Sharing and Automation for
Privacy Preserving Attack Neutralization

(H2020 833418)

D2.2.1 Privacy Requirements (M6)

Published by the SAPPAN Consortium

Dissemination Level: Public

H2020-SU-ICT-2018-2020 – Cybersecurity

Document control page

Document file: Deliverable D2.2.1 Privacy Requirements
Document version: 1.0
Document owner: Sebastian Schäfer (RWTH)

Work package: WP2
Task: T2.2 Privacy Requirements
Deliverable type: Report
Delivery month: M6
Document status: ☒ approved by the document owner for internal review
☒ approved for submission to the EC

Document History:

Version	Author(s)	Date	Summary of changes made
0.1	Sebastian Schäfer (RWTH), Arthur Drichel (RWTH)	20.09.2019	Outline and first draft
0.5	Sebastian Schäfer (RWTH), Arthur Drichel (RWTH)	25.10.2019	First complete version to collect feedback
0.6	Sebastian Schäfer (RWTH), Arthur Drichel (RWTH)	29.10.2019	Incorporated feedback
1.0	Sebastian Schäfer (RWTH), Arthur Drichel (RWTH)	30.10.2019	Finished version including all updates and feedback

Internal review history:

Reviewed by	Date	Summary of comments
Avikarsha Mandal (FIT)	27.10.2019	Include a scenario describing the whole process for privacy requirements on a specific use-case, editorial improvement, missing references
Tomas Jirsik (MU)	28.10.2019	Spell and grammar check

Executive Summary

The goal of this deliverable is to collect and describe privacy and sanitization requirements and apply them to the identified use-cases of SAPPAN. This is addressed from two sides: The end-user perspective and the organization perspective.

For the end-user perspective, we describe how data can be monitored, processed, and shared to be compliant to the GDPR. On the other side, organization specific requirements are dictated by individual policies. For example, a University might be willing to share data as long as it is compliant to the GDPR, but a commercial company might not be able to share any data related to customers.

In the first part of this document, we outline the key elements of the GDPR as well as a high level description of organization specific policies. Additionally, we introduce technologies and concepts related to privacy, e.g. anonymization techniques. In the second part, we briefly discuss privacy for SAPPAN architecture components related to data sharing. Finally, we describe the privacy requirements of a specific scenario in detail, categorize the data all identified use-cases for SAPPAN and discuss the privacy and sanitization requirements of these categories.

Contents

Privacy Requirements Gathering

In this chapter, we introduce the concepts that we use to define the privacy requirements for SAPPAN. We start with an overview of the GDPR. Afterwards, we describe additional organization specific requirements. Finally, we introduce some technologies and concepts that can be used to implement privacy requirements.

1.1 The General Data Protection Regulation for End-User Privacy

In this section, we describe the basics of the General Data Protection Regulation (GDPR) [1,2], which is the legal basis for the end-user privacy requirements of this project. The GDPR is a EU regulation regarding data protection of all citizens within the European Union (EU) which became legally binding on 25 May 2018. Moreover, it dictates in which way data of individuals has to be stored, processed, and protected. In the following, we introduce the key definitions, principles, user rights, and implications forming the basis of decisions we make regarding the privacy requirements of end users in this project. Note that these are neither complete nor written in strictly legal terms and only meant to summarize the key parts of the GDPR that are relevant in the context of SAPPAN.

1.1.1 Key Definitions

In this section, we outline the key definitions which are used in the following Sections describing the principles as well as the privacy requirements in the next Chapter.

1. **Personal data** describes any information directly related to an individual (see data subject).
2. **Data subject** is a natural person who can be identified directly or indirectly, e.g. via name, address, location, online identifiers, IP address, web cookies, and also by pseudonymous data if it is relatively easy to identify someone from it.
3. **Processing** describes the manual or automated performance of operations on personal data. This includes e.g. recording, alteration, sharing, retrieving, and storage of such data.
4. **Pseudonymization** describes an operation to process any personal data such that it cannot be linked to a data subject anymore, without additional information such as an encryption key or seed. If the additional information is kept separately with limited access, pseudonymization is a tool to ensure the privacy of personal information.
5. **Data Controller** is a natural or legal person or group deciding why and how personal data will be processed.
6. **Data Processor** is a natural or legal person or group processing data on behalf of the data controller.
7. **Recipient** is a natural or legal person or group to which personal data is disclosed.
8. **Consent** of the data subject describes a specific and freely given agreement of the data subject for processing his or her personal data.

1.1.2 General Data Protection Principles

In this section, we outline the key principles of the GDPR to give an overview of the rules according to which personal data has to be processed. The following points describe a basic guideline for the protection of personal data.

