlighted when the individuals' data are included in the event logs. According to the regulations such as the European General Data Protection Regulation (GDPR), organizations are compelled to respect the privacy of individuals while analyzing their data.

Together with the three mentioned types of process mining, different perspectives are also defined including *control-flow*, *organizational*, *case*, and *time* perspective [1]. The *control-flow perspective* focuses on activities and their order, which are often used by *process discovery* and *conformance checking* techniques. The *organizational perspective* focuses on resources and their relations, which are utilized by *social network discovery* techniques. The *case perspective* is focused on case-related attributes, and the *time perspective* is concerned with the time-related information, which can be exploited for *performance and bottleneck analyses.*

With respect to the main attributes of events, two different perspectives for privacy in process mining can be considered; *resource perspective* and *case perspective.* The *resource perspective* concerns the privacy rights of the individuals who perform activities, and the *case perspective* concerns the privacy rights of the individuals whose data are recorded and analyzed [10].

Figure 1 shows the general overview of privacy-related activities in process mining including Privacy-Preserving Data Publishing (PPDP), Privacy-Preserving Process Mining (PPPM), and Privacy Analysis (PrAn). PPDP tries to obscure the identity and/or sensitive data of individuals to preserve their privacy. PPDP techniques often apply one or more *anonymization operations*, e.g., *suppression*, *generalization*, etc., to provide the desired privacy requirements. PPPM intends to expand existing process mining algorithms to cope with intermediate results, so-called *abstractions* [6], generated by some PPDP techniques. Note that PPPM algorithms are inextricably linked with the corresponding PPDP approaches, and PPPM may refer to the entire privatization process, starting with an event log and finishing with process mining findings. PrAn, indicated with dashed lines in Figure 1, includes couple of activities: *risk analysis* and *utility analysis.* Both PrAn activities could be done for data and results. In this paper, we introduce a tool, named PC4PM,

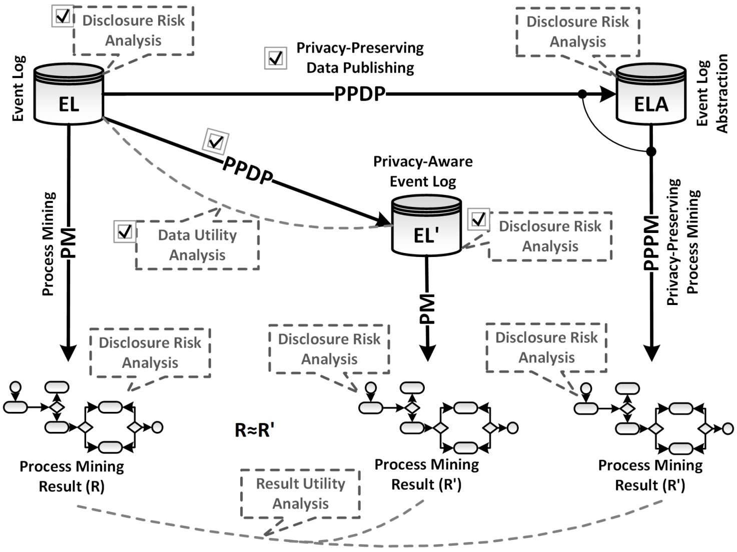


Figure 1: The general approach of privacy in process mining.

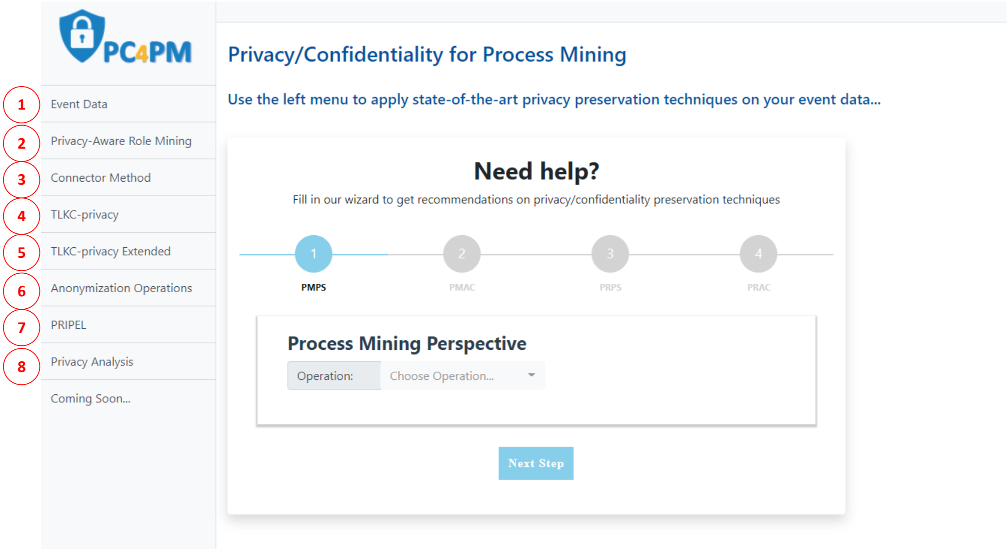
mainly focusing on the activities indicated by the check-boxes in Figure 1. PC4PM is the successor of the privacy tool introduced in [5], and it offers new privacy preservation techniques, privacy analysis, a set of anonymization operations, and user guidance that directs users to the right techniques based on their requirements.

PC4PM is a web-based application written in Python using the Django framework which mainly focuses on privacy-preserving data publishing and offers state-of-the-art privacy preservation techniques. Figure 2 shows the home page of the application where the main modules have been annotated on the left menu with numbers from 1 to 7. In the *event data* module, users are able to upload and manage standard XES[1] event logs and non-XES event logs which are called ELA (Event Log Abstraction). The *privacy-aware role mining* modules exploits the *decomposition method* for anonymously discovering roles (organizational structures) from event logs [4]. Using the *connector method*, users can securely discover their processes [11, 12]. The $TLKC$*-privacy* model and its extension provide group-based privacy guarantees for process mining [10]. The main anonymization operations introduced in [6] are also included as one of the main modules. Users can utilize the *privacy analysis* module to investigate disclosure risks and data utility [7]. Moreover, the *privacy metadata* proposed in [6] are also embedded in the offered privacy preservation techniques.

We define a *signature* to summarize different characteristics of each privacy preservation technique. The signature reflects the following information: *process mining perspective* (PmPs), *process mining activity* (PmAc), *privacy perspective* (PrPs), and *privacy activity* (PrAc). PmPs shows the process mining perspective(s) that a privacy technique focuses on, e.g., *control-flow.* PmAc shows that the utility of event data is preserved for which process mining activity(ies), e.g, *process discovery.* PrPs shows the privacy perspective of a privacy technique, i.e., *resource* or *case.* PrAc indicates the corresponding privacy-related activity(ies), i.e., PPDP, PPPM, or both. The signature is defined to guide users for selecting an appropriate privacy/confidentiality technique with respect to their requirements.

Each module of PC4PM is available individually as a Python package that could be installed and utilized in your Python scripts. The tool itself is available as a docker container that could be pulled and run locally on your machine. In the following, we first explain how to install the web-based application, then we describe the usage of each module.

[1] http://www.xes-standard.org/



Privacy/Confidentiality for Process Mining

Use the left menu to apply state-of-the-art privacy preservation techniques on your event data...

Need help?

Fill in our wizard to get recommendations on privacy/confidentiality preservation techniques

Process Mining Perspective

Operation: Choose Operation...

Figure 2: The home page of PC4PM. The modules in the left menu have been annotated with numbers from 1 to 8.

2 Installation and Source Code

Different privacy preservation techniques, anonymization operations, and the privacy analysis modules are available as separate GitHub repositories. To facilitate the usage and integration of the privacy preservation techniques, they are published as standard Python packages[2]. In the following, we provide links to the GitHub projects and name of the Python packages for the privacy techniques and anonymization operations:

- **Privacy-aware role mining**
  - GitHub: *https://github.com/m4jidRafiei/privacyAware-roleMining*
  - Python package: *pip install pp-role-mining*
- **Connector method**
  - GitHub: *https://github.com/m4jidRafiei/privacyAware-ConnectorMethod-DFG*
  - Python package: *pip install p-connector-dfg*
- $TLKC$**-privacy**
  - GitHub: *https://github.com/m4jidRafiei/TLKC-Privacy*
  - Python package: *pip install p-tlkc-privacy*
- $TLKC$**-privacy extended**
  - GitHub: *https://github.com/m4jidRafiei/TLKC-Privacy-Ext*
  - Python package: *pip install p-tlkc-privacy-ext*
- **Anonymization operations**
  - GitHub: *https://github.com/m4jidRafiei/PPDP-AnonOps*
  - Python package: *pip install ppdp-anonops*
- **PRIPEL (external library)**

[2] https://pypi.org/

  - GitHub: *https://github.com/samadeusfp/PRIPEL*
  - Python package: *pip install pp-pripel*

- **Privacy analysis**
  - GitHub: *https://github.com/m4jidRafiei/privacy-quantification*
  - Python package: *pip install p-privacy-qt*

- **FCB-anonymity**
  - GitHub: *https://github.com/m4jidRafiei/PP_CEDP*
  - Python package: *pip install pp-cedp*

- **Privacy metadata**
  - GitHub: *https://github.com/m4jidRafiei/privacy-metadata*
  - Python package: *pip install p-privacy-metadata*

Our infrastructure provides a hierarchy of usages such that users can use each technique independently. Moreover, they can use PC4PM which integrates a set of privacy preservation techniques as a stand-alone web-based application. The source code of PC4PM is available in a GitHub repository under *GNU General Public License v3.0*: *https://github.com/m4jidRafiei/PC4PM*. An executable release of PC4PM is also provided as a Docker container[3] which can simply be hosted and run by the users using the following commands:

```
docker pull m4jid/pc4pm
docker run -d -p 8000:8000 m4jid/pc4pm
```

Note that for using Docker, first you need to install Docker accourding to your operation system.[4] After running the docker, use your browser and enter the following address to run the web-based application: http://127.0.0.1:8000/

The scalability of the tool varies with respect to the privacy preservation technique and the size of the input event log. Based on our experiments, our tool can handle real-world event logs, e.g., the BPI challenge datasets[5].

