A privacy and security-aware blockchain based design for a digital certificates system

María Fernanda Molina
Gustavo Betarte
Carlos Luna
Universidad de la República, Facultad de Ingeniería,
PEDECIBA Informática,
Montevideo, Uruguay, 11300

Abstract

Blockchain technology supports building transparent and decentralized systems in which the executed transactions can be easily traceable. Suppose one such system is intended to manage and process personal data. In that case, complementary mechanisms are required that make it possible for the system to comply, for instance, with data protection regulations. This work studies the integration of off-chain capabilities in blockchain-based solutions. In particular, we have focused on mechanisms that support safely moving data or computational operations outside the core blockchain network. We have carried out a thorough analysis of the European data protection regulation and discussed the weaknesses and strengths regarding the security and privacy requirements established by that regulation of solutions built using traditional blockchain technology. As a direct consequence of this study, we have conceived, and present in this paper, a system architecture for the design of privacy-aware solutions that use that kind of technology and put forward a systematic approach for performing a security and privacy threat analysis of one such solution. We illustrate the use of the proposed methodological tools, presenting and discussing the high-level design and security and privacy assessment of a system that provides services to handle, store, and validate digital academic certificates.

Keywords: Blockchain, Off-chain, GDPR, personal data protection laws, design principles, security and privacy, threat analysis

1 Introduction

Blockchain is a disruptive and innovative technology [1]: it provides support to build decentralized systems where the participating nodes of the network process transactions without the need of a responsible intermediary or authority. Blockchain offers many strengths compared to traditional approaches, such as decentralization, transparency, and traceability. On the other side, this technology has some general weaknesses concerning scalability and performance issues, but most importantly, in our view, with confidentiality, immutability, and access control.

In this work, we elaborate on the approach that favors the incorporation of *off-chain* capabilities in blockchain-based solutions, moving data or computational operations outside the core blockchain network [2]. Current proposals of off-chain processes aim to leverage a blockchain solution by addressing some of the intrinsic functional weaknesses we have previously pointed out. Typical scenarios that are identified as requiring the use of off-chain solutions are those that, for their operational behavior, need to improve performance or cost calculation processing, to perform intermediate operations on the off-chain leaving the final transaction on the blockchain (off-chain Signatures Pattern [2]) or to perform a final complex calculation on the off-chain (challenge Response Pattern [2]).

We have carried out a thorough analysis of the GDPR [3] (General Data Protection Regulation) to determine the weaknesses and strengths of solutions built using blockchain technology regarding the security and privacy requirements established by that regulation. In particular, we study the different possible architectures of blockchain and off-chain technology and their advantages and disadvantages regarding protecting personal data regulations.

Of special interest was solving controversial subjects such as pseudonymization and data anonymization when using hash functions and public-key cryptography. In [4, 5, 6, 7, 8] several challenges the GDPR poses to solutions built using blockchain technology are presented and discussed. In those works, it is analyzed, for instance, the processing of pseudonymized and anonymized data and the potential privacy violations that might occur from the use of hash values and private/secret keys in blockchain solutions.

One of the main objectives of our work has been to develop a system architecture that helps system designers to select operational off-chain constructs that are integrated with traditional blockchain functionalities allow to build more secure and privacy-aware solutions. In [9], [10] and [11], different software architecture are proposed and discussed that have been conceived to make use of blockchain mechanisms to perform access control to data and auditing and to use off-chain solutions to storage and process personal data safely. We propose, additionally, to use the blockchain as an integrity network to validate data stored off-chain, using links to data stored on the off-chain network and the use of an integrity control check mechanism. In the first phase of our investigation, we have focused on developing two primary constructs of the system architecture: a high-level system architecture model and a use case model. The use case model consists of several use cases that cover the top services we understand can be used to build a blockchain and off-chain-based system compliant with well-established security and privacy requirements, particularly those established by the GDPR.

We also put forward a methodology to perform a security and privacy threat analysis of such solutions. We illustrate the use of the proposed methodological tools presenting and discussing the design and the security and privacy assessment of a system that provides services to handle, store and validate digital academic certificates. The problem was raised by SeCIU (Servicio Central de Informática Udelar), the IT services unit of the Universidad de la República, Uruguay.

1.1 Contributions

As a result of our investigation, we have presented and published three papers at three international conferences. We present in the paper [12] a condensed formulation of the GDPR-complying analysis, the integration of off-chain capabilities in blockchain-based solutions, and a summary of the system architecture described in this work. In the paper [13] we delve into the GDPR-complying analysis of blockchain and off-chain technologies and, in particular, its applicability in the Uruguayan laws of protection of personal data. In the work [14] in turn, we present in detail the system architecture for the design of privacy-aware solutions that are built using blockchain technology and the methodology developed for security and privacy threat analysis. We also illustrate the use of those methodological tools in designing and analyzing a specification of the digital academics certificate system.

The present work, based on the master thesis [15], integrates these published papers and delves into the study presented in them. In particular, in this paper, we put forward a detailed state of the art of blockchain and off-chain technologies and related security issues and provide a detailed explanation of different proposals for blockchain and off-chain solutions and guidelines to select off-chain models and their corresponding architectures and technologies, according to the problem to be addressed. In addition, in this paper, we provide a detailed description of the privacy and security-oriented blockchain design guidelines presented and the methodology for security and privacy threat modeling, explaining in detail all the use cases and security and privacy threats detected for the proof of concept. We illustrate the system architecture design and the security and privacy threat methodology on a realistic and not trivial digital certificates system. We believe that the proposed threat analysis methodology is a contribution that can be used in the analysis of systems that manipulate personal data. Regarding the risk analysis itself, we do not know of a study of that kind, except one of the fraudulent activities to make fake diplomas described in [16], being performed over systems that manage digital certificates.

The specification of the digital certificates system we present addresses privacy and GDPR compliance aspects that have not yet been considered. For example, in our design, the verification of certificates is mediated by an access control mechanism, unlike other proposals, where this function is enabled without such control. An exception is a system proposed in [17], which has been designed to be GDPR compliant, that requires the registration and authorization of third parties who wish to validate certificates.

This type of solution, which provides automated mechanisms to validate university degrees, is of particular interest to our university. Furthermore, it becomes of particular importance in the presence of national and international frauds related to the subject, like the ones that have recently been reported in [18].

1.2 Organization

The rest of this work is structured as follows. In Section 2 we present related works analyzed. Section 3 provides a primer on blockchain and off-chain concepts and technology. In Section 4 we analyze the European

data protection regulation and discuss how well blockchain and off-chain mechanisms adapt to provide support for building GDPR-compliant digital systems. In Section 5 we put forward a system architecture for hybrid blockchain and off-chain solutions, as well as a methodology for security and privacy threat analysis we have developed. Then, in Section 6 we provide the specification of the digital certificates system and discuss the results of the threat analysis carried out on that specification. We conclude and discuss further work in Section 7. Finally, in A, we present in detail the result of the threat analysis.

2 Related work

Several proposals exist for off-chain solutions, the most common ones being either an external datastore, an external server, or an external peer-to-peer network. In [19] two different off-chain models are proposed. The *distributed or channels* model consists of a group of equal nodes in a peer-to-peer network outside the blockchain, which is organized by pre-defined rules. The *commit-chains or centralized* model consists of a centralized system that receives and processes user requests and periodically responds to the chain. The result of the processing is transmitted to the blockchain, which in turn verifies the result before persisting it. To carry out that verification without disclosing confidential information, the use of zero-knowledge test and verification processes like zk-SNARKs [20] and zk-STARK [21] has been proposed.

An off-chain solution may also be conceived as a storage or external processing system or a hybrid one [22]. In [2] different off-chain processing patterns are proposed.

Some of the challenges concerning data protection requirements the GDPR poses to solutions built using blockchain technology are discussed in [4] and [5]. Those works analyze the processing of pseudonymized and anonymized data and the potential privacy violations from using hash values and private/secret keys in blockchain solutions. In [23] is analyzed the use of smart contracts to manage the agreements between a data owner and a service provider. In [24] the question of blockchain and GDPR compliancy is analyzed by comparing different opinions of legal and IT professionals. The authors put forward the advantages and disadvantages for civilians when using a blockchain to process their data. In [9], [10] and [11], different software architecture are proposed and discussed that have been conceived to make use of blockchain mechanisms to perform access control to data and auditing and to use off-chain solutions to the safe storage and process of personal data.

The methodology for security and privacy threat analysis we present in this work makes use of concepts and procedures present in STRIDE [25] and the CNIL [26] methodology for privacy risk analysis.

As to the design of systems that store and manage digital educational certificates, several solutions have been proposed to deal with this problem, with different degrees of development. In particular, we have analyzed Latin American solutions like the Brazilian RAP System [27] and the Argentine System BFA [28]. In [16] the solutions *Blockchain for Education* and *EduCTX* are discussed. All these systems propose a hybrid solution, using the blockchain to validate the certificates by storing a hash of the certificate and, in some cases, some additional information. Some of those solutions use public blockchain networks and others private ones, but none implement access control mechanisms to perform the certificate verification process. The candidate gives the employer his certificate (an act considered an implicit consent of personal data access). The employer validates the provided certificate with the blockchain, typically comparing hash values. However, performing this verification implies access to personal data since the hash of personal data constitutes personal information. Therefore the execution of that procedure should be authorized by the data owner, as it is also proposed in [17].

3 On blockchain and off-chain technologies

This section presents and analyzes blockchain and off-chain technologies, as well as a compendium of the leading security problems related to these technologies.

3.1 Blockchain concepts

Blockchain is a peer-to-peer system that builds blocks with no centralized authority. A blockchain network is composed of a set of transactions grouped in blocks. Each transaction is a unique cryptographically signed instruction representing the valid passage from one state to another. A transaction can be a message or a code (called *smart contract*) and can include payment for its execution.

The immutability of the chain is based on the fact that each block of the chain contains the hash of the head of the previous block. In this way, a change in a block implies a change in the whole chain.

Regarding the access control of users, blockchain networks can be categorized into two types: i) *Permissionless* the network is open to anyone willing to participate, so the level of distrust among the participants

is high. Examples of these networks are Bitcoin [1] and Ethereum; and ii) *Permissioned,* the system is a private network where the entry of new participants is controlled, such as, for instance, Hyperledger [29].

3.2 Off-chain concepts

Off-chain transactions occur on a cryptocurrency network that moves the value outside of the blockchain. The premises of blockchain are transaction validation, consensus protocols, and decentralization. However, these premises generate all transactions to be validated, processed and stored in all the network nodes, which causes an overhead of work. Additionally, there is a cost to carrying out transactions, which may be due to the computational cost of processing or network storage. Finally, the peer-to-peer system generates a confidentiality problem, where information is processed and stored on all network nodes.

