2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW) | 978-1-6654-4488-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/EDOCW52865.2021.00053

Sec4ML: An approach to support Cybersecurity Data Publishing for Machine Learning tasks

Madalena Lopes e Silva
Dept. de Sistemas e Computação
Instituto Militar de Engenharia (IME)
Rio de Janeiro, Brazil
https://orcid.org/0000-0001-7024-667X

Kelli de Faria Cordeiro
Centro de Análises de Sistemas Navais
Marinha do Brasil (MB)
Rio de Janeiro, Brazil
https://orcid.org/0000-0001-5161-8810

Maria Claudia Cavalcanti
Dept. de Sistemas e Computação
Instituto Militar de Engenharia (IME)
Rio de Janeiro, Brazil
https://orcid.org/0000-0003-4965-9941

***Abstract*—Despite the exponential growth of the World Wide Web since its creation, there are still few available datasets of cybersecurity incidents to be reused due to several issues, such as privacy-preserving concerns and data publication format standardization. As a result, the domain incidents analysis are precarious impacting on the Intrusion Detection Systems (IDS) development. The LOD (Linked Open Data) practices, which allows the sharing of data on the Web as a large and interconnected data graph, together with the FAIR (Findable, Accessible, Interoperable, and Reusable) principles, which guides the publication of data for reuse, can support the sharing of cybersecurity incidents datasets. Furthermore, anonymization techniques can be used to handle privacy concerns. Moreover, Machine Learning (ML) techniques can be used to improve IDS effectiveness. This article proposes the Sec4ML approach which supports the preparation of cybersecurity incident datasets for ML techniques using LOD practices and following FAIR principles, involving, among others, anonymization and preprocessing subprocesses, which are illustrated using data extracted from the UNSW-NB15 dataset.**

***Index Terms*—cybersecurity, anonymization, linked data, machine learning, artificial intelligence, FAIR principles**

I. Introduction

In the last years, we can notice that there has been a considerable increase in communications technologies usage, either of new ways of business through applications or in new platforms, as the Internet of Things (IoT), mobile apps, Industry 4.0, and social networks. Consequently, there has been an increase in data generated by these technologies. This new scenario brings new threats and vulnerabilities, such as data leakage or invasion of privacy.

Despite the growth of data generation and usage on the Web, there are still few available datasets of cybersecurity incidents to be reused. The main concerns about sharing cybersecurity data are privacy-preserving challenges and data publication format standardization. Thus, the lack of these data impacts negatively on the development and tuning of Intrusion Detection Systems (IDS) and Intrusion Prevention Systems (IPS).

The LOD (Linked Open Data[1]) cloud emerged to promote and accelerate data sharing on the web, as a large and interconnect data graph. The use of standard semantic resources and formats has facilitated the interlinking between data resources, providing a better way for publishing and reusing the data for research.

Lined up with the LOD, Wilkinson et al. proposed [6] the FAIR principles, which are detailed as a set of premises to improve the publishing process of research data and metadata. LOD and FAIR principles together can help to fill the lack of cybersecurity incidents data by providing an efficient way to face the privacy-preserving challenge and the standardization issues.

Alongside, Machine Learning (ML) techniques and models are evolving to facilitate and accelerate the discovery of knowledge, domain data analyses and highlight new data of inferences, specifically in the cybersecurity domain [24]. ML can be used to increase the IDS and IPS effectiveness. However, how to feed such ML algorithms with quality and up to date data?

To deal with this issue, this article proposes the Sec4ML approach, which supports the preparation of reusable cybersecurity incident datasets for ML algorithms following FAIR principles and involving, among others, anonymization and preprocessing of data. This approach addresses the limited availability of cybersecurity incident datasets, providing an environment to facilitate and motivate the creation of these datasets for publication, and consequently, increasing the number of datasets available for research.

The article is organized as follows. In Section II, some basic concepts were presented to facilitate the understanding of this work. In Section III, related works are analyzed from the point of view of the necessary characteristics to solve the identified problem. In Section IV, an overview of the approach is presented, as well as its architecture and process, detailing the Anonymization and Preprocessing subprocesses. Besides,

[1] https://lod-cloud.net/

978-1-6654-4488-0/21/$31.00 ©2021 IEEE
DOI 10.1109/EDOCW52865.2021.00053

it is illustrated using data extracted from the UNSW-NB15 dataset [36].

II. Background

To better understand the proposed approach, some important concepts are presented in this Section. First, it is introduced one of the most important initiatives in data publishing, the *Linked Open Data (LOD)*. Another relevant initiative proposes the *FAIR principles* which state some premises to promote the publishing, sharing, and reuse of research data and metadata. To achieve such premises in a research project, it is necessary to keep records about the whole publication workflow. These records are called *provenance metadata*, which is another important concept described in this section. A fourth concept to be define is related to the publication of research metadata in compliance with preserving the privacy (*privacy-preserving*) of individuals, resources, or organizations. Finally, some other concepts are briefly presented, related to *Machine Learning* (ML) and its relation with the development and tuning of cybersecurity incidents tools.

The LOD cloud initiative emerged due to the need for publishing structured data on the Web. The idea was to promote and evolve the Semantic Web, which adds a semantic layer that enables the understanding of structured or textual content by humans and machines. To participate in the LOD cloud, it is necessary to represent structured data using the Resource Description Framework (RDF[2]), which is a standard model for interconnecting web resources defined by the World Wide Consortium (W3C[3]). It consists of a set of datasets, which are expressed in RDF, where each data item is represented as a triple structure (subject, predicate, and object).

Even though with LOD facilities, other demands and difficulties are faced by researchers and developers when publishing data. Examples of such problems are the difficulty to find related datasets to connect to, and poorly interconnected datasets, i.e., links with poor semantics interconnecting datasets. The semantic enrichment, also called instance match, is a desirable characteristic in all LOD publishing projects that, when reached, makes possible to maximize the connectivity among datasets in LOD [9] [8]. This characteristic can be reached when similar data instances are linked by specific properties like owl:sameAs or, when similar classes and properties are linked by owl:equivalentClass and owl:equivalentProperty, respectively [7]. The FAIR principles (Findable, Accessible, Reusable and Interoperable) [6], describe a minimum set of data management requirements to deal with the reusability problem. These requirements are desirable characteristics that make the dataset more accessible, visible, and interoperable. They were proposed in a way that is independent of technology, that is, they can be applied regardless of the technology used. To reach FAIRness means to provide a FAIR compliant environment, in which it is possible to design processes that attend all FAIR principles (listed in Table I).

