On Blockchains and the General Data Protection Regulation

Luis-Daniel Ibáñez, Kieron O'Hara, and Elena Simperl

{l.d.ibanez|kmoh|e.simperl}@southampton.ac.uk
University of Southampton

1 Introduction

Blockchain technologies have emerged as a revolutionary way of executing business processes in a decentralised way. Spearheaded by their use as cryptocurrency - as a way to transact digital assets without the need of a bank - blockchains have attracted the attention of the general public and mainstream media. Beyond cryptocurrencies, large companies, entrepreneurs, and investors have recognized the potential of the underlying technologies as a game-changer in the way we formalise and manage digital relationships in the Web. According to the Worldwide Semiannual Blockchain Spending Guide published by IDC, blockchain technologies received $945M of investment in 2017, and are estimated to get a staggering $2.1 billion by the end of 2018.

In parallel, the European Union has recently introduced the General Data Protection Regulation (GDPR) to protect and empower all EU citizens data privacy and to reshape the way organizations across the region approach data privacy, with particular emphasis in reducing the power asymmetry between organisations that manage and exploit personal data, and the individual to which these data belongs. Unfortunately, despite the fact that from a philosophical perspective blockchains also pursue the empowering of individuals and reducing the influence of centralised parties, several scholars and practitioners have raised concerns about the conflict between the way blockchains work and some articles of the GDPR. This 'collision course', as some enthusiasts and developers have called, on the one hand generates anxiety about the possibility that GDPR might hinder European innovation in blockchain technologies, while on the other, opens potential opportunities in the use of blockchain technologies as a tool for enforcing GDPR.

In this paper, we review the legal and technological state of play of the GDPR-Blockchain relationship. Next, we analyse three interaction scenarios between data subjects and blockchain systems, and propose possible ways of achieving GDPR compliance by using state of the art technologies. Finally we review current efforts in the use of blockchains to enforce GDPR principles, in particular 'Data Protection by Design'

2 Brief introduction to Blockchain technologies

In the most general sense, a blockchain is an append-only list of sets of cryptographically signed records or *transactions* (called *blocks*) that a number of parties want to update. Each time a block is appended to the chain, is linked to the immediately precedent block, inducing an ordering of the records by their timestamp of inclusion in the blockchain. Blockchains also implicitly include two further elements, first, a set of *validation rules* that define the conditions for records and blocks to be included in the blockchain; second, an algorithm or *protocol* to enforce validation rules that is trusted by all parties that record data in the blockchain.

In the beginning, blockchains were created for particular transactions that have pre-established validation rules (e.g. transfer digital value). However, further platforms generalised the whole concept, enabling the definition of arbitrary transactions and validation rules in a programming language, that can be then deployed with the guarantee that the rules it defines will be enforced through the particular protocol that the platform implements. These platforms allowed the realisation of Nick Szabo's vision of *smart contracts*, a world where several kinds of contractual clauses (such as collateral, bonding, delineation of property rights, etc.) on assets that can be digitally controlled can be embedded in the hardware and software we deal with, reducing transaction costs imposed by either principals, third parties, or their tools. Following Szabo's terminology, transactions and validation rules that are defined in this way are now commonly known as *smart contracts*, and the systems to deploy and execute them are known as *smart contract platforms.*

The most illustrative example of a blockchain is a transaction ledger where members of a community record transfers of assets among them, for example, to record that Alice gave 10 coins to Bob or Bob gave 5 coins to Charlie. In this particular case, the main validity rule is that Alice currently owns at least the same amount of coins that she attempts to transfer. Concerning validation enforcement, the most straightforward solution is to trust a single *ledger keeper*, or in modern terms, a bank. From the point of view of Smart Contracts, a community can program an application to transfer coins, and deploy it in the server of one of the members (or use an external cloud provider) that becomes the *application keeper*, *i.e.*, in charge that the program runs as intended. However, entrusting a single central entity with such power has a number of potential risks, enumerated below. For the sake of simplicity we use the term *keeper* to refer both to ledger and application keepers.

1. Introduces a *Single point of failure*, that is, if the keeper goes offline, then no one can do any transaction.

2. Keepers commonly charge for their services. As single points of failure, they are in a position of power to increase their charges by threatening to block users from the service

3. How to be sure that the keeper does not manipulate the ledger or application to benefit certain members of the community or itself? What if it changes the validation rules?

4. For some use cases, it might be impossible to reach an agreement on who name as keeper

These issues motivated the development of *de-centralised* validation protocols that also provide stronger guarantees on the *immutability* of stored records[1], with the ideal of providing the members of the community with a mechanism to share the validation task among themselves. Such protocols are classified in two categories according to the knowledge of the identities of transacting parties: *permissioned* and *permissionless* blockchains.

In *permissioned blockchains* there is an external mechanism that allows the identification of the parties that want to add records to the blockchain and/or participate as validators. Protocols that can be applied to permissioned blockchains have been studied for decades in the field of Distributed Systems, where they are known as Byzantine Fault-Tolerant (BFT) protocols, and have witnessed renewed attention from both researchers and practitioners after their connection to blockchains was established. The most well-known platform in this realm is the Hyperledger Fabric[2], that enables the creation of permissioned blockchains with custom identification services and a choice of BFT protocols.

