Towards Using Public Blockchain in Information-Centric Networks: Challenges Imposed by the European Union’s General Data Protection Regulation

Dominik Schmelz  
Vienna University of Technology  
Industrial Software (INSO)  
Vienna, Austria  
dominik.schmelz@inso.tuwien.ac.at

Gerald Fischer  
Vienna University of Technology  
Industrial Software (INSO)  
Vienna, Austria  
gerald.fischer@inso.tuwien.ac.at

Phillip Niemeier  
Vienna University of Technology  
Industrial Software (INSO)  
Vienna, Austria  
phillip.niemeier@inso.tuwien.ac.at

Lei Zhu  
Vienna University of Technology  
Industrial Software (INSO)  
Vienna, Austria  
lei.zhu@inso.tuwien.ac.at

Thomas Grechenig  
Vienna University of Technology  
Industrial Software (INSO)  
Vienna, Austria  
thomas.grechenig@inso.tuwien.ac.at

Abstract—Advances in computer science have raised concerns about the privacy of personal information. Therefore solutions to address these concerns have to be found. Blockchain enables new approaches to solve privacy issues in distributed systems, but at the same time also raises new concerns with its openness and immutability. The European Union has taken steps towards addressing information privacy concerns and define rights of data subjects and obligations of controllers and processors of personal data. We will apply and discuss these in light of current Blockchain implementations. This will result in a guideline for GDPR compliant Blockchain developments in the future.

Keywords—Blockchain, Information-Centric Networking, Data protection, GDPR, Privacy

I. INTRODUCTION

Originally used in payment systems, the blockchain technology is increasingly used in other fields such as asset management, identity provider, insurance, and fund-raising. The so-called Proof of Work (PoW) requires participants to present a solution to a proposed challenge, and is used as a consensus algorithm in the blockchain technology as well as in preventing Distributed Denial of Service (DDoS) attacks, e-mail spam and blog comment spam. These and other use-cases are dealing with personal data and storing personal data in such chains. Protection of personal data has become increasingly important since advances in information technology have raised concerns about information privacy and its impacts [1]. Therefore researchers have to discuss information privacy issues, including technical solutions, to address these concerns. With the European Union regulation (EU) 2016/679 General Data Protection Regulation (GDPR), data protection has received a legal foundation concerning rights of natural persons (data subjects) whose data are processed [2, Art. 12-23]. With it also come obligations of controllers (who decide the purposes and means of the processing) and processors of personal data [2, Art. 24-43]. The GDPR provides a large number of measures to protect personal data, which companies process, against misuse. This includes, among other things, the ability of data subjects to prevent further processing of their data, not only by the controllers, but also by third parties. Furthermore, data processing should be made more transparent for the subjects by giving them a right to receive information about which data is processed for which purposes by whom and may at any time submit an application for modification or deletion of the data. The GDPR is applicable to all companies in the EU that process personal data, as well as third-country companies, if they offer services to EU citizens [2, Art. 3]. Compliance with the data protection measures is enforced by public authorities, who also have extensive rights of access to the processing activities of personal data. Failure to comply with the data protection measures will result in fines of up to 20 million euro or, in the case of a business, up to 4% of its total worldwide annual turnover. In case of a group of undertakings, the annual turnover of the entire group, not that of the individual legal entity is considered.

In large corporate groups, inter-organization information transmission asks for large-scale communication information systems. Blockchain technology has become a feasible technology for inter-organizational communication because of its fault tolerance, durability and attack resistance, authenticity, transparency and openness. For the GDPR on the other hand, an inter-organizational transfer of personal data from one legal entity to another constitutes a transfer of personal data and
underlies information obligations [2, Art. 13 (1) lit e] and obligations towards the implementation of organizational and technical requirements [2, Art. 16-22]. Special rules apply for the transfers of personal data to third countries or international organizations [2, Art. 44-50].

It can be shown that rather simple applications without privacy strategies, such as a blockchain-based collection of the mileage of vehicles, face several data protection problems. The basis of the communication regarding vehicles is a vehicle identification number (VIN) which in some cases “is directly related to the identity of the owner of the car who is in several cases identical with the driver” [3, point 7] and is therefore personal data [2, Art. 4 (1)]. Chapter IV will discuss this use case in greater detail.

Considering the aforementioned possible fines one the one hand and the advantages of the usage of a blockchain technology on the other hand, implementation of this technology must be checked against the GDPR. The identification of the roles defined by the GDPR and the arising obligations by these create a problem since they are highly distributed and may on some occasions not be under direct control of the originator of the data. The enforcement of these obligations creates further challenges, since the participants in a blockchain-based system might be anonymous.

The basis and technical contribution of this publication is a structured analysis of the two currently most used blockchain implementations, with regards to the requirements stated by the GDPR, answering the question whether they are compliant to the GDPR and what further research must be done.

