PIA REPORT

SECURITY AND PRIVACY

Marta Longo 202207985

Sara Táboas 202205101

Table of contents

6. Conclusion

1) Introduction

In assignment 3, we were asked to perform a PIA (Privacy Impact System) for a real-world project, called COP-MODE, which involved gathering data from a group of users through data collection campaigns.

Smart devices are consistently and constantly gathering data. By collecting, storing, analyzing and even sharing this data, we end up creating what we think is personal data but does not always belong to us. It is truly part of the businesses that collect it with our consent. Regardless of all the progress made in privacy regulation for the past years, it is still a fact that our data is controlled by companies. For this reason, the number of projects related to privacy management and regulation is increasing, projects such as COP-MODE.

The COP-MODE system is a research project of the University of Coimbra, University of Porto, University of Cambridge and INESC TEC, focused on providing better privacy to mobile devices and, hence, developing privacy-preserving mechanisms to empower users with control over their data.

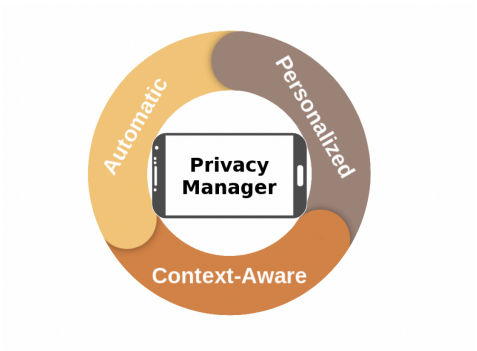

Figure 1. Objective of COP-MODE Project

2) **Describe/characterize the system setup that is the subject of this PIA.**

The project COP-MODE works on enhancing the privacy of mobile users by creating a **permission manager** that intercepts permission checks done by any app to the operative system and prompts the user to either accept or deny the permission at that specific time, while collecting contextual data. The answer to a particular permission for a given app is cached for 30 minutes to avoid constant prompts.

An anonymized version of the collected dataset is made available to researchers.

COP-MODE PROJECT

In order to develop the privacy manager, data needs to be collected. Several campaigns were run with volunteers that were willing to contribute to the project.

PROCESS:

1) The first step starts with the recruitment of participants. For this purpose, the COP-MODE team shall advertise the opening of a campaign through an email list, for example. A potential participant shall then download and install COP-MODE Apps Retriever (CM-AR) through the Play Store.
2) After downloading CM-AR from the Play Store, the participant shall run the app while following its instructions. CM-AR will send the participant's email and the list of installed apps (and respective permissions) to our server. (**Note** that by list of apps we refer to the name of each installed apps, that is, no other data or information belonging to the apps is collected). For instance, we collect that the participant has the application WhatsApp installed, but we don't have access to its messages or its contacts. The participant's email is collected only as a

communication medium for the duration of the campaign. At the end of the campaign we delete all emails such that the data is anonymized. The list of apps is used by the COP-MODE team to install them in a campaign's smartphone that is then lent to the participant for the duration of the campaign. Thus, the campaign smartphone will come pre-installed with the participant's personal apps.

3) As referred in step #2, after sending the list of apps to the server, the COP-MODE team will install these apps in a campaign's smartphone. This smartphone will additionally have pre-installed COP-MODE Naive Permission Manager (CM-NPM), a permission manager that will prompt the participant apps' permission requests. CM-NPM will additionally collect and send the data for/to the COP-MODE server. An email will be sent to the participant to pick up the smartphone and when receiving the device, the participant shall sign our data collecting agreement.
4) For the period of one week, the participant shall use the campaign smartphone as his main personal phone. Throughout this period, the permission manager CM-NPM will prompt the user to whether to allow or deny apps' access to resources. The participant's responses to these prompts will be collected as well as both user and device contextual data.
5) After the data collection period, the participant shall receive an email from the COP-MODE team, signaling the end of the campaign. The participant shall then return the smartphone and get his voucher as reward for his participation.

