TU DELFT

MASTER THESIS

Blockchain-based distributed tamper-proof filesystem using threshold encryption

Author:
Angela PLOMP

Supervisor:
Dr. Johan POUWELSE

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in the

Distributed Systems Group
Computer Science

May 17, 2018

Chapter 1

Introduction

1.1 Sensitive medical data

1.1.1 Barbie's hospitalization

In January 2018, Dutch reality star Samantha "Barbie" de Jong received acute medical care in the Haga Hospital in The Hague (De Telegraaf 2018). Her hospitalization was met with great interest from several media companies, who speculated about possible causes. A few weeks later, it turned out that an abnormally large number of Haga Hospital employees had looked into one partical patient's files: Ms. De Jong's. This news resparked a debate about the merits and risks of storing medical data the way we do.

1.1.2 Digitalized medical records

Medical health records, once stored on paper cards in the doctor's office, have moved towards digital files that can be shared between health care providers such as GP's, hospitals and specialized clinics. These records may contain extremely sensitive data: most people would not want others to know if they suffer from stigmatized illnesses like sexually transmittable diseases or mental disorders. In a more practical way, information about someone's medical history may for example have a negative effect on their chances of being hired for a job. In some parts of the world, medical identity theft is a problem. This is when a person uses another person's identity to fraudulently receive health care or prescription drugs. According to a study on medical identity theft from 2016, the last years showed an upward trend in the number of medical identity theft cases in the USA. The main causes for this identity theft are the stealing or abusing of credentials of family members, a data breach at a healthcare provider or the submission of credentials on a phishing page (Ponemon Institute 2016).

1.1.3 Data ownership

A central question with regard to this type of information is: Who owns the data? Many patients feel that they do not control access to their data, but would like to be able to access the data themselves, look at the history of data access and give or deny access permissions to healthcare providers (World Economic Forum 2012). The data is about them, so they feel they should have ultimate control over it. In a particularly bad case, patients that doubts the confidentiality of their records may not make completely honest disclosures, holding back potentially crucial information. On the other hand, the data has been collected and stored by the healthcare providers. They invest time and money into this process. Data ownership should not be seen as a

binary either/or choice. Moreover, the burden of coming up with policies and implementation of these policies lies on the health care provider (Kostkova et al. 2016).

1.1.4 Legal aspects of medical data

In May 2018, the General Data Protection Regulation (GDPR) will come into effect in the European Union, as a replacement of the Data Protection Directive (DPD) of 1995. The DPD already forced EU member states to take into account data protection on computers and other electronic devices (Calder 2016). The GDPR presents six principles that should be adhered to when collecting, storing and processing data. These mainly concern the proportionality of the data gathering for a certain goal and transparency of and consent for the use of the data.

Chapter 2

Problem statement

2.1 Barbie's medical records

The hospital where Ms. De Jong was treated for her medical problems, Haga Hospital, uses ChipSoft's HiX software for the storage and processing of patient's medical records (HagaZiekenhuis 2016). HiX does record the access of users to the digital files. During a routine check, the access to Ms. De Jong records by staff who were not treating her came to light. This violation of her privacy is deemed unacceptable by many people. In this research, my hope is to contribute to the development of a more secure medical record file system in which the patient's involvement is central. Ms. De Jong should easily have access to the log of persons who viewed the record herself. Additionally, she should know the exact contents of these records and agree with their storage.

2.2 Research goal

The goal of this master thesis project, is to research the possibilities of expanding patients' power over and knowledge about their medical records. This power consists of two parts:

1. Accountability on access;
2. Validation of EMR entries from both the health care provider and the patient.

In addition to this, the more traditional requirements for an EMR still stand. For example, the files should be kept secret for unauthorized people through strong encryption.

2.2.1 Accountability on access

Accountability on access means that a patient can verify who has accessed accessed a file, and when. There should be no way for someone to access the file without leaving a trace. When a patient questions the legitimacy of an access event, the person who looked into the file can be asked for an explanation. A recent paper that points out the lack of patient-centered transparency requirements for medical data systems, states that transparency is needed for accountability. The authors define ex-post transparency as *"enabling the patient to be informed or get informed about what happened to his/her medical and personal data"* (Spagnuelo and Lenzini 2016). In order to fulfill this ex-post transparency goal, a number of transparency requirements were formulated. When it comes to the relation between transparency and accountability, the most relevant of these requirements are:

1. The medical record system must provide the patient with accountability mechanisms.
2. The medical record system must provide the patient with evidence regarding permissions history for auditing purposes.
3. The medical record system must provide the patient with evidence of security breaches.

