2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)

When Blockchain Meets the Right to be Forgotten: Technology Versus Law in the Healthcare Industry

1st Aurelie Bayle
Université de Montpellie
Montpellier, France
given_name.family_name@etu.umontpellier.fr

2nd Mirko Koscina
École normale supérieure
Paris, France
given_name.family_name@ens.fr

3rd David Manset
Almerys
Clermont-Ferrand, France
given_name.family_name@g2s-group.com

4th Octavio Perez-Kempner
Almerys
Clermont-Ferrand, France
given_name.family_name@almerys.com

***Abstract*—In this work we propose a new blockchain model that ensure the GDPR compliance by handling references to the sensitive data and using metadata instead of manipulate private data directly within the blockchain. We accomplish this by defining a modular architecture that relies on strong cryptographic assumptions that provide the means to guarantee that the right to be forgotten is being well enforced.**

***Index Terms*—GDPR, blockchain, permissioned ledger, right to be forgotten, MyHealthMyData**

I. Introduction

During the last years, the popularity of the blockchain and cryptocurrencies has been increasing and has reached great notoriety not only in scientific and IT journals but also within the public sphere. Although there are many kinds of cryptocurrencies in circulation nowadays, the most popular is Bitcoin. Since Bitcoin began attracting the attention of the financial, security and IT communities, several other blockchain implementations have been appearing. One of these is Ethereum, which is a programmable blockchain [9]. Rather than the pre-defined operations in bitcoins, Ethereum allows users to create their own operations, serving as a platform for many different applications based on blockchain like cryptocurrencies, smart contracts, and decentralized file storage among others. This feature is possible because the Ethereum Virtual Machine (EVM) is a Turing Complete Machine. The EVM let the developers create their own applications giving them the freedom to design their own implementations for specific services. Although the blockchain in Ethereum is similar to Bitcoin, they have some differences. It is worth to mention that unlike Bitcoin, Ethereum's blocks contain a copy of the transaction list and the last EVM state. This makes possible to synchronize the execution of smart contracts among the network nodes, which is vital in Ethereum's case since every smart contract has to run the same at each node. Additionally, even though both use a Proof-of-Work to select the new block to append to the chain, Bitcoin is based on CPU consumption and Ethereum on memory.

On the other hand, The Linux Foundation has taken a different approach about blockchain. They have proposed a different a blockchain architecture based on flexible framework capable of developing networks tailored according to new business models. This framework is called Hyperledger and makes possible to develop new services and applications based on a permissioned ledger. The Hyperledger project consists of five blockchain frameworks: Fabric, Iroha, Sawtooth, Burrow, and Indy. Fabric, which is the most popular implementation, is a modular blockchain framework that gives the flexibility to change different components by plug-and-play. Moreover, their blockchain replication process between the nodes is cost-efficient and is capable of processing about 3.500 tps [1], thanks to their consensus algorithm based on Practical Byzantine Fault Tolerant (PBFT) algorithm [2]. This makes Hyperledger Fabric one of the best options for customizing a blockchain implementation. Based on this new business-oriented blockchain networks, the use of these technologies is turning from financial transactions to business process management. Any business-oriented solution involves management of confidential or private information. With the arrival of the new European General Data Protection Regulation (GDPR), the blockchain openness and the immutability is turning into a problem from the new regulation point of view.

Other models have been proposed based on the fact that encrypted data cannot be retrieved within a reasonable period of time without proper authorization and therefore concluding that it can be stored directly in the blockchain. This notion is not robust enough to guaranty the GDPR compliance for the right to be forgotten in systems where the data may be used for more extended periods of time. Stronger assumptions have to be made.

Our contribution is the proposal of a modular system where data providers and data consumers can interact with each other by using a blockchain to keep track of every interaction, in order to enforce the compliance of the GDPR through smart contracts, but without storing any sensitive data within the blockchain. Furthermore, each involved entity may assume different roles at the same time, and due to the nature that smart contracts are handled, greater flexibility can be achieved without compromising the sensitive information. Our proposal

978-1-5386-7325-6/18/$31.00 ©2018 IEEE
DOI 10.1109/WI.2018.00133

is contextualized for the Healthcare Industry and this work represent an abstraction of the system already implemented in MyHealthMyData project by the authors of this paper.

II. Personal data in the blockchain and the right to be forgotten

Blockchain and the European Data Protection Regulation, which came into force in May 2018, are currently two of the new key topics, always rising the same question about the Regulations application to the technology. More precisely, GDPR and blockchain are often mentioned with a potential clash between distributed ledgers and some principles or rights conferred by the Regulation. The most popular and debated of them presented as the biggest challenge for blockchains implementations in the Regulation scope may be the right to be forgotten. Effectively, the blockchain immutability allows considering that by design, anything cannot be deleted from the ledger. As much as analysis can be done, the first question to ask is how the blockchain would trigger the GDPR application and all the ensuing consequences and requirements.