TABLE I: The FAIR Guiding Principles [6]

	Principles
F	F1. (meta)data are assigned a globally unique and persistent identifier
	F2. Data are described with rich metadata (defined by R1 below)
	F3. Metadata clearly and explicitly include the identifier of the data it describes
	F4. (meta)data are registered or indexed in a searchable resource
A	A1. (meta)data are retrievable by their identifier using a standardized communications protocol
	A1.1 The protocol is open, free, and universally implementable
	A1.2 The protocol allows for an authentication and authorization procedure, where necessary
	A2. Metadata are accessible, even when the data are no longer available
I	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
	I2. (meta)data use vocabularies that follow FAIR principles
	I3. (meta)data include qualified references to other (meta)data
R	R1. Meta(data) are richly described with a plurality of accurate and relevant attributes
	R1.1. (meta)data are released with a clear and accessible data usage license
	R1.2. (meta)data are associated with detailed provenance
	R1.3. (meta)data meet domain-relevant community standards

Any publishing process, to comply with the FAIR principles, should capture and store provenance data. The Oxford English Dictionary [39] defines provenance as *the source or origin of an object*. In terms of data that result from computational processes or workflows, provenance include all kinds of data and events related to any procedure that generated this data.

This issue showed more importance in the last years because of its relevance within the research society. Herschel et al. [29] classify provenance capture in four ways: (a) provenance metadata, that is, the most general type of provenance, independent of models or access method; (b) information systems provenance, where information systems metadata are captured; (c) workflow provenance, where workflow metadata are captured and (d) data provenance, metadata process by systems data. These authors also pointed different forms of workflow provenance: (I) prospective, that handles the structure and static context of a given workflow; (II) retrospective, that, on the other hand, handles information about a given workflow execution, i.e., information available when running the workflow; and (III) evolution provenance, that is concerned with changes made between two versions of workflow. This paper will be concerned only with the retrospective provenance.

Even though these systems that capture provenance follow some of those above-mentioned FAIR premises, it is fundamental to preserve the privacy of stakeholders. The concept of preserving privacy consists of approaches and techniques designed to hide the identity and/or sensitive data of the data subject. There are several different techniques and approaches

[2] https://www.w3.org/RDF/

[3] https: //www.w3.org/tr/rdf-sparql-query/

that aim at minimizing the likelihood of reidentifying individuals and organizations, such as shows Fung [26].

Among the possible techniques to be applied to datasets to preserve the privacy of individuals or organizations involved, are the anonymization techniques. There are simple strategies such as data suppression, generalization, disturbance, and the generation of synthetic data.

Although research data may be published in the LOD cloud, privacy-preserved and associated to its provenance data, as the volume of data generated grows rapidly, it becomes impossible for humans to analyze such data. Machine Learning (ML) and Deep learning (DL) are some of the AI techniques that are largely used to support analysis of such large data. Specially for the cybersecurity domain, they are frequently and successfully used for tuning IDS prediction [21] [22]. ML techniques are more efficient than the conventional signature-based strategies because a small variation in attack pattern can bypass a signature-based IDS. Thus, as ML strategies can learn from traffic behavior, they are able to detect attack variants [23]. Public labeled datasets for ML techniques are usually prepared as follows: they are divided into two parts, training and test datasets. The first part is used for the learning phase, and the second part to evaluate the efficiency of an ML algorithm. Therefore, the more datasets are made available for training ML algorithms, the greater the chances of building better and more efficient systems for cybersecurity incident detection. Finally, it is important to highlight the difference between FAIRness related to FAIR principles and the fairness in ML algorithms. While the aforementioned concept is related to the collective knowledge about these principles, the second concept is related to providing a fair process, which means processing data without any kind of discrimination, such as gender or race issues. This work focuses on the first one.

III. Related Works

In the literature, we can observe some solutions for the data publishing process in LOD. The following works were found by systematic searches in the literature and by applying the snow bowling technique. In the work of Jacobsen et al. [11] the authors present a generic workflow for publishing data compliance with FAIR principles. The workflow is divided into three phases: (i) Pre-FAIRification, (ii) FAIRification, and (iii) Post-FAIRfication. Another relevant point highlighted is the differentiation of ”A” from FAIR, which does not mean ”open” but accessible in the way it should be, especially due to legal restrictions, preservation of privacy, national security, and protection of competitiveness in private organizations. The extent to which data will be opened for access, and not just made available for discovery, is defined by each publishing organization, under the aforementioned restrictions. This work, however, does not detail how each phase and step should be deployed, which makes it hard to implement to each domain.

In the paper of de Mendonça et al. [37], it was proposed an approach to collect, link, and publish provenance data of ETL steps implemented as workflows and, jointly, an ontology that foundation tasks data capture of given workflow execution. The privacy-preserving of both data and metadata, however, not was approached in this paper. In the same way, Rautenberg et al. [38] presented LODFLOW, another approach for linked data publishing, capturing metadata from tasks of a given workflow execution. This approach allows one to follow all steps in this process and capture the provenance metadata step by step. However, this work did not present an approach to privacy-preserving process and publishing linked open data.

A specific way of to handler cybersecurity data was exposed by Fahad et al. [10]. It is propose a solution to generate, integrate, and share security incidents over Supervisory Control and Data Acquisition (SCADA), system control and monitor industrial infrastructure functions. Adopted for the application of ML techniques, this framework uses the concept of clustering to apply noise addition algorithms according to the attribute type, grouping numerical attributes such as IDs, categorical or qualitative, and hierarchical IP addresses. Despite being the most complete work found as far as it was possible to search, in the sense of publishing anonymized data, issues of the capture of provenance and publication of data as linked open data were not addressed. Another way to deal with this issue is presented in the work of Salvadori et al. [27], Data Linking as a Service is an infrastructure for generating and publishing Linked Data on the Web. This infrastructure aims at facilitating the execution of necessary processes to properly publish high-quality linked data from distributed and heterogeneous datasets. It is included semantic enrichment, such as data linking, structure optimization, and the publication and applies Mining Association Rules techniques to identify patterns. Indeed, an existing Web resource is associated with other ones, according to a given association strategy.

Concerning data preparation for ML tasks, Moura et al. [34] propose an approach for the preparation of training and test datasets intended for ML classification algorithms. Based on PreProcessing Operators Ontology (PPO-O), the work introduces a preprocessing tool that supports scientists and Information Technology (IT) professionals in the choice of the used operators.