3 Event Data Management

The *event data management* module contains two tabs to upload and manage the event data that could be standard XES event logs or non-standard event data, called *Event Log Abstraction* (ELA) [6]. In this module, an event log can be set as the input for the privacy preservation techniques. Figure 3 shows the event data management module of PC4PM.

4 User Guidance Wizard

To help users find a low effort entry into the usage of our tool, a help wizard is integrated into the landing page of the application. A user can specify certain needs in four aspects of privacy for a request and the wizard highlights the most appropriate techniques to fullfill the request with. There are four areas of information which can be entered for a help request:

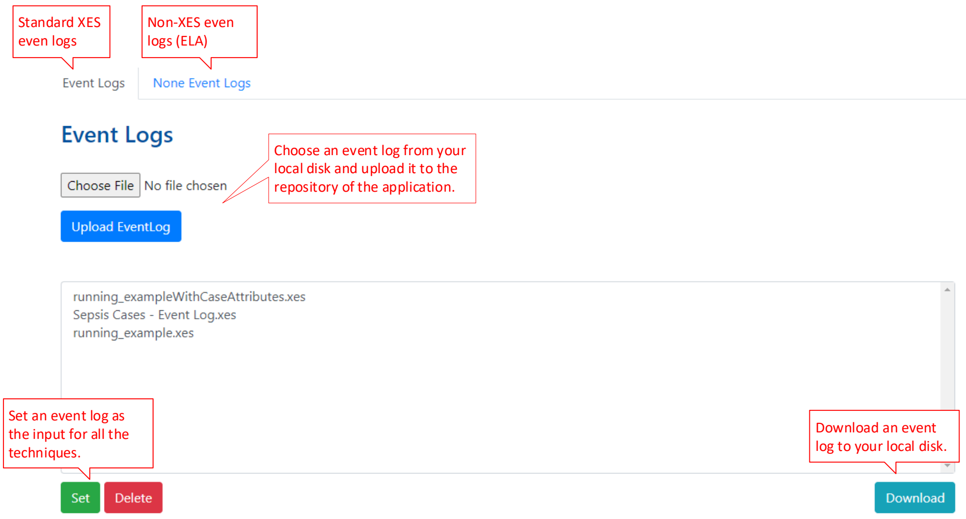
1. *Process Mining Perspective (PmPs)*: The process mining perspective(s) that a privacy technique focuses on, e.g., control-flow.

2. *Process Mining Activity (PmAc)*: Sets the process mining activity(ies), i.e, process discovery or role mining, which the utility of event data should be preserved for.

[3] https://hub.docker.com/r/m4jid/pc4pm

[4] https://docs.docker.com/get-docker/

[5] https://data.4tu.nl/



Standard XES even logs

Non-XES even logs (ELA)

Event Logs

Choose an event log from your local disk and upload it to the repository of the application.

running_exampleWithCaseAttributes.xes
Sepsis Cases - Event Log.xes
running_example.xes

Set an event log as the input for all the techniques.

Download an event log to your local disk.

Figure 3: The event data management module of the application.

3. *Privacy Perspective (PrPs)*: Sets the privacy perspective of a privacy technique, i.e., resource or case.

4. *Privacy Activity (PrAc)*: Indicates the corresponding privacy-related activity(ies), i.e., PPDP, PPPM, or PrAn.

After completing these settings, the links to the applications of the matching techniques are highlighted in the navigation menu. An example for a query would be selecting the following for the required information: Control-Flow, Time, and Case for *PmPs*, Process Discovery, Process Discovery Based on DFG, DFG Discovery, and Performance Analysis for *PmAc*, Case for *PrPs* and PPDP for *PrAc* would result in the *TLKC-privacy* and the *Anonymization operations* being highlighted.

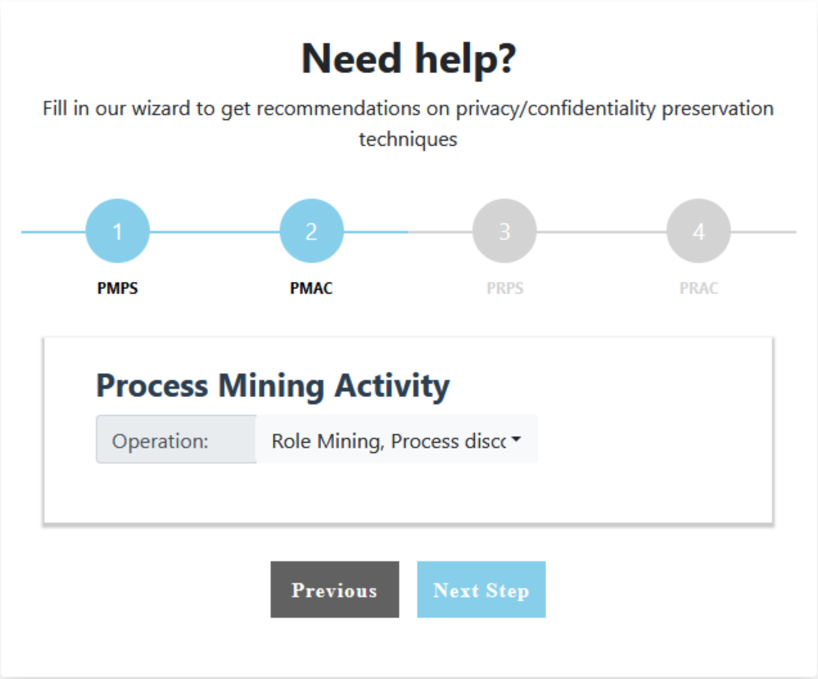
From this highlighting on, all applications are equipped with information tooltips for the configuration parameters of the corresponding technique. Once an application is selected, these tooltips give additional advice on how to use a technique and what the parameters a user can modify are used for. Hence, we provide different types of support measures to guide a user through the application and help to take effective measures to improve privacy.

5 Privacy-aware Role Mining

The *privacy-aware role mining* technique focuses on the organizational perspective of process mining. It provides anonymous *joint-activities* social network discovery that is used to identify roles in event logs. The goal is to identify roles without revealing *who performed what?*. The challenge is to mitigate frequency-based attacks which can be launched even against the encrypted data [4].

The main idea of the *privacy-aware role mining* technique is to decompose activities into other activities such that the *frequency* and *position* of activities get perturbed. However, the similarities between resources should remain as similar as possible. To this end, we need to determine the number of substitutions for each activity, and the way of distributing the frequency of the main activity among its substitutions. This technique uses three main approaches to decompose the main activities into substitutions including *fixed-value*, *selective*, and *frequency-based* [4]. The frequency of the main activity is uniformly distributed among its substitutions. The distribution could be done with respect to the resource, i.e., *resource-aware*, or without considering the resources. The latter is not recommended since it could result in different similarity values for the resources.

The *fixed-value* technique decompose all the activities of the event log to the fixed number of activities specified by users. In the *selective* technique, only the sensitive frequencies are targeted



Need help?

Fill in our wizard to get recommendations on privacy/confidentiality preservation techniques

Process Mining Activity

Operation: Role Mining, Process disc

Figure 4: The help wizard integrates into the landing page to provide guidance for new users.

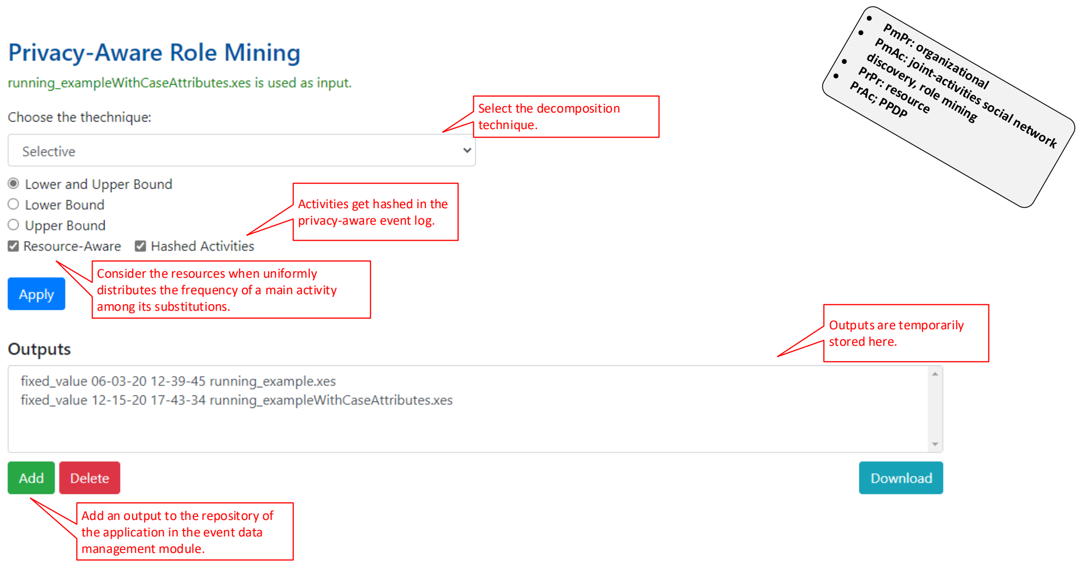
to get perturbed which can be specified by users. The *lower* and *upper* bound of frequencies can be chosen as the sensitive frequencies. Activities are categorized into the lower bound of frequencies if their frequency is below the *lower quartile* of the box plot of frequencies, and they are categorized into the upper bound of frequencies if their frequency is above the *upper quartile* of the box plot. In the *frequency-based* technique, the substitutions are allocated based on the relative frequencies of the main activities. A fixed value can also be added to the new frequencies.

After applying a technique, the privacy-aware event log in the XES format is provided in the corresponding *outputs* section. The generated event log preserves the data utility for mining roles from *resources* without exposing the set of activities performed by resources. Figure 5 shows the privacy-aware role mining module of PC4PM where the decomposition technique is set to *selective*. The gray box in the figure shows the signature of the technique which contains the process mining and privacy perspectives and activities of the technique.