1. Processing of personal data must be lawful, fair, and transparent to the data subject. This includes that the data subject has given consent to the processing of personal data. Also, the data subject must be informed about the purpose of data processing in a clear and transparent way. Additionally, data subjects have the right to withdraw their consent at all times and the data controller needs to keep documentary evidence of consent.
2. The purpose of data processing must be clearly specified and communicated to the data subject before the data is collected. The collected data must not be used for other purposes than specified.
3. For a given purpose, only the amount of data that is necessary to fulfill the purpose should be collected.
4. Additionally, the collected data should only be stored as long as necessary to fulfill the specified purpose.
5. The collected data must be kept accurate and up to date.
6. The integrity and confidentiality of stored and processed data must be guaranteed. This can be achieved by encryption or integrity protection algorithm. In addition, the access to stored data should be secured as much as possible, e.g. using end-to-end encryption and two-factor authentication.
7. The data controller must be able to demonstrate GDPR compliance according to the principles at all times (accountability). To be able to achieve that, everything regarding personal data (e.g. collecting, using, storing) should be documented including a reference to the person that is responsible for these actions. In case of larger scale or regular monitoring of data, a Data Protection Officer is required, who is responsible for ensuring GDPR compliance.
8. All staff that is working with personal data should be trained on these security and privacy policies. Also, access to personal data should be limited to only the persons who need it.

1.1.3 Data Protection Officer

In this section, we briefly describe the tasks of a Data Protection Officer (DPO).

In general, there is no need for *every* data controller to have a DPO. However, there are three criteria when it becomes necessary.

1. When personal data is processed by a public authority (except for courts acting in their judicial capacity).
2. When the organization's core activities require regular and systematic monitoring of data subjects on a large scale.
3. When the organization's core activities consist of large scale processing of data categories referenced in Article 9 of the GDPR (e.g. data with respect to ethnic origin, political opinions, religious beliefs, genetic or biometric data, health data), or data related to criminal convictions (Article 10 of the GDPR).

The tasks of a DPO include the following:

1. Inform and advice employees involved in the processing of personal data about the GDPR, including staff training.
2. Monitor compliance with the GDPR and assigning of responsibilities.
3. Provide advice when requested.
4. Cooperate with the supervisory authority and act as the contact point for potential issues.

1.1.4 Rights of Data Subjects

In this section, we outline the data subject's rights with respect to privacy. These rights need to be reflected in how personal data is handled.

1. Communication between the data controller and the data subject must be transparent and easily accessible.
2. The data subject must be provided with contact information of the data controller and, if existent, the data protection officer.
3. The data subject has the right to request the purpose of processing his or her personal data as well as the time period where the data is collected.
4. The data subject has the right for rectification and erasure of his or her personal data as well as restriction of the processing.
5. The data subject has the right to receive all personal data collected concerning him or her. The transmission of this data needs to be in a structured, commonly used and machine readable format.
6. The data subject has the right to object processing of his or her personal data.
7. The data subject has the right to not be subject of automated decision making based on his or her personal data if such processing results in legal effects concerning the data subject.

1.1.5 Obligations Data Controller and Data Processor

The principles and user rights imply several obligations for the data controller and the data processor in order to comply with the GDPR. In this section, we outline measures to achieve this.

1. All personal data needs to be maintained and stored in a structured way to be able to transfer it to the data subject on request.
2. The processor needs to implement mechanisms to restrict data processing of individuals or to delete all stored data.
3. All processing and decision making based on personal data needs to be recorded and made available on request to ensure transparency.
4. Personal data must be pseudonymized, e.g. using encryption. The additional information to revert this process, e.g. decryption keys, must be stored separately to the data.
5. In case of a personal data breach, the data subject must be informed within 72 hours after the breach. Also, the data controller is under the legal obligation to notify the supervisory authority.

1.2 Data Sanitization and Organization Specific Privacy Requirements

In the previous section, we described the key principles of the GDPR, which is the legal basis for the privacy of end-user data. In addition to that, the second aspect of privacy requirements for SAPPAN considers the perspective of protecting sensitive data of organizations taking part in the sharing of data. In contrast to end-user privacy, such additional requirements are not defined by law but by organization-specific policies. This implies, that each organization might have different requirements on what data can be shared.

Some organizations (e.g. universities) might be able to share raw data as long as it is anonymized in a GDPR compliant way. Other organizations might be able to share only detection models trained on their private data, but not the data itself. Yet other organizations might not be able to share any data, but can contribute to collaborative

classification and detection tasks, or share handling models that are independent of any personal data. In order to address the specified use-cases for SAPPAN, it is useful to develop solutions for multiple privacy levels. In the following, we present a list of cases where either sharing is not possible at all or further anonymization is needed before sharing, even if the information to share is already compliant to applicable law (GDPR). Most of these requirements are related to intellectual property which is not yet published, still in development, or used commercially.

1. Bachelor, master, and Ph.D. theses
2. Research and development purposes and projects
3. Analysis, validation, and testing of new concepts with technological partners
4. Data collected from customers
5. Technology used in commercial products

In general, the DPO of each organization will decide what kind of data will be shared. The DPO will mainly ensure compliance with applicable law, but he or she might also decide against sharing even if it is compliant with e.g. the GDPR.

1.3 Privacy Enhancing Technologies and Concepts

In the following, we present some technologies which can be used to enhance the privacy for individuals and organizations. The mentioned techniques can be applied in the context of SAPPAN, however, this list may not be complete and we do not limit us to the usage of these.