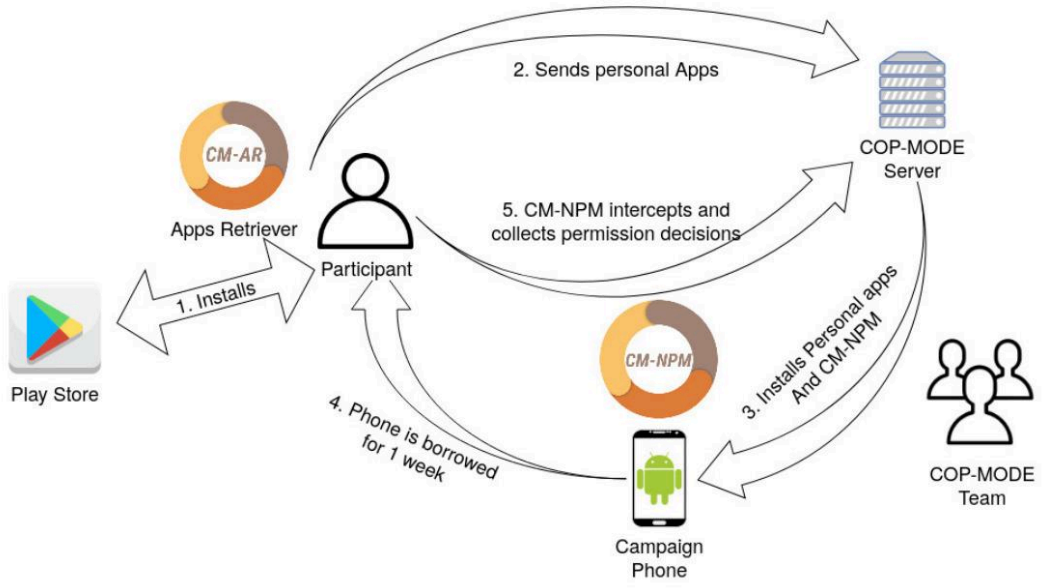

Figure 2. Process of COP-MODE Campaign

What PII is collected/processed?

- Email Address (valid for one week, corresponding to the campaign week);

Instantaneous Data:

- List of installed apps on the smartphone and their respective permissions;*
- Calendar entries (event identifiers, their location, and start and end dates);*
- Connection type (Wi-Fi, 4G);
- Device operation.

Continuous Data:

- Geographic location;*
- Nearby devices.*

Permission Data:

- App information (*name, version, category, visibility);
- User decision;
- User input.

(* data deemed sensitive)

Which PII principals are affected?

The data subjects provide personal information, which is processed and stored by the PII controller/processor. This includes **sensitive data** marked with a '*'.

What systems and processes are involved in the handling of PII?

- **Equipment and Software:** The use of a lent phone with pre-installed user applications and COP-MODE's permission manager, which collects necessary data for the project.
- **Data Collection:** The collection of various types of data, including installed apps, connectivity type, device context, calendar entries, location, neighbor devices identifiers, and permission prompts data.
- **Data Handling:** The storage and management of collected data by the COP-MODE team at the University of Coimbra, ensuring high privacy and security standards**.**
- **Data Sharing:** The potential sharing of anonymized and sanitized data with academic partners for research purposes, under strict privacy conditions.

These systems and processes are designed to respect the privacy of participants while collecting data for research on enhancing mobile device privacy through context-awareness. Participants have control over their data and can request its deletion at any time.

3) Initial risk assessment for a plain setup without any corrective measures

• Very high: These risks must be absolutely avoided or significantly reduced by implementing controls that reduce both their impact and likelihood. Recommended practice suggests that the organization implement independent controls of prevention (actions taken prior to a damaging event), protection (actions taken during a damaging event), and recovery (actions taken after a damaging event).