In summary, we find the following problems that can be solved with off-chain: i) scalability problems through throughput and transaction processing latency in blockchain; ii) cost problems derived from the payment made for transactions; iii) confidentiality problems, when it is required to process a calculation, the information must be exposed; and iv) requirement to store information outside the blockchain, either because of confidentiality or because of a large amount of data where a cost problem arises.

3.2.1 Off-chain processing

An off-chain solution may be conceived as a storage or external processing system or a hybrid one [22]. When processing a state change in a storage processing system, the stored information is received from an external node, which communicates the state change to the blockchain. The data is stored again in the external node upon entering the new state.A verification step should be performed to ensure the integrity of the data stored externally. An option to validate the data stored in the chain is to keep in the blockchain a hash value of the data that is externally stored. Additionally, it is necessary to implement an external storage access control system to preserve confidentiality. On the other hand, if the availability of the information stored is essential, it is pertinent to think about schemes where the data is stored redundantly. The Interplanetary File System (IPFS) [30], and SWARM [31] systems help with access control and availability issues.

In *off-chain computation* model, a part of the processing is performed on the off-chain. It is necessary to incorporate a verification process to ensure that the result returned is correct, as there is no integrity by design as in blockchain.

Finally, there are cases in which it may be necessary to use a *hybrid model*, using off-chain to store data and to perform computational process.

3.2.2 Off-chain architectures

In turn, we can find two types of architectures in off-chain: distributed or channels model and centralized or commit-chains model.

The *distributed or channels model* consists of a group of equal nodes in a peer-to-peer network outside the blockchain, which is organized by pre-defined rules, for example, through a smart contract. Generally, these systems require a unanimous consensus of the participants since only transactions approved and signed by all nodes are considered valid. A node can at any time dump the approved off-chain calculations into the blockchain. In the same way, if there is a dispute about the outcome of the off-chain, such as a dishonest participant trying to lie about it, honest participants can resolve the disagreement on the blockchain.

A potential risk in the distributed or channels model is the retention of funds when an initial fund is required to enter the system, and a malicious actor blocks them simply by not accepting the signature of the other participant. In addition, if off-chain is to protect the confidentiality of the information, this is compromised in a dispute. It is resolved in the blockchain, where all the nodes see the information.

On the other side, the *centralized or commit-chains model* consists of a centralized and not necessary reliable system that receives and processes user requests and periodically responds to the chain. This system performs the processing and returns the result along with proof that it is correct, which is verified by the blockchain, using mechanisms of zero-knowledge test and verification process, as is explained in Section 3.3. As it is a centralized system, an availability problem is added, which does not occur in the case of the channels model. Another difference of the channels model is that the establishment and state transition is accessible since it is a communication between a node and the external entity.

3.3 Test and verification processes

Testing and verification processes are necessary for a centralized off-chain architecture, where the third party that performs the processing and provides the result to the chain is not trusted. This type of verification is used by a third party to demonstrate that they have specific information or meet a particular requirement.

In short, it consists of a party (the prover) proving to another (the verifier) that a statement is true without revealing any information beyond the validity of the word itself. For example, given the hash of a random number, the prover could convince the verifier that there is indeed a number with this hash value without revealing what it is. This property is essential for confidentiality and privacy.

The prover, who performs the calculation, must send proof that the verifier will verify. For the test and verification process to be correct and efficient, it must meet the requirements described below. First i) the test sent (by the prover) must be short and non-interactive; that is, the prover must be able to send the test in a single message. Additionally, ii) the cost of verification must be independent of the computational cost of the problem to be solved. Ideally, the cost of verification should be less than solving the problem. Finally iii) the off-chain provider can use private information during the test generation, but the verifier will not learn or infer anything from the test obtained. This property is known as *zero-knowledge*.

Currently, the most efficient way to produce zero-knowledge, non-interactive tests that are short enough to be published on a blockchain is to have an initial configuration phase that generates a standard reference chain, shared by the prover and the verifier. For the system to work, it must be assumed that this standard reference chain was honestly generated. It can be trusted since the verification process will be compromised if a malicious user performs the initial configuration step.

In summary, these systems are based on the idea that the test and verification system is more efficient than the resolution of a complex calculation since it should be much easier to verify than to perform the computation. Researchers have made significant progress in this method in these recent years. However, the initial configuration costs and computational overhead for the cloud in the current state of technology make these methods unsuitable for most real-world applications.

3.4 Criteria for selecting off-chain models and architectures

As has been explained, there exist several proposals for off-chain solutions. In Table 1 we summarize those variants together with some guidelines to choose off-chain models and their corresponding architectures and technologies according to the problem to be addressed.

Off-chain models and implementation					
Problem	**Model**	**Architecture**	**Method**	**Technology**	**Tools**
Improve calculation processing	off-chain computation	channels/commit-chain	Delegate the computation work to the off-chain	Non-interactive and short verification process	Whisper Msj. Protocol/zk-SNARKs/zk-STARKs
Improve processing of a final calculation	off-chain computation	channels	Challenge-response system	Not required	does not apply
Reduce calculation processing costs	off-chain computation	channels/commit-chain	Delegate the computation work to the off-chain	Cost verification process more minor than the problem to solve	zk-SNARKs/zk-STARKs
Ensure the cost of intermediate transactions	off-chain computation	channels	Payment channel	Peer-to-peer system	Whisper Msj. Protocol
Reduce storage costs	off-chain storage	channels/commit-chain	External storage systems	Cost verification process less than the problem to solve	zk-SNARKs/zk-STARKs/BD: DHT, DDB
Ensure the confidentiality of information required to perform a calculation	off-chain computation	commit-chain	Delegate the computation work to the off-chain	Verification process zero-knowledge	zk-SNARKs/zk-STARKs
Ensure the confidentiality of intermediate transactions	off-chain computation	channels	Payment channel	Peer-to-peer system outside the blockchain	Whisper Msj protocol
Ensure the confidentiality of the information to be stored	off-chain storage	commit-chain	External storage systems	hash verification integrity/External storage access control system	Interplanetary File System (IPFS), SWARM DHT, DDB
Compliance with GDPR	hybrid model	commit-chain	External storage systems/Delegate computing work to off-chain	Zero-knowledge verification process/Hash verification integrity/External storage access control system	Zk-SNARKs/zk-STARKs
Ensure the confidentiality and availability of information to be stored	off-chain storage	channels	External storage systems	Zero-knowledge verification process/Hash verification integrity / External storage access control Peer-to-peer system	Interplanetary File System (IPFS), SWARM DHT, DDB

Table 1: Criteria for selecting off-chain models and architectures

3.5 Blockchain security issues

Blockchain security relies on public-key cryptography for identifying transactions and hash technology to provide guarantees of the immutability of the chain: each block of the chain contains the hash value of the head of the previous block. Therefore, a change in a block implies a difference in the whole chain. This architecture ensures integrity, immutability, and traceability by design. However, it presents some security problems that are described below.

1. *Confidentiality* Since all transactions must be validated and processed by all the network nodes, all the information necessary to perform this processing must be public, undermining the confidentiality of the data. For example, suppose a customer wants to prove that he has the token of a provider that publishes the hash of the tickets he provides. In that case, the customer can generate the hash of his token and thereby prove that it has one of the hashes published by the provider. However, to do this on the blockchain, he must perform the execution of the hash in the blockchain so that all the nodes perform the performance, thus revealing their token.

   Everything used in a smart contract is publicly visible, including local and status variables marked private. Defining something as private only prevents other contracts from accessing and modifying information, but this information will always be visible to everyone in the chain. Additionally, although the data can be stored in encrypted form, specific metadata is always available. Data can be inferred (through pattern recognition, for example), such as the type of activity and the volume associated with any public address of the network.

2. *Immutability* The impossibility of modifying smart contracts implies a problem of immutable bugs; that is, once a contract is published, it cannot be altered. This will be a problem if bugs are found in the code. Nevertheless, there are ways to use a new contract instead of the original one, similar to "updating" a contract. One way to update the code is to create an intelligent intermediary contract that will keep the smart contract address active. Therefore, all calls and transactions will be redirected to the active version with the delegate call function. That way, the same contract address will always be used, but that contract can execute a different smart contract code at the end. In addition, as will be seen in Section 3, immutability threatens compliance with personal data protection regulations.

3. *Key management risk* Blockchain is susceptible to theft of private keys and the control of the assets associated with external addresses being taken away. Digital assets could become unrecoverable in the case of theft of private keys, mainly due to the lack of an administrator or system controller [32]. In the world of blockchain, the possession of keys and ownership of content are synonyms, and the best way to obtain keys is to attack the weakest point in the chain, personal or cellular computers. It is essential to follow good wallet and critical management practices to mitigate this issue.

4. *Unpredictable state* When a user sends a transaction to the network to invoke a contract, he cannot be sure of the contract's state when the transaction is executed. In the interim, other transactions could alter the status of the contract since no order is guaranteed in the execution of the transactions. This happens especially in lottery systems.

5. *Transaction security issues* A risk related to transactions is the risk of double-spending, that is, spending the same money more than once. This can happen when some consensus protocols are used that allow for multiple simultaneous chains to be generated over some time until one is dropped [33]. Blockchain users protect themselves from this fraud by waiting for several confirmations when receiving payments, as transactions become more irreversible as the number of confirmations increases. Each time a new block is added to the chain, the verification is confirmed again. As a consensus, many users expect to have six confirmations before accepting a transaction as payment, to avoid the problem of double-spending.

6. *Consensus protocols security issues* Lottery systems may imply that several nodes propose a block at the same time. Block duplication is solved because miners use the criterion of attaching new blocks to the longer chain, so shorter chains are discarded. The system works fine if a coordinated group does not control the network.

A known security problem of the permissionless blockchain refers to consensus protocols and what would happen if the network is dominated by a majority (51 percent attack). These consensus protocols require most miners, to be honest. If a group of miners working together dominates more than 50 percent of the network, the network stops decentralized and becomes controlled by a group. In such a situation, it is even possible to alter transactional records or generate double-spend fraud.This has not happened so far, but it is not impossible for this risk to materialize, taking into account that the miners merge into groups to join their effort.

Additionally, the report "Majority Is Not Enough: Bitcoin Mining Is Vulnerable" [34] explains a method in which a group of non-honest miners works in coordination to obtain proportionally higher incomes than honest miners: subsequently, the income of the non-honest group (called selfish-mines) increases linearly with the size of the group. The system used for this is forcing honest miners to spend their mining on blocks destined not to be part of the blockchain, since they are in chains that will be discarded. Because of this, the selfish miners work on a private branch of the chain, while honest miners work on the public branch. The selfish-mines make their chain public when it is longer than the public one, so honest miners leave the chain they were working on, throwing away their work. This large-scale operation can be combined with a Sybil attack ([35]), where selfish-miners can use their power to invalidate transactions on the network, isolating honest nodes that will only receive information from non-honest nodes.