1.3.1 Aggregation

A simple technique to obfuscate individual data points is aggregation. Instead of transmitting single information one by one, it is often useful to aggregate data e.g. by calculating the average over all data items in order to hide individual data items within a database. This is of course not possible for every type of data items.

1.3.2 Data Anonymization

If data has to be stored and no mapping of specific data items to individuals is required, it is necessary to remove all Identifiable Attributes (IA). IA are attributes which directly reveal the identity to which a data item belongs. This, for instance, can be the IP addresses. If it is required to retain such a mapping, encryption or pseudonymization of the IA should be used and the mapping should be stored preferably encrypted an in different place. In the context of SAPPAN, especially pseudonymization of IP addresses will be relevant. For that, multiple solutions exist, e.g. the CryptoPan library or the tool Capsan).

However, simple removing the IAs might not be enough, e.g. (5-digit ZIP code, birth date, gender) uniquely identify 87% of the population in the U.S. Hence, it is required to generalize or suppress so-called Quasi-identifiers (QID) which are a set of attributes that can reveal the identity.

k-anonymity

k-anonymity [3] is a property of anonymized data which limits the success rate of de-anonymization.

For k-anonymity all IAs are removed and the QID's are changed such that at least k tuples have the same QID's values. Thereby, each data item within a dataset cannot be distinguished from at least k−1 other entries. Sensitive Attributes (SA) which can be used e.g. for data mining or machine learning are left unchanged.

However, this might not be enough when SA's within an equivalence class lack diversity. An attacker with background knowledge (e.g. she knows some of the SAs) could still de-anonymize a target. l-diversity and t-closeness are two approaches which enhance the protection of k-anonymity even more.

l-diversity

k-anonymity focuses on the identifying information and does not prevent privacy leaks on the sensitive part. In k-anonymity all sensitive values in one equivalence class could be the same, and therefore reveal the sensitive information. l-diversity [4] builds on k-anonymity and additionally requires that the distribution of a sensitive attribute in each equivalence class has at least l "well represented" values to protect against attribute disclosure.

t-closeness

l-diversity limits the information gain between a prior belief B0 of a sensitive attribute (before any knowledge of the database) and a final belief B2 (after examining the database and the relevant equivalence class) by requiring that P (the distribution in the equivalence class) has l-diversity.

An adversary could gain information between the prior belief B0 and a posterior belief B1 by examining the global distribution Q, which should be treated as public knowledge. If the information gain from B0 to B1 is large, it means that Q contains lots of new information. Since it is not possible to control the people's access to Q, t-closeness [5] aims to limit the information gain between the posterior belief B1 and the final belief B2 by limiting the difference between P and Q. The closer P and Q are, the closer B1 and B2 are.

Hence, t-closeness measures how close the distribution of values of an attribute in an equivalence class is, compared to the distribution of values of the attribute in the whole database. An equivalence class is said to have t-closeness, if the distance between the distribution of a sensitive attribute in a class and the distribution of the attribute in the whole database is no more than a threshold t.

1.3.3 Encryption

Data should be stored encrypted, thereby, in case of a security breach, leaked data cannot be interpreted without the encryption key. While this ensures data protection, encryption can be used to enhance the privacy of individuals and organizations for communication, data sharing, and distributed computations. Besides standard encryption algorithms (e.g. AES), there exist other concepts that can be used for privacy-preserving computation, which we briefly outline in the following.

Secure Multi-Party Computation (SMPC)

The goal of Secure Multi-Party Computation (SMPC) [6] is to provide methods for parties to jointly compute a function over their inputs while keeping those inputs private. Thereby, all parties can compute a correct result for a given function, that is equal to the output which would have been computed by a trusted third party, without disclosing private input data to any party.

For instance, a popular example for a method enabling secure computation for two parties is Yao's Garbled Circuits. However, to use this method it is required to convert the function to compute into a Boolean circuit. Such circuits can become huge for many functions.

Homomorphic Encryption (HE)

Homomorphic Encryption (HE) [7] allows standard operations (e.g. additions and multiplications) on encrypted messages. Thereby, a function can take encrypted instead of plaintext data as input and compute the result without the need for decryption. Thereby, a correct computation can be guaranteed without leaking any information on the input values or the result. The decryption of the result is possible using the same cryptographic scheme which was used to encrypt the inputs.

The Paillier cryptosystem is a popular example for an additively homomorphic encryption scheme.

Shamir's Secret Sharing

Shamir's secret sharing [8] allows a number of parties to split information among each other such that the information of one party is useless. Only if all parties collaborate and combine their shares it is possible to recover the split information.

Differential Privacy

Differential privacy [9] aims at ensuring that the outcome of any analysis of a database is equally likely, independent of whether any individual is part of the dataset or not. Thereby, an attacker is not able to distinguish whether an individual is included in the dataset based on the results of queries to the database. Differential privacy is a measured metric that can be used to provide indistinguishable outcomes of different entries. The general idea for achieving differential privacy is adding noise to the dataset. It is important that the noise follows a random and unpredictable pattern in a non-deterministic way. One possibility of adding noise is to apply the Laplace mechanism. Hence, differential privacy minimizes the risk that is incurred by joining a database.