In *permissionless blockchains* anyone can add records or contribute to validating transactions via a pseudonym, permissionless blockchain face the additional challenge of dealing with the possibility of a malicious party crafting several different pseudonyms to attempt to influence the validation protocol in its favour, a technique known as the *Sybil attack.* Almost all cryptocurrencies (Bitcoin, Litecoin, etc) belong to these category. The most popular smart contract platform, Ethereum also belongs to this category. Currently, the most battle-tested permissionless protocol is the so-called *Proof-of-work*[3]. In all BFT protocols a key step is the decision of which validator is going to validate the next set of transactions. In permissioned blockchains, this can be done either randomly, or following a pre-defined sequence, however, in the permissionless scenario, this fails due to the Sybil attack. Proof-of-work's key innovation is a novel kind of 'lottery' to assign the next validation turn: give it to the first validator that solves a cryptographic puzzle specifically designed

[1] As the concept of Blockchains popularised in the context of the quest for decentralisation, it is commonly assumed that a Blockchain's validity enforcement protocol is necessarily decentralised

[2] https://www.hyperledger.org/

[3] Also known as *Nakamoto consensus* by distributed systems researchers, following Satoshi Nakamoto's seminal paper describing Bitcoin

to be solved only by the means of brute force calculation[4] , thus, making its probability of being solved a function of the available computational power. This makes a Sybil attack impractical, as instead of generating more pseudonyms than honest participants, an attacker needs to have more computational power than the combination of all honest participants.

The absence of identifiable parties has immediate implications in privacy and data protection with respect to permissioned blockchains. On the one hand, it is good for privacy, as an individual can participate in a permissionless blockchain without revealing its identity, but on the other, one needs to agree with the fact that an unknown individual or organisation may process data input to the system. Also, note that the fact of contributing as validator in a blockchain (permissionless of permissioned) comes with the responsibility of processing transactions that do not belong to you.

3 Brief Introduction to GDPR

Any democracy which hopes to have a flourishing economic private sector must balance the twin goods of privacy and information flow (the latter includes freedom of speech and rights to information). In general, protecting privacy involves putting friction into the flow of information, while a freer flow will help both to create a more informed public and to support knowledge-based economic activity. The concept of *data protection* is meant to balance these two ideals which have a tendency to pull in opposite directions. Data protection laws give data subjects rights over the use of their personal data by others (for instance, to be told what data is held, to correct incorrect data, to have excessive data deleted, and to have their data processed according to principles of fairness), but also give data controllers rights to process data in certain circumstances (see below). Hence data protection law is *not* privacy law — it does not protect privacy *per se*, but it gives data subjects certain rights to control aspects of their data's processing according to their preferences about privacy and openness.

On 25th May, 2018, the European Union brought in the General Data Protection Regulation (GDPR), to update its previous data protection regime, which dated back to the Data Protection Directive (DPD) of 1995. The DPD, designed for an era of databases, had struggled with networked information spaces, and the GDPR was a much-needed update. Furthermore, not only has it attempted to catch up with technological change, it is a much more principled document, providing a new (tougher) framework for the processing of personal data. Its effects will reach out beyond the European Economic Area, to cover any organisation processing data about EU residents.

If the 99 articles were boiled down to a single word, it would be *accountability*. Organisations need to consider the risks involved in processing data, to remain vigilant against those risks, and to build data protection into all their processing practice ('data protection by design'), and will be liable to heavy fines if they neglect these duties. The GDPR sets out the six core DP principles. Processing should be lawful, fair and transparent; data should be collected for a specific purpose and not processed for incompatible purposes (purpose limitation); data held should be adequate and not excessive (data minimisation); data should be accurate and up to date; data should not be stored for longer than necessary; data should be processed securely.

The framework is meant to be prescriptive about how data is stored, and to give more control to data subjects about how to get hold of their data, to correct false data, or to get inappropriately held data deleted. Consent to process data now has to be unambiguous and for specific (named) purposes – catch-all clauses will no longer be sufficient. Organisations must draw up detailed Data Protection Impact Assessments to describe the impact of the processing on data protection, and must hold an inventory of all the personal data it holds. Organisations processing personal data have to appoint a data protection officer to monitor compliance, and to be a contact point for regulators and data subjects.

As with the DPD, the GDPR allows data processing under a specific set of circumstances, at least one of which must apply: the data subject has consented to the processing; processing is necessary for a contract to which the data subject is a party; processing is necessary to comply with a legal obligation; processing is necessary to protect the vital interests of the data subject or someone else; processing is necessary for a task in the public interest; processing is necessary for the legitimate interests of the data controller. In that case where personal data finds its way into a blockchain, perhaps the most likely and hopeful grounds for processing are that there is a contract or consent in play.

[4] The validator that solves the puzzle has proved to the network that he has *worked*, i.e., invested computational resources to get the turn

The data subject also has new rights, including a right to erasure, which could be a problem given the perpetual existence, even multiply hashed, of the full chain. There must be a ground for demanding erasure, however, and two (of the six) grounds may be challengeable. The first is that the personal data is no longer necessary for the purposes for which it was obtained. A possible counterargument is that it is and always was necessary to process the personal data in perpetuity, if it is to be processed at all in a blockchain, and so perpetual processing is always necessary for the purpose for which it was obtained. If personal data is to be put in a land registry (say), then it has to stay there forever because that is the point of a blockchain, and if it is removed then the blockchain is compromised, even if the land ownership is later transferred. The second is that consent was the only ground for processing, and it has been withdrawn. If consent has been gathered under the strict assumption that processing would be in a blockchain, so that therefore consent must be in perpetuity, then it is possible that this ground could be avoided too. A more potentially problematic ground is that the data has to be erased for compliance with a legal regulation.

Key definitions for blockchain include those for personal data and pseudonymous data. Personal data is any information relating to an identified or identifiable natural person – that is, someone who can be identified, directly or indirectly, from the data. The GDPR lists several types of identifiers (some new ones since the DPD), but identification need not be via an identifier. With respect to encrypted or hashed data, this is still personal data. However, it falls under a new class of personal data created by the GDPR, which is pseudonymised data. Pseudonymisation is the processing of personal data so that additional information is needed to effect identification, as long as the additional information is kept separately behind a technical or administrative firewall. Pseudonymisation does not prevent personal data being personal, but it gives the organisation more leeway for its processing, because the risks are correspondingly lower. All data in all blockchains is processed – processing includes storage, and also would include the processing done to the old blocks to mine new ones.