The broad GDPR scope requires processing of personal data in context of the activities of an establishment of a controller or a processor in the EU. This applies regardless of whether the processing takes place in the EU or not. It is also under the GDPR regulation, the processing of personal data of the data subjects who are in the EU by a data controller or data processor not established in the EU.

Considering the territorial scope, blockchain does not admit any borders. The ledgers participants are located anywhere in the world and including the EU area. On the other hand, the material scope is vaster than countered, insofar as data processing and personal data have large meaning. Indeed, data processing means any operation or set of operations which is performed on personal data or sets of personal data, whether or not by automated means, such as collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction, while personal data could gather any information relating to an identified or identifiable natural person [1].

In a blockchain system, transactions are initiated with a combination of private and public keys and with the last one being seen on the public ledger, this can lead to identifying the participant if that public key is used several times. Thus, a public key can be considered as personal data, by analogy with a decision of the European Court of Justice about the IP address [2]. Also, personal data may be included in a transaction with a hash, considered as pseudonymized data, and triggering GDPRs application too. Theoretically, the blockchain ecosystem should be under the scope of the GDPR, but consequently, it raises the question to know who has to be compliant.

The Regulation determines four different roles according to the text: the data controllers and the data processors, being responsible and having to demonstrate their compliance with; the data subjects concerned by the processing; and the third parties, authorized to process personal data under the authority of the controller or processor. Therefore, to comply with GDPR, data controllers and processors must be clearly identified, and here is the challenge. In a public blockchain configuration, based on a decentralized architecture where all the peer-to-peer network can add transactions without any control or authorization of a central authority, everyone could be considered as a controller because of his action, and at the same time as a processor because of the copy held in the computer. The situation seems easier in case of private schemes, clearly identifying an administrator. The blockchain is a new structure and architecture, not anticipated by the classic scheme considered in the GDPR. In the fiat world, there is always an identified data controller, also considered as a central authority, but within a blockchain public scheme, the absence of central authority is fundamental.

In a nutshell, the blockchain technology as a protocol cannot be well qualified as a data controller or processor. The responsibility is transferred to the people orbiting around the blockchain, considered as third party. Any actor (developer, miner, or simple reader), considered as a network third party, will be in charge of the compliance with data protection laws. Notably, it can concern all the exchanges proposing wallets (and who must comply with the recent KYC regulations [5]), all the project managers creating or using blockchain as a service with different use cases. To conclude, each blockchain or project involving that technology must be precisely analyzed to identify the obligations imposed, as well as the respect of data subjects rights. In the present case, if the blockchains third services are bound by Regulation, the likelihood of each right implementation in the protocol must be analyzed, with a particular focus on the right to be forgotten.

GDPRs primary goal is to give back power to the data subjects on their own data, but this is posing a massive challenge for blockchain projects and implementations. Although to compare with most rights, the right to be forgotten is not absolute, and many exceptions are listed in the third point of the Article 17 of the GDPR : to exercise the right of freedom of expression and information, for compliance with a legal obligation, for reasons of public interest in the area of public health, for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, or for the establishment, exercise or defence of legal claims.

On a legal point of view, the key feature of the blockchain technology seems to conflict with the right to be forgotten and requires a clear position of the Article 29 Working group (G29), or from the national data protection authorities with a common position about blockchain implementations facing the Regulation. However, in the meantime, it does not imply that

[1] Article 4 - REGULATION (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC - General Data Protection Regulation

[2] Judgment of 19 October 2016, *P. Breyer v. Bundesrepublik Deutschland*, C-582/14, EU:C:2016:779

blockchain projects cannot be compliant with the Regulation.