1.3.4 Privacy-Preserving Machine Learning

Differential Private Stochastic Gradient Descent (DPSGD)

Stochastic Gradient Descent (SGD) is the common technique for training machine learning classifiers. It follows an iterative approach to optimize an objective function. In each iteration an approximation of the actual gradient of the entire data set is computed

over a random subset of the data. A possibility to introduce differential privacy to SGD is to add noise to the update equation [10].

Private Aggregation of Teacher Ensembles (PATE)

The Private Aggregation of Teacher Ensembles (PATE) [11] approach provides strong privacy guarantees for sensitive training data used in machine learning. Fig. 1 provides an overview over this framework. In this approach, multiple models are trained using several disjoint datasets. These models may rely on sensitive data and are therefore not published. However, they are used as "teachers" to train a "student" model. The student learns its classification capabilities based on querying an aggregate teacher model using public data. Here, the student cannot access the individual teacher models or the underlying data. The aggregated teacher model answers the student's queries by the voting of all teacher models while adding the Laplacian distribution to each class. Thereby, differential privacy is assured for the privacy properties of the student model.

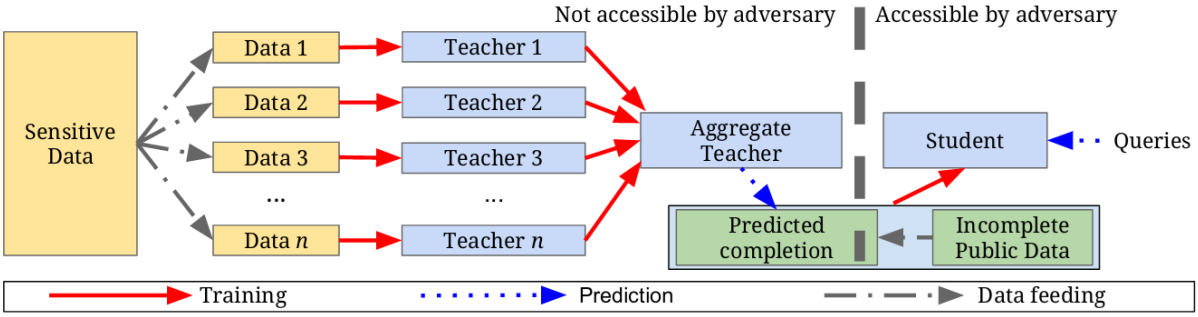

Figure 1: Overview of the PATE approach: (1) an ensemble of teachers is trained on disjoint subsets of the sensitive data, (2) a student model is trained on public data labeled using the ensemble. Source: [11]

Learning Anonymized Representations with Adversarial Neural Networks

Instead of using private sensitive data to train machine learning classifiers it might be possible for some use-cases to generate synthetic data that fits the statistical descriptions of the private input.

The authors of [12] use adversarial neural networks that aim at learning data representations that preserve the relevant parts for classification while dismissing private sensitive information. Such anonymized representations could be used for privacy-preserving sharing of data.

1.4 Private Sharing of Data in the Context of Machine Learning

In SAPPAN we share various types of data in order to enhance individual intrusion detection capabilities and to reduce response times. Thereby, the classification accuracies of detection models can be improved due to the increased amount of data which can be used for training. However, the shared data might reveal sensitive information about individuals or organizations. Hence, it might be desirable to

anonymize data before sharing. Moreover, some organizations (e.g. companies) might not be willed to share specific types of data while others (e.g. universities) are. Therefore, it is important to define different possibilities of information sharing which assure different levels of privacy. In the following, we shortly present four different types of knowledge sharing, which account for different levels of privacy.

Sharing of anonymized data

The simplest method in achieving a certain level of privacy is by removing all Identifiable Attributes which directly enable mapping to individual data points. However, as presented in Section 1.3.2, it also might be required to generalize or suppress the Quasi-identifiers.

Sharing of trained models

If the data used to train machine learning models is confidential, it might be possible to share the trained classifier instead of the data itself. However, it might still be possible to infer from the model to the utilized training data at least to a certain degree. Here, differential privacy could be applied in order to hinder re-identification attacks.

Teacher-Student Models

As described in the PATE approach, it is possible to train a local student classification model based on teacher models which have been trained on sensitive data. Thereby, the student model is trained solely using queries to an aggregated teacher which responses with the noisy voting of all teachers. By adding noise to the votes, differential privacy can be guaranteed. In contrast to sharing pre-trained models, here, the model parameters, which can be intellectual property, are not shared.

Federated Learning

Federated learning [13] can be used to improve a public machine learning model using private input data in a privacy-preserving manner. In this scenario, a model which is trained on public data is obtained by multiple parties. Each party improves the current model by feeding private data to the model. The changes to the model are summarized by a focused update which can be shared to the source of the public model. All updates of all participating parties are then averaged and applied to the shared model in order to improve it. With this procedure, the whole private training data sets remain locally at the sides of the participating parties.

2 Privacy Requirements of the SAPPAN Architecture and Use-Cases

In this chapter, we describe the privacy requirements of the SAPPAN architecture as well as the identified use-cases. In the first section, we outline the parts of the SAPPAN architecture that are involved in sharing, storing, and processing of data related to individuals as well as organizations that are part of SAPPAN. In the next section, we describe and classify what types of data is needed for each identified use-case and specify the corresponding privacy requirements.