The other important definition is that of the data controller (this is little changed from the DPD). The controller is the point of accountability in the GDPR, and it is the person, body or agency which (perhaps jointly) determines the purposes and means of processing. Those actually processing the data on behalf of the controller are called data processors, and have fewer responsibilities. There is a question, with a blockchain, as to who is the data controller. We might assume that the miners are data processors, with the controller being someone (who?) in charge of the blockchain. However, another interpretation is that each locally-stored version of a block that contains personal data (i.e. each node in the P2P network) should be treated separately, and the miner determines why and how its own local version of the block is processed. In that case, each miner is a data controller, and each bears the full responsibility for the processing of its own block. A third interpretation is that the blockchain is a single entity, but that each miner is a data controller jointly with all the others, determining why and how the data are processed.

Not all can be read into the letter of the law. Much will depend on how case law evolves in the courts, to see whether, e.g., ideas about what constitutes pseudonymous data will be interpreted widely or narrowly. Furthermore, the DP regulators will have an effect too; for instance, Britain's ICO has generally been concerned with proportionality, while France's CNIL takes a much harder line.

4 State of Play

The GDPR was developed in the context of a world where business models based on collecting user's personal data in exchange of gratis services, and then monetize knowledge and analytics extracted from it thrive [1]. These models have been extremely successful for generating revenue, but raise several concerns about data protection. Take for example the recent Facebook-Cambridge Analytica scandal[5], where personal data was transferred to a third party without data subjects knowing when and for what. GDPR aims at solving the problem by providing the legal ground for making businesses profiting from personal data accountable for how they process it and exploit it.

The motivation of blockchains is quite similar: replace central entities controlling (and ultimately exploiting) everyone's transactions and data, for decentralised validation of transactions where is much harder for someone to gain control to the detriment of others. Furthermore, it has

[5] https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html

been claimed that through a Blockchain system, data subjects can be empowered to exert more control on their data (an idea that we develop in section 6). Unfortunately, some of the technical foundations of current decentralised protocols are in conflict with some of the GDPR postulates. This is because GDPR aims to *regulate* the world of centralised data control, whereas the aim of blockchain is to *challenge* it; hence GDPR retains the centralised roles and makes them accountable, but it does *not* imagine, or leave room for, any alternatives. Hence it may have the unintended consequence of stifling decentralising innovation. Based on previous work by researchers [2, 3] and practitioners (that raise their voice in interviews and blog posts, e.g., [4, 5] we list below the most critical conflicts between the GDPR's conservative assumptions and blockchain.

GDPR roles in blockchain networks, as explained in section 3, GDPR defines the roles of data controllers and processors to assign responsibilities. These roles are designed to fit a centralised scenario where a single company collects personal data and defines how is processed (controller) that may outsource part of the processing to third parties (processors). Blockchains are by definition run by several parties, making difficult to assign roles. In permissioned blockchains, there are two possibilities, for the case where a community decided together the validation rules that the blockchain implements, one can consider them all to fall under the definition of *Joint data controllers*[6], where all share the responsibility of compliance. If a permissioned blockchain accepts the contribution of validators that did not participate in the definition of the validation rules, these fall under the definition of processors. The fact that in permissioned blockchains the identities of participants are known simplifies matters, as 'paper contracts' could be signed before granting access to the blockchain, and in case of problem, the off-chain legal system can notify and prosecute accordingly. Unfortunately, for permissionless blockchains, the situation is quite unclear. In the case of Bitcoin, validation rules were initially defined by the core developers (the initial members of the blockchain) and further participants limited to apply those rules, creating a similar joint-controller/processor assignment that in the permissioned case. However, since long ago, core developers don't participate in the network as validators, and it is also possible for validators to update validation rules, provided they reach enough consensus among them[7]. This fact suggests that a permissionless blockchain closer to an scheme where all participants are potential joint-controllers. To complicate matters the pseudonymous nature of participants makes impossible to individually address any of them, and raises concerns about the transfer of data to third countries[8], as a validator might be located in a foreign jurisdiction.

Hashes and public keys as personal data. Permissionless blockchain protocols heavily rely on hashing and public/private key cryptography. Hashes enable the efficient storage of transactions in a format that enables their verification, while public/private key cryptography provides a mean to validate the sender and the receiver of the transaction[9]. A priori, permissioned protocols don't need these techniques, but many of them integrate them for efficiency reasons, and to improve the privacy of exchanged messages. This leads to the following question: Is a hash of personal data, anonymised data? This relates to the possibility of identifying a data subject by personal data *stored* the blockchain. Clearly, storing it plainly will allow anyone with access to the blockchain to it, but if hashes are stored instead, is it enough? ; The second question is what about public/private keys?, in many blockchains, at least the public part of it needs to be accessible to validators, the ways to manage them change dramatically according if they are considered personal data or not.

For the case of hashes, Opinion 05/2014 of article 29 working party considered that hashing is a pseudonymisation technique as a hash still allows 'linkability' to an identity through comparison with the original document. Therefore, applications using hashes need to be cautious of where they store the original documents containing personal data. If one document gets leaked, and an attacker knows it was stored in a blockchain, it is fairly easy to recover to which transactions it belongs.