7. *Smarts contract security issues* The paper [36] presents a study of the best-known vulnerabilities of smart contracts in Ethereum, classifying them according to whether it is a vulnerability of the Solidity programming language, the EVM, or the blockchain technology. In [37] known attacks on Ethereum smart contracts can be found; within these, one of the best-known attacks is known as the DAO attack.

   A critical security risk related to smart contracts is the *oracle* (a participant external to the blockchain that provides data) that many smart contracts use to function. If malicious agents attacked these oracles, they could potentially corrupt the network by generating a chain effect across the entire network.

   Another security problem is the difficulty of *smart contracts verification*, which is how to verify that a smart contract effectively does what it says it does and does not perform other operations or that its execution has no unwanted side effects. One way to mitigate this problem is to provide information or rely on standards that can give users confidence in the characteristics of the published smart contract. An example of standards to which a smart contract can adhere is the Ethereum improvement proposal (EIP) [38], which is a design document of a smart contract that gives information to the Ethereum community or describes a new function for Ethereum, its processes or environment. The author of the EIP is responsible for generating consensus within the community and documenting dissenting opinions. Another method to mitigate this problem is to look for mechanisms to verify the published code compared to the source code, either to make our contract more reliable or to rely on contracts created by third parties. Some initiatives seek to verify that the published code matches the source code, such as Etherscan [39]. The objective is to trust the contracts that have been verified because it is possible to see the source code and verify that the published bytecode matches the source code. Finally, some initiatives work to verify the Solidity source code formally, or the EVM compilation, to demonstrate that it has no programming vulnerabilities or to prove that a contract complies with his high-level specification. A general problem with these initiatives is that most smart contracts do not publish their source code.

   The Solidity programming language also has security problems, like i) *call to the unknown*, which refers to the fact that some Solidity primitives may have the effect of invoking the fallback function that can be used for attacks. ii) *Gasless send*, that is, problems arising from an óut-of-gaséxception when using the send function. There may be vulnerabilities arising from the mishandling of exceptions in Solidity, called iii) *exception disorders*. iv) *Type casts* are vulnerabilities derived from the abuse of some exceptions of type of data in Solidity. v) *Reentrancy*, also known as the ŕecursive callv́ulnerability. The fallback mechanism can generate the recursive call of non-recursive functions. Another problem is vi) *keeping secrets*, since the blockchain is public, everyone can inspect the content of a transaction and infer the value of the fields, even if it is declared private. To ensure that a field remains secret, it is necessary to use cryptographic techniques.

3.6 Off-chain security issues

The use of off-chain to store information introduces problems already solved in the blockchain.

1. *Off-chain integrity issues* When storing or processing data outside the blockchain, integrity is not guaranteed by default as in blockchain. It is not possible to change a block transaction without changing the entire network. One way to ensure the integrity of stored data is to keep a reference and a hash value of the information stored externally in the blockchain. In the case of the computational process, it could be used zero-knowledge test and verification processes, as was described in Section 3.3.

2. *Off-chain availability Issues* In the blockchain, data availability is guaranteed by storing the information in all the network nodes. When data is stored outside the blockchain, this property is challenging to ensure since there is a single point of failure. Thus, solutions such as IPFS [30] and SWARM [31] have been proposed, where information is stored in a decentralized and redundant manner.

3. *Off-chain traceability issues* Unlike blockchain, where the audit log is natural since it is guaranteed the immutability and integrity of the stored data, off-chain auxiliary mechanisms must be considered to register an audit trail of the accesses and changes made.

4 Blockchain constructs, data protection principles and requirements

In this section we analyze how compatible blockchain is with personal data protection regulations like GDPR [3], the European Data Protection Regulation.

4.1 GDPR scope, roles and responsibilities

According to *Article 4* of GDPR, personal data includes all the data that is or can be assigned to a natural person. This consists of identifying people, such as public keys used in blockchain. GDPR defines the following roles:

- *Data Controller*: it is responsible for the processing of information and the appointment of the processor role (*Article 28*). A data controller can process the data collected using its processes. It can also work with a third party or an external service to process the data collected. Even in this situation, the data controller will not transfer control of the data to the third-party service, as it will be responsible for specifying how the external services will use and process the data.

- *Data Processor*: a data processor processes the data that is provided to him by the data controller, but he does not own the data he processes or controls. The data processor cannot change the purpose and data use, since the data controller defines this.

- *Representative*: is a natural or legal person designated by the controller or processor that represents the controller or processor about their respective obligations under this regulation.

- *Owner*: is the owner of personal information.

- *Recipient*: a natural or legal person, public authority, agency, or another body, to which the personal data are disclosed, whether a third party or not.

Who can assume these roles in a blockchain network is an issue that is under discussion, and it depends mainly on the type of blockchain network (permissionless or permissioned, as discussed in Section 2).

Beyond the assignment of roles, part of the responsibilities of these actors can be automated through a smart contract, which achieves: i) treat personal data only following documented instructions; ii) ensures that the people authorized to process personal data have agreed to respect confidentiality or are subject to a confidentiality obligation; and iii) deletes or returns all personal data once the provision of the processing services is finished and delete existing copies.

4.2 Shared responsibility

Regarding the responsibility of processing personal data in a blockchain network, the cases of permissionless and permissioned networks should be considered separately. In the case of closed networks, it is easier to define who is responsible, so it is better to use this type of network whenever possible. This role cannot be applied to a person or institution in an open blockchain network. Still, it can be analyzed if it is admissible in the legal field, considering the shared responsibility of all or some of the blockchain members. This is considered in the GDPR in its *Article 26*, as joint controllers for the treatment, if the following is fulfilled *"Where two or more controllers jointly determine the purposes and means of processing, they shall be joint controllers"*.

The article [40] analyzes the possible responsibilities of the different roles involved in a blockchain network, like *network developers*, who are part of the volunteer who adjusts the protocol using the blockchain, *nodes*, who sign and send transactions to the blockchain network through a node, *miners*, who execute the transactions sent by the nodes, and *smart contract developers*, which are companies or developers that create smart contracts used on the network.

The report refers to the document *"Blockchain and the GDPR of the European Union Blockchain Observatory and Forum"* [5], that dismisses the responsibility of developers, with the understanding that *"they volunteer to work on an open-source project and, in many cases, they do not receive direct compensation for their efforts and, in essence, they simply create a useful tool, they do not prescribe how this tool should be used"*. As for network users, who sign and send transactions to the blockchain network through a node, it is stated in this document that if they send personal data to the blockchain as part of commercial activity, they are more likely to be considered data controllers. However, suppose they send their data for personal use, for example, to buy or sell cryptocurrencies. In that case, they are likely to be subject to the family exemption of the GDPR and cannot be considered data controllers. An argument in favor of using permissionless networks is that the user gives his consent by choosing to use the network. However, the GDPR stipulates that the consent must be specific and unambiguous, implying an active granting of permission, not a passive one. Similarly, it could be argued that when initiating a transaction, a user is assuming a contractual obligation with the platform. Still, here it is also a passive act without explicit terms. However, *"Commission Nationale de l'Informatique et des Libertés"* (CNIL) [41] notes that the nodes, which have the right to write in the chain, can be considered data controllers, arguing that blockchain nodes define the purposes (objectives pursued by the processing) and the means (data format, use of blockchain technology, etc.) of processing. On the contrary, the CNIL does not consider the miners a data controller since they only validate transactions and do not decide on the purpose and means of processing.

The CNIL has indicated, concerning smart contracts, that software developers can be external providers but, if they actively participate in data processing, they can also be joint processors or controllers, depending on their role in determining the purposes of the processing.

As the background of responsibility, in Google Spain, the Court of Justice [5] emphasized the need to *"guarantee, through a broad definition of the concept of controller, effective and complete protection of stakeholders"*. Consequently, the Google search engine operator was qualified as a data controller even though it did not exercise control over personal data published on third-party websites.

What is clear is that we are in a grey area where, in some cases, it will not be possible to identify a controller. The issue of legality is more direct in the context of an authorized private network since it is possible to require each participant of the network to accept specific terms and conditions before being granted access to the network.

4.3 Confidentiality of personal data

It is difficult to ensure confidentiality in blockchain since data is available to all network members. As an alternative, it is possible to store data outside the blockchain, in an off-chain network, or use technologies like Trusted Execution Environments (TEEs), described in [42, 43]. There are blockchain networks that use this technology, such as Secret Network [44] or Hyperledger Avalon [45].

Additionally, *Recital 30 of GDPR* states that persons may be associated with online identifiers such as internet protocol addresses. As discussed in [40], blockchain public/private keys can be considered an identifier of a person and therefore are considered personal data. Public keys can also reveal a transactions pattern that could identify an individual user. Related to this, report [40] mentions a judgment on April 2014 of Digital Rights Ireland [46], where it was considered that metadata (such as location data or IP addresses) could also be personal data since *"those data, taken as a whole, may allow exact conclusions to be drawn concerning the private lives of the persons whose data has been retained, such as the habits of everyday life, permanent or temporary places of residence, daily or other movements, the activities carried out, the social relationships of those persons and the social environments frequented by them"*. Public keys cannot be obfuscated or deleted in the blockchain since they identify transactions. Related to this, the document on blockchain and the GDPR of the *European Union Blockchain Observatory and Forum* [5] indicates that one way of trying to avoid traceability is to use a new pair of keys for each transaction to prevent them from being linked to a common owner. The objective is to mitigate the risk of the owner of a key being revealed, that the binding could show all other transactions that belonged to the same owner. An alternative is to use decentralized digital identities technologies to protect identities, like the tools provided by Hyperledger Indy [47].

Similarly, hash values used in blockchain can be considered personal data, and [5] warns that when using a hash function, it is necessary to be aware that patterns that could allow traceability are not being created. This is due to the risk of bonding, which refers to situations in which pattern analysis enables an analyzer to

discover information about a particular individual. The cited paper gives an example of an application that performs purchase or sale transactions that publishes a hash with the buyer's address on a blockchain to record each transaction. In this case, if the registered hash is the same each time a user orders a transaction, the times and frequency of each user's transactions can be easily inferred. If simple information is saved, the outputs of a hash can be guessed from known inputs. An example mentioned in the report [40], is given by Edward Felten of the *American Consumer Protection of the Federal Trade Commission* [48], showed that it is pretty easy to establish someone's identity based on the hash functions that are derived from social security numbers, doing brute force work on the possible social security numbers for a country (about a billion in the United States). To increase the entropy of the information used to make a hash, it is suggested to use a single-use salt model that generates a separate random element for each message. The format of the extended message suggested for processing a hash would be as seen in Figure 1.