2.1 Privacy related to SAPPAN Architecture Components

The specification of SAPPANs architecture is described in detail in Deliverable D2.4.1, hence we will only shortly introduce the components that have explicit requirements on privacy. For that, we describe where and how the privacy requirements of the use-cases are reflected in the architecture. This applies mostly to the components that are directly related to sharing of data.

2.1.1 Anonymizer and Sanitizer

Whenever any kind of data is transferred between two organizations or shared with all organizations, the components Anonymizer and Sanitizer are involved. Both need to be available locally at each organization participating in sharing of data and they implement functionality to transform data into a state which does not violate any privacy requirements. For that, input and output format need to be standardized (e.g. PCAP, NetFlow, Linux Syslog).

The Anonymizer will be used to ensure that all data is compliant to the GDPR before sharing. To do that, all personal identifiers (e.g. IP addresses, mail addresses, location data, hostnames) need to be either removed or replaced with a pseudonym which is only reversible by the sharing organization. In case a detection task is performed on the shared data, the latter should be used such that the sharing organization can match potential alerts. The Anonymizer will be implemented in the same way for each participating organization because it enforces legal regulations.

The Sanitizer will be used to ensure those additional requirements as mentioned in Section 1.2. These reflect organization specific policies, e.g. how customer data needs to be modified before it can be shared. Because of that, the Sanitizer will differ for each organization as opposed to the Anonymizer.

2.1.2 Intelligence Provider and Feature Provider

The components of the SAPPAN architecture that are responsible for data are called Provider Proxies. There exist two types, namely the Intelligence Provider and the Feature Provider. The first handles sharing of any kind of intelligence (e.g. detection models, detection rules, response handling playbooks) between the organizations via the SAPPAN sharing system. The latter handles transfer of features for collaborative learning or detection tasks (e.g. labeled datasets, PCAP files) via end-to-end encrypted peer-to-peer connections.

When data is shared or sent to another party, the corresponding Provider Proxy always uses the Anonymizer and Sanitizer to process the data. This ensures that all data leaving an organization environment is compliant to the GDPR as well as to organization specific policies. The specifics on how these components need to handle different types of data are described in Chapter 3 in more detail.

2.1.3 Intelligence Consumer and Feature Consumer

The counterparts of the Provider components are the Intelligence Consumer and Feature Consumer, respectively. According to the architecture, the compliance to privacy requirements of all data reaching the consumer components is ensured by the Anonymizer and Sanitizer. Besides ensuring the end user privacy for the data it is shared, we also consider the information what kind of intelligence each organization is interested in as private. For that, we introduced the Filter Manager component in the architecture. In general, all information that is shared via the Intelligence Provider to the

SAPPAN sharing system will be forwarded to all organizations. Using the Filter Manager, each organization can specify what types of intelligence are useful for them. Since each organization has its own Filter Manager instance, the sharing system does not learn anything about the interests of the organizations.

2.1.4 Internal Databases

The SAPPAN architecture includes several database components. These databases are used for storing machine learning datasets, machine learning metadata, filtering rules, detection metadata, intelligence, provenance tracking, system configuration information, and administrative information of the sharing system. In general, access to these databases needs to be restricted, such that only humans or technical components that require the data can access it, e.g. using encryption and authentication.

2.2 Privacy Requirements on Data from Use-Cases

In this section, we first describe the whole process of describing privacy requirements of one specific use-case. We then, aggregate the identified use-cases for SAPPAN based on the data that is used and discuss requirements for anonymization and sanitization for each category. Second, we describe our methodology in more detail. Afterwards, we present the aggregated use-cases and categories.

2.2.1 Complete Privacy Requirements of a specific Use-Case

In the following, we describe the complete process of specifying the privacy requirements for the use case SAPPAN-selected 7 (Domain Generation Algorithm (DGA) detection). We discuss how the data needs to be monitored, stored, and shared and present multiple levels of privacy for each step. The following table (from deliverable D2.1.1) describes the use-case in detail.

Descriptive name/goal	**Detection, assessment and handling of infected hosts or IoT devices by malware that uses Domain Generation Algorithms (DGAs)**
Team member	SOC analyst
Unique use case ID	SAPPAN-selected 7
Steps	1. The classifiers detect an algorithmically generated domain name 2. The analyst confirms or disproves the incident. The triage can be omitted in case of high-confidence classification. 3. Network traffic for the infected host is blocked until malware is removed. 4. Malicious DNS queries are attributed to a malware by another classifier. 5. Based on the detected malware, handling steps are recommended. 6. The analyst confirms or modifies the handling steps. 7. Intelligence of known malicious domain names and DGAs (e.g. DGArchive) is updated and shared
Data sources	DNS NX-traffic for detection, DNS queries and IP addresses for response action
Tools	• Machine learning classifiers on NX traffic for detection of algorithmically generated domain names • Multi-class classification model to attribute malicious queries to malware which generated it

Monitoring and sharing of data for training

Before the detection of malicious NX-traffic can start, the machine learning classifiers need to be trained. The training can either be done locally or in a collaborative way using the input of multiple organizations. For training purposes, only NX responses need to be monitored and stored, and other DNS traffic can be discarded. The monitored traffic is then either completely labeled as benign (if the network is considered clean), or labeled manually using existing detection methods. For the labeling process, the IP addresses corresponding to the NX traffic might be needed, but the training data which is stored afterwards only consists of labelled NX domains without any personal information. Hence, no anonymization is needed. However, even NX traffic alone might include information that is considered confidential by some organizations (e.g. typos in queries for websites). Therefore, it might be necessary for some organizations to perform sanitization.