With respect to public keys, [6] traces a parallel with Case C-582/14 of the Court of Justice of the European Union (CJEU) *Patrick Breyer vs. Bundesrepublik Deutschland* on 19 October 2016, relating to dynamic IP addresses. In this case The CJEU decided that dynamic IP addresses collected by an online media service provider only constitute personal data if the possibility to

[6] Article 26

[7] Introducing then the governance scheme of the blockchain into the equation

[8] Chapter 5 of GDPR

[9] Back to the coin transfer example in section 2, one can imagine that coins are locked into lockboxes that open with cryptographic keys, by using the key in the transaction, Alice can prove to validators that she owns the coins that she intend to transfer to Bob. Coins are then put in a fresh lockbox that only Bob can open with a different key

combine the address with data necessary to identify the user of a website held by a third party (i.e. user's internet service provider) constitutes a mean "likely reasonably to be used to identify" the individual or by a third party, with the caveat that "if the identification of the data subject was prohibited by law or practically impossible on account of the fact that it requires a disproportionate effort in terms of time, cost and man-power, so that the risk of identification appears in reality to be insignificant"

For blockchains and public keys, this means that attention needs to be paid to where the data that may be used to link a public key with its owner, in a similar fashion to precautions required when storing hashes. In the initial design of permissionless blockchains, users themselves should be in charge of generating the public/private key pairs to be used and take responsibility for safeguarding the private parts. Under this assumption, and following the caveat of C-582/14, the public key won't be considered personal data. However, many users rely on centralised services to manage their keys, creating a vector of attack, as the central server holds the data required to identify users from public keys[10]. In permissionless blockchains, that rely on external, often centralised, identity services, one needs to pay attention to who controls those services, and if a situation similar to C-582/14 could arise.

Principles of personal data processing Current blockchain protocols rely on a full copy of the ledger being replicated at least at each validating node to effectively validate the general state of the ledger. Some researchers consider this feature as profoundly at odds with some of the data protection principles laid out in article 5 of the GDPR. We proceed to review each of these principles on the light of their relationship to blockchains.

The first principle is the *Fair and lawful processing* of personal data, that translates to having legitimate reasons to collect it, be transparent about the intended usage and make sure nothing unlawful is done with it. We consider the first two easy to achieve in a blockchain scenario, as in most cases, validation rules are openly available, or they can be easily communicated to end users. Regarding unlawfulness, there is a question about *control* on data processors to make them accountable. In a centralised scenario, one can't never be sure that the processor is doing something unlawful with personal data[11], but if something ever pops up, they can be identified and prosecuted accordingly. For permissioned blockchains, if one assumes that validators in a permissioned blockchain are joint data controllers, one can easily fall back to the centralised case covered by GDPR. However, in permissionless blockchains, where identities cannot be established, accountability on individual processors is harder.

The second principle is *purpose specification*, that translates to specify the purpose or purposes for which personal data is obtained, and that anything done with the data must be compatible with this. Similar to fair and lawful processing, is easy to inform users of the purposes, but, not that easy to control that any outsider is profiting from the public availability of a permissionless blockchain to execute a processing beyond the original purposes stated at collection time.

The third principle is the one of *adequacy* that states that collected personal data shall be adequate, relevant and not excessive in relation to the purpose or purposes for which they are processed. We consider this principle independent of the use of blockchains, that is, the question of adequacy is the same when a data controller stores into a centralised database or in a blockchain. Consortiums running permissioned blockchains and designers of permissionless blockchains need to correctly assess if data required for validation rules is adequate, in the same way a centralised data controller would for data mining process.

The fourth and fifth principles, *accuracy* and *retention* refer respectively to the accuracy and requirement that personal data needs to be up-to-date, and personal data not being retained longer than is necessary for the purpose it was obtained for. These principles are linked to two of the rights introduced by GDPR: the *right to erasure* and the *right to amendment*, that allow data subjects to request the amendment or erasure of their personal data under certain circumstances. Blockchains, by design, retain the whole history of transactions as part of their strategy to guarantee decentralised validation. In the same tonic, it was also made very hard to delete or amend already committed transactions. Current blockchain protocols take advantage of the chain structure to optimize the way the ledger is constructed and validated, as such, it is extremely hard to delete or update a transaction in the chain. The latter has been often marketed as a feature of

[10] In the case of cryptocurrencies, this is even riskier. Recall that the public/private key is what a cryptocurrency network accepts as prove of ownership of coins. If private keys are stolen, digital assets are stolen too. The japanese cryptocurrency exchange site Coincheck lost $500M of clients' assets to a hack.

[11] c.f. Facebook and the Cambridge Analytica scandal

data stored in blockchains to bec *immutability.* Immutability is desirable in some cases, but, if in the context of personal data, it is in direct conflict with the rights of erasure and amendedment.

Proposed solutions from the technical side: A large amount of work from developers and technologists, especially in the context of permissionless blockchains, has been directed at the problems of (i) how to avoid the identification of users from the analysis of blockchain transactions, and (ii) how to obscure transaction contents from outsiders and validators, with a minimal tradeoff in the complexity of the validation protocol. Indeed, the requirement imposed by current protocols of every validating node requiring to store a full copy of the whole ledger, gives the opportunity to attackers to join the blockchain as validators, download a copy of the blockchain, and apply data mining techniques to try to identify users. Permissioned blockchains are no exempt of this danger, as one party can use the same techniques to try to infer private information, like how many times other members of the network have transacted. We note that most works target the preservation of the *privacy* of transactors, while the more general problem of data protection has only been considered recently.

Bitcoin developers suggest to always generate a new public/private key pair for each transaction, as reuse of the same creates patterns that are easier to detect by mining techniques (see for example the work of [7] for Bitcoin and of [8] for Ethereum). Other researchers have pointed out that the analysis of the messages exchanged in the network enables the linking of public keys to IP addresses, leading to identification. The naive solution of hiding behind an anonymity network like TOR has been considered by some researchers as insufficient in the case of Bitcoin-like networks[9], even motivating proposals about the complete re-designing of how messages are exchanged [10].