These requirements guide the design of an EMR system that center the patient's need of privacy and power over their own data. Thus, the system proposed in this paper will be evaluated against these criteria.

2.2.2 Validation of EMR entries

Validation of EMR entries means that an entry becomes official only when both the patient and the health care provider have agreed to the entry. This is similar to a person sending a registered letter and the recipient signing for delivery. The patient cannot claim not to know the content of the entry.

2.3 Research question

Taking the aforementioned considerations into account, the research question for this thesis project is as follows:
R: *"How can an Electronic Medical Record (EMR) system be designed, that guarantees accountability on access and validation on entry addition?'*

This question can be split into two subquestions:
R1: *"How can accountability on access be guaranteed in an EMR?"*
R2: *"How can entries be validated by a patient as well as a healthcare provider in an EMR?"*

In Chapter 2: Previous work, the existing literature on these topics is explored. A possible solution is proposed in Chapter 3.

Chapter 3

Previous work

3.1 Blockchain

Considering that we are looking for a system that ensures that access to it is being logged in a tamper-proof way, a technology that comes to mind is blockchain. Blockchain emerged in 2008 with the implementation of the first cryptocurrency, Bitcoin. Essentially, blockchain is a peer-to-peer distributed ledger, which can only be updated via consensus (Nakamoto 2008). It runs as a layer on top of TCP/IP. Blockchains can be public, private or semi-private. Anyone can participate in a public (or permissionless) blockchain: all participants hold a copy of the ledger but none of the participants actually own the ledger. This ensures the decentralized nature of the blockchain. A private blockchain is open only to an organization or consortium. Semi-private blockchains are a combination of a public and private part (Bashir 2017). A block minimally consists of:

1. The hash of the previous block;
2. A nonce (number used only once);
3. A bundle of transactions.

The first block in a blockchain is called the genesis block. This is hardcoded at the time the blockchain was started. To add a block to the blockchain, all nodes must agree on a single version of truth. There are roughly two categories of consensus mechanisms: Proof- and leader-based or Byzantine fault tolerance-based Bitcoin uses the proof-of work consensus mechanism to prove that enough computational resources have been spent in order to propose an addition to the blockchain. Nodes can compete with each other to be selected in proportion to their computing capacity. For Bitcoin, the proof-of-work requirement is as follows: H (N || P_{hash} || Tx || Tx || . . . Tx) < Target. N represents a nonce, P_{hash} is the hash value of the previous block and Tx are the transactions in the proposed block. The hash value of these concatenated fields should be smaller than the set Target for difficulty. This problem cannot be solved with a smart algorithm: it must be brute forced. A major quality of this system is the effectiveness against Sybil attacks as a result of the high costs of creating pseudonymous identities (Vukolić 2015). A drawback is that it is (obviously) computationally intensive, and therefore uses much energy, which is an unnecessary strain on the environment. The proof-of-stake algorithm uses the stake that a user has in the system, for example invested time, to trust that the benefits of performing malicious activities would not outweigh the benefits of staying in the system as a trusted member (Kiayias et al. 2017).
Deposit-based consensus requires putting in a deposit before proposing a block to be added to the blockchain. In case the block is rejected by others, the user loses its deposit (Solat 2017). Reputation-based mechanisms let members elect a leader

node, based on the reputation it has built on the network. When a transaction is added to a block, it should be clear who has performed this transaction. Particularly in the medical use case, any access to the EMR should be linked to an identity. A digital signature confirms the identity, under the condition that such a signature can be verified but cannot be forged. Digital signatures can be issued using different algorithms. Bitcoin uses the Elliptic Curve Digital Signature Algorithm (ECDSA). Adding a block to the blockchain is done through the following consensus algorithm (Nakamoto 2008): new transactions are broadcast to all nodes; each node collects transactions into a block; in each round, a random node (selected by the proof-of-work) gets to broadcast its block; other nodes accept the block if and only if all transactions in it are valid; nodes express their acceptance of the block by including its hash in the next block they create. As a rule of thumb, a block is 'permanently' added if it has been in in the blockchain for six rounds. The probability of another version of the blockchain, not containing this particular block, becoming longer and thus the official blockchain, is negligible. Because every block contains a hash pointer to the previous block, one can access the previous information, but also verify that it has not changed. Tampering is evident because the hash of the changed information would change, too. A binary tree with hash pointers is called a Merkle tree. Advantages of Merkle trees are: a Merkle tree can hold many items, but one just needs to remember the root hash one can verify membership of the tree in just O(log n) time and space (Szydlo 2004). Athough data can be stored in a blockchain directly, a blockchain is not suitable to store large amounts of data. This is why many blockchain-based systems use a distributed has table (DHT).