Table II presents a comparison among the selected works, pointing to some necessary characteristics to solve the evidenced problem. It is worth to notice that characteristics such as Anonymization, FAIRness and ML Preprocessing are less attended by the selected related works. Therefore, in the present work we present the Sec4ML approach that intends to cover all the required characteristics.

IV. Sec4ML Approach

The Sec4ML was conceived to support scenarios where researchers need to prepare cybersecurity data and metadata for reuse, following the FAIR principles. With such support, it will increase the number of cybersecurity incidents' datasets for research, and consequently, contribute to the tuning of ML algorithms. When adopting the Sec4ML approach, privacy-preserving issues are addressed along with the preparation of the dataset for ML tasks.

TABLE II: Related Works Comparison

Related Works	Characteristics						
	1	2	3	4	5	6	7
Salvadori et al.	✓	✓	X	✓	X	X	X
Fahad et al.	X	X	X	X	✓	X	X
de Mendonça et al.	✓	✓	✓	✓	X	X	X
Rautenberg et al.	✓	✓	✓	✓	X	X	X
Jacobsen et al.	X	X	X	X	X	✓	X
Moura et al.	X	✓	✓	X	X	X	✓
Sec4ML (this paper)	✓	✓	✓	✓	✓	✓	✓

Caption: 1- Linked Data 2- Ontology 3- Provenance
4 - Semantic Enrichment 5 - Anonymization
6 - FAIR 7 - Preprocessing ML

The Sec4ML architecture is presented in the Section IV-A. Section IV-B shows two macro processes of the approach. Section IV-B1 details the anonymization subprocess. In the same way, Section IV-B2 explains the use of preprocessing operators. Finally, Section IV-C highlights how anonymization and preprocessing operators work, using an example.

A. Architecture

In the mentioned scenario, Sec4ML distinguishes four main roles: (i) the controller, who is responsible for the source data, and who will be able to access the data, for example, through cybersecurity log tools; (ii) the operator, who is responsible for performing the data transformation tasks; (iii) the provider, who is responsible for publishing and/or updating published data; and (iv) the consumer, who may be any researcher or developer who is interested in reusing the available datasets.

Figure 1 shows the Sec4ML architecture, which is organized in two main parts: back-end and front-end. All those roles interact with the components of the front-end, while all data resources are at the back-end. The *Controller* has access to the source data (*Tabular Datasets*) through the *Processor* module. The *Operator* will transform data also through this module. On the other hand, the *Provider* manages the publication of the transformed data (*Graph databases*) through the *Publisher* module. Both modules, *Processor* and *Publisher* interact with the *Provenance Manager*, providing provenance metadata, which feeds the *Provenance data* repository. Finally, the *Consumer* is able to access and reuse not only the transformed data, but also the provenance data, through a variety of interfaces, such as FDP endpoints, Dashboards and SPARQL endpoints.

The *Cybersecurity Dataset Processor* works as a workflow of tasks which will be detailed in the next subsection. This module guides the *Operator* on transforming the raw data into anonymized and preprocessed data for ML tasks. While these transformations are underway the *Provenance Manager* is able to register them. This provides retrospective provenance data, which enables the reproducibility of those transformations, guaranteeing the detailed explanation about the transformed data, as well as the individual re-identification, in case of anonymized data, for investigation purposes. The *Provenance Manager* component is responsible for capturing workflow metadata at each step, along the whole execution of the workflow. These captured data makes it possible to reproduce a part of or all workflow execution. Finally, all the provenance and transformed data are available in the RDF format.

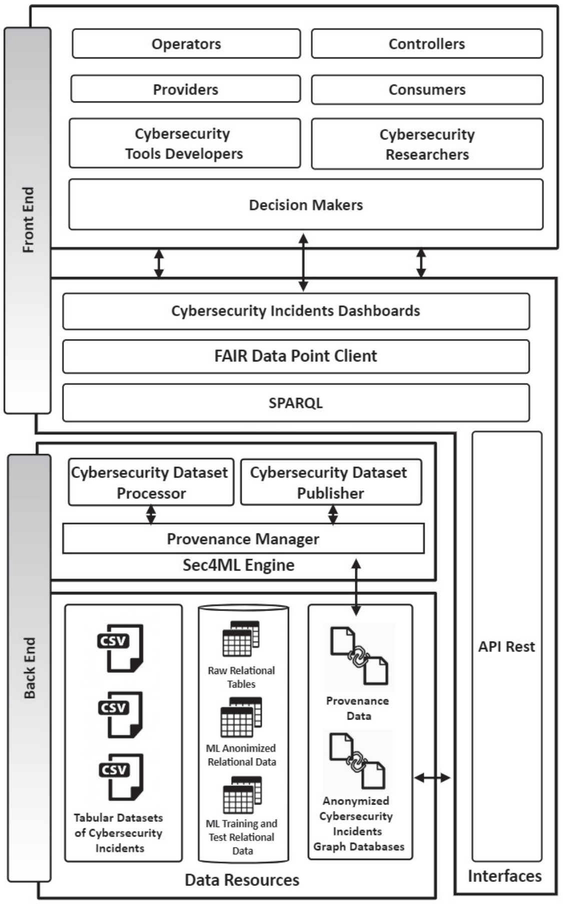
Front End
Operators
Controllers
Providers
Consumers
Cybersecurity Tools Developers
Cybersecurity Researchers
Decision Makers
Cybersecurity Incidents Dashboards
FAIR Data Point Client
SPARQL
Back End
Cybersecurity Dataset Processor
Cybersecurity Dataset Publisher
Provenance Manager
Sec4ML Engine
CSV
CSV
CSV
Tabular Datasets of Cybersecurity Incidents
Raw Relational Tables
ML Anonimized Relational Data
ML Training and Test Relational Data
Provenance Data
Anonymized Cybersecurity Incidents Graph Databases
Data Resources
API Rest
Interfaces

Fig. 1: Sec4ML Architecture

B. Processes

As already mentioned, the Sec4ML architecture back-end is formed of three main components that work together as two macroprocesses, shown in Figure 2. The *Processing and Publication* process is responsible for the collection, transformation, and publication of cybersecurity data; and the *Provenance Data Catching* process, which is responsible for capturing the provenance data. This process can occur at any time, always when anyone will need to look up data.

Those subprocesses of the three macroprocesses aforementioned will see more detail in the following Subsections.