Further efforts look at integrating advanced cryptography techniques into validation protocols to obscure potentially private information. For example, *Ring Signatures* is a technique that enable the signing of a transaction by several parties, in such a way that an outsider can determine that one of them generated the transaction, but is unable to tell which one did. By carefully selecting parties from the network to create a 'ring', a protocol can effectively protect the identity of the sender while still allowing validators to validate transactions.

Other solutions go one step further and tackle the problem of obscuring the content of transactions, while still allowing validators to validate them. In this space, a combination of ring signatures with other cryptographic techniques called *Ring Confidential Transactions* [11] was implemented in the Monero cryptocurrency. An alternative is the use of *zero-knowledge proofs.* A zero-knowledge proof is a general method by which one party (the prover) can prove to another party (the verifier) that she knows a value x, without conveying any information apart from the fact that she knows the value x. In the context of blockchains, this maps to the problem of a transaction sender wanting to prove to a validator that her transaction complies with the validation rules without revealing the values being transacted, or who is the recipient. For example, succinct non-interactive zero-knowledge proofs (Zk-snarks) [12] have been successfully implemented within the Z-Cash cryptocurrency.

For achieving the same goal for Smart Contract platforms, the Enigma system [13] leverages the concept of *Secure Multi-Party Computation* (SMPC) a subfield of cryptography that studies methods for parties to jointly compute a function over their inputs while keeping those inputs private. For example, a group of people could provide access to their salary, and together compute the average wage of the group without revealing the exact value of their salaries. Enigma combines a blockchain network for providing verification of hashes with a SMPC network where the data that correspond to hashes is split between different nodes, coordinated by a protocol that allows them to compute functions together without leaking information to other nodes. Specifically, no single party ever has access to data in its entirety; instead, every party has a meaningless (i.e., seemingly random) piece of it. Programmers can then specify which parts of the computation should be executed publicly (in the blockchain) and which in the SMPC, creating the concept of *Private Smart Contracts.*

Note that, although most of these techniques have been developed for permissionless blockchains, they can be easily adapted to the permissioned case.

Proposed solutions from the legal side: Despite technicians' activity in terms of strengthening the pseudonymisation capabilities of blockchain systems, there are very few technical papers specifically tailored to accommodate the principles of erasure and minimisation, perhaps due to the belief that if a high enough level of anonymisation of personal data within blockchain systems is achieved, the GDPR could be side-stepped from its very beginning[12]. Lawyers picked up the

[12] Note that outside many of the cited researchers and practitioners work outside Europe, where data protection

gauntlet and have put forward some recommendations. [3], for example, plays it safe and advocates to store personal data off-chain, having in the chain only a hash that could be used to link to an encrypted database where full data is stored. From the technical side, [14] identified the use of this technique in the wild, naming it "Hashing-Out", as opposed to "Max-compression", when one tries to compress data as much as possible to make it fit in the blockchain[13]. From a legal point of view, hashing-out greatly simplifies matters, as if personal data is stored in a database under the control of an identifiable data controller, compliance with GDPR principles is much easier. However, in a way, this might be considered a betrayal to the decentralisation principle of blockchains, as a certain degree of control on data remains in the hands of a single centralised party. If a failure (intentional or not), happens, then data is irremediably lost, as it cannot be reconstructed from the hash. Furthermore, an availability failure (again, intentional or not) can block the entire data processing, bringing us back to square one of the problem that motivated blockchains (Section 2) [14]

Hashing-out is also an alternative for implementing the right for erasure. If a request to be forgotten is received, one only needs to erase any off-chain data that could be used to identify the subject if linked to the in-chain hash. A similar procedure can be used to implement the right of amendment: first, the incorrect record is 'erased', then, the amended data is added to the chain in a new transaction[15]. From the technical side, the use of *chameleon hashes* to devise *Redactable Blockchains*[15] has been explored. Informally, a chameleon hash is a hash that contains a digital 'trapdoor'. The knowledge of how to open the 'trapdoor' enables the 'breaking' of the hash. To amend a transaction, one goes to the block that contains it and uses the trapdoor to 'open' it, regenerate the block with the amendment and stitch it back to the chain in the original position. The knowledge of the trapdoor can be given to a trusted third party, or added as a primitive to the protocol to enable its decentralised execution. Note that, despite the fact that a party could try to redact a blockchain in its favour, it is still needed that all others accept the redacted version. Unfortunately, to be redactable a blockchain needs to include chameleon hashes since its inception making impossible to add redactability to existing Blockchains.

Lawyers have also noted that the right to erasure is not an absolute right, and that the concept of erasure leaves room for interpretation. This open the door to the following alternative solution: when an erasure or amendment is requested, append to the blockchain a transaction that contains a reference to the one that is being erased/amended that semantically invalidates it. However, the applicability of such a solution depends on the significance of erroneous data being still visible, even if the blockchain attests its amendment. Imagine for example that a blockchain is used to store data about sexual offenders[16], due to a mistake, a record of someone that has not committed such a crime appears in the blockchain. This citizen invokes his right to amendment, and a transaction on the blockchain is pushed such that the record is 'invalidated'. Is this enough? Or the fact that the wrong record is still there poses a risk for the citizen? The situation is similar to two documented cases that involved Google's search engine. Two person that were convicted in the UK for different crimes asked Google to remove links to news articles that reported their convictions from its index[17]. A court ruled favourably in one case because considered the crime not serious enough, while ruled against the second for the opposite reason, declaring the information to be relevant for the public. In this case, the court took responsibility for balancing the public's right to access the historical record, with the potential impact on the person. From a technical point of view, this ruling means that for the particular use case of using a blockchain to archive convictions, a *hard* delete functionality is required.