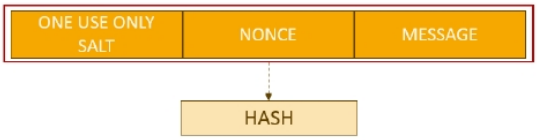

Figure 1: Format of the extended message for processing a hash [49]

4.4 Deletion or modification of personal data

Some of the rights included in the new regulations force the data controller to delete or modify data. This includes *Article 16 - Right of rectification*, *Article 18 - Right to limit the treatment of the GDPR* and *Article 23 - Limitations of the GDPR*. However, it is impossible to delete or modify the data in a blockchain once it resides in the chain. One way to modify data is to introduce a new block that updates a block's previous data.

Given the immutability of the data stored in blockchain, it is essential to determine what can be considered acceptable for personal data protection regarding data erasing. An alternative to erase the information is to process the data instead of deleting it so that it becomes anonymized and is no longer within the scope of the GDPR. In the case of hashes, this is equivalent to eliminating the original data from which the hash was formed. Regarding anonymization techniques, the work [40] concludes that the GDPR takes a risk-based approach, as it takes into account not only current technology but also future one. *Recital 26* of the GDPR states that for anonymization techniques, it should be considered, which is the available technology at the time of the processing. In that respect, Article 29 of Working Group[4] postulates that the possible advancement of technology should be considered in the period in which personal data will be stored. In the case of blockchain, this period is undefined, so in [40] it is argued that any data should be considered as personal data since it cannot reasonably be assumed that identification will remain anonymous in the future.

Despite these arguments, there are antecedents where anonymization has been accepted as a form of erasure. CNIL, in an article on blockchain [41], agrees with the deletion of original data as a method of deletion, even if the hash remains in the blockchain.

4.5 Weaknesses and strengths of blockchain and off-chain solutions regarding personal data protection regulations

Considering the previous discussion, we understand that the data protection requirements defined by GDPR can hardly be addressed by blockchain technology and the concept of a ledger stored by all network participants.

As a summary, we can indicate as breaches of blockchain permissionless against the personal data protection laws, issues like i) lack of assigned managers, ii) lack of control over access to information, iii) lack of control over the regulation in the use of information and the fact that iv) there is no territorial control. And in general, for blockchain technology, there is i) difficulty in guaranteeing the confidentiality of the information, ii) difficulty to delete or correct information and to iii) compliance with the right to a human intervention goes against the automation proposed by blockchain.

On the other side, blockchain technology meets by design some of the requirements of these data protection laws, such as the right of access to data by the owners (*Articles 12 to 15 and Article 20 of the right to data portability*), traceability, encryption, and hash techniques as security by design mechanisms and finally transparency in the data processing. The audit log is natural in blockchain since is guaranteed the

immutability and integrity of the stored information. This meets by design requeriments like *Article 30 - Registration of treatment activities* and *Article 15 - Right of access of the interested party.*

5 Privacy and security oriented blockchain design guidelines

Given the analysis carried out in the previous sections, we propose considering a hybrid model in which off-chain is used to store personal data and blockchain to ensure integrity, traceability, and access control. Personal data should not be stored on the blockchain but instead register a link to access the data residing in off-chain storage.

Given the functional characteristics of the blockchain technology, it has been suggested, for instance in [10], to use those functionalities to build embedded in the system i) an *access control network* that stores the access control policy and, upon request, performs the corresponding authentication and authorization tasks, and ii) an *audit network* that records all actions that have been performed, such as authorizations, access requests, modification, and deletion of data. It must be registered, for instance, who accessed the data and both approved and denied access.

We propose to carry out the integrity control through blockchain technology as one more service provided by this network. This can be done by storing a link of the information stored in the off-chain, along with a hash of this information.

As was explained in [11], we propose the use of a Gateway for external communication towards the blockchain and off-chain. As was proposed in [10], it could be used as a Security Token Service as an authorization mechanism. Once validated access through the access control network, the blockchain can send a token that the off-chain must validate before delivering information.

5.1 The use case model

This model consists of eight use cases covering the top services we understand can build a blockchain and off-chain-based system compliant with well-established security and privacy requirements.

We consider the following actors as requestors of services provided by and to the system: *Data owner* (DO), the user, owner of personal data; *Data controller* (DC), responsible for data processing; *Data processor* (DP), a data controller can delegate processing tasks to a data processor and the *Receiver (recipient)*, a third party who wants to access personal data. These actors must be registered in the system to perform authentication and authorization operations. For this, different solutions can be used, for example, using a system of certificates with a private-public key to authenticate data owner and a user and password system for recipients.

These use cases are illustrated in Figure 2 and are formally specified by the sequence diagrams shown in Figure 3 and described next.

1. *Register personal data* To register personal data, either the DC or the DP must first obtain authorization from the DO. Once the registration is authorized, the DC or the DP sends the data off-chain through the Gateway. The off-chain network communicates with the integrity network blockchain to store an integrity control (a hash of the personal data).

2. *Grant Access* The DO and the DC or DP must authorize a recipient's access to personal data. This access is registered in the blockchain access policy network. For a recipient to access the information, the DO and the DC or DP must authorize the request and send this authorization to the blockchain to update the access policy. This implies, therefore, registering the recipient in the system. It should be noted that the off-chain does not intervene in this flow since the data is neither accessed nor modified.

3. *Revoke Accesse* This process is done by updating the access policy stored on the blockchain. The blockchain validates the identification and access authorization before processing the requested change.

4. *Data access* Access to data is made through a request to the Gateway, which validates access through the access control blockchain network. The access control blockchain network returns the Gateway a token and the link to the data stored in the off-chain network. After that, the Gateway sends this data to the off-chain, validating the token, asking the blockchain, and returning the data to the Gateway.

5. *Verify data* This use case refers to validating the integrity of data. The user has data that he wants to validate against the system to verify its integrity. This can be done by comparing the hash of the original data against the hash stored in the blockchain network. As a hash of personal data is personal data by itself, access to it must be authorized previously. Once approved, the integrity blockchain performs the integrity control, comparing the stored hash with the one presented by the user and with no need to consult the off-chain.

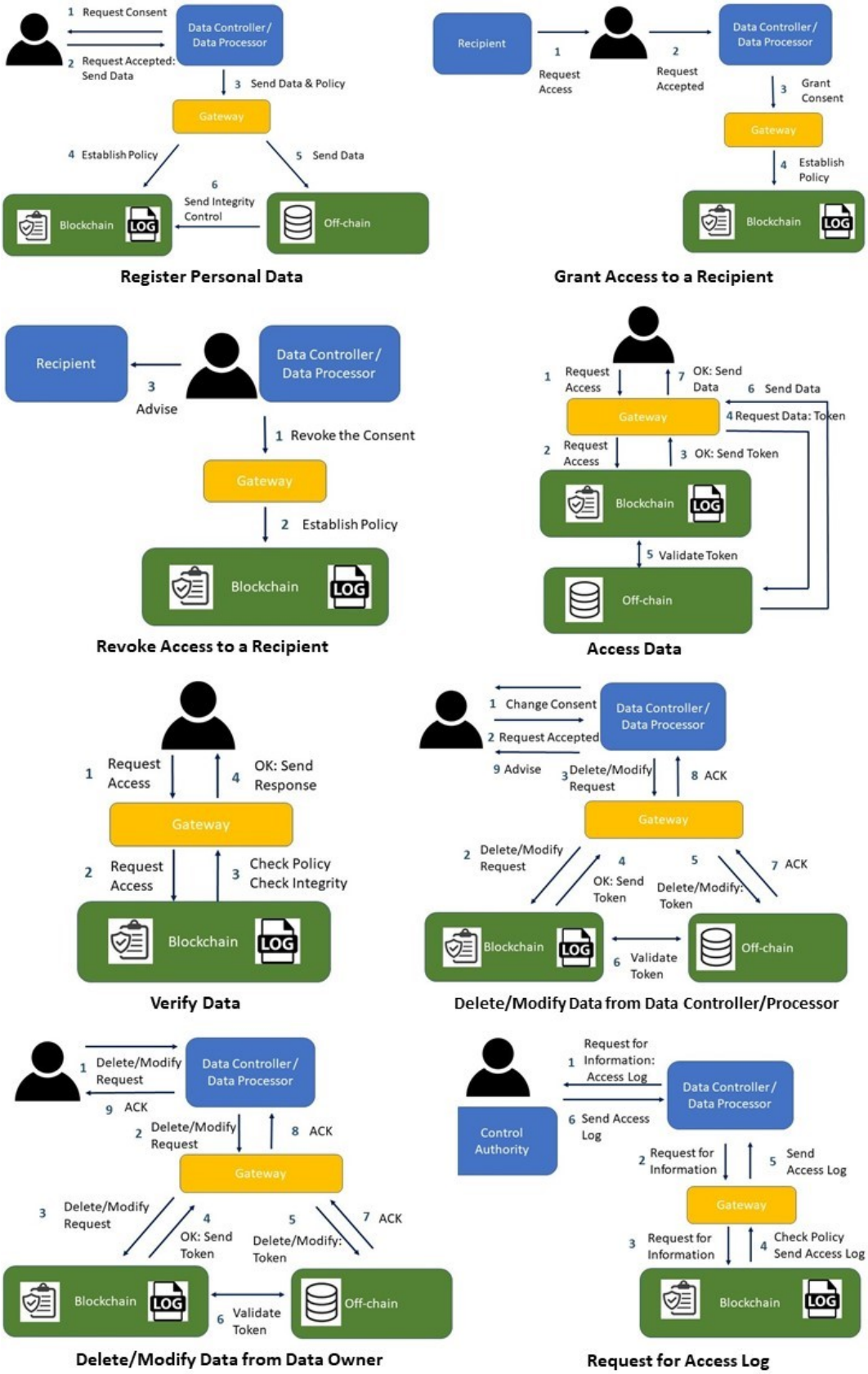

Figure 2: Use case diagrams

6. *Delete/Modify Data* The DO, DC, or DP may need to modify or delete data. First, they must inform or request authorization among themselves and send the request to the Gateway. Once again, the Gateway verifies the policy against the access control network, blockchain, using a token service. If