The paper is organized as follows. Firstly, we present the background for current data processing on the blockchain and discuss the presence and categories of personal data therein (Sect. III and IV). Then, we describe the roles in these applications and who is accountable (Sect. V). After that, the findings and implications of these are presented and compared (Sect. VI). Finally, we reflect on the findings, draw the conclusion and list future work (Sect. VII).

II. RELATED WORK

Blockchain is often used synonymously for the block-based storage system and the inter-node peer-to-peer communication powering distributed ledgers. The blockchain in the means of the storage system, stores records in so-called blocks, including a cryptographic signature of the preceding block [4]. This guarantees that the chain has not been modified and that it can be used like a ledger. Distributing it among several nodes (most commonly via a peer-to-peer protocol) makes it a distributed ledger. These nodes usually distribute the blocks across multiple sites, countries, or institutions. The distribution of the information makes the system fault tolerant, but bears certain data privacy, scalability, and interoperability issues [5].

For a detailed analysis of the data protection properties of blockchain implementations, the two currently largest blockchain implementations with respect to market capitalization, namely Bitcoin and Ethereum, have been chosen for this analysis. Bitcoin was the first widespread blockchain implementation published by Nakamoto [6]. Since then the Bitcoin Foundation has been working on the further development of the reference client named Bitcoin Core. Bitcoin itself, like many other blockchain implementations, uses a pseudonymization mechanism to achieve privacy, namely cryptographic public-key identifications also known as bitcoin addresses.

Many studies have shown potential de-anonymization attacks on Bitcoin [7] [8] [9] [10]. It has also been shown that bitcoin addresses can be mapped to IP addresses with a high probability [11] [12]. Implications of this connection will be discussed in section III.

Ethereum [4] added several functionalities to the blockchain feature set. Most remarkably, a Turing-complete programming language and the option to arbitrarily store persistence data which enables the definition of smart contracts. Since the initial release in 2015 several attacks on smart contracts have been performed [13]. Existing de-anonymization attacks on Ethereum are possible under certain circumstances [14].

Apart from cryptocurrencies as applications of the blockchain technology, more general approaches to decentralized blockchain privacy exist [15]. While anonymity in blockchains is desired, according to [16], only pseudonymity is guaranteed in the aforementioned implementations.

The introduction of a General Data Protection Regulation had the goal of replacing the EU Data Protection Directive 95/46/EC [17], which was adopted in 1995. The idea was to unify and consolidate the data protection laws. In 2016, the EU regulation (EU) 2016/679 (GDPR) entered into force and has been enforced since May 2018. One core feature of the new regulation is the right to be forgotten [18] which means that a natural person has the right that his/her data is anonymized or deleted after the identification is not needed anymore for the purpose of the processing. Another more general feature is privacy by design which means that systems should be designed in a way to minimize the amount of personal data processed [19]. Whether this can be achieved with blockchain technology will be discussed in section VI.

III. PERSONAL DATA ON THE BLOCKCHAIN

The GDPR defines personal data as “any information relating to an identified or identifiable natural person (data subject): an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors […]” [2, Art. 4]. This means that all information that can lead to the direct or indirect identification of a natural person is considered personal data.

In blockchain technology, personal data can either be processed during the execution of the relevant protocol or as a payload within a transaction. Profiling [2, Art. 22] might be possible by analyzing multiple transactions, for example by behavior pattern clustering [20] but was not in the scope of our research regarding data protection.

All considered protocols contain indirect identifiers that relate to a natural person, since the idea of a value transferring
chain is to only allow a holder of a certain private key to access the value that has been transferred or stored. The question is, whether it is possible to follow such a relation from the identifier to a natural person with certainty. Bitcoin, for example, had a "send to IP address" functionality that allowed clients to transfer Bitcoin from an address to an IP address. The Court of Justice of the European Union (CJEU) recently ruled that IP addresses can be personal data in some contexts [21]. Despite the fact that this type of transaction has been removed [22] in 2011 it shows how the processing of personal data was directly used in a payment transaction.

Other functionalities in the Bitcoin protocol transfer value directly to public keys (pay to public key) or hashes of public keys called addresses (pay to public key hash). Hashing creates a directly related value which cannot be reversed easily. In any case, it is considered a pseudonymization technique by the Article 29 Working Party [23]. Furthermore, the pay to public key hash transaction hides the actual address only until the funding is used as an input in another transaction. It therefore provides only little privacy enhancement.