Another possibility for erasure relies on the use of a sufficiently strong encryption scheme for data stored in the blockchain, then, erasure can be achieved by destroying the encryption key. However, despite the mathematical proofs defining the limits of encryption schemes, the legal definition of what 'sufficiently encrypted' means is still under debate. Much will depend on how much of the dataset is encrypted; if only identifiers have been encrypted and some of the rest of

directives are less stringent

[13] However, they treat the general problem of attaching any type of metadata to blockchain transactions, and did not consider what happens if such data qualifies as personal

[14] What the centralised party cannot do is modify the data without being noticed, as the verification against the hash stored in the blockchain would put him in evidence

[15] It needs to be noted that the previous hash, which is still considered personal data, will remain in the blockchain, but pointing to 'nowhere'

[16] that in many countries already exists in a centralised form

[17] Google loses 'right to be forgotten' case, BBC news 13/04/2018 http://www.bbc.co.uk/news/technology-43752344

the data is in clear, then reidentification is more likely via jigsaw ID techniques. In the absence of case law the Article 29 Working Party's anonymisation opinion already cited (05/2014) is perhaps the most germane, even if it predates GDPR. This says "The sole implementation of a semantic translation of personal data, as happens with key-coding, does not eliminate the possibility to restore the data back to their original structure — either by applying the algorithm in the opposite way, or by brute force attacks, depending on the nature of the schemes, or as a result of a data breach. State-of-the-art encryption can ensure that data is protected to a higher degree, i.e. it is unintelligible for entities that ignore the decryption key, but it does not necessarily result in anonymisation. For as long as the key or the original data are available (even in the case of a trusted third party, contractually bound to provide secure key escrow service), the possibility to identify a data subject is not eliminated. This does put a lot of significance upon the key's existence, even behind firewalls, and so the destruction of both the original data and the key (given a completely encrypted dataset) would seem to put the data controller in a strong position. However, the opinion also draws specific attention to the possibility of a brute force attack, which destruction of the key will not solve (although if the space of possibilities is so large as to suggest that a brute force attack is not a 'means reasonably likely to be used' by an intruder, this will boost the controller's defence). The controller may also have to keep the risk of reidentification under review as technology evolves[18]. For example, quantum computers could make a brute force attack possible.

Erasure for efficiency reasons The problem of data erasure been tackled by some Blockchain developers, although not in connection with data protection, but for efficiency reasons. Recall that a copy of the ledger being kept by the Blockchain network needs to be kept (at least) by all validators, therefore, as transactions are added, concerns about the ledger size and its impact on validators storage and may appear. This has led to the natural appearance of hashing-out strategies like the *Lightning Network*, where Bitcoin transactions occur in network of payment channels before being commited to Bitcoin, or methods to delete some data like Ethereum's *contract self-destruction*[19] and *State-tree pruning*[20].

Self-destruction is a special instruction in Ethereum that allows a smart contract to be deleted from the current state of the blockchain, however, the code of the contract and the history of any variable update (that are modelled as transactions) can be retrieved from the history of previous states registered in the blockchain, making it an incomplete solution for implementing the erasure principle . State-tree pruning is a strategy for eliminating some data of Ethereum's blockchain, resembling the garbage collection functionality of programming languages. As such, the choice of what is pruned is driven by technical reasons (unused or very old records) instead of responding to a *human* demand.

Amendment in smart contracts Smart contract platforms enable the execution of arbitrary logic, where the execution is carried out in a decentralised manner according to the underlying blockchain validation rules. As such, smart contracts have a certain degree of control over their own internal state. If one of these variables happens to hold personal data, one can implement logic to enables its amendment. Note however that, at least in existing smart contract platforms like Ethereum, the previous value can be retrieved from the transaction history, as each state transition in the contract is modelled as a transaction, thus, registered in the Blockchain.

5 Possible ways to ensure GDPR compliance

In this section, we outline several strategies that could be adopted towards compliance with the GDPR. To ease the analysis, we define the three most common scenarios of how a data subject interacts with a blockchain, then, for each scenario, we proceed to propose a possible role assignment, and enumerate applicable strategies for data minimisation and right to erasure and amendment.

Our scenarios are inspired in the categorisation of blockchain systems of [6], that separates them in blockchains developed with an specific purpose (further differentiating between the ones designed to store personal data, and the ones designed to store/process non personal data), and *non-specialised* blockchains that can store and process any kind of data, roughly matching the concept of Smart Contract Platforms.

[18] Thanks to Alison Knight for discussions on this point.

[19] http://solidity.readthedocs.io/en/latest/introduction-to-smart-contracts.html#self-destruct

[20] https://blog.ethereum.org/2015/06/26/state-tree-pruning/

Scenario 1: An individual interacts directly with a permissionless blockchain For example, an individual that buys and exchange cryptocurrency without intervention from a third party. From a role perspective, there is no escape from the fact that no data controller can be identified, making impossible to make someone accountable for any GDPR complaint. As pointed by out by [6], it is likely that the onus of compliance will need to be put on the users themselves, through terms of use that: a) prohibit posting of certain kinds of personal data; and b) require users to have consent or another legal basis for processing. Designers and developers of permissionless blockchains might consider to implement techniques described in section 4, like zero-knowledge proofs, to offer at least partial guarantees to individuals. However, it has to be noted that, depending on the particular governance scheme used by the blockchain, changes that make the blockchain *less* compliant might be introduced.