• High: These risks should be avoided or reduced by implementing controls that reduce the impact and/or likelihood, as appropriate. For example, the matrix has a high-risk entry for a very high impact and low likelihood; in this case, the emphasis is on reducing impact. The emphasis for these risks should be on prevention if the impact is relatively high and the likelihood is relatively low and on recovery if the impact is relatively low and the likelihood is relatively high.

• Moderate: The approach for moderate risk is essentially the same as for high risk. The difference is that moderate risks are of lesser priority, and the organization may choose to devote fewer resources to addressing them.

• Low: The organization may be willing to accept these risks without further control implementation, especially if the treatment of other security or privacy risks also reduce this risk.

• Very low: The organization may be willing to accept these risks because further attempts at reduction would not be cost-effective

Risks	**Likelihood**	**Severity**	**Risk Level**
Illegitimate access to data	High	Very high	Maximum
Unwanted modification of data	Very high	High	Maximum
Data disappearance	Very high	High	Maximum

Figure 3. Risk Assessment of the PIA Software without the corrective measures

Illegitimate access to data

- **Likelihood:** The likelihood of this threat happening is high, given that there are no safety measures in place when data is sent from the smartphone to the COP-MODE server. Someone unauthorized could intercept this transfer and access the data illegitimately.
- **Severity:** The severity of this threat is very high, as unauthorized access to the data could result in data being modified, deleted, or illegitimately shared to harm the project team and/or the individuals whose data was used for the project.

Unwanted modification of data

- **Likelihood:** The likelihood of this threat happening is very high, due to the fact that there is a considerable number of people involved in the project, which increases the probability of data being changed. Additionally, as the project involves data treatment, there are multiple points of interaction where data can be altered unintentionally.
- **Severity:** The severity of this threat happening is high, considering that it can harm the data used in the project, leading to errors in analysis, decision-making, or reporting, impacting the quality of the project's outcomes. It will, also, require, additional resources/time to recover the modified data. This would affect the project reputation and effectiveness.

Data disappearance

- **Likelihood:** The likelihood of this threat happening is very high because a researcher working with the dataset may accidentally delete the data, which can occur quite frequently. Additionally, if the data is accessed illegitimately, it could be deleted, impacting the reputation of the project and those involved.
- **Severity:** The severity of this threat is high because there may be chances of recovering the data or not. This could lead to a misrepresentation of the data collected, to inconsequent decisions being made by the team and will require additional resources/time.