In case an organization is not willing to share (even sanitized) NX traffic, it is still possible for them to participate in collaborative training. For this, different methods for more privacy preserving machine learning can be used, as described in Section 1.4. Besides sharing of anonymized data, the following methods can be used to achieve different levels of privacy.

1. **Sharing of trained models:** The organization trains models locally based on their data and shares the resulting models. These can either be used by other organizations or combined with other models to a global model.
2. **Teacher-student models:** This method does not even require to share trained models. The organization only needs to provide a classification service with its model, such that another organization can train a global model based on the classification results.
3. **Federated learning:** For this method, the organization only updates an already existing model by retraining it with their own data.

Monitoring and sharing of data for detection

For DGA detection, only NX domains are needed. However, if the detection produces an alert, the organization needs to link the malicious domain back to its source. Hence, the IP addresses producing the NX traffic need to be monitored as well. However, IP addresses are personal identifiers and not sharable. In the case of collaborative detection, two levels of privacy can be used.

1. **Sharing of NX domains with pseudonyms:** In this case, the IP addresses corresponding to NX domains are pseudonymized before sharing. This is done, such that only the organization owning the data can reverse the pseudonymization to link possible alerts back to IP addresses. However, other organizations can still see which NX domains were queried by the same source (without knowing who the source is).
2. **Sharing of NX domains only:** In the second case, the organization only shares NX domains. In case of an alert, the organization needs to search for the source, for example, using a log file. This requires more computational effort, but is also more privacy preserving.

Sharing of gathered intelligence

When a NX domain was classified as malicious (either using local or collaborative detection), it can be used as new intelligence. Because the intelligence only consists of a malicious domain it can be shared without anonymization, because no personal information needs to be included. However, there might be cases where an organization does not want to share this intelligence, e.g. to not disclose a possible infection within their network.

2.2.2 Methodology for General Privacy Requirements

In the following, we describe the process of gathering general privacy requirements based on the data in all remaining use cases. As a first step, we extract all different data types from the use-cases (defined in Deliverable D2.2.1) and add them to a table together with the corresponding use-case identifier. Additionally, we add a type tag, information whether the data is sharable, and if anonymization is required. As a next step, we categorize all types that are similar with respect to privacy and sanitization requirements. Finally, we describe the data types of each category and discuss whether the data in this category requires anonymization and sanitization. We also describe how the anonymization and sanitization can be achieved and present the possibility to introduce multiple levels of privacy. This enables us to classify each use-case into one or more categories.

Note, that at this stage of the project it is still unclear how the tasks required for the use-cases will be solved. Some use-cases will include collaborative learning or detection methods. The development of these approaches will be done in future steps. Hence, the description of privacy requirements for the use-cases can only be done in a more general way.

2.2.3 Data types within Use-Cases

In the following, we present the table (aggregated from the identified use-cases of Deliverable D2.2.1) including all different data types including the use-cases which use the data. For each data type, we include information whether the data is sharable and if it requires anonymization. G denotes the generalizes use-cases and S the selected use-cases.

Data	Use-Case	Type	Sharable	Anonymization
Third-party data types (reputation lists, OSINT data)	G1, S9	intelligence	yes	no
SIEM events	G1	event	yes	yes
Operating system events (e.g. security events logs, P*owershell* logs)	G1	log	no	-
Low-level end point events (e.g. process creation, module loading, file system access, network connections)	G1	log	no	-
End point protection events (e.g. malware detections)	G1	event	yes	yes
Corrective action	G1-3,6-13.15	handling information	yes	yes