6 Connector Method

The *connector method* focuses on the *control-flow* perspective of process mining and implements an encryption-based method for discovering directly-follows graphs [11, 12]. It breaks the traces down into the collection of directly-follow relations which are securely stored in a data structure. After applying the method, the privacy-aware event data are provided in the corresponding *outputs* section as an XML file with the ELA format [6]. Figure 6 shows a part of an ELA as output. The *header* section contains metadata containing *name of the original event log* (if exists), *name of the method*, and *desired analyses* which could be performed given this abstraction of the event log.

Note that the connector method eliminates the concept of trace, i.e., the sequence of activities which is very useful to mitigate many attacks that are possible based on little information about traces, e.g., the length of a trace, a specific sequence of activities, etc. However, it also *restricts* the analyses which could be done based on traces. We say "restricts" and not "eliminates" because users are still able to selectively restore trace of specific cases on-demand using the information stored in the *connector* field. The *connector* filed contains the encrypted value of the (activity, previous-activity) pairs that can be used to reconstruct the original trace of a case.



Privacy-Aware Role Mining

running_exampleWithCaseAttributes.xes is used as input.

Choose the thechnique:

Selective

● Lower and Upper Bound
○ Lower Bound
○ Upper Bound
☑ Resource-Aware ☑ Hashed Activities

Apply

Outputs

fixed_value 06-03-20 12-39-45 running_example.xes
fixed_value 12-15-20 17-43-34 running_exampleWithCaseAttributes.xes

Add Delete Download

Select the decomposition technique.

Activities get hashed in the privacy-aware event log.

Consider the resources when uniformly distributes the frequency of a main activity among its substitutions.

Outputs are temporarily stored here.

Add an output to the repository of the application in the event data management module.

- PmPr: organizational
- PmAc: joint-activities social network discovery, role mining
- PrPr: resource
- PrAc; PPDP

Figure 5: The privacy-aware role mining module of the application. The gray box is the signature of the technique.

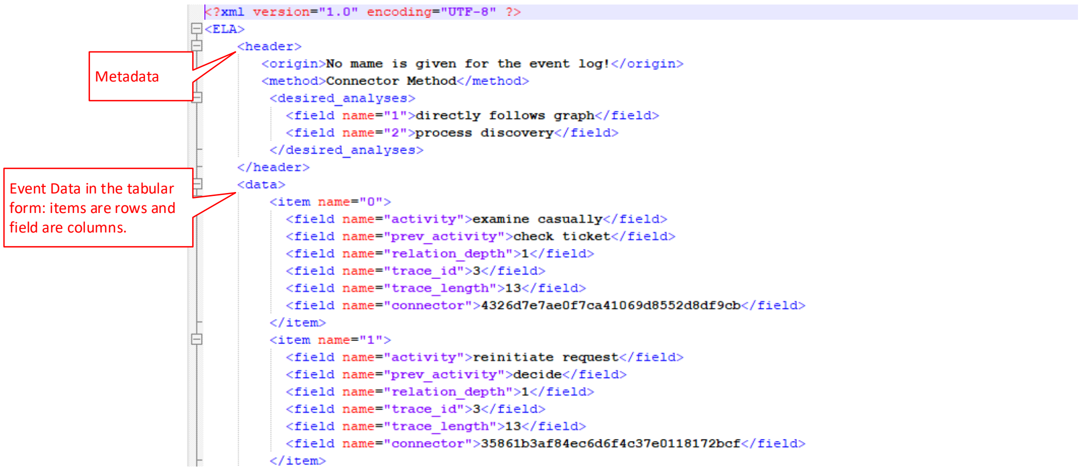
As can be seen in Figure 7, users are able to specify the *encryption key* and *encryption method* which are used to encrypt the value of connector filed. Moreover, one can specify the extra information to store in the resulting event log abstraction. The gray box in the figure shows the signature of the technique.

7 $TLKC$-privacy Model

The $TLKC$-*privacy* module focuses on the *control-flow*, *time*, and *case* perspectives. It implements the $TLKC$-privacy model for process mining [10] that provides group-based privacy guarantees against case and attribute *linkage attacks*. *Case linkage* attacks aim to single out a case based on some background knowledge. *Attribute linkage* attacks try to realize *sensitive attributes* belonging to a case. It assumes four main types of background knowledge: *set*, *multiset*, *sequence*, and *relative*. The *set* type of background knowledge assumes that an adversary knows a subset of activities having been performed for the case, and this information can lead to the case and/or attribute linkage attacks. The *multiset* type of background knowledge assumes that the adversary knows a sub-multiset of activities having been done for the victim case. The *sequence* type of background knowledge assumes that the adversary knows a subsequence of activities having been performed for the case. The *relative* type extends the *sequence* type with the relative time differences information such that we assume the adversary knows the sequence of activities and also the relative time differences between activities.

In the $TLKC$-privacy model, T refers to the accuracy of timestamps in the privacy-aware event log, L refers to the power of background knowledge, K refers to the k in the k-anonymity definition, and C refers to the bound of confidence regarding the sensitive attribute values in an equivalence class. The power of background knowledge refers to the size of candidates of background knowledge with respect to the assumed type, e.g., $\{a, b\}$ is a candidate of set background knowledge with size 2. The accuracy of timestamps could be generalize to *seconds*, *minutes*, *hours*, and *days*. The timestamps of an original event log get generalized before applying the privacy technique.

$TLKC$-privacy guarantees that assuming a specific type of background knowledge and an integer value L as the size of background knowledge, each case falls into a group of K cases that all have the same values for the assumed type and size of background knowledge. Moreover, it guarantees that the cases of the same group have different values for the sensitive attribute(s) in such a way that the adversary cannot infer the sensitive value of each case with a confidence


Metadata
Event Data in the tabular form: items are rows and field are columns.