Dataset Transformation and Generation process that includes three subprocesses: *Data Collecting*, perform the data collection either from IDS logs or other tools that generate cybersecurity events records; *Data Treatment* performs data normalization or cleaning by the researchers and *Dataset Generation* that concatenate and group records to create a dataset. In the third subprocess, the data is described with metadata, creating some minimal structure for describing the data, such as a data catalog or data dictionary. It allows that the process meets the FAIR principles **F1 - (meta)data are assigned a globally unique and persistent identifier, F2 - Data are described with rich metadata**, and **F3 - Metadata clearly and explicitly include the identifier of the data**. The principle F1 is reached through of Digital Object Identifier (DOI) assignment or another persistent identifier. On the other hand, the F2 principle is also addressed by the addition of rich metadata to data itself, providing rich search results. Finally, the F3 principle is addressed by the addition of data identifiers to metadata. At the end of this process, we have the dataset itself, in tabular format and a data description structure. In the case of it choosing a public dataset, only the first task is performed.

The *Data Anonymization* process is performed by a set of activities to apply anonymization treatment on dataset attributes, keeping them useful for research. Its specific tasks are detailed in Section IV-B1.

The macroprocess *Processing and Publication* is constituted by the following processes:

The *Data Preprocessing* process is composed by two subprocesses: *Data Extraction*, responsible for extracting data from the tabular dataset and inserting it in a relational database, used as staging database, and *Data Preprocessing* submits the dataset to preprocessing operators, applying adaptations and modifications over the data aiming to make it ready to apply ML techniques. More details of preprocessing operators will see in Section IV-B2.

The *Data Triplification* process is composed by the following two subprocesses. *Data Triplification*, that transforms relational data, within the staging database, in a structure of triples, making it possible to publish it as an RDF resource. The published triplified data, visualized as a graph database, compose the LOD. *Data Triplification* addresses the FAIR principle **I1 - (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation**. *Semantic Alignment* may create references from triplified data for other resources in LOD if it exists. *Semantic Alignment* subprocess also address a FAIR principle **I3 - (meta)data include qualified references to other (meta)data**.

Throughout the *Metadata and Data Publication* at the end of all previous subprocesses, the triplified data are published in a triplestore software, like GraphDB, Amazon Neptune and Virtuoso. The captured provenance data are published in a Fair Data Point (FDP[4]). This publication allows that the process meets the FAIR principle **F4 – Meta(data) are registered or indexed in a searchable resource**.

This subprocess follows those recommendations recently published recommendations by Research Data Alliance (RDA) from Research Metadata Schemas Workgroup [5] to perform the *Metadata and Data Publication* task where are suggested actions, best practices, and tools that should be followed for metadata capture, store, and publishing.

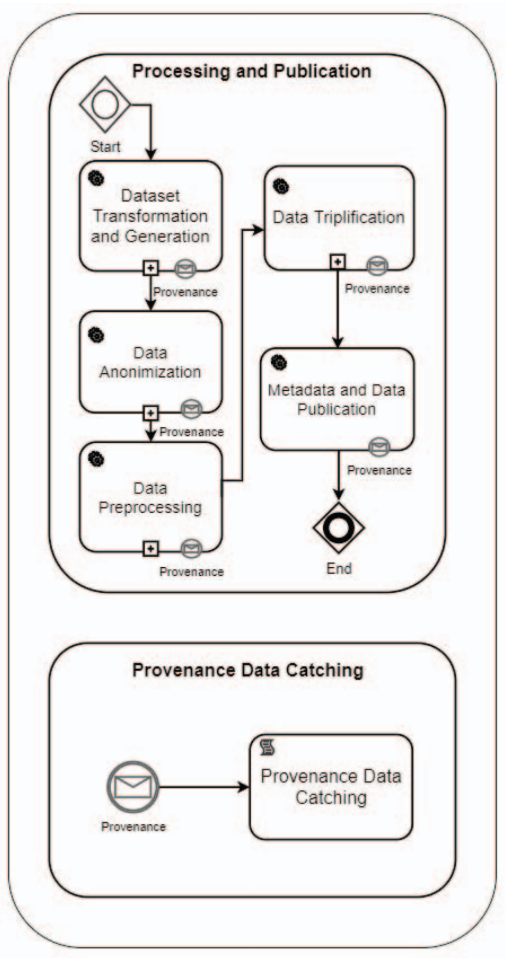

Fig. 2: Sec4ML Process. Note that the Provenance message is captured along the whole workflow and delivered to the Provenance Data Catching task.

Once the data is published in one data repository mentioned, it is possible, through some available interfaces on the WEB, such as SPARQL queries, API Rest, or other tool, query the data, addressing the FAIR principles **A1 - (meta)data are retrievable by their identifier using a standardized communications protocol, A1.1 - The protocol is open, free, and universally implementable**, and **A2 - Metadata are accessible, even when the data are no longer available.** However, only with queries from FDP, it can addresses the principle **A1.2 - The protocol allows for an authentication and authorization procedure, where necessary**, by means of user authentication.

Throughout the whole process, a provenance data capture is performed to make data tracking, reuse, and response to legal issues. This provenance data includes workflow and retrospective provenance, both data necessary to, in the future, provide good results for scientists or researchers in the search

[4] https://www.fairdatapoint.org/

[5] https://www.rd-alliance.org/groups/research-metadata-schemas-wg

for reuse-ready data. The captured metadata will be modelled by reusing semantic resources and vocabularies aimed at provenance data. Thus, this subprocess, for example, captures provenance data from datasets such as the source of the dataset, date of its generation, version, if is labeled or not. It also captures provenance data from the workflow such as date/hour of execution, name of the operator, executed subprocess identification, among others.

1) Anonymization: One of the most critical processes of the Sec4ML approach is the *Anonymization* subprocess, depicted in Figure 3. It was designed to be as flexible as possible, allowing parameter choices that adjust the execution flow. Firstly, *Data Extration* task retrieves data from the staging database. In sequence, it is verified if the dataset contains identifier attributes, which is usual. In this case, it is applied symmetric cryptography routines with suitable algorithms according to the nature of the attributes. For instance, for an attribute that contains IP addresses, AES prefix-preserving algorithms and truncation algorithms could be applied, while for MAC addresses, hash-based algorithms could be used [35].

Attributes quasi-identifiers must also be anonymized (*Quasi-identifiers Data Treatment* task). There is a verification if quasi-identifier attributes are present. If yes, some privacy-preserving techniques must be applied. Finally, it is checked if there are sensitive attributes, which can receive an anonymization treatment.