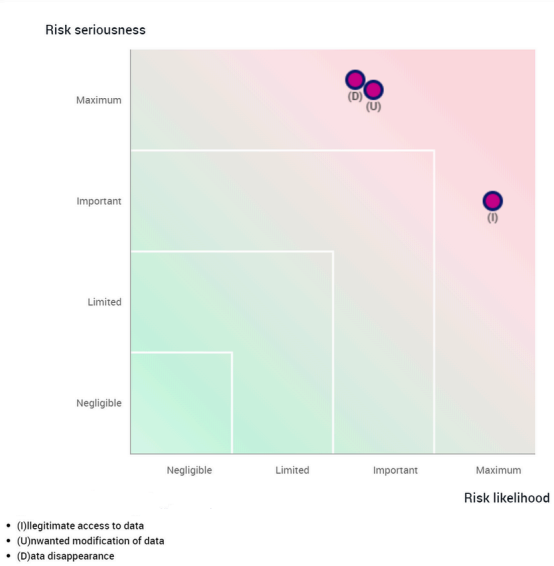

Figure 4. Risk Mapping of the risks mentioned above

4) A plan of action (correction measures) to address the identified risks

Risks	Mitigations
Illegitimate access to data	**Cryptography:** Encrypting the data can protect it from unauthorized access. This ensures that even if someone intercepts the data, they cannot decipher it without the encryption key. For encrypting data in this project, the most appropriate encryption algorithm is AES. AES is ideal for encrypting large datasets because it uses symmetric keys, making it efficient for both encryption and decryption. By choosing a secure key length and restricting its access to only authorized team members, AES can provide security to the data in the dataset. The AES algorithm can, also, be used to encrypt data during transmission. We can combine cryptography with **Data Masking.** **Data Masking and Anonymization**: Data masking involves replacing sensitive data with altered values, which makes it difficult for unauthorized individuals to extract meaningful information from the dataset. Data anonymization involves removing or altering personal data to the point where individuals can no longer be identified. **Network Segmentation**: Dividing our network into segments can limit the spread of unauthorized access. This prevents attackers from moving laterally within our network and accessing sensitive data. **Access Control**: Implement strict access controls to ensure that only authorized team members can access the data. We can use **MFA** (**Multi-Factor Authentication)**, which requires users to provide multiple forms of verification, such as a password and a code, before granting access to a system or application. **Access Privileges:** Regularly review and update access privileges. **Data Minimization:** Collect and store only the data necessary for your operations. This reduces the risk of unauthorized linkage and ensures compliance with privacy regulations. For example, we could collect location data at a less granular level, such as city-level or regional-level data, instead of precise GPS coordinates.

Unwanted modification of data and data disappearance	**Data Backup and Recovery:** If data is regularly backed up, in the event of an undesired modification in the dataset, we will always have a copy of the original data stored. This allows the recovery of information to its state before the unwanted modifications occurred. **Version Control Systems:** These systems track changes in data and maintain a history of the modifications. In the context of the COP-MODE project, alterations are frequent and systems like this reduce the time spent recovering data. VCS are able to identify the source of unwanted changes and revert them to previous versions. **Digital Signatures:** They can also be used to reduce the impact of unwanted modification of data and data disappearance, by providing means to verify the integrity of data by generating a unique cryptographic fingerprint of the data and encrypting it with the signer's (that can be a member of the research team) private key. This allows team members to verify that the data has not been altered since it was signed, providing assurance of its integrity. **Team members monitoring**: The monitoring process involves team members submitting detailed “commit” messages each time they change the dataset somehow, providing a clear description of the modifications made. Therefore, if there is any change without this message, it would be identified properly as unwanted. **Human Authentication:** Whenever there's an undesired modification to the data, it requires approval from more than one authenticated member by the server. This prevents unauthorized alterations by unauthenticated users but relies on the human factor, which can delay the process.
Data leak through eavesdropping of communications between smartphones and the project server.	**Secure Server Configurations: Virtual Private Networks (VPNs) -** A VPN would establish a secure and encrypted connection between the user’s smartphone and the COP-MODE server. They use encryption algorithms to secure data in transit, requiring an encryption key to read any information obtained by interception. Additionally, when using a VPN, the user’s IP address is masked by the VPN server’s IP address, enhancing anonymity. **Secure Communication Protocols:** Protocols like HTTPS and TLS (Transport Layer Security)

	are considered secure for communications between smartphones and servers and its implementation is a good practice for projects like COP-MODE. Firstly, they use encryption to protect data as it travels between smartphones and servers. They also ensure that communication between smartphones and the server is established with the intended server. Besides this, these protocols mitigate several common attacks, such as Man-in-the-Middle attacks, Hijacking and Packet Sniffing and guarantee the maintenance of data integrity. **Cryptography:** Encrypting the data can protect it from unauthorized access. This ensures that even if someone intercepts the data, they cannot decipher it without the encryption key. We proposed for this issue the AES algorithm, that uses symmetric keys and it's appropriate to large datasets. **Certificate-Based Authentication:** Implement certificate-based authentication to verify the identity of both the smartphones and the project server. This ensures that only authorized devices and servers can communicate with each other, mitigating risks like Man-in-the-Middle attacks.
Data leak from unauthorized server/data access	**Access Controls:** Implementing measures of access control ensures that only authorized users can access sensitive data based on their roles and responsibilities, while providing an additional layer of security for remote access. **Role-Based Access Control (RBAC) -** RBAC limits access to server resources and data based on a user's role within a project. **Multi-Factor Authentication (MFA) -** MFA requires multi-factor authentication for remote server access. **Secure Server Configurations: Virtual Private Network (VPNs) -** A VPN would establish a secure and encrypted connection between the user's smartphone and the COP-MODE server. They use encryption algorithms to secure data in transit, requiring an encryption key to read any information obtained by interception. Additionally, when using a VPN, the user's IP address is masked by the VPN server's IP address, enhancing anonymity. **Monitor Server Access:** Implement IDS (Intrusion Detection Systems - tools designed to detect unauthorized access, malicious activity, or policy violations within a computer network