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<ELA>
    <header>
       <origin>No mame is given for the event log!</origin>
       <method>Connector Method</method>
        <desired_analyses>
          <field name="1">directly follows graph</field>
          <field name="2">process discovery</field>
        </desired_analyses>
    </header>
    <data>
        <item name="0">
          <field name="activity">examine casually</field>
          <field name="prev_activity">check ticket</field>
          <field name="relation_depth">1</field>
          <field name="trace_id">3</field>
          <field name="trace_length">13</field>
          <field name="connector">4326d7e7ae0f7ca41069d8552d8df9cb</field>
        </item>
        <item name="1">
          <field name="activity">reinitiate request</field>
          <field name="prev_activity">decide</field>
          <field name="relation_depth">1</field>
          <field name="trace_id">3</field>
          <field name="trace_length">13</field>
          <field name="connector">35861b3af84ec6d6f4c37e0118172bcf</field>
        </item>
```

Figure 6: A part of an event log abstraction resulting from the connector method.

higher than C. Applying this method results in a privacy-aware event log in the XES format that preserves data utility for process discovery and performance analysis. Figure 8 shows the $TLKC$-privacy module of PC4PM. Again, the gray box in the figure shows the signature of the technique. Note that Θ is used as the frequency threshold to discover the set of *maximal frequent traces* (MFT) which is used as the utility measure by the technique.

8 $TLKC$-privacy Model (Extended Version)

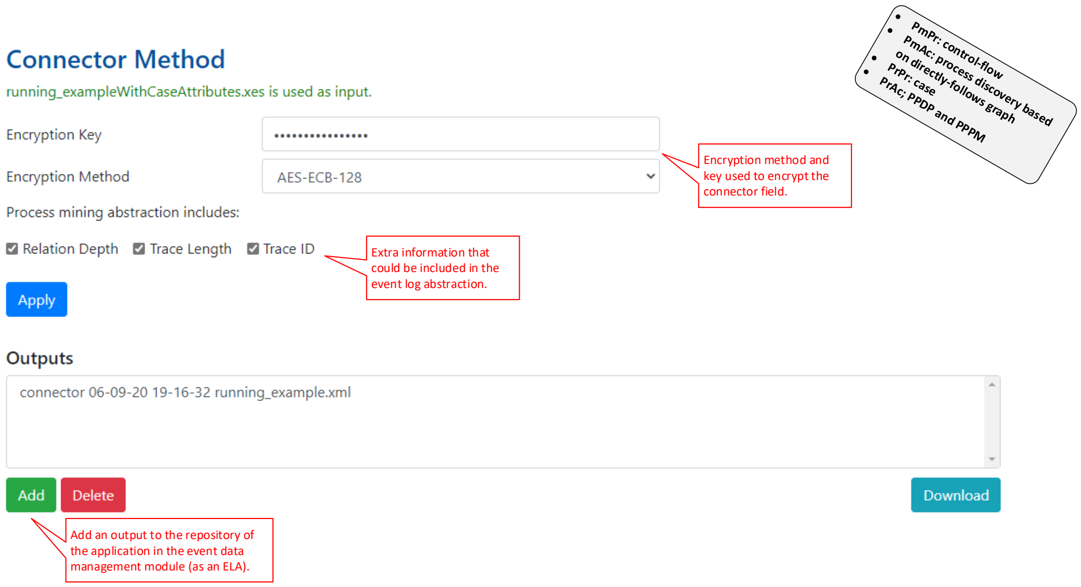
In [8], we propose the extended version of $TLKC$-privacy (Section 7) covering all the main perspectives of process mining including *control-flow*, *time*, *case*, and *organizational* perspectives. It empowers the adjustability of the proposed technique by adding new parameters to adjust privacy guarantees and the loss of accuracy. Moreover, it defines a new utility measure to tackle the drawbacks of the $TLKC$-privacy model.

Figure 9 shows the extended version of $TLKC$-privacy module in PC4PM. The process mining perspective that the model focuses on is specified by the *background knowledge type* and *background knowledge attribute* (the orange text below the type and attribute of background knowledge is changed when the user selects the type and attribute). The extended version of $TLKC$-privacy provides the same type of guarantees as the main algorithm for more aspects of event data. The main algorithm focuses on the activity attribute for analyzing the possible attacks. However, the extended version focuses on both activities and resources. In the extended version, users are able to adjust the importance of privacy and utility using the parameters α and β. Note that β indicates a negative effect on the utility, i.e., a higher value for *beta* results is removing events having a higher negative effect on the utility.

9 Anonymization Operations

Anonymization operations are the main functions which are employed by privacy preservation techniques to provide the desired privacy requirements. The anonymization operations alone do not provide privacy guarantees such as k-anonymity. They may have any number of parameters and can be applied at the case or event level, and the target of the operation could be a case, an event, or attributes of such an object. In [6], the main anonymization operations in process mining are listed as follows:

- *suppression* (*sup*)
- *addition* (*add*)



Connector Method

running_exampleWithCaseAttributes.xes is used as input.

Encryption Key: ••••••••••••••••

Encryption Method: AES-ECB-128

Process mining abstraction includes:

☑ Relation Depth ☑ Trace Length ☑ Trace ID

Outputs

connector 06-09-20 19-16-32 running_example.xml

- PmPr: control-flow
- PmAc: process discovery based on directly-follows graph
- PrPr: case
- PrAc; PPDP and PPPM

Encryption method and key used to encrypt the connector field.

Extra information that could be included in the event log abstraction.

Add an output to the repository of the application in the event data management module (as an ELA).

Figure 7: The connector method of the application. The gray box is the signature of the technique.

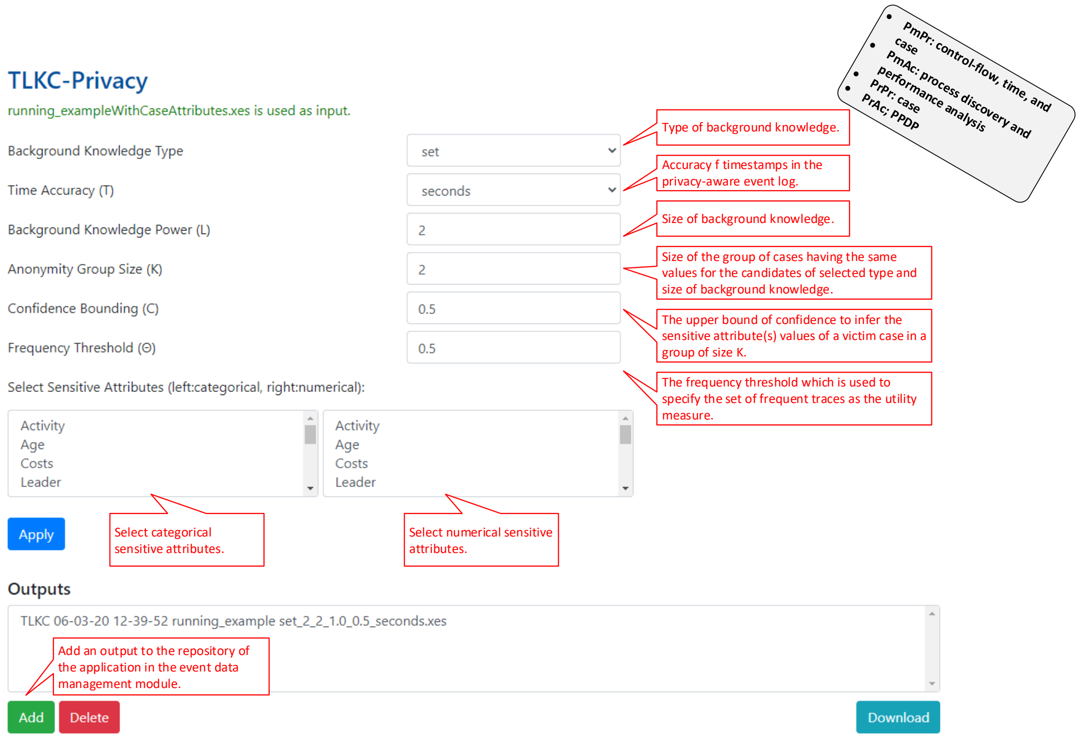
- *substitution* (*sub*)
- *condensation* (*con*)
- *swapping* (*swa*)
- *generalization* (*gen*)
- *cryptography* (*cry*).

The *anonymization operations* module of PC4PM implements the above-mentioned anonymization operations. Figure 10 shows the main page of the anonymization operations module in the tool. As can be seen, it has four main sections including *operation selection*, *operation configuration*, *operation ordering*, and *outputs*.

1. *Operation selection*: lets users select type and level of operation.
2. *Operation configuration*: depending on the selected operation this box provides certain configuration options, specifying the details of the anonymization operation, as well as possible filters. Note that all changes to a configuration need to be saved by pressing the “Save” button.
3. *Operation ordering and apply*: the ordering of the operations in this list is the order they are applied to the selected log. This order can be changed by selecting an entry and using the “Up” / “Down” buttons below the list. Execute the configured operations on the selected log by pressing the “Apply Operations” button.
4. *Outputs*: box shows newly created logs, offering the choice of downloading or deleting it as well as adding the new log to the log storage.

9.1 Addition

The addition operation is meant to inject new events into the traces of a log. As the events should match the specific log selected, they can be designed manually by hitting the green button “Create New” which opens the event designer dialog. Within the designer event attributes can de added/deleted using the “ADD NEW” or “Delete” buttons. Note that manually specifying



- PmPr: control-flow, time, and case
- PmAc: process discovery and performance analysis
- PrPr: case
- PrAc; PPDP

TLKC-Privacy

running_exampleWithCaseAttributes.xes is used as input.

Background Knowledge Type: set — Type of background knowledge.

Time Accuracy (T): seconds — Accuracy f timestamps in the privacy-aware event log.

Background Knowledge Power (L): 2 — Size of background knowledge.

Anonymity Group Size (K): 2 — Size of the group of cases having the same values for the candidates of selected type and size of background knowledge.

Confidence Bounding (C): 0.5 — The upper bound of confidence to infer the sensitive attribute(s) values of a victim case in a group of size K.

Frequency Threshold (Θ): 0.5 — The frequency threshold which is used to specify the set of frequent traces as the utility measure.

Select Sensitive Attributes (left:categorical, right:numerical):

Activity Age Costs Leader	Activity Age Costs Leader
Select categorical sensitive attributes.	Select numerical sensitive attributes.

Outputs

TLKC 06-03-20 12-39-52 running_example set_2_2_1.0_0.5_seconds.xes

Add an output to the repository of the application in the event data management module.

Figure 8: The $TLKC$-privacy module of the application. The gray box is the signature of the technique.

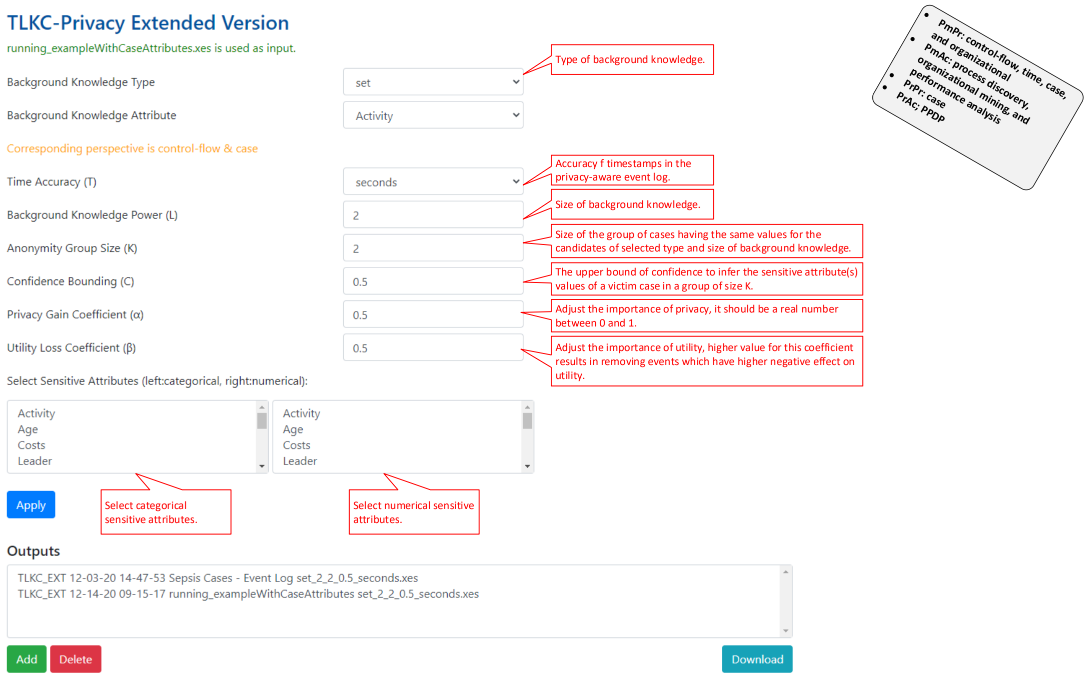
timestamp attributes is currently not supported. They will get generated randomly with a deviation of 1 to 3600 seconds according to the surrounding events in the trace. Once created, the customized events can be chosen to be applied by the addition operation in the last configuration row "Events to add". Further, they can be edited or deleted using the corresponding buttons in the first row of configuration.