As long as information is directly connected or connectable to a data subject it is considered personal information. Pseudonymization, such as the aforementioned hashing, is a data protection measure [2, Art. 32] but the resulting pseudonymized data is still considered personal data. Only anonymized data that cannot be connected to a person is not personal data [2, Recital 28]. Blockchain links the natural persons to the outputs of transactions with a private and public key mechanism where the public key is published openly. Therefore most blockchain implementations including Bitcoin contain personal information within the protocol.

```solidity
pragma solidity ^0.4.20;

contract Mileage {
    mapping(string => uint) private mileages;

    function getMileage(string _vin) constant public returns (uint) {
        return mileages[_vin];
    }

    function setMileage(string _vin, uint _newMileage) public {
        if (_newMileage >= mileages[_vin])
            mileages[_vin] = _newMileage;
    }
}
```

Listing 1. Simple smart contract that stores the mileage of a vehicle

The openness and availability make the blockchain feasible for several other applications including those which store personal information (e.g. named). Listing 1 shows an example of an Ethereum Smart Contract allowing anyone to store the mileage of vehicles, identified by a vehicle identification number (VIN). The purpose of this use-case is that you have a publicly accessible immutable mileage database which might be useful against odometer fraud. This is by no means a production-ready contract and should only be used for demonstration purposes. It shows the essential behavior of a state of the smart contract and the way information is processed. VINs are, according to the European Data Protection Supervisor, private data [3, point 7]. Therefore this contract processes private information. In case of private data in the payload, each node in the network processes this personal data.

This example can also be implemented in other applications of the blockchain technology. For example, on the Bitcoin blockchain, an unspendable transaction (OP_RETURN) could be created to store the information, namely a hash of the VIN (or several properties of the vehicle) and the mileage.

In summary this use-case can be GDPR compliant if all users agree to store their data in a distributed and publicly accessible ledger, but as soon the ownership of the vehicle changes, new users (which might not agree to have their data stored in this fashion) will not be able to remove their data from the blockchain, since the history will be (depending on the implementation of the particular blockchain) visible for an undefined amount of time.

After showing that private data is being processed on blockchains the question remains if this states a processing according to the GDPR. This will be covered in the next sections.

IV. DATA PROCESSING ON THE BLOCKCHAIN

Processing is defined by the GDPR as any operation (including retrieval, transmission, and storage) of personal data [2, Art. 4 (2)].

Several different data processings are done by different legal entities in the network. Figure 1 shows a data flow centered abstraction. A natural person uses a client, which is a software, to access the blockchain, which is used to transfer a transaction (tx) (1) to the Network over an internet service provider (ISP). For data processing abstraction purposes the network consists of two entity-groups, namely miners and nodes. Nodes store the full blockchain and transmit blocks and therefore transactions within the network (2). Miners generate new blocks by processing the transactions (among other things). A service provider is a generalization of an entity that knows the identity (ID) of the natural person and offers some service that creates or receives transactions (3) on the blockchain. An example would be an exchange or payment provider. This shows that entities process the data without directly knowing the identity of the natural person, and others knowing the identity. In some contexts, it could be argued that an address in a transaction relates to the natural person and is therefore personal data.

The Court of Justice of the European Union (CJEU) published its judgment in Case C-582/14 (Breyer) [21] that dynamic IP addresses are to be seen as personal data because they are linkable to the identity of an individual. It was argued that there exists a party (in this case the ISP) that knows the connection between the IP address and the natural person, and interpreting Directive 95/47/EC with reference to Recital 26, it is to be seen as personal data. One could conclude that this also applies to Bitcoin addresses.

If this is the case, then all Pay-to-Pubkey and Pay-to-PubkeyHash transactions, which currently are about 82% of the unspent transaction set of Bitcoin [24] contain private data which would be processed if they are spent.
From a technical perspective, a blockchain network consists of nodes sending and receiving messages. From a data protection perspective, there are several legal entities within and around the network processing personal data.

Taking a detailed look at the legal entities in Figure 1 one can see that there are legal entities within the network itself, namely miners (validating transactions and creating blocks) and full nodes (that save all blocks and send and receive transactions). The service provider works as a kind of gateway that provides access to the blockchain network. Service providers, for example for the aforementioned exchanges, require identification (b) of the data subject or even Know Your Customer (KYC) verification of the identity.

These KYC requirements result from regulative obligations of official legal entities or corporate bodies towards the supervisory bodies in different countries. The actual trends show the globally aligned view that blockchain products are to be categorized as investment products or payment products, thus they have to be regulated as such [25].

To comply with Anti Money Laundering (AML) law, legal entities are obligated to legitimate persons active on the financial markets (e.g. buying or selling products on exchanges). Anonymity for the buyers or sellers is not an attribute welcomed by the supervisory bodies and certainly is not compliant with AML laws. In addition to the AML obligations and anti-terrorist financing prevention, monetary transactions must be traceable:

“There should be traceability of all transactions and process flows; proper authentication is needed in all communications between the entities involved (i.e. the TPP, the account servicing PSP, the merchant/payee) also to prove which entity was responsible for which part of the process in the event of repudiation, operational problems, security incidents and/or fraud.” [26]

This is another regulative obligation forcing exchanges to conduct the KYC process in order to register buyers and sellers on the market.