3.2 Blockchain-based EMR systems

3.2.1 Scientific work

This research would definitely not be the first to incorporate blockchain into a EMR system. A white paper from Ekblaw identifies interoperability challenges between healthcare provider systems as a major barrier towards effective data sharing. They designed a public key cryptography-based blockchain structure that could be applied to create append-only, immutable, timestamped EMRs (Ekblaw et al. 2016). The block content consists of information about data ownership and viewership permissions. Zyskind & Nathan proposed a model called OpenPDS for an information system in which a mechanism for returning computations on the data is included: return answers instead of data itself. The contribution of this paper is twofold: Combination of blockchain and off-blockchain storage to construct a personal data management platform focused on privacy; Perform trusted computing on blockchain-handled data. The proposed systems treats users as the owners of their data and provides them with data transparency and fine-grained access control. A rough sketch of the functionality of the system is as follows: A users installs the application on a smartphone. Data collected on the phone is encrypted using a shared encryption key and sent to the blockchain. The blockchain routes it to an off-blockchain key-value store using a DHT, only retaining a SHA-256 hash pointer. Anyone wanting to access the data can send a request to the blockchain, which in turn verifies the digital signature of the requester as well as the listed permissions for this user (Zyskind, Nathan, et al. 2015). Assuming that users manage their keys in a secure manner, the system provides security and privacy. An adversary cannot really learn interesting information from the blockchain itself, because it only stores hash pointers. Even if it would control a large amount of nodes, the raw data is still

encrypted using a key that none of the nodes possess. Adversaries are prevented from posing as a user because of the digitally-signed transactions and the decentralized nature of blockchain. In 2016, Xiao Yue presented a fairly similar system called the Healthcare Data Gateway app. It is a combination of a traditional database and a gateway. Personal electronic medical data is managed by a blockchain. All data requests are evaluated for permission. In case of a granted permission, secure multiparty computation (sMPC) is used to process patient data without risking patient privacy (Yue et al. 2016). Enigma is a computation platform proposed by Zyskind. Their paper states that blockchain can neither handle privacy nor heavy computations. Enigma can be connected to an existing blockchain. The goal of the platform is to facilitate developers to build privacy-by-design, decentralized applications without using a trusted third party (Zyskind, Nathan, and Pentland 2015). Just like most blockchain-based systems, it uses a DHT that stores references to the data. sMPC is used by splitting date between nodes and performing computation on these nodes without transferring any information from one node to another. Each node has a piece of seemingly random data, that is useless on its own. In general, sMPC systems are based on secret sharing. This is a category of threshold cryptosystems, in which a secret s is divided into n parts, and at least t shares are required to reconstruct s. Such a system is written as a (t, n) threshold system. Shamir's secret sharing scheme is a famous example of a secret sharing scheme, which uses polynomial interpolation. The Enigma platform provides an API which facilitates the uses of a sharing scheme based on Shamir's scheme. In total, there are three decentralized databases in the system: the public ledger, the DHT and the sMPC database. Nodes are compensated for their computational resources via computation fees.

3.2.2 Startups and industry-based projects

Several startups and government- or industry-based projects have come up in the last few years on the subject of blockchain in healthcare. These range from conceptual frameworks to functioning prototypes. A few Dutch projects are listed here.

Mijn Zorg Log

Mijn Zorg Log is a smartphone app, developed by the Dutch National Health Care Institute. The app can be used by people who receive home care to log the hours that the home help spent at their house and the nature of the care. The home care provider can then verify these hours and use them for their administration.

MedMij

MedMij is a framework that consists of agreements about how medical data should be exchanged in a blockchain-based healthcare application. It is therefore not a working product. Health care providers that want to develop a digital healthcare application, can hire a MedMij-certified vendor to implement a compliant system.