Figure 11 shows the configuration section of the addition operation. Three adding options are available which can be chosen using the Addition-Operation (Config.-Row 2):

- "Add new event as first in trace"
- "Add new event as last in trace"
- "Add new event at random position"

These operations will always be performed if the checkbox on Config.-Row 3 is not checked. If the checkbox is selected, a filter can be applied, offering a selection to which traces of the selected events are added. The filter consists of four parts: *match-type*, *attribute*, *operator*, and *value of the attribute*. The *match-type* could be one of the following types:

- "Apply condition on case attributes": The events will be added to a trace iff the case contains the selected filter attribute and this attribute matches the given value.
- "Trace Length": The events will be added to a trace iff the number of events in the trace matches the given value.
- "Match on first event in trace needed": Addition will only take place iff the first event in the trace contains the selected filter attribute and this attribute matches the given value.
- "Match on last event in trace needed": The last event in the trace needs to meet the filter specifications in order to apply the addition operation.



TLKC-Privacy Extended Version

running_exampleWithCaseAttributes.xes is used as input.

Background Knowledge Type: set

Background Knowledge Attribute: Activity

Corresponding perspective is control-flow & case

Time Accuracy (T): seconds

Background Knowledge Power (L): 2

Anonymity Group Size (K): 2

Confidence Bounding (C): 0.5

Privacy Gain Coefficient (α): 0.5

Utility Loss Coefficient (β): 0.5

Select Sensitive Attributes (left:categorical, right:numerical):

Activity
Age
Costs
Leader

Outputs

TLKC_EXT 12-03-20 14-47-53 Sepsis Cases - Event Log set_2_2_0.5_seconds.xes
TLKC_EXT 12-14-20 09-15-17 running_exampleWithCaseAttributes set_2_2_0.5_seconds.xes

Type of background knowledge.

Accuracy f timestamps in the privacy-aware event log.

Size of background knowledge.

Size of the group of cases having the same values for the candidates of selected type and size of background knowledge.

The upper bound of confidence to infer the sensitive attribute(s) values of a victim case in a group of size K.

Adjust the importance of privacy, it should be a real number between 0 and 1.

Adjust the importance of utility, higher value for this coefficient results in removing events which have higher negative effect on utility.

Select categorical sensitive attributes.

Select numerical sensitive attributes.

- PmPr: control-flow, time, case, and organizational
- PmAc: process discovery, organizational mining, and performance analysis
- PrPr: case
- PrAc; PPDP

Figure 9: The extended version of $TLKC$-privacy module of the application. The gray box is the signature of the technique.

- "Match on any event in trace needed": At least one event in the trace, position doe not matter, needs to meet the filter conditions.
- "Match on all events in trace needed": All events in the trace need to meet the filter conditions.

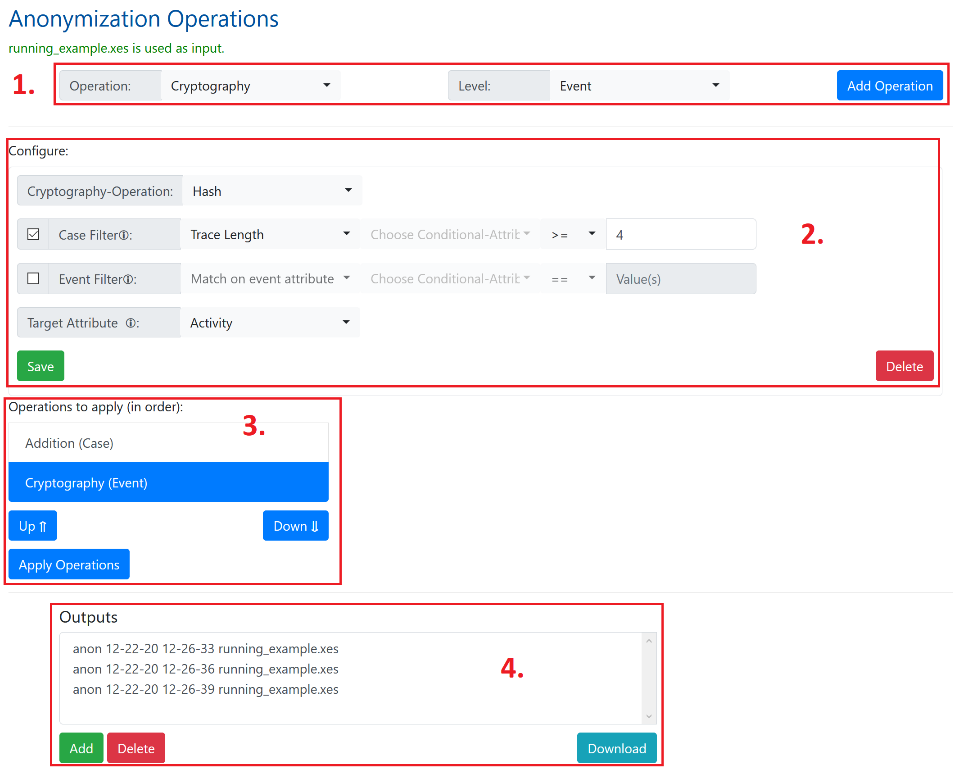
Following the sample from Figure 11, the events "CustomEvent" and "TestEvent" will be added in the beginning of a trace if and only if the first event in the trace contains an attribute "Age" with value larger or equal to 50.

9.2 Condensation

Condensation first condenses the cases into similar clusters based on the sensitive attribute values, then in each cluster, the sensitive attribute values are replaced with a collective statistical value, e.g., mean or mode, of the cluster. Figure 12 shows the configuration options for the condensation operation. Three different clustering techniques can be chosen:

- "kMeans": The attribute data of all cases (or events, depending on which level was selected when adding this operation) will be separated into the specified amount of clusters using the kMeans clustering technique (Config.-Row 2 "Num. of clusters"). If there have been attributes with categorical instead of numerical data, the data of these columns will first be OneHot-Encoded, then the resulting vectors will be mapped onto a normalization into the interval [0,1]. The "Weight Designer" in Config.-Row 4 is not available for this option!
- "kModes": For clustering numerical and categorical data a kModes approach is available, utilizing the python package "kmodes"[6], which implements a basic k-modes approach with randomly initialized cluster centroids and a matrix dissimilarity measure by Ng et al.[3]. This clustering approach will consider both the target attribute and some descriptive attributes. The "Weight Designer" in Config.-Row 4 is not available for this option!

[6] https://github.com/nicodv/kmodes



Anonymization Operations

running_example.xes is used as input.

1. Operation: Cryptography — Level: Event

2. Configure:

Cryptography-Operation: Hash

☑ Case Filter: Trace Length — Choose Conditional-Attrib — >= — 4

☐ Event Filter: Match on event attribute — Choose Conditional-Attrib — == — Value(s)

Target Attribute: Activity

3. Operations to apply (in order):

Addition (Case)

Cryptography (Event)

4. Outputs

anon 12-22-20 12-26-33 running_example.xes
anon 12-22-20 12-26-36 running_example.xes
anon 12-22-20 12-26-39 running_example.xes

Figure 10: Anonymization Operations - Interface

- “Weighted Euclidian Distance”: This clustering techniques are applicable for numerical and categorical data. Categorical data will be OneHot-Encoded, then the resulting vectors will be mapped onto a normalization into the interval [0,1]. Following this step, cluster centroids will be randomly chosen and the (weighted) distance between these centroids and all other cases/events will be calculated using the following equation where A and B are the vectors of case/event attributes and w is the weight of each attribute in the vector. All cases/events are assigned to the cluster with the lowest distance to the cluster's centroid.

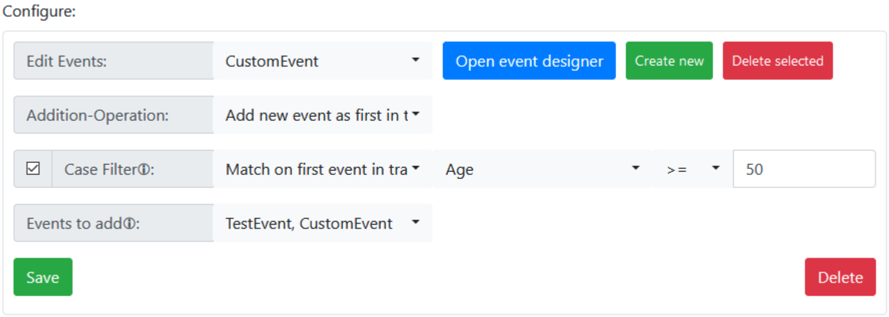
$$d(A,B,w) = \sqrt{\sum_i w_i(A_i - B_i)^2} \tag{1}$$

After the clusters have been built, one of three assignment methods can be chosen:

- “Mode”: Determines the mode for all the available values of the target attribute in a cluster, and overwrites all attribute instances in the cluster with this value.

- “Mean”: Determines the mean for all the available values of the target attribute in a cluster and overwrites all attribute instances in the cluster with this value. (Only applicable if the target attribute has been stored as a numerical value in the XES file).

- “Median”: Determines the median for all the available values of the target attribute in a cluster and overwrites all attribute instances in the cluster with this value. (Only applicable if the target attribute has been stored as a numeric value in the XES file).



Configure:

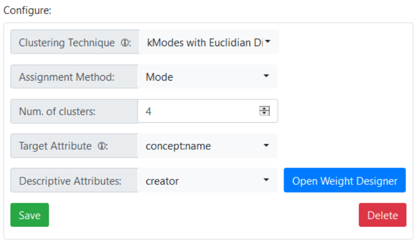
Edit Events: CustomEvent

Addition-Operation: Add new event as first in t

☑ Case Filter①: Match on first event in tra | Age | >= | 50

Events to add①: TestEvent, CustomEvent

Figure 11: Addition - Configuration



Configure:

Clustering Technique ①: kModes with Euclidian Di

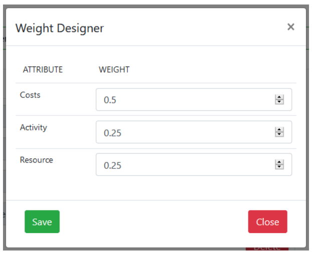
Assignment Method: Mode

Num. of clusters: 4

Target Attribute ①: concept:name

Descriptive Attributes: creator

Figure 12: Condensation - Configuration



Weight Designer

ATTRIBUTE	WEIGHT
Costs	0.5
Activity	0.25
Resource	0.25

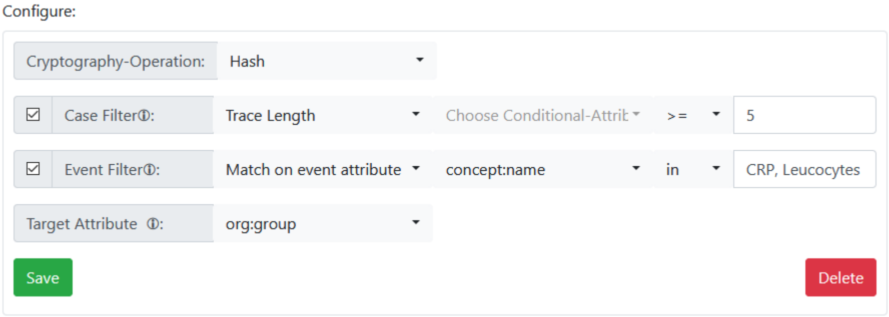