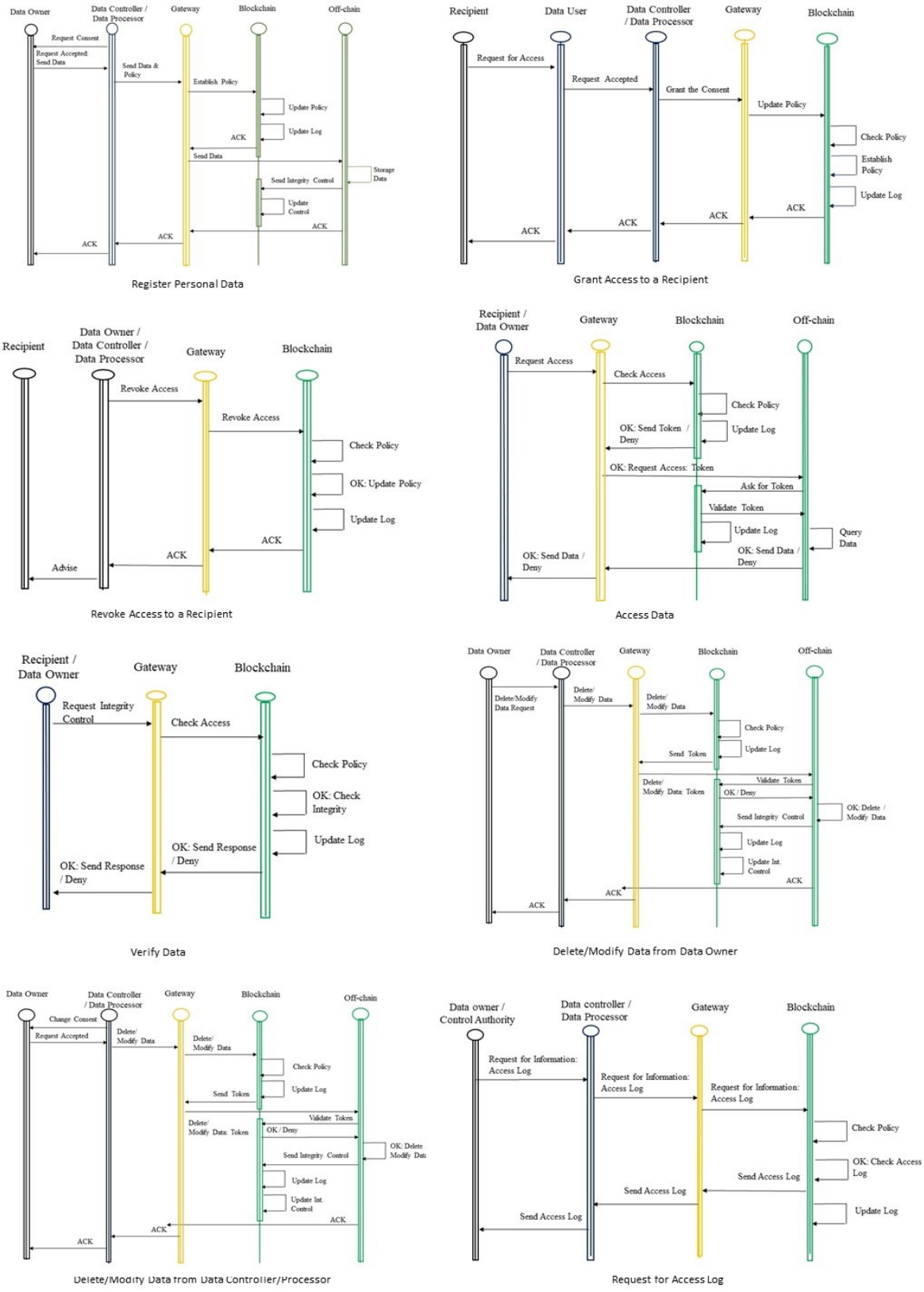

Figure 3: Sequence diagrams

the token is validated, the off-chain network changes and adjusts the integrity check hash.

7. *Request for Access Log* The data owner or authority can request data access history. After verifying the authorization against the access policy network, the blockchain network retrieves this data.

5.1.1 Some implementation considerations

In addition to the design constructs just described, it is also important to evaluate which network models will be adopted and other implementation considerations.

As previously discussed in Section 3, the creation of a permissioned blockchain network is recommended to handle personal data. In turn, using an off-chain with commit-chain architecture network is recommended. Still, being a centralized and non-distributed system implies designing solutions to ensure system availability.

As already mentioned, a hash value of data could be stored in the blockchain as an integrity control mechanism. Thus, the proposed validation compares the hash of the data stored in the blockchain with the data stored in the off-chain. To do this safely, namely avoiding the hash reversibility problem, it is necessary to consider the recommendations designed to increase the entropy level of possible hashes. That implies, for instance, introducing random elements and taking into account the identifiers linked to a hash. Related to this, it is essential to choose the certificates carefully and public keys to identify the system's actors, for example, analyzing if it is applicable to use a new pair of keys for each transaction to avoid traceability.

5.2 A methodology for security and privacy threat modelling

To validate the defined architecture, we propose a security and privacy threat analysis methodology to apply to specific use cases that use this high-level architecture.

There exist several methodological proposals for the development of threat models. We shall follow the one proposed by OWASP [25], which consists of a process with six steps: identify security objectives, profile and decompose the application, identify threats, identify vulnerabilities and rank the threats.

First, it is necessary to identify the *security and privacy properties* that should be guaranteed the system satisfies. This could be done using the CI4AM security mechanism [25], which group security objectives into categories of confidentiality, integrity, availability, authentication, authorization, auditing, and management.

Next is necessary to *profile* to understand better the use of the application, the technology to be used, system actors, etc., and *decompose* the application for what is suggested to use the Threat Modelling tool of Microsoft [50].

The next step is the *threat analysis* that consists of determining, characterizing, and classifying threats that may affect the analyzed application from the point of view of security and privacy. It is possible to identify the system's vulnerabilities from the identified threats, which can be exploited by the threats analyzed. Finally, the danger can be ranked to help define what actions to take.

To achieve both a security and a privacy analysis, we shall make combined use of the well-known Microsoft's STRIDE methodology [50] and the risk methodology defined by CNIL for privacy risk management [26]. STRIDE considers threats strictly from a security point of view, categorizing them into spoofing, tampering, repudiation, information disclosure, denial of service, and elevation of privilege. These threats are related to authentication, integrity, non-repudiation, confidentiality, availability, and authorization security properties.

CNIL considers threat groups related to processes, like unavailability of legal procedures and processing changes. It also includes threat groups related to personal data, like illegitimate access to personal data (confidentiality), unwanted changes in personal data (integrity), and disappearance of personal data (availability).

Some of the threats related to *unavailability of legal processes* that we consider affect GDPR compliance are legal problems derived from: i) *lack of responsibility for the treatment of personal data*; ii) *lack of processes to ensure the veracity or the consent of the treatment*, iii) *lack of processes for the correct treatment of data*, iv) *lack of processes to delete data*, and v) *failures in the information provided to the user*. Related to change in processing threats, we consider legal problems derived from vi) *failures in the processes to ensure the integrity or the purpose of the treatment*, vii) *failures in the processes to delete data*, and viii) *deficiency of controls over the treatment*.

6 High level design of a privacy-aware digital academic certificates system

The IT services unit of the Universidad de la República, Uruguay, is in the process of studying alternatives for implementing and deploying a system that provides functionalities to handle, store and validate university certificates, like degree diplomas, using blockchain technology. The main objective of the solution is that (graduated) students can obtain their digital titles or certificates and deliver them to third parties (for example, potential employers), which could validate the authenticity and legitimacy of those certificates using that same system. The use of blockchain technology is considered a possible solution to validate the integrity of certificates, supported by the immutability property of the system.

6.1 Existing solutions

Several solutions have been proposed to deal with this problem. One of the systems analyzed by SeCIU is the *Brazilian RAP System*. In Brazil, a regulation sets two years from March 2019 for universities to implement a diploma verification system. Universities are working to generate digital diplomas using signed XML documents and blockchain. RAP system [27] proposes to use the public networks Ethereum and Bitcoin to provide existence and integrity controls and off-chain networks to preserve digital documents. Additionally, an authentication module performs data validation with the blockchain network and retrieves the information from the off-chain network. Another analyzed system is the *Argentine BFA System*. This system was developed by the University of Córdoba and based on the BFA (Blockchain Federal Argentina) [28]. BFA is based on Ethereum technology and works under a permission blockchain model. Once the University records are validated, a digital document is saved and digitally signed by the teacher and stored on the blockchain. In BFA, no documents or files are stored within the blockchain; only the hashes of those documents are saved.

The *Blockchain for Education* (BFE) [51] initiative consists of a network of universities that have implemented a hierarchy system to add Universities to the system and to validate the signing certificates. There is an accreditation authority responsible for authorizing the entry of other universities to the system, and certification authorities, which are the universities that belong to the system, are responsible for signing certificates and storing them in the blockchain. BFE uses a public network Ethereum so, like adding any transaction on this network, adding certificates comes at a cost, which is a disadvantage of the system. It also uses a smart contract for identity management and another smart contract to manage and store certificates in the blockchain.

EduCTX [52] is a system that proposes a global higher education credit platform based on the concept of the European Credit Transfer and Accumulation System (ECTS). It constitutes a globally trusted, decentralized higher education credit and grading system based on a permissioned blockchain network of higher education institutions (HEI). In this system, every time a student completes a course or saves an exam, his HEI will transfer the appropriate number of ECTX tokens to his blockchain address. When an organization wants to verify the student's course obligation completion, the student has to send his blockchain address and redeem the script to the verifier organization. Then the organization checks the amount of ECTX tokens against the blockchain, which represents the student's academic credit achievements.

In [17] the authors propose the use of independent private blockchain networks connected to a core blockchain to create a system integrating multiple universities. Like the one we present in this article, the system requires authentication to validate certificates and uses the core blockchain network for three services: a control access service, a data service, and a log service. The work also conducts a comparative investigation of various solutions that use blockchain to validate university certificates, analyzing, in particular, the compliance of these solutions with the GDPR.

6.2 The design of our solution

We have specified the expected behavior of the system using the GDPR-compliant constructs of the system architecture introduced in Section 4. We shall present and discuss the most relevant design decisions in what follows.

To apply the proposed high-level architecture, first, we mapped the actors of the system to the institutions and individuals of the problem's domain: i) *SeCIU* is the *Data controller* (DC), the responsible for the personal data handled by the system; ii) the *School Registry Offices* are the *Data processor* (DP), working as delegates assigned by the SeCIU; iii) the *Students* are the *Data owner* (DO), the owners of personal data; and finally iv) the *Receiver (recipient)* are the employers or institutions that want to validate a certificate.

As it was explained in Section 4, personal data, which in this case are the universities certificates, are stored in an off-chain network. The audit, access control, and verification operations are conducted through a blockchain network, which can be accessed through a Gateway by the School Registry Offices, students, and authorized third parties. A hash of each certificate is stored in the blockchain to validate the certificates. The validation system compares this hash with the certificate's hash presented by the student. As was discussed in Section 3, hash values should be considered personal data as they are pseudo-anonymized data. In this respect, the authorities of the Universidad de la República decided to formally ask the Uruguayan Data Protection Agency, the *URCDP* (*Unidad Reguladora y de Control de Datos Personales*), whether it is valid, from the legal point of view, to implement a system where the verification of a degree diploma is public. The URCDP responded that such verification requires the consent of the individual that earned the certificate and therefore cannot be provided as an open and public function (Dictum 9-2018, Expediente 2017-2-10-000394) [53]. Therefore, certificate verification should not be a public function, and access to that operation should be controlled.