2) Preprocessing: Even though datasets were prepared with privacy-preserving guarantees of the involved entities, indeed, they are not ready yet to be submitted to ML algorithms. Before publishing them, it is necessary to submit the dataset to preprocessing tasks. The *Preprocessing* subprocess unfolds in three tasks: *Data Extration*, *Data Mapping* and *Data Preprocessing*. *Data Extration* retrieves the records from the staging database. *Data Mapping* task performs a semantic correlation between attributes' names and ontologies that can be reused, following the dataset structure semantics and using the most appropriate terms. This task addresses the FAIR principle **I2. (meta)data use vocabularies that follow FAIR principles** searching and creating semantic correlations with data in LOD. It also addresses the principle **R1.3. (meta)data meet domain-relevant community standards** working with patterns and vocabularies more widespread within the scientific community. Although it is not detailed in the present work, this task may involve many sub-tasks and may use market available tools.

Concerning the preparation of data for ML techniques, the *Data Preprocessing* subprocess applies KDD operators defined in PPO-O [34], such as data selection, cleaning, outliers removing, dimensionality reduction, normalization, among others. In that work, Moura et al. define and implement eight activities that correspond to KDD operators.

Typically, preprocessing data for ML algorithms can to need more than one round to reach the desired result. Besides, each operator must be applied according to a particular sequence. In [34], the authors propose an assistant to guide the user on the preprocessing operator execution. Based on the guidelines of such assistant, and using the same example data shown in Table IV, the application of preprocessing operators was performed in two subsequent rounds, and their results are shown in Tables V and VI, respectively.

In the first round the *Data Coding* operator was applied (*OrdinalEncoder()* function) to codify qualitative attributes such as *proto* and *state*. The values of *service* attribute were already removed in the *Anonimyzation* task, but the attribute was still present in the dataset schema. In the *Preprocessing* task the *service* attribute was removed by the *Column Selection* operator (*DropQualitativeColumn()* function). Attributes such as *sport*, *dsport*, *dur* and *dbytes*, received imputed data to replace zero values with average by applying the *Data Missing Imputation* operator (*ImputationAverage()* function). After this, all attribute values of the *sport* and *dsport* attributes were replaced by normalized values applying standardization, using *Data Normalization* operator (*StandardScaler()* function). Finally, the attribute class *attack_cat* has its NULL values replaced by ”Unknown” by the application of the *Data Missing Imputation* operator (*ImputationUnknown()* function).

In the second round, the attribute *attack_cat* was modified by the *Data Coding* operator (*OrdinalEncoder()* function) to codify its values. At the end, the attribute *label* was removed from the dataset by the application of the *Column Selection* operator (*DropQualitativeColumn()* function), because it was only a flag for the classifier attribute *attack_cat*.

Other operators can be applied as required, such as *Data Outlier Treatment* to remove outliers, *Oversampling* to correct majority class, *Undersampling* to correct minority class, and *Holdout* to generate training and test datasets. However, these operators were not used in this article due to the reduced size of the example data.

C. Application Example

To show how the process performs the anonymization transformations and the application of the KDD preoperators, it was chosen a public labeled network dataset. Among the public cybersecurity analyzed datasets according to the proposed eleven criteria [40], was choose the UNSW-NB15 Network Dataset (raw data) [36], a cybersecurity incidents dataset generated synthetically by Australian Centre for Cyber Security (ACCS)[6]. It was created with a combination of real and synthetical attack activities. Thus, Table III shows a few tuples and attributes of this dataset. Among its 49 attributes, we used just a few in the application example:

- *srcip*: source IP address;
- *sport*: source port that originated the event registered;
- *dstip*: destination IP address;
- *dsport*: destination port related to the registered event;
- *proto*: protocol used in the registered event;
- *state*: state related to the registered event;
- *dur*: total duration of the event;
- *sbytes*: number of bytes involved in the event from the source to destination;

[6] http://www.accs.unsw.adfa.edu.au/

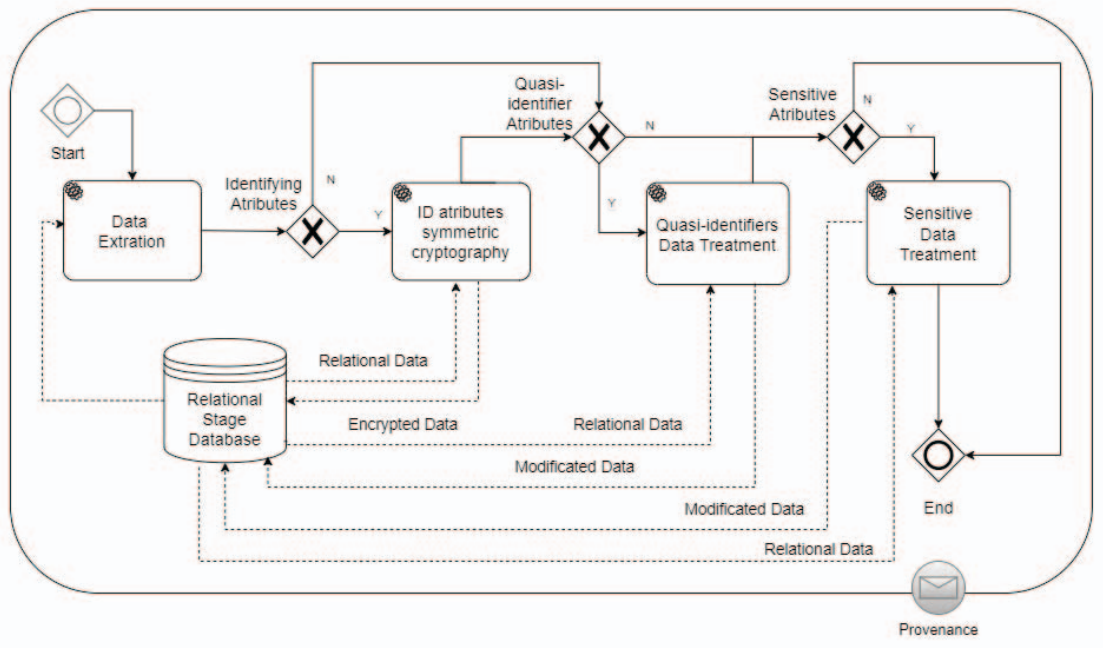

Fig. 3: Anonymization Task

- *dbytes*: number of bytes involved in event from destination to source;
- *attack_cat*: category of event, either if it is a real activity or a synthetic attack; and
- *label*: binary flag, turned on (1) when the registered event is attack records or turned off (0) when the event is a normal activity.