Figure 13: Condensation - Weight Designer

9.3 Cryptography

For the cryptography operation two different techniques are available:

- Hash: The target attributes' value is hashed using the Ripemd160 algorithm and a built-in salt, preventing simple rainbow table attacks.
- Encrypt: The target attributes' value will be encrypted using the python "Fernet" library (AES 128-Bit in CBC mode using PKCS7 padding[7]), currently using a built in default password.

The operation can also be applied conditionally when the conditional checkbox is set. As Figure 14 shows, the hashing operation would only be applied to events in traces with at least 5 events and iff the "concept:name" attribute is either "CRP" or "Leucocytes". Whether the target attribute is a case or event attribute is determined by the level selected when the cryptography operation is added.



Configure:

Cryptography-Operation: Hash

☑ Case Filter①: Trace Length | Choose Conditional-Attrit | >= | 5

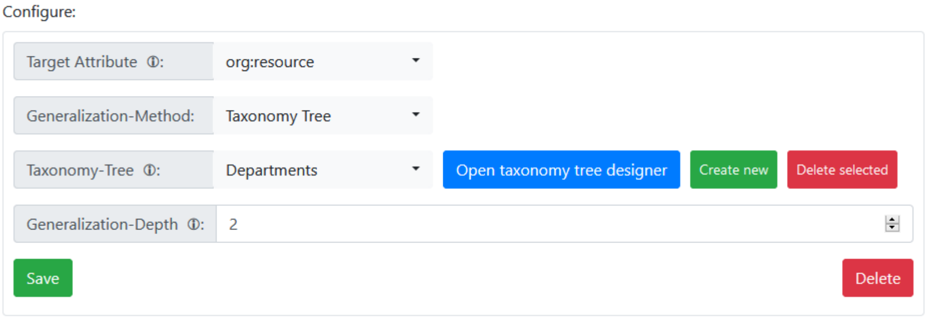
☑ Event Filter①: Match on event attribute | concept:name | in | CRP, Leucocytes

Target Attribute ①: org:group

Figure 14: Cryptography - Configuration

[7] https://cryptography.io/en/latest/fernet.html

9.4 Generalization


Configure:

Target Attribute ⓘ: org:resource

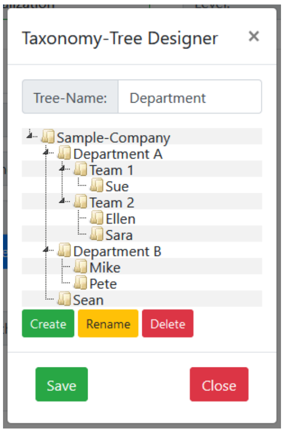
Generalization-Method: Taxonomy Tree

Taxonomy-Tree ⓘ: Departments

Generalization-Depth ⓘ: 2

Figure 15: Generalization - Taxonomy Tree Configuration

The generalization operation consists of two different approaches. (1) an approach using taxonomy trees for generalization as shown in Figure 15. (2) a generalization mode targeting the timestamp attribute of events (Figure 17). Using the taxonomy tree approach, a target attribute and a generalization depth can be specified. After the creation of a taxonomy tree, using the designer shown in Figure 16 (Config.-Row 3), a generalization depth can be specified. This depth represents the steps taken from the leaf node upwards during the generalization (e.g. an attributes value is "Sara" and the depth is given as 2, the resulting value for the attribute would be "Department A"). If for a certain leaf value, the depth exceeds the actual possible steps up to the root node, the top most value possible is chosen, which in this example would be "Sample-Company".


Taxonomy-Tree Designer

Tree-Name: Department

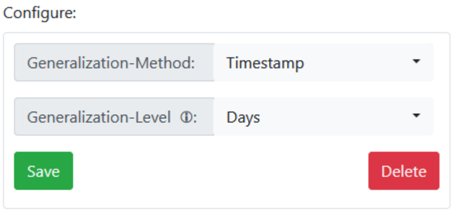
- Sample-Company
  - Department A
    - Team 1
      - Sue
    - Team 2
      - Ellen
      - Sara
  - Department B
    - Mike
    - Pete
  - Sean

Figure 16: Taxonomy Tree Designer

As this generalization is targeted against categorical values, a different generalization mode is offered for the "org:timestamp" attribute. By setting a generalization level, like "Days" in Figure 17, all time-values below this level will be pruned. Given the timestamp "07.12.2020 12:34:56.789", the possible results with respect to the levels are shown in Table 1:

Level	Result
Seconds	2020.12.07 12:34:56.000
Minutes	2020.12.07 12:34:00.000
Hours	2020.12.07 12:00:00.000
Days	2020.12.07 00:00:00.000
Months	2020.12.01 00:00:00.000
Years	2020.01.01 00:00:00.000

Table 1: Generalization - Timestamp Levels


Configure:

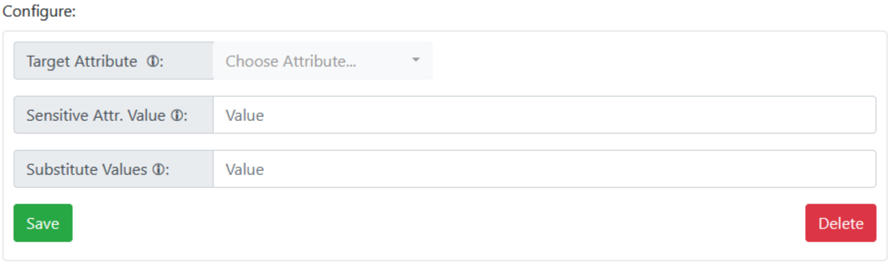
Generalization-Method: Timestamp

Generalization-Level ⓘ: Days

Figure 17: Generalization - Timestamp Configuration

9.5 Substitution

Replaces the sensitive attribute values, which can be entered comma-separated, of the selected target attribute with a random value, taken from the same attribute of other events. If there are substitute values provided, these values will be taken and distributed randomly instead of taking the non-sensitive values from the log. After the substitution, the sensitive value will not appear anymore as a target attribute value in the resulting event log.



Configure:

Target Attribute ⓘ: Choose Attribute...

Sensitive Attr. Value ⓘ: Value

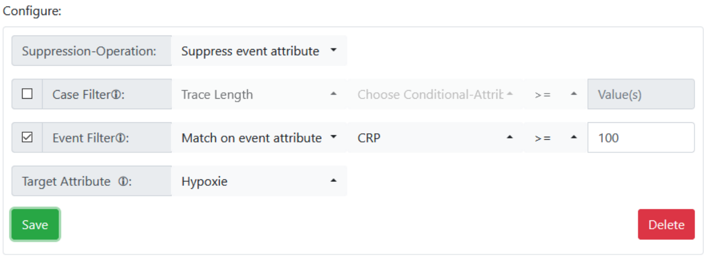
Substitute Values ⓘ: Value

Figure 18: Substitution - Configuration

9.6 Suppression

Depending on the operation level chosen when adding the suppression operation there are two different actions to apply to the log:

- Suppress case/event entirely: The conditional filter has to be set for this mode to work. If a case/event contains the specified conditional attribute and its value equals the filter value, the case/event is removed from the log/trace. Note that no target attribute is required for this mode, so Config.-Row 4 is not available.
- Suppress case/event attribute: This mode sets the value of the targeted case/event attribute to "None". If no filter is applied, all instances of the target attribute will be replaced. Otherwise, only cases/events matching the filter will be altered.



Configure:

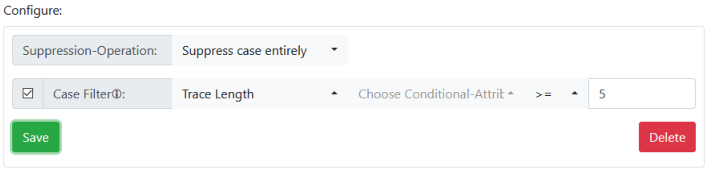
Suppression-Operation: Suppress event attribute

□ Case Filterⓘ: Trace Length | Choose Conditional-Attrit | >= | Value(s)

☑ Event Filterⓘ: Match on event attribute | CRP | >= | 100

Target Attribute ⓘ: Hypoxie

Figure 19: Suppression - Configuration



Configure:

Suppression-Operation: Suppress case entirely

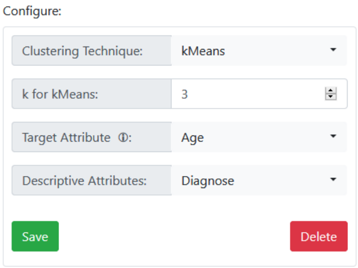
☑ Case Filterⓘ: Trace Length | Choose Conditional-Attrit | >= | 5

Figure 20: Suppression - Trace Length

9.7 Swapping

Swapping aims to anonymize data by exchanging the values of a sensitive attribute between individual cases. The individual cases which are chosen to exchange the sensitive attribute values are supposed to have similar sensitive attribute values. Therefore, cases need to be clustered into clusters with similar sensitive attribute values. The cases for swapping in the same cluster are done randomly. The swapping operation supports the same clustering techniques as the condensation operation in Section 9.2. Therefore, by using a kMeans based approach with categorical values, they will be OneHot-Encoded and normalized to the interval [0,1] to apply the kMeans clustering.

After applying the clustering on the selected attribute, a new value out of all available attribute values in the according cluster is chosen randomly and assigned to the case/event attribute. This random selection may as well include the original attribute value.



Configure:

Clustering Technique: kMeans

k for kMeans: 3

Target Attribute ⓘ: Age

Descriptive Attributes: Diagnose

Figure 21: Swapping - Configuration

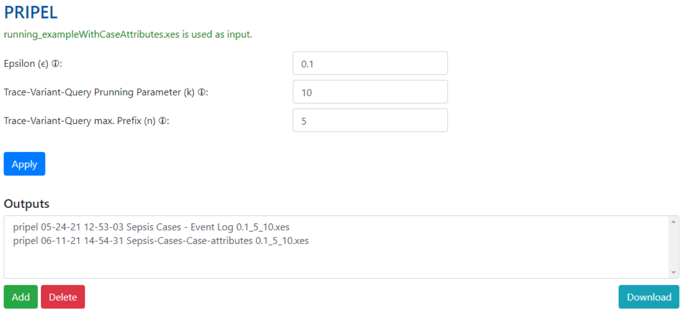
10 PRIPEL

To demonstate the flexibility of the PC4PM application, this module implements an external package for the *PRIPEL* (Privacy-preserving event log publishing with contextual information) technique, proposed in [2]. In contrast to approaches providing privacy guarantees for an entire event log while focusing on a process' control-flow, PRIPEL was designed to provide privacy a level of individual cases instead. The technique attempts to preserve contextual information, e.g., case attribute values, to allow for a lager variety of later analyses.

To use the technique the user has to select an event log first, and set the parameters of the algorithm using the interface shown in Fig. 22.

The epsilon (ϵ) parameter is a floating point number, specifying the strength of the differential privacy guarantee. Parameter value n specifies the maximal length of the prefix of considered traces, used for the trace variant query of the PRIPEL technique. Last, the parameter k is defined as the pruning parameter of the trace variant query. It is used to control the number of explored potential activity sequences in the trace variant query. Here, a higher value allows to take only more common prefixes into consideration and therefore reduce the runtime of the technique on the cost of the results data utility.

