Designing an Electronic Medical Record system with tamper-proof logging and information on a need-to-know basis

Research proposal

Angela Plomp
4105419

Introduction

Sensitive information in 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.

Ownership of the data

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).

Information on a need-to-know-basis

A user may want to give some relevant information to a healthcare provider, but keep non-relevant information that is also stored in the record private. Take a look at the following example. An aspiring soldier is being subject to a health check in order to make sure that (s)he fits the requirements. Next to a physical fitness test, the doctor wants to check whether the aspiring soldier has suffered from any perasonality disorder at any time during the past 10 years. A simple 'yes' or 'no' is everything the doctor needs: whether this concerns a paranoid disorder, borderline disorder or antisocial disorder is irrelevant, because they all make the applicant unfit for the job. Therefore, obtaining the exact nature of the condition is an unnecessary privacy breach (Lindell & Pinkas, 2009).

Accountability

When a healthcare provider accesses the data in an EMR, it should be possible to trace this action back to the responsible person. In order to ensure accountability, the access should be logged in such a way that it cannot be tampered with. A healthcare provider should not be able to hide the access or give a false name or timestamp. Ideally, one should not be able to alter an entry in the logs at all (Zyskind & Nathan, 2015).

Research question

Taking the aforementioned considerations into account, the research question for this thesis project is as follows:

“How can an Electronic Medical Record (EMR) system be designed, that guarantees accountability on access and provides information on a need-to-know basis?”

After exploring related work on this topic in section 2, this research question will be decomposed into subquestions in section 3. A timeline is presented in section 4, a short indication of testing and evaluation in section 5 and finally an overview of risks and mitigation in section 6.

Related work

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 permission-less) 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:

- The hash of the previous block
- A nonce (number used only once)
- 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 (Bashir, 2017): 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:

$H (N \,||\, P\_hash \,||\, Tx \,||\, Tx \,||\, \ldots \, Tx) < Target$

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 (Bentov et al., 2014).

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, 2014)

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).

Blockchain-based EMR systems

This research would definitely not be the first to incorporate blockchain into a EMR system.

A white paper from Ekblaw et al. (2016) 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. The block content consists of information about data ownership and viewership permissions.

Zyskind & Nathan (2015) 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. This paper is probably the closest related to the proposed research. 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.

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 and in case of a positive permission, secure multiparty computation (sMPC) is used to process patient data without risking patient privacy.

Enigma is a computation platform proposed by Zyskind et al. (2015). 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. 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. Samir’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 Sharmir’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.

Research question

The research question is as follows:

R: How can an Electronic Medical Record (EMR) system be designed, that guarantees accountability on access and provides information on a need-to-know basis?

The goal of this thesis project is to design an EMR system and provide a basic but functional proof-of-concept.

Decomposition of the research question:

- *R1: How can accountability on access be guaranteed in an EMR system?*
- *R2: How can an EMH record provide the minimal needed information?*

R1: How can accountability on access be guaranteed in an EMR system?

In the previous section on related work, we have seen that blockchain is an attractive solution for achieving a tamper-proof log that enables accountability in a system. However, blockchain is not a one-size-fits-all technology. One must give an answer to the following questions:

- *Should a public or (semi)private blockchain be used?* Considering that the system would be used by patients and health care providers, a permissioned blockchain would seem sensible.
- *Which consensus algorithm should be used?* One of the mechanisms like proof-of-work or proof-of-stake should be chosen. This also depends on the type of blockchain that we chose in the previous question.
- *What kind of off-blockchain storage solution should be used?* A blockchain is not suitable for storage of sensitive information. Therefore, we should investigate ways to store the data itself in a secure manner and let the blockchain link to it, for example using a DHT.

R2: How can an EMH record provide the minimal needed information?