3.3 Identities and signatures

3.3.1 Self-sovereign identities

Accountability on access can only be established when it is guaranteed that the person accessing or modifying the file is indeed the person who is recorded as doing so.

This means that we will need a solid identification and authentication method for the file system. Traditionally, this goal has been attained by using username/password systems. There are several drawbacks to this system. It provides a terrible user experience for many people, especially if they have to memorize a large amount of passwords and change them regularly. This sometimes leads to irresponsible password behaviour (Adams and Sasse 1999). Another issue is that a user has to create a new identity for each application. These identities only exist within the context of each specific website or application, leading to great volumes of data duplication (Tobin and Reed 2016).

3.3.2 Digital signatures

As paperwork has been replaced by digital entries, digital signatures have taken over the role of traditional signatures. A digital signature provides proof of the integrity of the authorship, because anyone can verify that the signature is based on the author's public key. On the other hand, only the person who creates the message should be able to generate a valid signature. In general, the steps to create a digital signature are as follows:

1. The signature algorithm is a function of the signer's private key k_{pr}. Hence, only one person can sign a message x, assuming that the private keys are kept secret.

2. The message x is an input to the signature algorithm as well, to make sure that the signature is related to the message and cannot be re-used.

3. A digital signature algorithm is run with the right inputs, which yields signature s. Then, s is appended to x and the pair (x, s) can be sent.

Digital signatures can be created using a range of different algorithms, based on for example prime factorization (RSA-based signatures) or the discrete logarithm problem (ElGamal-based signatures) or on the elliptic curve discrete logarithm problem.

3.3.3 Elliptic Curve Digital Signature Algorithm

Elliptic curves have some advantages over RSA and discrete logarithm-based schemes. One of these advantages is that a small key length provides the same security as other schemes, but with a shorter processing time. The Elliptic Curve Digital Signature Algorithm (ECDSA) is defined over prime fields as well as over Galois fields. Here, the procedures for the more popular version over prime fields are given (Paar and Pelzl 2009).

1. For key generation, an elliptic curve E is chosen with modulus p, coefficients a and b and a point A which generates a cyclic group of prime order q. Choose a random integer d such that $0 < d < q$. Compute the new point $B = dA$.
   $k_{pub} = (p, a, b, q, A, B)$
   $k_{pr} = (d)$

2. In order to generate a signature, an integer such that $0 < k_E < q$ is chosen as an ephemeral key. Compute $R = k_E A$. Let $r = x_R$ (the x-coordinate of point R) and compute the signature $s \equiv (h(x) + d \cdot r)k_E^{-1} \bmod q$.

The main analytical attack against ECDSA, assuming that the parameters are chosen correctly, is trying to solve the elliptic curve discrete logarithm problem. Considering that this is an NP-complete problem, it is extremely unrealistic to solve this in time.

3.3.4 Elliptic curve threshold signatures

Similarly to the threshold encryption schemes discussed before, threshold cryptography can be applied to digital signatures. A scheme to achieve this was first presented in 1992 by Desmedt & Frankel. This method was based on the RSA signature scheme (Desmedt and Frankel 1991). Since then, many papers have been published presenting threshold signature schemes. One of them was a robust Elliptic Curve threshold DSA scheme. For this project, the focus will be on Elliptic Curve threshold signature schemes, because of the previously mentioned advantages. Specifically, a scheme is needed which is fit to execute on a distributed system.

3.3.5 Threshold ECDSA in a fully distributed system

In 2015, researchers at the Worcester Polytechnic institute presented a fully distributed signature system for threshold ECDSA, named *Nephele* (Green and Eisenbarth 2015). This system is mainly built to protect the key from side-channel attacks and is designed in such a way that a private key never even needs to appear in memory. The key generation as well as the signature generation algorithm is fully distributed. It also allows for fully distributed key re-sharing.

Key generation

The private key is chosen by all the nodes together using Joint Random Secret Sharing (JRSS). In this technique, each node chooses a random local secret value and shares it with the group, using Shamir's Secret Sharing (Gennaro et al. 1996). Every node adds all the shares together (including its own), resulting in the joint random secret share. Just one of the nodes needs to introduce randomness to keep the joint secret unknown.

Chapter 4

Design choices

4.1 Blockchain choices

There are several ready-to-use blockchain libraries available that could be used for this project.