Scenario 2: Applications that use permissionless blockchains as backend In this case, a data subject interacts with an application that uses a permissionless blockchain as backend, *e.g.*, a set of Ethereum Smart Contracts. Role-wise, the owners of the intermediary application can be identified as data controllers, as they are the ones that decide what personal data is collected from the subject, and what parts of it are stored and/or processed in the permissionless blockchain. As such, they are responsible of inform the user about the fact that some of their personal data will be stored inside a blockchain, and what are the strategies in place to hash it, encrypt it, and protect it.

Currently, the only way to guarantee GDPR compliance in this scenario is to hash out any personal data to a server controlled by someone that will be identified as a data controller. To decrease the possibility of pre-image attacks, state of the art *salting* techniques can be used. A *salt* is a random string that is concatenated to data to be encrypted, kept under the control of the data processor. Also, any table that matches pseudonyms (public keys) generated on behalf of data subjects required by the smart contract, to their identities needs also to be stored off-blockchain. In use cases where actual processing (instead of only storage) of personal data is encoded in a smart contract and a hash cannot be used, a possibility is the use of multi-party private computation schemes, however, although the encryption provided by the scheme might be considered enough for GDPR purposes, further research is needed to verify that the level of decentralisation offered by the additional network correspond to expectations.

Scenario 3: Permissioned blockchains This case can be separated in further two, in a similar manner to the difference between scenarios 1 and 2: the first subcase is an individual that joins on his own volition a consortium to run a permissioned blockchain. In this case, the individual needs to agree with the fact that others might process any data he inputs[21], and that he/she will be responsible of validating other's transactions, with the corresponding responsibility in case of a data breach or misuse. A possible role assignment in this case is that all members are joint data controllers. As such, they need to agree in a set of terms where they lay out how they are going to respond to requests of members related to any of the GDPR rights.

The second subcase is when a permissioned blockchain consortium offers services to end-users, storing their personal data in their blockchain. Again, the simplest way to be in good terms with the GDPR a is that members of the consortium declare themselves as joint data controllers.

In any case, the most important ingredient for compliance is common sense. Similar to what a centralised data processor should ask itself regarding personal data, we propose the following checklist for data processors that intend to use Blockchains as part of their technology stack.

1. What personal data will be collected? What part of it will be stored or processed in a blockchain?

2. What processing personal data will undergo in a blockchain? What is the advantage of decentralising that process?

3. What type of blockchain will be used? What is known about the validators in it? Can they be bound to a contractual agreement?

4. If data is going to enter the blockchain encrypted or hashed, who holds the keys (or links to original data)?

[21] Yet again, the implementation of strategies discussed in section 4 might provide some guarantees about how much the others are effectively able to see

5. If the blockchain intends to support a certain type of transaction, are the identities of transactors inferrable from the contents of the blockchain? If pseudonyms are used, who holds the 'linking table' back to data subjects?

6 Blockchains as data protection by design enablers

Article 25 of the GDPR outlines the general obligation for data controllers and processors to implement technical and organisational measures to show that that the integration of data protection into processing activities. Up to this point, we have only considered conflicts and mitigation possibilities between GDPR and blockchains, but recall that blockchains' technologies *leit motivs* include resolving imbalances between centralised intermediaries and the users that depend on them, and providing a mechanism to guarantee trust among parties that not completely trust each other. In the GDPR scenario, data subjects do not completely trust data controllers and are now entitled to more control on how their data is processed. Data controllers and processors would like to have a mechanism to prove to data subjects and data protection authorities that their processing complied with what the data subject granted. Under this line of thought, blockchains may turn to be a powerful tool to implement the postulates of article 25, instead of a headache for technicians and lawyers.

A first possibility in this space is the design of a blockchain where transactions represent transfer of 'data access rights' from data subjects to data controllers, and from data controllers to data processors. A data subject may also include in such transactions an structured description of what processing she allows for her data. The storage of this 'Consent 2.0' in a blockchain ensures it can be verified by all actors in the context of that particular processing. The development of languages and formalisms to capture consent was discussed in a recent W3C workshop on 'Data Privacy Controls and Vocabularies'[22], with the support of the Horizon2020 funded research project Special[23].

Structuring consent information in a machine-readable format allows their input to automated agents, ensuring . In particular, some data processes could be encoded as Smart Contracts, whose trace of execution is registered in the blockchain. Through this, data controllers and processors can demonstrate that processing complies exactly with both the purpose specification and data subject's consent definition. For the processes that can also be executed in a multi-party private computation scheme like Enigma [13] or with homomorphic encryption algorithms that have been studied in the context of Big Data, one can also ensure that further data processors process pieces of data on which they can't reconstruct the original data. This path is being explored by Horizon2020 funded research and innovation actions MyHealthMyData[24] and BodyPass[25] in the context of the exploitation of medical data.

7 Conclusion

In this paper, we have provided an overview of the tensions between Blockchain technologies and the General Data Protection Regulation, reviewed the legal and technological efforts aimed at understanding and mitigating possible conflicts, outlined possible solutions to common interaction scenarios between data subjects and blockchains, based on existing technologies, and described the potential of blockchains to serve as enablers of data protection.

GDPR and blockchains share the motivation of empowering individuals and reduce the assimmetry between them and the organisations that process their data and their transactions. However, some of the technological advancements that make possible decentralisation in blockchains require data processing and storage to be distributed among members of a community or outsourced to unknown individuals or organisations, raising issues on the assignment of GDPR responsibilities, on how to avoid misuse of personal data, and how to comply with the rights of erasure and amendment. Permissioned blockchains appear to be easier to accommodate to GDPR, as their members can be designated joint data controllers, taking responsibility for data protection. On the other hand, in permissionless blockchains, where anonymity is at the core of the design and communities

[22] https://www.w3.org/2018/vocabws/

[23] https://www.specialprivacy.eu/

[24] http://www.myhealthmydata.eu/

[25] http://bodypass.eu/

around them are supranational, it is not possible to guarantee the accountability of every data processor.