As an output the package produces a XES file, which is again listed in the already shown output section of the application.



PRIPEL

running_exampleWithCaseAttributes.xes is used as input.

Epsilon (ϵ) ⓘ: 0.1

Trace-Variant-Query Prunning Parameter (k) ⓘ: 10

Trace-Variant-Query max. Prefix (n) ⓘ: 5

Outputs

pripel 05-24-21 12-53-03 Sepsis Cases - Event Log 0.1_5_10.xes
pripel 06-11-21 14-54-31 Sepsis-Cases-Case-attributes 0.1_5_10.xes

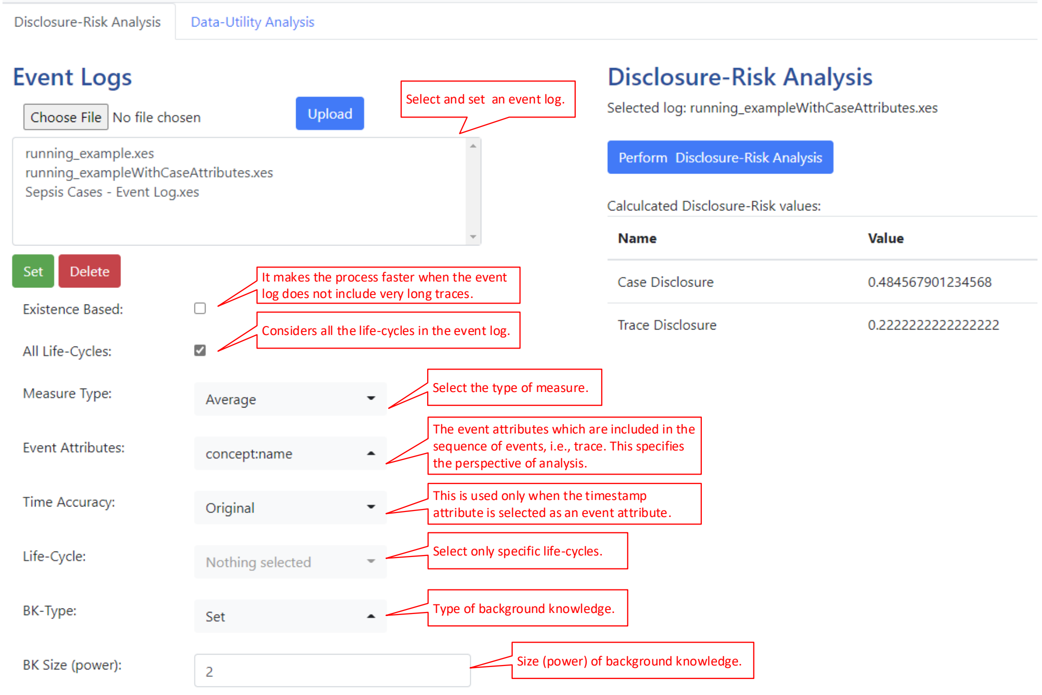
Figure 22: The parameter adjustment UI for the PRIPEL module.

The concept of mounting additional Django applications into the PC4PM application allows for an easy extension as shown with this technique. Future upcoming techniques can be implemented in a similar way.

11 Privacy Analysis

This module implements the techniques for analyzing privacy of event logs proposed in [7]. It quantifies *disclosure risks*, *data utility*, and *FCB-anonymity* [9]. Two measures are provided to quantify the disclosure risks associated with event logs: *identity (case) disclosure* and *attribute (trace) disclosure*.

The *case disclosure* is calculated based on the uniqueness of cases based on the candidates of background knowledge for the given type and size of background knowledge. Figure 23 shows the *disclosure risk analysis* module of PC4PM.



Event Logs

No file chosen

running_example.xes
running_exampleWithCaseAttributes.xes
Sepsis Cases - Event Log.xes

Existence Based: □

All Life-Cycles: ☑

Measure Type: Average

Event Attributes: concept:name

Time Accuracy: Original

Life-Cycle: Nothing selected

BK-Type: Set

BK Size (power): 2

Select and set an event log.

It makes the process faster when the event log does not include very long traces.

Considers all the life-cycles in the event log.

Select the type of measure.

The event attributes which are included in the sequence of events, i.e., trace. This specifies the perspective of analysis.

This is used only when the timestamp attribute is selected as an event attribute.

Select only specific life-cycles.

Type of background knowledge.

Size (power) of background knowledge.

Disclosure-Risk Analysis

Selected log: running_exampleWithCaseAttributes.xes

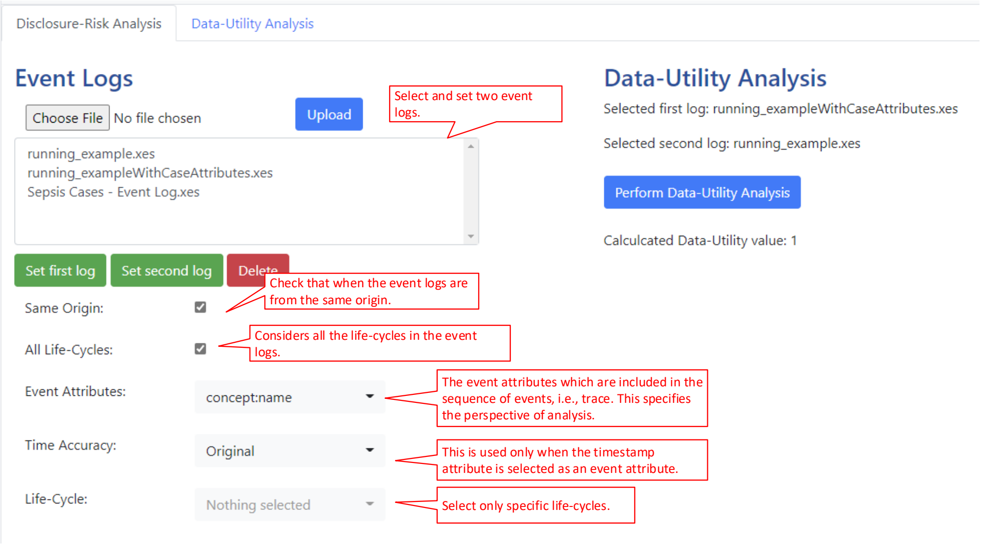
Calculcated Disclosure-Risk values:

Name	Value
Case Disclosure	0.484567901234568
Trace Disclosure	0.2222222222222222

Figure 23: The disclosure risk analysis module of the application.

The *trace disclosure* is obtained based on the entropy of trace variants which are categorized in the same group given the type and size of background knowledge. Intuitively, lower values for the trace disclosure indicate higher uncertainty for understanding the trace which belongs to a victim case. Two types of measurement are available: *average* and *worst-case*. When the *measure type* is average, the average of disclosure risks obtained for the different candidates of background knowledge is considered as the final risk value of the event log. When the *measure type* is worst-case, the worst value of disclosure risks obtained for the different candidates of background knowledge is considered as the final risk value of the event log, i.e., the largest value. The focus perspective of analysis is specified using the "Event Attributes" drop-down list, i.e., the input event log is simplified onto the event attributes specified by this parameter. This also specifies the concept of trace, i.e., a trace is considered as a sequence of events containing the event attributes specified by users in the "Event Attributes" drop-down list. The *time accuracy* could be *original*, *seconds*, *minutes*, *hours*, or *days*. This parameter is used only when timestamps are selected as one of the event attributes. The *life-cycle* also needs to be specified, or one can instead check the "All Life-Cycles" checkbox.

The *data utility* is calculated based on the *earth mover's distance* between two event logs. Namely, an original event log and a privacy-aware even log. The distance is obtained based on the distribution of trace variants. Figure 24 shows the *data utility analysis* module of PC4PM. Similar to risk analysis, a trace is considered as a sequence of events containing the event attributes specified by users in the "Event Attributes" drop-down list. Note that if the two event logs selected for calculating the data utility are from the same origin, i.e., they have the same universe of values for their attributes, “Same Origin” has to be checked (this is often the case).



Event Logs

No file chosen

running_example.xes
running_exampleWithCaseAttributes.xes
Sepsis Cases - Event Log.xes

Same Origin: ☑

All Life-Cycles: ☑

Event Attributes: concept:name

Time Accuracy: Original

Life-Cycle: Nothing selected

Select and set two event logs.

Check that when the event logs are from the same origin.

Considers all the life-cycles in the event logs.

The event attributes which are included in the sequence of events, i.e., trace. This specifies the perspective of analysis.

This is used only when the timestamp attribute is selected as an event attribute.

Select only specific life-cycles.

Data-Utility Analysis

Selected first log: running_exampleWithCaseAttributes.xes

Selected second log: running_example.xes

Calculcated Data-Utility value: 1

Figure 24: The data utility analysis module of the application.

The FCB-anonymity evaluates the risks of privatized event logs when they are published continuously. It evaluates the risk when the so-called correspondence knowledge [9] is available for an adversary. Currently, the module considers two event logs and reports the following information: (1) k-anonymity of the first and second event logs, (2) k-anonymity of the first event log after launching the F-attack, so-called F-anonymity, (3) k-anonymity of the second event log after launching the C-attack, so-called C-anonymity, and (4) k-anonymity of the second event log after launching the B-attack, so-called B-anonymity.

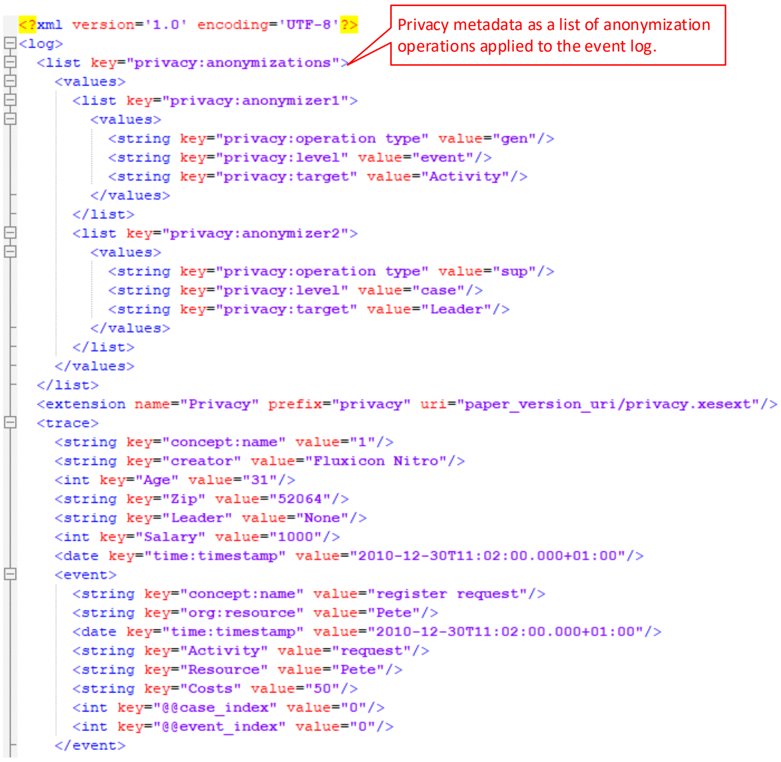
12 Privacy Metadata

The privacy metadata proposed in [6] are also supported by PC4PM such that all the privacy-aware (anonymized) event logs contain privacy metadata. Since almost all the privacy preservation techniques apply a sequence of the anonymization operations discussed in Section 9, the privacy metadata is formed based on these operations. The privacy metadata are added as an extension to the XES standard of event logs. The privacy metadata of an standard XES event log contains three main attributes: *operation type*, *level*, and *target*. These attributes indicate the type, level, and target of an applied anonymization operation, respectively. For instance, Figure 25 shows part of a privacy-aware event log where first a generalization technique has been applied at the level of events and the target was “Activity”, then a suppression technique has been applied at the level of cases and the target was *Leader*.

References

[1] van der Aalst, W.M.P.: Process Mining - Data Science in Action, Second Edition. Springer (2016). https://doi.org/10.1007/978-3-662-49851-4

[2] Fahrenkrog-Petersen, S.A., van der Aa, H., Weidlich, M.: PRIPEL: privacy-preserving event log publishing including contextual information. In: Fahland, D., Ghidini, C., Becker, J., Dumas, M. (eds.) Business Process Management - 18th International Conference, BPM 2020, Seville, Spain, September 13-18, 2020, Proceedings. Lecture Notes in Computer Science, vol.



Privacy metadata as a list of anonymization operations applied to the event log.