	or system) and IPS (Intrusion Prevention Systems - tools to prevent identified threats) to detect and answer any data breach or unauthorized access. **Cryptography:** Encrypting the data can protect it from unauthorized access. This ensures that even if someone intercepts the data, they cannot decipher it without the encryption key. For encrypting data in this project, the most appropriate encryption algorithm is AES. AES is ideal for encrypting large datasets because it uses symmetric keys, making it efficient for both encryption and decryption. By choosing a secure key length and restricting its access to only authorized team members, AES can provide security to the data in the dataset. The AES algorithm can, also, be used to encrypt data during transmission. We can combine cryptography with **Data Masking.**
Data at rest linkage. The data in the server should not be linkable to an individual by third parties. I.e. the email address must be protected (not clear text), but the PII collector/processor must still be able to link it to the data in the server (e.g. for deletion in case the participant requests it).	**Cryptography:** Encrypting PII, such as email address, at rest to ensure that it cannot be easily linked to individuals by third parties. We could again use AES, since it works better for large amounts of data. **Data Anonymization:** Anonymizing data helps ensure it cannot be linked to specific individuals. Mask or obscure certain parts of the data to prevent direct identification. For example, for the list of apps we could group by categories (entertainment, health, finances, etc) or for the geo location, we could convert precise geographic coordinates into less precise representations, such as zip codes or neighborhood names. **Access Controls: Role-Based Access Control (RBAC) -** Ensures that only authorized users have access to the server and data, according to their role and responsibility in the project. **Multi-Factor Authentication (MFA) -** Requires multi-factor authentication for remote server access. Third parties once linked to the project have to be given credentials to access data, otherwise they could access all the data they wanted, even the sensitive data.
Sensitive information leakage, in particular the names of the	**Data Anonymization:** Instead of using the raw application names, we could anonymize them by grouping them into broader categories. For example, instead of storing specific application

applications that are collected	names like "Facebook" or "Instagram," we would categorize them as "Social Media" applications. This reduces the granularity of the data while still providing valuable insights into the types of applications used. **Cryptography:** Encrypting the application names at rest and in transit ensures that unauthorized parties cannot access or read them, using AES because for the reasons mentioned above. **Secure Data Processing:** Ensures that data processing complies with privacy regulations mentioned on the data collection agreement. **Access Control**: Implement strict access controls to ensure that only authorized team members can access the data. We can use **MFA** (**Multi-Factor Authentication)**, which requires users to provide multiple forms of verification, such as a password and a code, before granting access to a system or application. **Access Privileges:** Regularly review and update access privileges.

Figure 5. Mitigations of all the risks

5) Risk assessment for an improved setup with application of corrective measures. Assessment/evaluation of the evolution of the risks with the corrective measures.