Figure 4 depicts the data flow diagram (DFD) of the system, detailing the data and control flow that takes place.

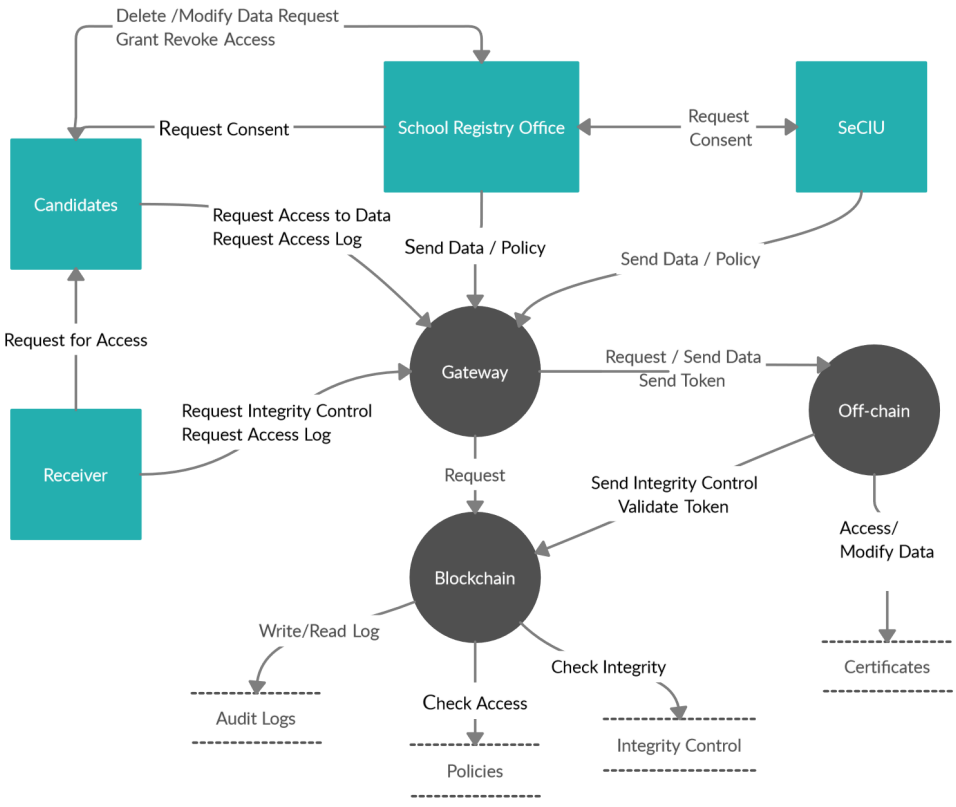

Figure 4: Data Flow Diagram of the system

We have specified the behavior of the digital certificates system using refined versions of the generic use cases introduced in the use case model discussed in Section 5.1. In [14] we present and discuss the specification of some of the processes of the system. In [54] we provide the complete specification of the system.

6.3 Security and privacy threat modeling of the system

In what follows, we describe the results obtained from applying the proposed threat analysis methodology.

We have identified the following security objectives: i) protect certificates against unauthorized access and allow verification process only to authorized persons; ii) certificates must not be modified without authorization and must be verifiable using a validation process; iii) availability of the certificate verification system and the certificates; iv) all participants, SecIU, School Registry Offices, students, and employers must be identified and authenticated; v) students must authorize companies or third parties who want their certificates to be validated, and only students can see their certificates. Additionally, the SeCIU or the School Registry Offices must authorize changes in the access policy requested by the students. Finally, no person can access the information if the system did not previously authorize it; vi) all accesses, requested permissions, and unauthorized access requests changes must be recorded; vii) the certificates are managed and stored by the SIU, and the Gateway concentrates the interaction with the blockchain and off-chain, and ultimately viii) the system processes must comply with GDPR requirements, like those related to responsibility for data processing, deletion, and traceability of personal data and integrity and the consent of the treatment.

We have used the Microsoft Threat Modeling Tool [25] to carry out the risk analysis. As a result, we identified 110 threats related to the STRIDE categories and 18 threats related to blockchain and off-chain technology issues described in Section 2 and to use cases described in [16]. Additionally, we have also detected 10 threats related to privacy issues.

Figure 5 illustrates the diagram generated by the Microsoft Threat Modeling Tool, from which security threats are identified.

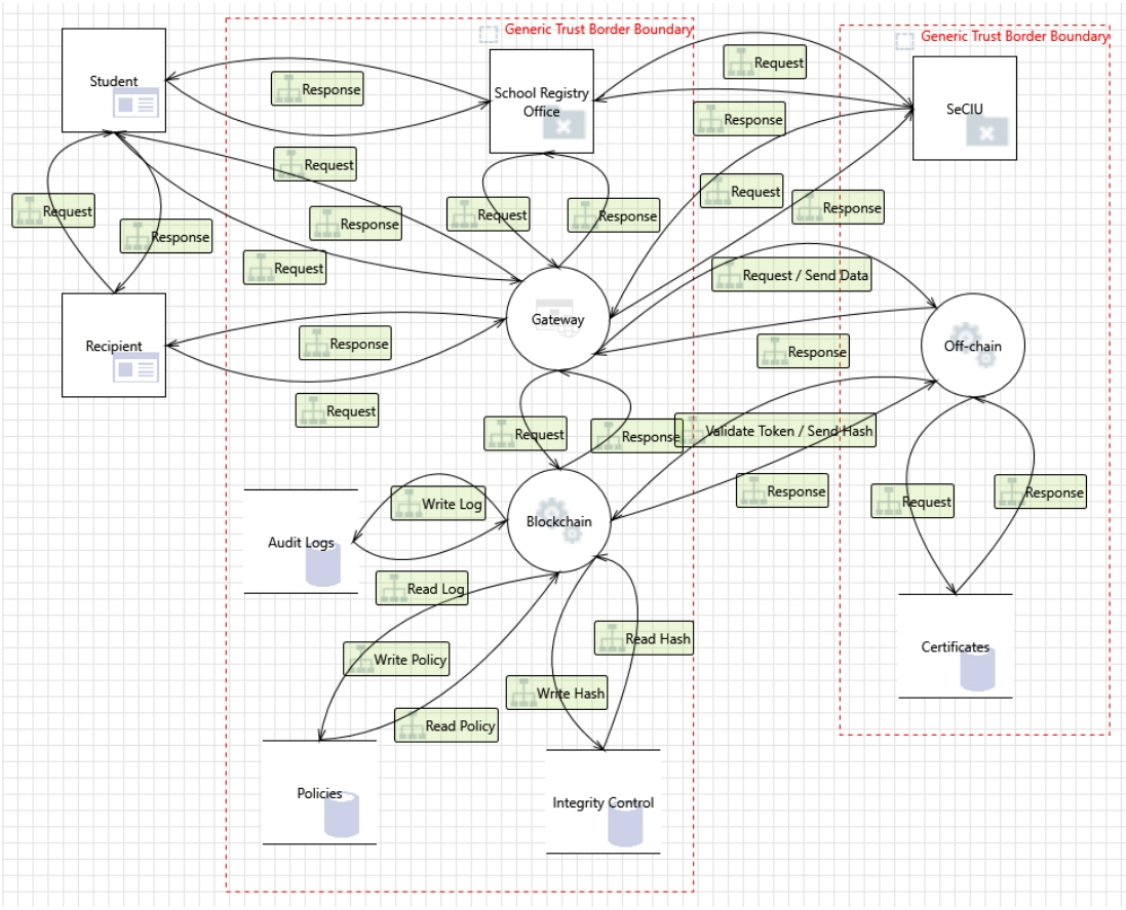

Figure 5: Threat risk analysis

6.4 Mitigation measures

In A we provide a catalog of the threats to the specified system that have been identified. First, we present those threats that can be mitigated by design, which amount to 56% of the total identified. Then, we display the threats that can be mitigated by implementing mechanisms that guarantee, among others, traffic encryption, protection of user credentials, control measures for personnel of School Registry Offices, and the use of hash functions. Out of the total threats detected, six are considered not applicable, and two are unmitigated threats that concern Denial of Service.

7 Conclusion and further work

We have identified and discussed the challenges the European regulation on data protection poses to designing and implementing software systems that manage personal data. In particular, we are interested in those systems built using blockchain technology. To analyze this, we first conducted a study of blockchain and off-chain technologies to understand the weaknesses and strengths of these solutions regarding personal data protection regulations.

The first conclusion we made is that intrinsic characteristics of blockchain make this technology incompatible with personal data protection regulations, so it is necessary to incorporate other components, such as the off-chain constructs we have presented and discussed in this work. On the other hand, blockchain technology meets by design some of the requirements of personal data protection regulations, which helps meet the needs of security by design or privacy by default. In this sense, we proposed using blockchain to perform access control, audit, and integrity control operations while personal data is stored on an off-chain network.

As a result of the analysis we have carried out, we put forward a high-level system architecture to specify the behavior and services of a system that integrates blockchain and off-chain functionalities, complying with the requirements of the personal data protection regulations. The system's two main components are a software architecture model and a use case model, which have been conceived to support the construction of systems that are personal data protection regulations compliant by design. Based on these models,

we proposed an architecture that incorporates some of the concepts presented, adding the integrity check functionality, even as a standalone functionality, and we also suggest some implementation considerations that should be taken into account since there are requirements in personal data protection regulations that are not satisfied by design. We proposed a model that embodies eight use cases, including the verification operation that does not require access to the data itself but requires authorization. This operation is essential for the completeness of the proof of concept we have carried out since, as pointed out in Section 5, the Uruguayan Data Protection Agency has indicated that access to the verification process should be controlled. According to the personal data protection regulations requirements, we analyzed the relationship between the owner, the data controller, and the data processor.

We have also put forward a methodology for performing security and privacy threat analysis of systems of that kind, combining the STRIDE methodology and the risk methodology defined by CNIL for privacy risk management.

We illustrated the use of the proposed high-level system and the threat analysis methodology on a realistic and not trivial digital certificates system. The design constructs are expressive enough to specify the required functionalities of that system. Additionally, due to the analysis, we have carried out, we also observed that about half of the threats identified are mitigated by design.