```xml
<?xml version='1.0' encoding='UTF-8'?>
<log>
  <list key="privacy:anonymizations">
    <values>
      <list key="privacy:anonymizer1">
        <values>
          <string key="privacy:operation type" value="gen"/>
          <string key="privacy:level" value="event"/>
          <string key="privacy:target" value="Activity"/>
        </values>
      </list>
      <list key="privacy:anonymizer2">
        <values>
          <string key="privacy:operation type" value="sup"/>
          <string key="privacy:level" value="case"/>
          <string key="privacy:target" value="Leader"/>
        </values>
      </list>
    </values>
  </list>
  <extension name="Privacy" prefix="privacy" uri="paper_version_uri/privacy.xesext"/>
  <trace>
    <string key="concept:name" value="1"/>
    <string key="creator" value="Fluxicon Nitro"/>
    <int key="Age" value="31"/>
    <string key="Zip" value="52064"/>
    <string key="Leader" value="None"/>
    <int key="Salary" value="1000"/>
    <date key="time:timestamp" value="2010-12-30T11:02:00.000+01:00"/>
    <event>
      <string key="concept:name" value="register request"/>
      <string key="org:resource" value="Pete"/>
      <date key="time:timestamp" value="2010-12-30T11:02:00.000+01:00"/>
      <string key="Activity" value="request"/>
      <string key="Resource" value="Pete"/>
      <string key="Costs" value="50"/>
      <int key="@@case_index" value="0"/>
      <int key="@@event_index" value="0"/>
    </event>
```

Figure 25: A sample of privacy metadata added to a standard XES event log.

12168, pp. 111–128. Springer (2020). https://doi.org/10.1007/978-3-030-58666-9_7, `https://doi.org/10.1007/978-3-030-58666-9_7`

[3] Ng, M., Li, M.J., Huang, J., He, Z.: On the impact of dissimilarity measure in k-modes clustering algorithm. IEEE transactions on pattern analysis and machine intelligence **29**, 503–7 (04 2007). https://doi.org/10.1109/TPAMI.2007.53

[4] Rafiei, M., van der Aalst, W.M.P.: Mining roles from event logs while preserving privacy. In: Francescomarino, C.D., Dijkman, R.M., Zdun, U. (eds.) Business Process Management Workshops - BPM 2019 International Workshops, Vienna, Austria, September 1-6, 2019, Revised Selected Papers. Lecture Notes in Business Information Processing, vol. 362, pp. 676–689. Springer (2019). https://doi.org/10.1007/978-3-030-37453-2_54

[5] Rafiei, M., van der Aalst, W.M.P.: Practical aspect of privacy-preserving data publishing in process mining. In: Proceedings of the Best Dissertation Award, Doctoral Consortium, and Demonstration & Resources Track at BPM 2020 co-located with the 18th International Conference on Business Process Management (BPM 2020). CEUR-WS.org (2020)

[6] Rafiei, M., van der Aalst, W.M.P.: Privacy-preserving data publishing in process mining. In: Fahland, D., Ghidini, C., Becker, J., Dumas, M. (eds.) Business Process Management Forum - BPM Forum 2020, Seville, Spain, September 13-18, 2020, Proceedings. Lecture Notes in Business Information Processing, vol. 392, pp. 122–138. Springer (2020). https://doi.org/10.1007/978-3-030-58638-6_8

[7] Rafiei, M., van der Aalst, W.M.P.: Towards quantifying privacy in process mining. In: International Conference on Process Mining - ICPM 2020 International Workshops, Padua, Italy, October 4-9, 2020 (2020)

[8] Rafiei, M., van der Aalst, W.M.P.: Group-based privacy preservation techniques for process mining. CoRR **abs/2105.11983** (2021), `https://arxiv.org/abs/2105.11983`

[9] Rafiei, M., van der Aalst, W.M.P.: Privacy-preserving continuous event data publishing. CoRR **abs/2105.11991** (2021), `https://arxiv.org/abs/2105.11991`

[10] Rafiei, M., Wagner, M., van der Aalst, W.M.P.: TLKC-privacy model for process mining. In: Dalpiaz, F., Zdravkovic, J., Loucopoulos, P. (eds.) Research Challenges in Information Science - 14th International Conference, RCIS 2020, Limassol, Cyprus, September 23-25, 2020, Proceedings. Lecture Notes in Business Information Processing, vol. 385, pp. 398–416. Springer (2020). https://doi.org/10.1007/978-3-030-50316-1_24

[11] Rafiei, M., von Waldthausen, L., van der Aalst, W.M.P.: Ensuring confidentiality in process mining. In: Ceravolo, P., López, M.T.G., van Keulen, M. (eds.) Proceedings of the 8th International Symposium on Data-driven Process Discovery and Analysis (SIMPDA 2018), Seville, Spain, December 13-14, 2018. CEUR Workshop Proceedings, vol. 2270, pp. 3–17. CEUR-WS.org (2018)

[12] Rafiei, M., von Waldthausen, L., van der Aalst, W.M.P.: Supporting confidentiality in process mining using abstraction and encryption. In: Ceravolo, P., van Keulen, M., López, M.T.G. (eds.) Data-Driven Process Discovery and Analysis - 8th IFIP WG 2.6 International Symposium, SIMPDA 2018, Seville, Spain, December 13-14, 2018, and 9th International Symposium, SIMPDA 2019, Bled, Slovenia, September 8, 2019, Revised Selected Papers. Lecture Notes in Business Information Processing, vol. 379, pp. 101–123. Springer (2019). https://doi.org/10.1007/978-3-030-46633-6_6