Also, the Internet Service Provider (ISP) knows the identity (a) of the data subject. Service providers not receiving the identity directly usually receive at least an IP address that makes the client identifiable. A subject can directly transmit a transaction via a node (1) that is connected to the blockchain network or indirectly via a service provider (3). Service providers can also identify the subject first and then receive a transaction over a node (1, 3).

The question for each party within this processing is whether the party is a processor, controller or if there is a joint responsibility. The ISP normally constitutes a controller when IP addresses are concerned, but in the context of the blockchain transaction only relays the information without inspecting or changing the information and would therefore arguably constitute a processor. The miners have the role of processors that verify the information and receive a reward for this verification. From a general view, they do this for the purpose of securing the network. From the perspective of one transaction, they verify the transaction and clear the transaction. The nodes receive the transactions and blocks, verify and transfer them, and eventually store them which is considered processing.

From a general view, they do this to have valid information spread over the network. From the perspective of one transaction, they validate it and are responsible to make the transaction known to other parties. The question remains if a node “determines the purposes and means of the processing” with respect to the definition of a controller according to [2, Art 4 No 7] or whether nodes do it jointly. Since a node decides if a transaction is valid and to whom it is propagated it will most likely be seen as a controller himself.

Knowing the data being processed and the roles from a data protection perspective, it is important to find out what impacts...
these have on the network and each participant.

VI. IMPLICATIONS ON THE BLOCKCHAIN AND ITS USERS

The GDPR describes privacy by default and privacy by design as a requirement to be implemented into systems processing private data [2, Art. 25]. Privacy by default means that the strictest settings for privacy should be applied by default. Privacy by design requires the implementation of data minimization principles meaning that only the needed personal information is being processed (including receiving and sending). In a blockchain network, a transaction is not only distributed between those who are involved in a transaction but, due to the mechanics of a blockchain, to all nodes. Furthermore, it not only sends needed information to a node but distributes all information in the network which also contradicts data minimization.

Another issue is that the period of time in which this data is being processed is not defined. The GDPR requires a definition of the time period after which private data will be deleted and generally a deletion of information after the purpose of the processing is expired (with exceptions). The most common blockchain technologies analyzed here do not allow the deletion of any transaction which also contradicts the right to be forgotten (deletion) of a data subject. Other rights of the data subject like the right to rectification are also not possible since transactions cannot be changed after they have been transmitted. On the other hand, the right to information and portability are implemented since everyone can access the blockchain. The controller has the obligation to keep records of all processing activities besides a data protection impact assessment and other obligations. Each node would therefore have to comply with all rights of the subject and obligations of the GDPR.

Further obligations arise if the personal data is transferred outside of the European Union (to third countries) including checks if the country implements sufficient data protection regulations. The nature of blockchain does not allow the transfer to be restricted only to European Union countries. Blocks are always sent to all nodes whether they are within or outside of the European Union.

Since personal data is transferred in the protocol as well as in the payload (as described in section III) the GDPR applies for the data processing and the aforementioned obligations arise. A solution to that problem can be complete anonymity in the case of protocol information. This would include untraceable payment transactions and anonymization of relaying addresses. The KYC and AML laws mentioned in section V contradict the efforts of alternative blockchain implementations trying to reach this goal (e.g. Monero [27] and Zcash [28]). Therefore the use of these might solve issues with the GDPR, but will probably raise other legal issues.

In the case of personal data in the payload, the information must be unreadable to any party that is not allowed to process the personal information. Some [29] try to mitigate this issue by disconnecting the chain from the public internet (private blockchain) which generally does not solve the problem, but minimizes it since all processors are usually known. Others [15] try to only store hashes or other non-invertible derivations of the clear-text on-chain and store the actual private data off-chain (e.g. in separate centralized databases). Hashing is considered a pseudonymization technique by the Article 29 Working Party [23] and therefore the hashed private data on the blockchain is still considered private data.

VII. CONCLUSION AND FUTURE WORK

Blockchain technology implements high availability and transparency at the cost of data protection. It was shown that several ground principles of the GDPR contradict the fundamentals of blockchain technology. Consequences for all participants in such a network are possible. Since the GDPR has only been effective for a short time, we have to wait until the European Justice Court decides if there is a legal basis to this in a blockchain network as implemented today. The possibility of significant fines should motivate all participants to act quickly. Implications for the network are unpredictable, neither is it foreseeable whether an application with compliance to the GDPR could be implemented on current blockchain candidates.

Further research should clarify the requirements needed to be implemented by a GDPR compliant blockchain. Techniques for further improvement of data protection on the blockchain will be needed for this to be achievable.