Risks	Likelihood	Severity	Risk Level
Illegitimate access to data	Moderate	Moderate	Important
Unwanted modification of data	High	Low	Limited

Data disappearance	High	Low	Limited
Data leak through eavesdropping of communications between smartphones and the project server.	Moderate	Moderate	Important
Data leak from unauthorized server/data access.	Low	Moderate	Limited
Data at rest linkage. The data in the server should not be linkable to an individual by third parties. I.e. the email address must be protected (not clear text), but the PII collector/processor must still be able to link it to the data in the server (e.g. for deletion in case the participant requests it).	Moderate	Moderate	Important
Sensitive information leakage, in particular the names of the applications that are collected.	High	Low	Limited

Figure 6. Risk Assessment with the corrective measures

Illegitimate access to data:

- **Likelihood:** The likelihood of this threat is moderate because with access control, privileges and data minimization we are reducing the data collected and restricting its access to only authorized people, which reduces the likelihood of data being accessed illegitimately.
- **Severity:** The severity of this threat is moderate since the data is encrypted and anonymous. This reduces the likelihood of the attacker perceiving the data and using it to their advantage without having at least some knowledge of cryptography algorithms and understanding of the data context.

Unwanted modification of data:

- **Likelihood:** The likelihood of this threat slightly decreases to high because of team members monitoring, human authentication and digital signatures. This helps reduce the unwanted modification of data but it is still likely to happen.

- **Severity:** The severity of this threat drastically decreases to low since we implemented mitigations that secure the data to a point where it is easy to access the data before any changes in the dataset.

Data disappearance:

- **Likelihood:** The likelihood of this threat slightly decreases (to high) because of team members monitoring, human authentication and digital signatures. This helps reduce the unwanted modification of data but it is still likely to happen.
- **Severity:** The severity of this threat drastically decreases to low since we implemented mitigations that secure the data to a point where it is easy to access the data before any eliminated data.

Data leak through eavesdropping of communications between smartphones and the project server:

- **Likelihood:** The likelihood of this threat can be considered moderate because with communication protocols, such as HTTPS, and server configurations, as VPNs, we can assure that most connections between smartphones and the server remain secure, avoiding interceptions.
- **Severity:** The severity of this threat is moderate because the data is encrypted and anonymized. In the event of a data breach, the dataset will be relatively protected. However, the gravity of the situation is serious because someone with knowledge of encryption and anonymization techniques can still decipher the data.

Data leak from unauthorized server/data access:

- **Likelihood:** The likelihood of this threat can be considered low, because, by restricting access with RBAC and MFA and using server configurations, such as VPNs, it is extremely difficult to access the server and the dataset.
- **Severity:** The severity of this threat can be considered moderate. Since the data is encrypted, it makes it difficult for the unauthorized individual to perceive the data.

Data at rest linkage. The data in the server should not be linkable to an individual by third parties:

- **Likelihood:** The likelihood of this threat is moderate due to the access controls implemented, such as RBAC and MFA, that prevent access to the sensitive data that can be linked to an individual by third parties.
- **Severity:** The severity can be considered moderate because while the data is protected to some extent, there's still a risk of potential linkage by determined attackers or parties with advanced capabilities.

Sensitive information leakage, in particular the names of the applications that are collected:

- **Likelihood:** The likelihood of this threat is high because of Secure Data Processing, Access Control and Access Privileges. This helps reduce the unauthorized access of the server but it is still likely to happen.
- **Severity:** The severity of this threat is assessed as low because the mitigations we implemented protect and mask the sensitive data. If this leakage occurs, data is safe.

6) Conclusion

A robust privacy management system is crucial for research projects like COP-MODE, which collect and process various types of Personally Identifiable Information (PII). The implementation of a comprehensive Privacy Impact Assessment (PIA) identified several significant risks, including illegitimate data access, unwanted data modification, data disappearance, communication eavesdropping, and sensitive information leakage.

Our team proposed corrective measures to mitigate these risks, with a focus on encryption, data anonymization, network segmentation, and multi-factor authentication. The use of cryptographic techniques is essential for protecting data both in transit and at rest. Additionally, measures such as VPNs, HTTPS, and secure server configurations contribute to reducing the likelihood of eavesdropping and unauthorized server access.

Overall, the corrective measures proposed and implemented by our project team create a solid framework for protecting participant privacy. However, ongoing risk assessment and adjustment of security practices are essential to maintain a secure and privacy-respecting environment for data collection and processing.