4.1.1 TrustChain

Researchers at TU Delft developed TrustChain, a scalable blockchain with an emphasis on resilience against one of the primary challenges in permissionless blockchains: Sybil attacks (Otte 2017). A Sybil attack takes place when an adversary forges many fake identities to gain a larger influence of that system than it should actually have (Douceur 2002). The author states that when there is no central trusted authority to assert the one-on-one correspondence between an entity and its identity, it is practically impossible to distinguish identities. This poses a fundamental problem for permissionless blockchains, because they are fully decentralized.

4.2 Digital signature algorithm

Threshold ECDSA. Benefits (key length etc). Choice of elliptic curve.

Bibliography

Adams, A. and M. A. Sasse (1999). "Users are not the enemy". In: *Communications of the ACM* 42.12, pp. 40–46.

Bashir, I. (2017). *Mastering Blockchain*. Packt Publishing Ltd.

Calder, A. (2016). *EU GDPR A Pocket Guide*. IT Governance Ltd.

De Telegraaf (2018). "Barbie met spoed naar ziekenhuis gebracht". In: *De Telegraaf*.

Desmedt, Y. and Y. Frankel (1991). "Shared generation of authenticators and signatures". In: *Annual International Cryptology Conference*. Springer, pp. 457–469.

Douceur, John R (2002). "The sybil attack". In: *International workshop on peer-to-peer systems*. Springer, pp. 251–260.

Ekblaw, A. et al. (2016). "A Case Study for Blockchain in Healthcare:"MedRec" prototype for electronic health records and medical research data". In: *Proceedings of IEEE Open & Big Data Conference*. Vol. 13, p. 13.

Gennaro, R. et al. (1996). "Robust threshold DSS signatures". In: *International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, pp. 354–371.

Green, Marc and Thomas Eisenbarth (2015). "Strength in Numbers: Threshold ECDSA to Protect Keys in the Cloud". In: *IACR Cryptology ePrint Archive* 2015, p. 1169.

HagaZiekenhuis (2016). "HagaZiekenhuis stapt succesvol over naar EPD HiX". In:

Kiayias, A . et al. (2017). "Ouroboros: A provably secure proof-of-stake blockchain protocol". In: *Annual International Cryptology Conference*. Springer, pp. 357–388.

Kostkova, P . et al. (2016). "Who owns the data? Open data for healthcare". In: *Frontiers in public health* 4, p. 7.

Nakamoto, S. (2008). "Bitcoin: A peer-to-peer electronic cash system". In:

Otte, P. et al. (2017). "TrustChain: A Sybil-resistant scalable blockchain". In: *Future Generation Computer Systems*.

Paar, Christof and Jan Pelzl (2009). *Understanding cryptography: a textbook for students and practitioners*. Springer Science & Business Media.

Ponemon Institute (2016). *Sixth Annual Study on Privacy and Security of Healthcare Data*.

Solat, S. (2017). "RDV: Register, Deposit, Vote: a full decentralized consensus algorithm for blockchain based networks". In: *arXiv preprint arXiv:1707.05091*.

Spagnuelo, Dayana and Gabriele Lenzini (2016). "Patient-centred transparency requirements for medical data sharing systems". In: *New Advances in Information Systems and Technologies*. Springer, pp. 1073–1083.

Szydlo, M. (2004). "Merkle tree traversal in log space and time". In: *International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, pp. 541–554.

Tobin, A. and D. Reed (2016). "The Inevitable Rise of Self-Sovereign Identity". In: *The Sovrin Foundation*.

Vukolić, M. (2015). "The quest for scalable blockchain fabric: Proof-of-work vs. BFT replication". In: *International Workshop on Open Problems in Network Security*. Springer, pp. 112–125.

World Economic Forum (2012). *Rethinking personal data: A new lens for strengthening trust*.

Yue, X. et al. (2016). "Healthcare data gateways: found healthcare intelligence on blockchain with novel privacy risk control". In: *Journal of medical systems* 40.10, p. 218.

Zyskind, G., O. Nathan, and A. Pentland (2015). "Enigma: Decentralized computation platform with guaranteed privacy". In: *arXiv preprint arXiv:1506.03471*.

Zyskind, G., O. Nathan, et al. (2015). "Decentralizing privacy: Using blockchain to protect personal data". In: *Security and Privacy Workshops (SPW), 2015 IEEE*. IEEE, pp. 180–184.