Blacklisted host IP or domain name or URL	G2	network data	yes	no
DHCP logs	G2,9,10	log	no	-
Computer type (server, workstation, printer, router,...)	G2,5,9,10	additional data for assessment	yes	no
Computer OS including version and patch level	G2,9,10	additional data for assessment	yes	yes
LDAP user info	G2,9,10	additional data for assessment	no	-
User information	G2,3,9-11	additional data for assessment	no	-
Organizational unit	G2,9,10	additional data for assessment	no	-
Organization unit distance[N1]	G2,4	additional data for assessment	yes	no
Alerts log	G2,9,10	log	no	-
Surrounding/related traffic	G2,3,9-13 S3,4,5	network traffic	yes	yes
Host log	G2,9,10 S3 S6,8,9	log	no	-
Process which caused anomaly	G2,9,10	additional data for assessment	yes	no
Malware type	G2,9,10, 12	additional data for assessment	yes	no
Host information	G3,11	additional data for assessment	no	-
Phishing e-mail	G3,4 S1,6	additional data for assessment	yes	yes
Phishing URL	G3 S1,2	additional data for assessment	yes	no
SMTP information	G3 S6	network data	yes	no
Category of phishing	G4	intelligence	yes	no
Blacklisted domain list	G4	intelligence	yes	no
Number of targeted employees	G4	additional data for assessment	yes	no
Flow data	G5 S3,4	network data	yes	yes
Alerts	G5	event	yes	yes[N2]
History of blacklisted host IPs	G5	intelligence	yes	no
DNS traffic	G5,14 S2,7	network traffic	no[N7]	-
Inventory	G5 S6	additional data for assessment	no	-
Behavioral pattern	G5	intelligence	yes[N3]	no
Network monitoring alert system rules	G5	intelligence	yes[N4]	no
Data from UEBA analytics	G5	intelligence	yes	yes
Incoming and outgoing network traffic (unencrypted and encrypted) of servers	G6 S8	network traffic	no	-

Best practices of other organizations	G6-8,14,15	intelligence	yes	no
External malicious IP addresses	G6,14,15	intelligence	yes	no
Internal IP addresses with mapping to user	G6,14,15 S7	data for corrective action	no	-
Login attempts in access logs	G7	log	yes	yes
LDAP/AD, host logs	G7 S6,7	log	yes	yes
HTTP requests	G8	network data	yes	no
Application information	G8,12	additional data for assessment	yes	depends
File hash	G8	additional data for assessment	yes	no
Network distance	G9,10	additional data for assessment	yes	no
Dictionary enumeration attempts	G11	additional data for assessment	yes	no[N5]
Adversary movement patterns	G11	intelligence	yes	yes[N6]
Vulnerability information	G12	additional data for assessment	yes	no
Malware sample	G12	additional data for assessment	yes	no
Adversary movement patterns	G12	intelligence	yes	yes[N6]
Network/Firewall logs	G13 S3,4	log	yes	yes
Initial Point of Compromise information	G13	additional data for assessment	yes	yes
Adversary network profile	G13	intelligence	yes	no
Benign public key certificates	G14 S2	additional data for assessment	no	-
Malicious public key certificates	G13,14 S2	additional data for assessment	yes	no
Benign TLS client hello messages	G14 S2	additional data for assessment	no	-
Malicious TLS client hello messages	G14 S2	additional data for assessment	yes	yes
Suspicious domain, host	G14	network data	yes	no
TCP traffic	G15	network traffic	yes	yes
UDP traffic	G15	network traffic	yes	yes
Session numbers	G15	additional data for assessment	yes	no
Local anomaly detection system non-functional requirements and specification	G16	additional data for assessment	yes	yes
Local anomaly detection system functional requirements and specification	G16	additional data for assessment	yes	yes
Organizational IT infrastructure description	G16	additional data for assessment	no	-
Reports on the key performance indicators for SOC	G16	additional data for assessment	yes	yes

Reports on the key performance indicators for local anomaly detection system	G16	additional data for assessment	yes	yes

[N1]Organization unit distance can be used to discriminate between targeted and random phishing campaigns.

[N2]Contains identifiers that might be considered personal, such as IPs, domains, e-mail addresses.

[N3]If the behavioral pattern is not specific for the given network, it makes sense to share it.

[N4]If the rule does not contain an IP address, it can be shared without restrictions.

[N5]Only if the dictionary is not personalized (generated per user).

[N6]For use with machine learning models.

[N7]Sharing of extracted features or whole machine learning models might be possible.

2.2.4 Privacy Requirements of Data Type Categories

In the following, we present the table of data categories derived from the use-cases. Based on this, we describe the individual privacy requirements for each category.

Data	Sharable	Anony-mization	Sanitiza-tion
Intelligence (e.g. black/white/reputation lists, OSINT data, UEBA analytics data, adversary profiles, behavioral patterns, alert rules, malware information)	yes	depends	yes
Alerts / SIEM / OS level / Low level Events	depends	depends	yes
Additional data for assessment (e.g. Organizational/Network/User/Host/Application or logs)	depends	depends	yes
Network traffic/flows and data included in or derived from network traffic (e.g. domain names, URLs, public key certificates, TLS handshakes)	depends	depends	yes
Detection and response decisions/best practices/corrective action	yes	depends	yes

Intelligence

This category includes all types of data that can directly be used for detection. This ranges from basic reputation lists (e.g. for IP addresses, domains, or mail addresses) to rule based patterns or machine learning models for detection.

In general, this type of data is already processed and derived by either human operators or local detection systems. By design it should not include personal information, but more general information on how to detect malicious behavior. Hence, anonymization should not be necessary.

Sanitization needs to be applied if the intelligence contains information that is considered confidential by the organization. For example, this can apply to detection rules used in commercial detection systems which are not meant to be public. Another example is rules generated by specific security tools, where the organization does not want to make public how they protect their infrastructure, in order to make attacks more difficult.

Alerts and Events

This category includes alerts and events produced by SIEM (Security Information and Event Management), operating systems, or detection systems. This kind of data often includes personal information, e.g. the source of an alert (IP addresses or other identifiers).