Table IV shows the same tuples, although some attributes were anonymized. The *srcip* and *dstip* attributes were submitted *ID attributes symmetric cryptography* task, where it was used the AES symmetric algorithm with a 128 length key, generating cipher values in hexadecimal. *dsport* attribute was treated by the execution of the *Quasi-identifiers data treatment* task, which applied the noise addition technique, except for zero values that were handled in the following *Preprocessing* subprocess.

The *service* attribute was treated by the execution of the *Sensitive Data Treatment*. One of the techniques of sensitive attributes handle is suppression. The column suppression operation was executed on this attribute, removing all values from the above-mentioned dataset. Thus, it is possible to generate a new version of the dataset without loss of utility and statistical features, which is required for data that will be reused for research and cybersecurity tool development.

V. Conclusion

In recent years, it was possible to observe a huge increase in data generation by new IT technologies. Together with this wave arises also a new reality, with different threats and vulnerabilities that can also to originate a data leak or invasion of privacy. However, resources and tools designed to combat threats do not has become better at the same speed of challenges within the cybersecurity domain.

At the same time, ML models and concepts emerged facilitating and accelerating the acquisition and learning of knowledge. Some labeled or ready-to-apply cybersecurity datasets have emerged for ML tasks. In this context, there are still many challenges to face, such as cultural issues and ready-made tools development to generate anonymous cybersecurity datasets. Due to these issues, there is a limited availability of these datasets, and consequently, the need to increase the number of datasets available for research.

This article proposes the Sec4ML approach which aims to fill this gap, making it possible to prepare, anonymize, and publish data and metadata of cybersecurity incidents for KDD tasks and, at the same time, generate datasets reusable ready for consumers of the FAIRness Web of Data. It is intended to support users on generating and publishing anonymized cybersecurity data. The approach implementation is in progress. Finally, as future works, it can be considered the improvement of this approach, not only on detailing the specification, but also on covering more ML tasks, besides dataset classification.

References

[1] F. T. de Oliveira, M. C. Cavalcanti, e R. M. Salles, “Towards effective reproducible botnet detection methods through scientific workflow management systems”, p. 14.

[2] G. B. de Figueiredo, J. L. R. Moreira, K. de Faria Cordeiro, e M. L. M. Campos, “Aligning DMBOK and Open Government with the FAIR Data Principles”, in Advances in Conceptual Modeling, Cham, 2019, p. 13–22.

[3] S. E. Coull e E. E. Kenneally, “A Qualitative Risk Assessment Framework for Sharing Computer Network Data”, SSRN Journal, 2012, doi: 10.2139/ssrn.2032315.

TABLE III: Original Data (UNSW-NB15 Dataset tuples)

srcip	sport	dstip	dsport	proto	state	dur	sbytes	dbytes	service	attack_cat	Label
59.166.0.0	1390	149.171.126.6	53	udp	CON	0.001055	132	164	dns	NULL	0
59.166.0.3	49664	149.171.126.0	53	udp	CON	0.001169	146	178	dns	NULL	0
59.166.0.6	2142	149.171.126.4	53	udp	CON	0.001134	132	164	dns	NULL	0
10.40.182.3	0	10.40.182.3	0	arp	INT	0	46	0	NULL	NULL	0
59.166.0.5	40726	149.171.126.6	53	udp	CON	0.001126	146	178	dns	NULL	0
59.166.0.7	12660	149.171.126.4	53	udp	CON	0.001167	132	164	dns	NULL	0
175.45.176.2	23357	149.171.126.16	80	tcp	FIN	0.240139	918	25552	http	Exploits	1
175.45.176.0	13284	149.171.126.16	80	tcp	FIN	2.39039	1362	268	http	Reconnaissance	1
175.45.176.0	14324	149.171.126.13	502	tcp	FIN	0.214066	468	268	NULL	DoS	1
175.45.176.2	15985	149.171.126.17	80	tcp	FIN	0.534953	1710	115722	http	Generic	1

[4] Department of Homeland Security. ”The Information Marketplace for Policy and Analysis of Cyber-risk Trust (IMPACT)”. 2016. Retrieved June 4, 2021 from https://www.impactcybertrust.org/.

[5] , Berners-Lee, Tim and Hendler, James and Lassila, Ora, ”The Semantic Web: A New Form of Web Content That is Meaningful to Computers Will Unleash a Revolution of New Possibilities”, Scientific American Journal, 2001.

[6] M. D. Wilkinson et al., “The FAIR Guiding Principles for scientific data management and stewardship”, Sci Data, vol. 3, nº 1, p. 160018, dez. 2016, doi: 10.1038/sdata.2016.18.

[7] J. Schaible, T. Gottron, e A. Scherp, “TermPicker: Enabling the Reuse of Vocabulary Terms by Exploiting Data from the Linked Open Data Cloud - An Extended Technical Report”, arXiv:1512.05685 [cs], jan. 2016, Acessado: nov. 04, 2020. [Online]. Disponível em: http://arxiv.org/abs/1512.05685

[8] J. Schaible, T. Gottron, e A. Scherp, “Survey on Common Strategies of Vocabulary Reuse in Linked Open Data Modeling”, in The Semantic Web: Trends and Challenges, vol. 8465, V. Presutti, C. d’Amato, F. Gandon, M. d’Aquin, S. Staab, e A. Tordai, Orgs. Cham: Springer International Publishing, 2014, p. 457–472. doi: 10.1007/978-3-319-07443-6_31.

[9] A. Assi, H. Mcheick, e W. Dhifli, “Data linking over RDF knowledge graphs: A survey”, Concurrency Computat Pract Exper, vol. 32, nº 19, out. 2020, doi: 10.1002/cpe.5746.

[10] A. Fahad, Z. Tari, A. Almalawi, A. Goscinski, I. Khalil, e A. Mahmood, “PPFSCADA: Privacy preserving framework for SCADA data publishing”, Future Generation Computer Systems, vol. 37, p. 496–511, jul. 2014, doi: 10.1016/j.future.2014.03.002.

[11] A. Jacobsen et al., “A Generic Workflow for the Data FAIRification Process”, Data Intelligence, vol. 2, nº 1–2, p. 56–65, jan. 2020, doi: 10.1162/dint_a_00028.

[12] H. Inoue, T. Amagasa, e H. Kitagawa, “An ETL Framework for Online Analytical Processing of Linked Open Data”, in Web-Age Information Management, vol. 7923, J. Wang, H. Xiong, Y. Ishikawa, J. Xu, e J. Zhou, Orgs. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, p. 111–117. doi: 10.1007/978-3-642-38562-9_12.