In the related work section, we have seen multiple ways to retrieve only the needed answer to a question instead of raw data, especially sMPC. The selected method should fit the following criteria: it should not leak data and it should be correct. Zero-knowledge proofs could be a suitable way of verifying the correctness of the data. Literature study should be conducted in order to make an informed decision on the most suitable implementation.

After answering these questions, a proof-of-concept of the system will be built. This will most likely be done in Java or Python because of the personal preferences and experience of the programmer.

Timeline

A thesis project comprises of nine months full-time research activity. Although it is difficult to make an accurate estimation of the time that each activity will take to completion, here is a rough sketch of the activity distribution during the set time:

- **First month:** Literature study. The goal is to find appropriate literature about the subject and acquire basic knowledge on blockchain technology and existing EMR systems.
- **Second month:** Literature study, with a focus on deepening knowledge about what has been done on this topic and what possibilities for improvement can be identified. Write chapter on related work.
- **Third month:** Answering R1 and starting with the design of the system.
- **Fourth month:** Answering R2 and working on the design of the system. Finalize research objective of thesis and write problem description chapter of thesis.
- **Fifth month:** Finishing the design of the system.
- **Sixth month:** Building the proof-of-concept and writing the thesis.
- **Seventh month:** Theoretical model, algorithm validation.
- **Eighth month:** Writing thesis.
- **Ninth month:** Writing thesis and preparing for the thesis defense.

Testing and evaluation

Because the personal interests of the researcher are mostly on the theoretical side, the bulk of this research will consist of the design of an EMR system instead of its implementation. The algorithms will be evaluated using the proof-of-concept, to check for the security and correctness of the resulting computations.

During the course of the project, medical professionals will be approached to check the correctness of the assumptions and the relevance of the problem statements.

Risks and mitigation

The risks involved in this research are very small, for the following reasons.

First of all, no human test subjects will be used, eliminating the risks of physical harm.

Second, no real data will be gathered. When the proof-of-concept system will be tested, dummy data will be used to fill in the ‘patient records’. This way, the burden of privacy regulation is avoided. The non-profit organization OpenMRS provides a test data set for medical records which consists of over 5000 patients (OpenMRS, 2017).

Of course, if the system would be used in real life, it should be thoroughly revised because the consequences of a bug in the system could potentially endanger the privacy of the user. However, the proof-of-concept is explicitly only meant as a demonstration and not as a ready-to-use system.

References

Bashir, I. (2017). Mastering blockchain. Packt Publishing.

Bentov, I., Lee, C., Mizrahi, A., & Rosenfeld, M. (2014). Proof of Activity: Extending Bitcoin's Proof of Work via Proof of Stake [Extended Abstract] y. *ACM SIGMETRICS Performance Evaluation Review*, *42*(3), 34-37.

Ekblaw, A., Azaria, A., Halamka, J. D., & Lippman, A. (2016, August). 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*.

Kostkova, P., Brewer, H., de Lusignan, S., Fottrell, E., Goldacre, B., Hart, G., ... & Ross, E. (2016). Who owns the data? Open data for healthcare. *Frontiers in public health*, *4*.

Lindell, Y., & Pinkas, B. (2009). Secure multiparty computation for privacy-preserving data mining. *Journal of Privacy and Confidentiality*, *1*(1), 5.

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

OpenMRS (2017) Demo data. Retrieved from https://wiki.openmrs.org/display/RES/Demo+Data.

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

Szydlo, M. (2004, January). Merkle tree traversal in log space and time. In *Eurocrypt* (Vol. 3027, pp. 541-554).

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

World Economic Forum (2012). *Rethinking personal data: Strengthening trust.* Retrieved from http://www3.weforum.org/docs/WEF_IT_RethinkingPersonalData_Report_2012.pdf.

Yue, X., Wang, H., Jin, D., Li, M., & Jiang, W. (2016). Healthcare data gateways: found healthcare intelligence on blockchain with novel privacy risk control. *Journal of medical systems*, *40*(10), 218.

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

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