To be able to share this information, it is necessary to anonymize it. It might be possible to replace all identifiers with pseudonyms. However, then it is still possible to link alerts or events from the same source. A more privacy-preserving approach is to simply remove these identifiers. However, if this data is used for collaborative detection, it is harder for the organization owning the data to link the detection results back to the source. Also, in case the data is shared for collaborative learning, the additional information which alerts or events were triggered by the same source is valuable.

However, this category of data might be classified as confidential by many organizations, because it includes critical information related to the organization's infrastructure. In most cases, this data will not be used for sharing but only for local detection and response actions.

Additional Contextual Data for assessment

This category includes additional contextual data that is used to assess alerts, e.g. by a SOC analyst. This data can contain information that allows for re-identification of individuals, but also data which is not sensitive, for example public key certificates or the number of open sessions. However, in most cases, it includes personal information, for example, information about the software that is running on the device that produced the alert and which person was operating it.
For the data to be useful locally, the included personal information enables to link alerts to a person or system. This data should only be kept as long as necessary. If this information contains data that exposes personal information, it can only be shared if it is anonymized.
Additionally, it might be required to perform data sanitization in order to ensure compliance with organizational policies. For example, an organization might not want to share what software is used by their employees.

Network data

This category includes data from various protocols or web technologies, for example NetFlows, DNS traffic, TLS handshakes, URLs, and Certificates. This data will be used locally as well as for collaborative training and detection.

In most cases, this data contains personal information, for example IP addresses, domains, and mail addresses. To be able to share this data, it is necessary to either remove these identifiers completely or to replace them with pseudonyms. For example, DNS traffic will be used to detect bots using Domain Generation Algorithms (DGA). In case a global model is used for collaborative detection based on shared DNS packets, the IP addresses are only needed to link a potential alert to its source. However, it suffices if the corresponding organization does this locally, while the global detection only classifies the DNS packet. This is an example where different privacy levels for

sharing can be specified. If the shared DNS packets contain pseudonymized IP addresses, it requires less computational effort to link an alert back to its source. However, if the DNS packets are shared without identifiers, the alert can still be linked to its source, but with additional effort (e.g. by searching network logs for the corresponding domain) and in favor of more privacy.

Since most data in this category contains critical information about the infrastructure of an organization, it is necessary to sanitize this data.

Detection and response actions

This category includes information about handling alerts, for example how to remove detected malware.
In most cases, this information is general enough such that it does not need anonymization. It comprises lists of actions on how to handle certain events which are not specific for individual users. In case this data includes an IP address (e.g. if it is a corrective action for a specific host), it has to be anonymized such that it only describes how to handle such an incident.
However, commercial companies like security vendors might not be willing to share information about how to handle alerts. It will depend on the organization, whether this can be shared. A non-commercial organization will most likely be more willing to share such information.

3 Summary

In this document, we first described the principles of the GDPR as a basis for end-user privacy requirements. We outlined additional requirements for sanitization to ensure that organization specific policies can be satisfied when data is shared. Next, we presented privacy enhancing technologies as well as approaches that allow collaborative learning of machine learning models in a privacy-preserving way. These can be used to achieve different levels of privacy to meet the requirements of different organizations. Finally, we categorized the SAPPAN use-cases based on used data and the privacy level they require.

4 References

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

[2] https://gdpr.eu

[3] L. Sweeney. “k-anonymity: A model for protecting privacy”. In: International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2002, pp. 557–570.

[4] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. “l-diversity: privacy beyond k-anonymity”. In: 22nd International Conference on Data Engineering (ICDE'06) 2006, pp. 24-24

[5] N., Li., T. Li, and S. Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity." In: IEEE 23rd International Conference on Data Engineering. IEEE, 2007.

[6] C., Dwork, and A., Roth. "The algorithmic foundations of differential privacy." Foundations and Trends in Theoretical Computer Science. 2014, pp. 211-407.

[7] O., Goldreich. “Foundations of cryptography: basic applications”. Cambridge University Press. 2004

[8] J., Katz, Y., Lindell. “Introduction to modern cryptography”. Chapman and Hall/CRC, 2014.

[9] A., Shamir. “How to share a secret.” In: Communications of the ACM, 22(11), 1979, pp. 612-613.

[10] S. Song, K. Chaudhuri, and A. D. Sarwate. “Stochastic gradient descent with differentially private updates”. In: Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE. IEEE. 2013, pp. 245–248.

[11] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar. “Semi-supervised knowledge transfer for deep learning from private training data”. In: Proceedings of the 5th International Conference on Learning Representations, Toulon, France. 2016

[12] C. Feutry, P. Piantanida, Y. Bengio, and P. Duhamel. “Learning anonymized representations with adversarial neural networks”. In: arXiv preprint arXiv:1802.09386 2018.

[13] J., Konečný, H. B., McMahan, B., F. X., Yu, P., Richtárik, A. T., Suresh, and D. Bacon. “Federated learning: Strategies for improving communication efficiency”. In: arXiv preprint arXiv:1610.05492. 2016.