[13] I. Salvadori, A. Huf, e F. Siqueira, “Data Linking as a Service: An Infrastructure for Generating and Publishing Linked Data on the Web”, in 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC), Madrid, Spain, jul. 2020, p. 262–271. doi: 10.1109/COMPSAC48688.2020.00042.

[14] S. Rautenberg, I. Ermilov, E. Marx, S. Auer, e A.-C. N. Ngomo, “LODFlow: a workflow management system for linked data processing”, in Proceedings of the 11th International Conference on Semantic Systems, Vienna Austria, set. 2015, p. 137–144. doi: 10.1145/2814864.2814882.

[15] R. R. de Mendonça, “ETL4LinkedProv: Managing Multigranular Linked Data Provenance”, vol. 7, nº 2. 2016, p. 16.

[16] Nuzzolese, Andrea Giovanni, et al. ”Semion: A Smart Triplification Tool.” EKAW (Posters and Demos). 2010.

[17] R. Pang, M. Allman, V. Paxson, e J. Lee, “The Devil and Packet Trace Anonymization”, ACM SIGCOMM Computer Communication Review, vol. 36, nº 1, p. 10, 2006.

[18] Y. J. Chew, S. Y. Ooi, K.-S. Wong, e Y. H. Pang, “Privacy Preserving of IP Address through Truncation Method in Network-based Intrusion Detection System”, in Proceedings of the 2019 8th International Conference on Software and Computer Applications, Penang Malaysia, fev. 2019, p. 569–573.

[19] J. Fan, J. Xu, M. H. Ammar, e S. B. Moon, “Prefix-preserving IP address anonymization: measurement-based security evaluation and a new cryptography-based scheme”, Computer Networks, vol. 46, nº 2, p. 253–272, out. 2004.

[20] P. Norvig, S. J. Russell, ”Inteligência artificial”, Elsevier, 2013.

[21] Y. Xin et al., “Machine Learning and Deep Learning Methods for Cybersecurity”, IEEE Access, vol. 6, p. 35365–35381, 2018.

[22] W. Seo e W. Pak, “Real-Time Network Intrusion Prevention System Based on Hybrid Machine Learning”, IEEE Access, vol. 9, p. 46386–46397, 2021, doi: 10.1109/ACCESS.2021.3066620.

[23] K. Shaukat, S. Luo, V. Varadharajan, I. A. Hameed, e M. Xu, “A Survey on Machine Learning Techniques for Cyber Security in the Last Decade”, IEEE Access, vol. 8, p. 222310–222354, 2020.

[24] D. Dasgupta, Z. Akhtar, e S. Sen, “Machine learning in cybersecurity: a comprehensive survey”, Journal of Defense Modeling Simulation, p. 154851292095127, set. 2020.

[25] F. Oliveira, M. Cavalcanti, and R. Salles. ”Towards effective reproducible botnet detection methods through scientific workflow management systems”, in Anais do XXXV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos, Belém, 2017.

[26] B. C. M. Fung, Org., Introduction to privacy-preserving data publishing: concepts and techniques. Boca Raton, Fla.: Chapman Hall/CRC, 2011.

[27] I. Salvadori, A. Huf, e F. Siqueira, “Data Linking as a Service: An Infrastructure for Generating and Publishing Linked Data on the Web”, in 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC), Madrid, Spain, jul. 2020, p. 262–271.

[28] S. Staab e R. Studer, Orgs., Handbook on Ontologies. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

[29] M. Herschel, R. Diestelkämper, e H. Ben Lahmar, “A survey on provenance: What for? What form? What from?”, The VLDB Journal, vol. 26, nº 6, p. 881–906, dez. 2017.

[30] G. Fisk, C. Ardi, N. Pickett, J. Heidemann, M. Fisk, e C. Papadopoulos, “Privacy Principles for Sharing Cyber Security Data”, in 2015 IEEE Security and Privacy Workshops, San Jose, CA, maio 2015, p. 193–197.

[31] R. Goldschmidt e E. Passos, Data mining: um guia Prático. 2005. Acessado: ago. 02, 2021. [Online]. Disponível em: http://www.sciencedirect.com/science/book/9788535218770

[32] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From Data Mining to Knowledge Discovery in Databases”, in AI Magazine, Volume 17, Number 3, 1996, p. 18.

[33] I. H. Witten, E.Frank, ”Data mining : practical machine learning tools and techniques”p. cm. – Elsevier, 2011.

[34] L. D. A. L. Moura, M. A. A. da Silva, K. de Faria Cordeiro, M. C. Cavalcanti, ”A Well-founded Ontology to Support the Preparation of Training and Test Datasets”. 2021.

[35] N. V. Dijkhuizen e J. V. D. Ham, “A Survey of Network Traffic Anonymisation Techniques and Implementations”, ACM Comput. Surv., vol. 51, nº 3, p. 1–27, jul. 2018.

[36] N. Moustafa e J. Slay, “UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)”, in 2015 Military Communications and Information Systems Conference (MilCIS), Canberra, Australia, nov. 2015, p. 1–6.

[37] R. R. de Mendonça, S. M. S. da Cruz, M. L. M. Campos, “ETL4LinkedProv: Managing Multigranular Linked Data Provenance”, in J. of Inf. and Data Manag., vol. 7, nº 2, p. 16. 2016.

[38] S. Rautenberg, I. Ermilov, E. Marx, S. Auer, e A.-C. N. Ngomo, “LODFlow: a workflow management system for linked data processing”,

in Proc. of the 11th Int. Conf. on Semantic Systems, Vienna Austria, set. 2015, p. 137–144.

[39] J. Freire, D. Koop, E. Santos, e C. Silva, “Provenance for Computational Tasks: A Survey”, Comput. Sci. Eng., vol. 10, nº 3, p. 11–21, maio 2008.

[40] D. Gumusbas, T. Yldrm, A. Genovese, e F. Scotti, “A Comprehensive Survey of Databases and Deep Learning Methods for Cybersecurity and Intrusion Detection Systems”, IEEE Systems Journal, p. 1–15, 2020, doi: 10.1109/JSYST.2020.2992966.