III. Technical Approach

A Blockchain is a distributed database which is based on records organized as a chain of blocks. A peer-to-peer network performs the management, updates and the operations of the database. One of the main characteristics of blockchain is their resistance to malicious modifications. This security level is achieved by using block timestamp and hash pointers that link the last block of the chain to the previous one. The blockchain design is such that any modification made on a block compels the regeneration of the following blocks in the chain, determining an exhaustive process which is extremely difficult to achieve. The state replications and updates to the blockchain are based on a consensus algorithm. This ensures that any update of the main chain will be performed by an honest node. The process to select the honest node that will have the right to add a new block will depend on the kind of blockchain implementation. The most popular consensus technique in blockchain is Proof-of-Work, which corresponds to solving a cryptographic puzzle [6]. Other alternatives to consensus schemes are based on the agreement between the network peers in a democratic scheme [2] or according to their assets [4]. The main principle of blockchain is to create a new database model that is maintained by a network of nodes instead of being fully allocated in a central server. Each node has a local version of the chain, and the process to update it is defined by a consensus protocol that ensures that nobody can change or delete a value previously recorded. This principle makes blockchain technology suitable to be used to record data for accountability, financial transaction settlement, system logs, and any other application where history must be maintained immutable. Nevertheless, databases are used in a wide range of applications in banking, telecommunication, healthcare industry, government, NGOs, among others. Hence, the new challenge for the blockchain technology is how to manage private information in a decentralized open database specially designed to keep their records immutable and allow everybody to read it.

From the privacy-preserving point of view, blockchain technology can be protected by using multiple cryptographic protocols. With these mathematical functions, we can hide information from anybody that is not allowed to have access to the data. One of the most famous cases of privacy-preserving blockchain implementation is Zcash. Their model considers private transactions by using a homomorphic encryption scheme and a novel consensus algorithm based on Zero-Knowledge SNARK [7]. Although Zcash has proposed a secure scheme to protect the data stored in the blockchain, this is not enough to comply with GDPR. The new European regulation defines that any private information protected under a scheme that allows retrieving the private data in some way is considered as pseudo-anonymized data. In this case, the system must give the user the option to be forgotten from the platform.

This is a big challenge for any blockchain-based service, considering that the main principle of the design is not to allow removing records previously stored. By now we have seen that the right to be forgotten has been solved in projects like MyHeathMyData (MHMD) avoiding the recording of private data into the blockchain. In the case of MHMD, the platform allows the data access to hospitals, research centers, pharmaceutical, among others; within a network of the healthcare institutions. Here, a blockchain platform is being used as a decentralized system for controlling, monitoring, and enforcing the GDPR guidelines during the data sharing lifecycle. Under this model, MHMD records information about the data treatment, keeping the private data inside of a central server at the data controller facilities. Finally, the business logic implementation and the traceability is achieved by recording metadata that can be mapped with the private data by using a particular mapping function that is also hosted outside of the blockchain. Hence, the right to be forgotten is enforced by removing the link between the blockchain and the private data in the mapping function.

Another alternative that complies with GDPR and the right to be forgotten is the approach purposed by BCDiploma. They presented an alternative to solve the issue by eliminating the way back that any cryptographic algorithm has, the secret key. By destroying the secret key, we can make extremely difficult decrypt the ciphered text. However, we cannot state that level of security needed to recover the private data is so high that the data can be considered anonymized after the secret key destruction. This is because a trial and error approach of finding out the correct key is always possible. Moreover, most popular asymmetric encryption algorithms can be broken by using quantum computing, so this also has to be taken into account when designing a blockchain infrastructure aimed at being compliant with GDPR.

IV. Our GDPR Blockchain Model

In order to address the main difficulties that GDPR has regarding the management of sensitive data, we propose in this article an abstraction of our MyHealthMyData model. Let's say that each member of the blockchain network is a Data Controller and a Data Consumer, and they can switch their role according to the activity that they would like to perform in the system. Now, consider that the network members are part of a consortium and are connected through a private blockchain network that is responsible keeping the tracking of the data life cycle and to orchestrate the secure data sharing process.

In our model, each Data Controller has a local data catalogue with the list of his available data items. The catalogue only keeps metadata in order to comply with GDPR. These data catalogues feed the central catalogue that shows the data available in the whole network. Each member of the network is a blockchain node that interacts with Data Controller/Data Consumer's servers through a blockchain driver that triggers transactions and listen the blockchain events. Moreover, in our blockchain model the private data is always stored in the Data Controller's facility and never in the blockchain. However, to

keep the tracking of the data life cycle, each action on a data item (injected or requested) is recorded into the blockchain by using a hash value of the data item. This hash value is mapped by using an off-chain mapping database (inside of the Data Controller's facility) that link the hash value to the data item.