As future work, it remains to complete the implementation of the prototype to assess the adequacy of the proposed high-level system architecture. An identified challenge is the conception of the mechanisms to register and authenticate the third parties that shall interact with the digital certificates system. In particular, as SECIU would prefer not to provide and manage those mechanisms, a decentralized solution must be considered. How well one such solution would integrate with the rest of the system, preserving the privacy requirements requires further study.

Related to the threat analysis methodology, it could be interesting to define a set of generic threats on blockchain and off-chain technology and personal data protection problems, as the threats analyzed are specific to the proof of concept use case.

References

[1] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," 2009. [Online]. Available: http://www.bitcoin.org/bitcoin.pdf

[2] J. Eberhardt and S. Tai, "On or off the blockchain? insights on off-chaining computation and data," in *Service-Oriented and Cloud Computing*, F. De Paoli, S. Schulte, and E. Broch Johnsen, Eds. Cham: Springer International Publishing, 2017, pp. 3–15.

[3] European Parliament and of the council, "Regulation (eu) 2016/679," *Official Journal of the European Union.* , 2016, uRL https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[4] Article 29 Working Party, "Working group on data protection - opinion 05/2014 on anonymization techniques," 2014.

[5] T. Lyons, L. Courcelas, and K. Timsit, "Blockchain and the gdpr. a thematic report prepared by the european union blockchain observatory and forumtitle," *European Union blockchain observatory and forum*, 2018.

[6] L.-D. Ibanez, K. O'Hara, and E. Simperl, "On blockchains and the general data protection regulation," Project Report, July 2018. [Online]. Available: https://eprints.soton.ac.uk/422879/

[7] G. P. Freund, P. B. Fagundes, and D. Macedo, "An analysis of blockchain and gdpr under the data lifecycle perspective," *Mobile Networks and Applications*, pp. 1–11, 2020.

[8] U. Tatar, Y. Gokce, and B. Nussbaum, "Law versus technology: Blockchain, gdpr, and tough tradeoffs," *Computer Law & Security Review*, vol. 38, p. 105454, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0267364920300595

[9] C. Li, B. Palanisamy, and R. Xu, "Scalable and privacy-preserving design of on/off-chain smart contracts," in *35th IEEE International Conference on Data Engineering Workshops, ICDE Workshops 2019, Macao, China, April 8-12, 2019*. IEEE, 2019, pp. 7–12. [Online]. Available: https://doi.org/10.1109/ICDEW.2019.00-43

[10] N. B. Truong, K. Sun, G. M. Lee, and Y. Guo, "Gdpr-compliant personal data management: A blockchain-based solution," *CoRR*, vol. abs/1904.03038, 2019. [Online]. Available: http://arxiv.org/abs/1904.03038

[11] C. Molina-Jiménez, I. Sfyrakis, E. Solaiman, I. Ng, M. W. Wong, A. Chun, and J. Crowcroft, "Implementation of smart contracts using hybrid architectures with on and off-blockchain components," in *8th IEEE International Symposium on Cloud and Service Computing, SC2 2018, Paris, France, November 18-21, 2018*. IEEE, 2018, pp. 83–90. [Online]. Available: https://doi.org/10.1109/SC2.2018.00018

[12] F. Molina, G. Betarte, and C. Luna, "Design principles for constructing GDPR-compliant blockchain solutions," in *2021 IEEE/ACM 4th International Workshop on Emerging Trends in Software Engineering for Blockchain (WETSEB), ICSE 2021*, 2021.

[13] ——, "On the compliance of blockchain technology with data protection regulations," in *2021 IEEE URUCON*, 2021, pp. 556–560. [Online]. Available: https://ieeexplore.ieee.org/document/9647141

[14] ——, "Privacy aware blockchain solutions: Design and threat analysis," in *XXIV Ibero-American Conference on Software Engineering, CibSE 2021*, 2021.

[15] F. Molina, "Constructing privacy aware blockchain solutions: design guidelines and threat analysis techniques," Master's thesis, 2021. [Online]. Available: https://hdl.handle.net/20.500.12008/30599

[16] R. H. Saye, "Potential of blockchain technology to solve fake diploma problem," 2019, URL https://jyx.jyu.fi/bitstream/handle/123456789/64817/1/URN:NBN:fi:jyu-201906253406.pdf.

[17] C. Delgado-von Eitzen, L. Anido-Rifon, and M. J. Fernandez-Iglesias, "Blockchain for the scalable issuance and verification of private academic information," in *2021 International Conference on Advanced Learning Technologies (ICALT)*, 2021, pp. 436–438.

[18] "Certificate fraud news," 2021, URL https://www.montevideo.com.uy/Noticias/Siete-condenados-por-estafa-en-caso-de-venta-de-titulos-universitarios-falsos-en-Colonia-uc792929.

[19] L. Gudgeon, P. Moreno-Sanchez, S. Roos, P. McCorry, and A. Gervais, "Sok: Off the chain transactions," *IACR Cryptol. ePrint Arch.*, vol. 2019, p. 360, 2019.

[20] C. Reitwiener, "zksnarks in a nutshell," 2016, URL https://blog.ethereum.org/2016/12/05/zksnarks-in-a-nutshell/.

[21] J. Eberhardt and S. Tai, "Zokrates - scalable privacy-preserving off-chain computations," in *2018 IEEE iThings and IEEE GreenCom and IEEE CPSCom and IEEE SmartData*, 2018, pp. 1084–1091.

[22] J. Eberhardt and J. Heiss, "Off-chaining models and approaches to off-chain computations," in *Proceedings of the 2nd Workshop on Scalable and Resilient Infrastructures for Distributed Ledgers*, ser. SERIAL18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 7–12.

[23] C. Daudén-Esmel, J. CastellÃ -Roca, A. Viejo, and J. Domingo-Ferrer, "Lightweight blockchain-based platform for gdpr-compliant personal data management," in *2021 IEEE 5th International Conference on Cryptography, Security and Privacy (CSP)*, 2021, pp. 68–73.

[24] M. Poelman and S. Iqbal, "Investigating the compliance of the gdpr: Processing personal data on a blockchain," in *2021 IEEE 5th International Conference on Cryptography, Security and Privacy (CSP)*, 2021, pp. 38–44.

[25] OWASP, "Owasp threat modeling," 2020, URL https://owasp.org/www-pdf-archive/AdvancedThreatModeling.pdf.

[26] CNIIL, "Methodology for privacy risk management," 2012, URL https://www.pdpjournals.com/docs/887983.pdf.

[27] R. C. et al., "Uso não financeiro de blockchain: Um estudo de casao sobre o registro, autenticacao e preservacao de documentos digitais academicos," in *Anais do I Workshop em Blockchain: Teoria, Tecnologias e Aplicacoes*. Porto Alegre, RS, Brasil: SBC, 2018. [Online]. Available: https://sol.sbc.org.br/index.php/wblockchain/article/view/2356

[28] BFA.ar, "Blockchain federal argentina," 2020, URL https://bfa.ar/.

[29] Hyperledger.org, "Hyperledger," 2020, URL https://www.hyperledger.org/.

[30] IPFS, "IPFS," 2020, URL https://ipfs.io/.

[31] SWARM, "SWARM," 2020, URL https://www.swarm.fund/.

[32] P. Santhana and A. Biswa, "Blockchain risk management risk functions need to play an active role in shaping blockchain strategy," 2017, URL https://www2.deloitte.com/content/dam/Deloitte/us/Documents/financial-services/us-fsi-blockchain-risk-management.pdf.

[33] U. W. Chohan, "The double spending problem and cryptocurrencies," *Information Systems & Economics eJournal.* , 2017, uRL: https://ssrn.com/abstract=3090174.

[34] E. G. S. Ittay Eyal, "Majority is not enough: Bitcoin mining is vulnerable," 2018, URL https://doi.org/10.1145/3212998.

[35] B. Wiki, "Bitcoin weaknesses," 2020, URL https://en.bitcoin.it/wiki/Weaknesses.

[36] T. C. Nicola Atzei, Massimo Bartoletti, "A survey of attacks on ethereum smart contracts," *Universita degli Studi di Cagliari, Cagliari, Italy*, 2016.

[37] "Ethereum wiki," 2021, URL https://eth.wiki/en/howto/smart-contract-safety.

[38] "Ethereum improvement proposals (eips)," 2021, URL https://eips.ethereum.org/.

[39] Etherscan, "The ethereum blockchain explorer," 2021, URL https://etherscan.io/.

[40] European Parliamentary Research Service, "Blockchain and the general data protection regulation. can distributed ledgers be squared with european data protection law?" 2019.

[41] Commission Nationale de l'Informatique et des Libertés, "Blockchain and the gdpr: Solutions for a responsible use of the blockchain in the context of personal data," 2018, URL https://www.cnil.fr/en/blockchain-and-gdpr-solutions-responsible-use-blockchain-context-personal-data.

[42] Enigma, "Enigma project," 2020, URL https://medium.com/@EnigmaMPC.

[43] F. Zhang, W. He, R. Cheng, J. Kos, N. Hynes, N. M. Johnson, A. Juels, A. Miller, and D. Song, "The ekiden platform for confidentiality-preserving, trustworthy, and performant smart contracts," *IEEE Secur. Priv.*, vol. 18, no. 3, pp. 17–27, 2020. [Online]. Available: https://doi.org/10.1109/MSEC.2020.2976984

[44] "Secret network," 2021, URL https://scrt.network.

[45] "Hyperledger avalon," 2021, URL https://www.hyperledger.org/use/avalon.

[46] "Digital rights ireland ltd, cases c 293/12 y c 594/1," 2014, URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62012CJ0293.

[47] Hyperledger.org, "Hyperleger indy," 2021, URL https://www.hyperledger.org/use/hyperledger-indy.

[48] E. Felten, "Does hashing make data anonymous?" 2012, URL https://www.ftc.gov/news-events/blogs/techftc/2012/04/does-hashing-make-data-anonymous.

[49] Agencia Española de Protección de Datos, "Introduction to the hash function as personal data pseudonymisation tecnhnique," *Agencia Española de Protección de Datos*, 2019.

[50] STRIDE, "Stride methodology," 2020, URL https://docs.microsoft.com/en-us/archive/msdn-magazine/2006/november/uncover-security-design-flaws-using-the-stride-approach.

[51] W. Prinz, S. Kolvenbach, and R. Ruland, "Blockchain for education: Lifelong learning passport," *ERCIM News*, vol. 2020, no. 120, 2020. [Online]. Available: https://dl.eusset.eu/handle/20.500.12015/3163

[52] EDUCTX.org, "EDUCTX – blockchain Lab:UM," 2020, URL http://www.eductx.org/.

[53] URCDP, "Dictum 9-2018, exp. 2017-2-10-000394," 2018.