TABLE IV: Anonymized Data (UNSW-NB15 Dataset tuples)

srcip[a]	sport	dstip[a]	dsport[b]	proto	state	dur	sbytes	dbytes	service[c]	attack_cat	Label
f0ef4b92207ed6a19bc5bf737449090a	1390	d6a78de1053951029fe17d2fafbf63dc	64	udp	CON	0.001055	132	164	*REMOVED*	NULL	0
76a9eedcc8c7f93672d4867e8d0ce18f	49664	038cb8b62827d30a23360c12fdad801c	64	udp	CON	0.001169	146	178	*REMOVED*	NULL	0
eacfb9eb5bdeb42519949def45e2d46a	2142	1ce73b5285825252bdcd8908c7e0f650	64	udp	CON	0.001134	132	164	*REMOVED*	NULL	0
53ec5cb7fe15f1fc7fd79dd5f1c81ad8	0	53ec5cb7fe15f1fc7fd79dd5f1c81ad8	0	arp	INT	0	46	0	*REMOVED*	NULL	0
11c76906f69e573cff39c6e8d76dbded	40726	d6a78de1053951029fe17d2fafbf63dc	64	udp	CON	0.001126	146	178	*REMOVED*	NULL	0
fc892446e2871bb474fbe0611fd6aa9b	12660	1ce73b5285825252bdcd8908c7e0f650	64	udp	CON	0.001167	132	164	*REMOVED*	NULL	0
d6c03904efb0c29310ba8db0a6047271	23357	9b627a55e3a25f17e7f4fe7bd8ae5c14	91	tcp	FIN	0.240139	918	25552	*REMOVED*	Exploits	1
053ad4f007bc578dbcae52a628285755	13284	9b627a55e3a25f17e7f4fe7bd8ae5c14	91	tcp	FIN	2.39039	1362	268	*REMOVED*	Reconnaissance	1
053ad4f007bc578dbcae52a628285755	14324	17090328565ce63ee9d71823c1d0edf0	513	tcp	FIN	0.214066	468	268	*REMOVED*	DoS	1
d6c03904efb0c29310ba8db0a6047271	15985	04f226e4e07b8c9468f3fbb7b44d4a04	91	tcp	FIN	0.534953	1710	115722	*REMOVED*	Generic	1

TABLE V: Preprocessed Data for ML - 1stRound (UNSW-NB15 Dataset tuples)

srcip	sport	dstip	dsport	proto	state	dur	sbytes	dbytes	attack_cat	Label
f0ef4b92207ed6a19bc5bf737449090a	-1.1499	d6a78de1053951029fe17d2fafbf63dc	-0.4158	1	1	0.001055	132	164	Unknown	0
76a9eedcc8c7f93672d4867e8d0ce18f	1.9865	038cb8b62827d30a23360c12fdad801c	-0.4158	1	1	0.001169	146	178	Unknown	0
eacfb9eb5bdeb42519949def45e2d46a	-1.101	1ce73b5285825252bdcd8908c7e0f650	-0.4158	1	1	0.001134	132	164	Unknown	0
53ec5cb7fe15f1fc7fd79dd5f1c81ad8	-0.1128	53ec5cb7fe15f1fc7fd79dd5f1c81ad8	-0.0837	3	2	0,338406	46	14265	Unknown	0
11c76906f69e573cff39c6e8d76dbded	1.4058	d6a78de1053951029fe17d2fafbf63dc	-0.4158	1	1	0.001126	146	178	Unknown	0
fc892446e2871bb474fbe0611fd6aa9b	-0.4177	1ce73b5285825252bdcd8908c7e0f650	-0.4158	1	1	0.001167	132	164	Unknown	0
d6c03904efb0c29310ba8db0a6047271	0.2773	9b627a55e3a25f17e7f4fe7bd8ae5c14	0.2209	2	3	0.240139	918	25552	Exploits	1
053ad4f007bc578dbcae52a628285755	-0.3771	9b627a55e3a25f17e7f4fe7bd8ae5c14	0.2209	2	3	2.39039	1362	268	Reconnaissance	1
053ad4f007bc578dbcae52a628285755	-0.3096	17090328565ce63ee9d71823c1d0edf0	2.8253	2	3	0.214066	468	268	DoS	1
d6c03904efb0c29310ba8db0a6047271	-0.2016	04f226e4e07b8c9468f3fbb7b44d4a04	0.2209	2	3	0.534953	1710	115722	Generic	1

TABLE VI: Preprocessed Data for ML - 2ndRound (UNSW-NB15 Dataset tuples)

srcip	sport	dstip	dsport	proto	state	dur	sbytes	dbytes	attack_cat
f0ef4b92207ed6a19bc5bf737449090a	-1.1499	d6a78de1053951029fe17d2fafbf63dc	-0.4158	1	1	0.001055	132	164	0
76a9eedcc8c7f93672d4867e8d0ce18f	1.9865	038cb8b62827d30a23360c12fdad801c	-0.4158	1	1	0.001169	146	178	0
eacfb9eb5bdeb42519949def45e2d46a	-1.101	1ce73b5285825252bdcd8908c7e0f650	-0.4158	1	1	0.001134	132	164	0
53ec5cb7fe15f1fc7fd79dd5f1c81ad8	-0.1128	53ec5cb7fe15f1fc7fd79dd5f1c81ad8	0.0837	3	2	0.338406	46	14265	0
11c76906f69e573cff39c6e8d76dbded	1.4058	d6a78de1053951029fe17d2fafbf63dc	-0.4158	1	1	0.001126	146	178	0
fc892446e2871bb474fbe0611fd6aa9b	-0.4177	1ce73b5285825252bdcd8908c7e0f650	-0.4158	1	1	0.001167	132	164	0
d6c03904efb0c29310ba8db0a6047271	0.2773	9b627a55e3a25f17e7f4fe7bd8ae5c14	-0.2209	2	3	0.240139	918	25552	1
053ad4f007bc578dbcae52a628285755	-0.3771	9b627a55e3a25f17e7f4fe7bd8ae5c14	-0.2209	2	3	2.39039	1362	268	2
053ad4f007bc578dbcae52a628285755	-0.3096	17090328565ce63ee9d71823c1d0edf0	2.8253	2	3	0.214066	468	268	3
d6c03904efb0c29310ba8db0a6047271	-0.2016	04f226e4e07b8c9468f3fbb7b44d4a04	-0.2209	2	3	0.534953	1710	115722	4

[a] Subprocess *ID Attributes Symmetric Cryptography - Task AES symmetric algorithm with 128 key length*
[b] Subprocess *Quasi-identifiers Data Treatement - Task Noise Adding*
[c] Subprocess *Sensitive Data Supression - Task Supression*