The first step of the data life cycle is to make available the data items to the consortium members. Each the data item to be injected into the local catalogue (also called contract) is indexed and then referenced in the blockchain by storing the hash value of the indexed data items. With this, the blockchain maintains the records of the available data and its history associated without needs to record the private data according to GDPR. This process consists in generates a transaction with the tuple (key,value) [3], where key is called *bcDataItemContractId* and corresponds to the name used to identify in the blockchain the hash of the data item. The value correspond to the *dataItemIdHash* that guarantees the referenced to the local mapping database. The *dataItemIdHash* corresponds to *dataItemIdHash = hash (encrypt (dataItemSymmetricKey, dataItemId || bcDataItemContractId || contractId))*, where

- dataItemId || bcDataItemContractId || contractId is the byte concatenation of the dataItemId (identifier of the data item at the provider's database), bcDataItemContractId (key value identifying the tuple) and contractId (identifier of the contract for the given data item). The first two items link the data item to the specific reference in the blockchain whereas the contractId is also included to avoid the correlation of the same data item with different contracts. Recall that the same data item can be used with different contracts.
- Symmetric encryption is used to protect the dataItemId as long as the key is not compromised.
- dataItemSymmetricKey is the symmetric key name used and is its value is known only by the Data Controller.
- hash(key, content) is a cryptographic hash function that ensures that the correctness of the information can be checked by authorized entities but that it cannot be reversed to retrieve such information.

Once the data is available in the system, a Data Consumer can request data items. This process is triggered by issuing a data item request called "study". The study may involves multiple data items that implies to perform different queries over different bcDataItemContractId references. Thus, making an efficient matching between a list of bcDataItemContractId and a study definition becomes crucial. In our model, each bcDataItemContractId has an ordering inside the blockchain. They are referenced by using a bitmap where each bit references one bcDataItemContractId (a simple integer counter). A bit with the value of 1 means this bcDataItemContractId is involved in that study. With this approach, if 1 billion bcDataItemContractId are indexed into the system, the uncompressed bitmap is 125 megabytes (1 billion divided by 8). A study will not involve all the data items, only few of them, which means that this bitmap can be compressed into few megabytes by using LZMA algorithm [8]. This approach allows us to maintain the track of the whole list of bcDataItemContractId (the data item and contract couples) involved in a study. Finally, the bitmap is stored in the mapping database off-chain. This allows us to remove the link between the data item used in a study and the value stored in the blockchain by modifying the bitmap directly in the mapping database. With this method we reach the right to be forgotten enforce by GDPR.

Now, for a given study a compressed bitmap of the bcDataItemContractId references has to be created in order to define it. The compressed bitmap is stored in the mapping database indexed by a study identifier. At this point the study definition (which is the binary representation of the definition of the study written by the data consumer) with its associated compressed bitmap are hashed together (following the same approach presented before) and stored into the blockchain along with the corresponding identifier of the associated contract for that study.

Finally, the right to be forgotten can be implemented by deleting the field in the bitmap pear each study that corresponds to the data item belonging to the user that is asking to be forgot from the system. This process is easy to implement due to is performed on the mapping database that is off the blockchain.

V. Conclusions

In this article we introduced our GDPR compliance blockchain model that provides the following properties to manage a private data sharing process:

- Traceability of the data life cycle
- Only a data provider can find out where its data was used.
- If two different studies provide the same results, the two hashes will be different.
- In case of an audit, the data provider can show which data items were involved in the study. This can be done because the hash into the blockchain can be checked against the copy held by the data provider, which in turn can check the data items and confirm their usage.
- Efficient implementation of the right to be forgotten.

Although our model is GDPR compliance, the mechanism to implement the right to be forgotten relies on the central based infrastructure of the Data Controller. Thus, as a future work, we can use cryptographic primitives to implement new models of privacy preserving smart contracts.

References

[1] Elli Androulaki, Artem Barger, Vita Bortnikov, Christian Cachin, Konstantinos Christidis, Angelo De Caro, David Enyeart, Christopher Ferris, Gennady Laventman, Yacov Manevich, et al. Hyperledger fabric: A distributed operating system for permissioned blockchains. *arXiv preprint arXiv:1801.10228*, 2018.

[2] Miguel Castro, Barbara Liskov, et al. Practical byzantine fault tolerance. In *OSDI*, volume 99, pages 173–186, 1999.

[3] IBM. Hyperledger fabric - transaction flow, 2018. [Online; accessed 14-September-2018].

[4] Sunny King and Scott Nadal. Ppcoin: Peer-to-peer crypto-currency with proof-of-stake. *self-published paper, August*, 19, 2012.

[5] Daniel Mulligan. Know your customer regulations and the international banking system: towards a general self-regulatory regime. *Fordham Int'l LJ*, 22:2324, 1998.
[6] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. 2008.
[7] P Peterson. Anatomy of a zcash transaction. *z. cash*, 2016.
[8] Wikipedia contributors. Lempelzivmarkov chain algorithm — Wikipedia, the free encyclopedia, 2018. [Online; accessed 14-September-2018].
[9] Gavin Wood. Ethereum: A secure decentralised generalised transaction ledger. *Ethereum Project Yellow Paper*, 151:1–32, 2014.