[54] F. Molina, G. Betarte, and C. Luna, "A blockchain based and GDPR-compliant design of a system for digital education certificates," 2020, URL https://arxiv.org/abs/2010.12980.

A Threat analysis of the academic certificates system

The following tables summarize the detected threats, classified according to whether the threats can be mitigated by design, by implementation, are not applicable, or are unmitigated threats. Details, as well as possible mitigations, are discussed in detail in [15].

A.1 Threats mitigated by design

Threats mitigated by design		
Category	**Title**	**Description**
CNIL - Change in processing	Failures in the processes to ensure the veracity of the treatment	Changes in the process affect the veracity of the information processed, or the data is not processed according to the defined purpose
CNIL - Change in processing	Deficiency of controls over the treatment	The treatment of the information is not as expected due to lack of controls over it.
CNIL - Change in processing	Failures in the processes to delete data	Claims because the information is not deleted correctly, following the agreed treatment of the information
CNIL - Unavailability of Legal Processes	Lack of responsibility for the treatment of personal data	No one assumes responsibility for a claim by a student for the treatment of their information
CNIL - Unavailability of Legal Processes	Lack of processes to ensure the veracity of the treatment	In a claim, the veracity of the treatment cannot be demonstrated
CNIL - Unavailability of Legal Processes	Failures in the information provided to the user	Incorrect information is provided to the student, such as access information to their data
Elevation of Privileges	An adversary may bypass critical steps or perform actions on behalf of other users (victims) due to improper validation logic	Failure to restrict the privileges and access rights to the application to individuals who require the privileges or access rights may result into the unauthorized use of data due to inappropriate rights settings and validation.
Elevation of Privileges	An adversary can gain unauthorized access to the database due to loose authorization rules	Database access should be configured with roles and privilege based on least privilege and need to know principle.
Elevation of Privileges	An adversary can gain unauthorized access to the database due to lack of network access protection	If there is no restriction at network or host firewall level to access the database, then anyone can attempt to connect to the database from an unauthorized location
Information Disclosure	An adversary can gain access to certain pages or the site as a whole.	Robots.txt is often found in the site's root directory and exists to regulate the bots that crawl the site.
Information Disclosure	An adversary can gain access to sensitive data by performing SQL injection	SQL injection is an attack in which malicious code is inserted into strings that are later passed to an instance of SQL Server for parsing and execution.
Information Disclosure	An adversary can gain access to sensitive PII or HBI data in database	Additional controls like Transparent Data Encryption, Column Level Encryption, EKM, etc., provide additional protection mechanisms to high-value PII or HBI data.
Spoofing	An adversary may spoof SeCIU and gain access to Web Application	If proper authentication is not in place, an adversary can spoof a source process or external entity and gain unauthorized access to the Web Application
Spoofing	An adversary may spoof School Registry Offices and gain access to Web Application	If proper authentication is not in place, an adversary can spoof a source process or external entity and gain unauthorized access to the Web Application
Spoofing	An adversary may spoof Blockchain and gain access to Web Application	If proper authentication is not in place, an adversary can spoof a source process or external entity and gain unauthorized access to the Web Application
Spoofing	An adversary can create a fake website and launch phishing attacks	Phishing is attempted to obtain sensitive information such as usernames, passwords, and credit card details, often for malicious reasons, by masquerading as a Web Server which is a trustworthy entity in electronic communication
Spoofing	An adversary may spoof Browser and gain access to Web Application	If proper authentication is not in place, an adversary can spoof a source process or external entity and gain unauthorized access to the Web Application
Spoofing	An adversary may spoof Off-chain and gain access to Web Application	If proper authentication is not in place, an adversary can spoof a source process or external entity and gain unauthorized access to the Web Application
Tampering	An adversary can gain access to sensitive data by performing SQL injection through Web App	SQL injection is an attack in which malicious code is inserted into strings that are later passed to an instance of SQL Server for parsing and execution.
Tampering	An adversary can gain access to sensitive data stored in Web App's config files	An adversary can gain access to the config files. and if sensitive data is stored in it, it would be compromised.
Tampering	An adversary may leverage the lack of monitoring systems and trigger anomalous traffic to the database	An adversary may leverage the lack of intrusion detection and prevention of abnormal database activities and trigger anomalous traffic to the database
Tampering	An adversary can tamper critical database and deny the action	An adversary can tamper essential securable of database and deny the action
Tampering	Fraudulent activity[16]	Student adds the certificate to his/her CV although (s)he did not attend or complete the degree there (Grolleau et al., 2008).
Tampering	Fraudulent Activity 2 [16]	Student makes a counterfeit paper copy of the original paper copy (Grolleau et al., 2008) by using similar paper and special security features used in the original paper certificate (Sharadeshe jal sonoder, 2015, Lancaster, 2017).
Tampering	Fraudulent Activity 5 [16]	Student hacks the university grading system and changes grades

Tampering	Fraudulent Activity 10 [16]	Corrupt officials change the students academic information like enrolment date, course completion date, and grade in study data management system to prove the validity of the counterfeit certificate (Ekushey Television ETV, 2014; Mir-Jabbar, 2017).
Tampering	Blockchain Security Issues - Immutability	The data stored in the Blockchain cannot be modified if it needs to be rectified.
Tampering	Blockchain Security Issues - Unpredictable state	When a user sends a transaction to the network to invoke a contract, he cannot be sure of the contract's state when the transaction is executed.
Tampering	Blockchain Security Issues - Transaction security issues	Risk of double-spending - Process something more than once.
Tampering	Off-chain Blockchain Security Issues - Integrity	The information stored in the off-chain could be altered without being detected.
Tampering	Off-chain Blockchain Security Issues - Traceability	In off-chain audit is not resolved by design. Therefore, accesses or modifications could be made without leaving a trace.

Table 2: Threats mitigated by design

A.2 Threats mitigated by implementation

Threats mitigated by implementation		
Category	**Title**	**Description**
CNIL - Unavailability of Legal Processes	Lack of processes to ensure the consent of the treatment	In a claim, it cannot be demonstrated that permission is given to the processing of the information
Information Disclosure	An adversary can reverse weakly encrypted, or hashed content	An adversary can reverse weakly encrypted or hashed content
Information Disclosure	An adversary can gain access to sensitive information through error messages	An adversary can gain access to sensitive data such as the following, through verbose error messages - Server names - Connection strings - Usernames - Passwords - SQL procedures - Details of dynamic SQL failures - Stack trace and lines of code - Variables stored in memory - Drive and folder locations - Application install points - Host configuration settings - Other internal application details
Information Disclosure	An adversary may gain access to sensitive data from log files	An adversary may gain access to sensitive data from log files
Information Disclosure	An adversary can reverse weakly encrypted, or hashed content	An adversary can reverse weakly encrypted or hashed content
Information Disclosure	An adversary may gain access to sensitive data from uncleared browser cache	An adversary may gain access to sensitive data from uncleared browser cache
Information Disclosure	An adversary can gain access to sensitive data by sniffing traffic to Web Application	An adversary may conduct man in the middle attack and downgrade TLS connection to clear text protocol, or force browser communication to pass through a proxy server that he controls. This may happen because the application may use mixed content or HTTP Strict Transport Security policy is not ensured.
Information Disclosure	An adversary may gain access to unmasked sensitive data such as certificates	An adversary may gain access to unmasked sensitive data such as certificates.
Information Disclosure	An adversary may gain access to unmasked sensitive data such as credit card numbers	An adversary may gain access to unmasked sensitive data such as credit card numbers because sensitive data is displayed on the user screen with no mask.
Information Disclosure	An adversary can gain access to sensitive data by performing SQL injection	SQL injection is an attack in which malicious code is inserted into strings that are later passed to an instance of SQL Server for parsing and execution.
Spoofing	An adversary can create a fake website and launch phishing attacks	Phishing is attempted to obtain sensitive information such as usernames, passwords, and credit card details (and sometimes, indirectly, money), often for malicious reasons, by masquerading as a Web Server which is a trustworthy entity in electronic communication
Spoofing	An adversary can steal sensitive data like user credentials	Attackers can exploit weaknesses in the system to steal user credentials.
Spoofing	An adversary can spoof the target web application due to insecure TLS certificate configuration	Ensure that TLS certificate parameters are configured with correct values
Spoofing	Attackers can steal user session cookies due to insecure cookie attributes	The session cookies are the identifier by which the server knows the current user's identity for each incoming request. If the attacker can steal the user token, he would access all user data and perform all actions on behalf of the user.
Spoofing	An adversary can access a user's session due to insecure coding practices.	The session cookies are the identifier by which the server knows the current user's identity for each incoming request. If the attacker can steal the user token, he would access all user data and perform all actions on behalf of the user.
Spoofing	An adversary can access a user's session due to improper logout and timeout.	The session cookies are the identifier by which the server knows the current user's identity for each incoming request. If the attacker can steal the user token, he would access all user data and perform all actions on behalf of the user.

Spoofing	Blockchain Security Issues - Key management risk.	Blockchain is susceptible to theft of private keys and the control of the assets associated with external addresses being taken away.
Tampering	An adversary can access sensitive data stored in Web App's config files.	An adversary can gain access to the config files. and if sensitive data is stored in it, it would be compromised.
Tampering	An attacker steals messages off the network and replays them to steal a user's session	An attacker steals messages off the network and replays them to steal a user's session
Tampering	An adversary can deface the target web application by injecting malicious code or uploading harmful files	Website defacement is an attack on a website where the attacker changes the site's visual appearance or a web page.

Table 3: Threats mitigated by implementation

A.3 Not applicable threats

Threats that does not apply		
Category	**Title**	**Description**
Tampering	Fraudulent Activity 3 [16]	Student buys certificate from non-accredited university or diploma mill.
Tampering	Fraudulent Activity 4 [16]	Student uses misleading translated copy of the actual document.
Tampering	Fraudulent Activity 6 [16]	Student uses work certificate or life experience and then converts that into academic credit with the support of corrupt officials in an accredited university degree.
Tampering	Fraudulent Activity 7 [16]	Corrupt teacher takes an unofficial fee to assure the passing grade without submitting the assignments or required studies done.
Tampering	Fraudulent Activity 8 [16]	Teachers are sometimes biased and grade students higher than their performance, such as on exam paper.
Tampering	Fraudulent Activity 9 [16]	Fraud syndicate has links with corrupt officials to store fake certificate data in the university database.

Table 4: Non applicable threats

A.4 Unmitigated Threats

Threats not mitigated		
Category	**Title**	**Description**
Denial of Service	An adversary can act on behalf of another user due to lack of controls against cross-domain requests	Failure to restrict requests originating from third-party domains may result in unauthorized actions or access of data.
Denial of service	Off-chain Blockchain Security Issues - Availability	In an off-chain commit model, availability is not guaranteed.

Table 5: Unmitigated Threats