Two types of solutions are being actively developed to help achieve compliance: first, integrating different families of cryptographical functions and private computation schemes to blockchains to enable decentralised validation without reveling the contents of transactions. the rationale is that if an outsourced data processor cannot decrypt what it is manipulating, then there is no data protection problems. Other types of hashes could be used to provide erasure and redaction capabilities. Second, use blockchains as decentralised *verification* machines that operate on hashes of data, with real data and the means to link it to the hash being stored under the responsibility of a designated data controller. Similarly to the previous solution, the rationale is that if the outsourced validators only see a hash and can't decrypt the actual data, then data is protected. From an erasure perspective, once the original data is deleted, the hash remaining in the blockchain cannot be linked to the data subject.

However, some questions remain open about the legal validity and the real extent of proposed solutions. Technologists, lawyers, and policy makers to sit on the same table to ensure that we achieve the right balance between guaranteeing EU citizen's rights and not hindering the undeniable innovation that blockchains bring, possibly by preparing an extension of the regulation applicable to the particular scenarios brought by blockchain technologies.

- What is a *sufficient* level of encryption for personal data to be stored in a blockchain? At what of the available level of encryption we can say that brute force would not be 'reasonably likely to be used', thus, enabling the implementation of personal data erasure through key erasure? In our opinion, current state of the art algorithms should be enough, and even considering the possible realisation of quantum computers, the cryptography community is already developing *post-quantum* solutions.

- The same question as above can be asked for hashing. Under which conditions can we confirm that eliminating the original data (and the used salt) is sufficient to guarantee erasure?

- One of the use cases of blockchains is to store an immutable, verifiable copy of a piece of data. This makes them ideal for archiving records where maximum transparency is desired (in this case, that a single party couldn't be trusted for keeping the,). However, this transparency comes at the price of disallowing updates and erasure. In what cases this would be desirable? Or do any blockchain intended to be used for this purpose needs to include mechanisms for redactability?

- Permissionless blockchains have become supranational entities, with their own decentralised governance. The fact that participants remain pseudonymous complicates matters when its time to assign levels of accountability. How to accommodate these new types of entities into the European legal framework? Is it desirable to let end-users decide to enter blockchains where part of their rights cannot be ensured?

Disclaimer: This paper was prepared for the EU Blockchain Observatory & Forum, an initiative of the European Commission, Directorate-General of Communications Networks, Content & Technology.

The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the Commission. The Commission does not guarantee the accuracy of the data included in this study. Neither the Commission nor any person acting on the Commission's behalf may be held responsible for the use which may be made of the information contained therein.

References

[1] A. Esteve, "The business of personal data: Google, Facebook, and privacy issues in the EU and the USA," *International Data Privacy Law*, vol. 7, pp. 36–47, Feb. 2017.

[2] P. De Filippi, "The Interplay between Decentralization and Privacy: The Case of Blockchain Technologies," SSRN Scholarly Paper, Social Science Research Network, Rochester, NY, Sept. 2016.

[3] M. Finck, “Blockchains and Data Protection in the European Union,” *SSRN Electronic Journal*, 2017.

[4] V. Buterin, “Privacy on the Blockchain.” https://blog.ethereum.org/2016/01/15/privacy-on-the-blockchain/, Jan. 2016.

[5] “How new eu privacy laws will impact blockchain: Expert take.” https://cointelegraph.com/news/how-new-eu-privacy-laws-will-impact-blockchain-expert-take, Mar. 2018.

[6] W. Maxwell and J. Salmon, “A guide to blockchain and data protection.” https://www.hlengage.com/_uploads/downloads/5425GuidetoblockchainV9FORWEB.pdf, 2017.

[7] J. D. Nick, “Data-driven De-Anonymization in Bitcoin,” Master’s thesis, Computer Systems Institute - ETH Zurich, 2015.

[8] N. Nchinda, “Exploring Pseudonimity on Ethereum.” https://media.consensys.net/exploring-pseudonimity-on-ethereum-dda257019eb4, July 2016.

[9] A. Biryukov and I. Pustogarov, “Bitcoin over Tor isn’t a Good Idea,” in *2015 IEEE Symposium on Security and Privacy*, pp. 122–134, May 2015.

[10] S. Bojja Venkatakrishnan, G. Fanti, and P. Viswanath, “Dandelion: Redesigning the Bitcoin Network for Anonymity,” *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 1, pp. 22:1–22:34, June 2017.

[11] S. Noether and A. Mackenzie, “Ring Confidential Transactions,” *Ledger*, vol. 1, pp. 1–18, Dec. 2016.

[12] E. Ben-Sasson, A. Chiesa, E. Tromer, and M. Virza, “Succinct non-interactive zero knowledge for a von neumann architecture,” in *23rd USENIX Security Symposium (USENIX Security 14)*, (San Diego, CA), pp. 781–796, USENIX Association, 2014.

[13] G. Zyskind, O. Nathan, and A. Pentland, “Enigma: Decentralized Computation Platform with Guaranteed Privacy,” tech. rep., Enigma.IO.

[14] L.-D. Ibáñez, H. Fryer, and E. Simperl, “Attaching Semantic Metadata to Cryptocurrency Transactions,” in *Workshop on Decentralising the Semantic Web*, 2017.

[15] G. Ateniese, B. Magri, D. Venturi, and E. Andrade, “Redactable Blockchain; or Rewriting History in Bitcoin and Friends,” in *2017 IEEE European Symposium on Security and Privacy (EuroS P)*, pp. 111–126, Apr. 2017.