ESSAY

Optimizing Disease Surveillance by Reporting on the Blockchain

Flávio C Coelho

Correspondence: fccoelho@fgv.br
Applied Mathematics School, Getulio Vargas Foundation, Praia de Botafogo, 190, Rio de Janeiro, Brazil
Full list of author information is available at the end of the article

Abstract

Disease surveillance, especially for infectious diseases, is a complex and inefficient process. Here we propose an optimized, blockchain-based monitoring and reporting system which can achieve all the desired features of an ideal surveillance system while maintaining costs down and being transparent and robust. We describe the technical specifications of such a a system and discuss possibilities for its implementation. Together with a token based incentive system, it is possible to rewards data quality as well as build a marketplace for data analysis which will help finance the surveillance system. Finally, the impact of the adoption of distributed ledger technology for disease surveillance is discussed.

Keywords: Blockchain; DAG; Disease Surveillance; Merkle tree

Introduction

Technologies commonly known as blockchains are implementations of distributed ledgers for the safekeeping of information, in a manner which does away with the need for centralized management and are censorship resistant. These features are achieved by means of cryptography and sophisticated consensus algorithms which guarantees the consistency of the distributed records.

These Technologies are finding a growing number of applications in many different areas[2]. The first successful application of a distributed ledger system was a digital currency, Bitcoin, which was proposed in 2008[7], and has since spread all over the world.

Blockchain technology has started to attract the attention in health management circles. An obvious application is the management of electronic health records [5]. Peterson et al. [9] argue that blockchains can solve all healthcare data exchange problems.

Disease surveillance, especially for infectious diseases, is a complex and inefficient process. This is due to the fact that such systems involve a large number of independent agents which must report to a centralized information system. The lack of a clear set of incentives, it can be a challenge to keep the information flow timely and accurate.

Surveillance systems are also typically composed of multiple sub-systems or stages, each focusing on a different level of Information management and analysis[8]. In this paper we will discuss only issues related of case monitoring and reporting, which is usually the first stage in any surveillance system.

Disease monitoring systems must aggregate data coming from a large network of agents which must report disease cases according to a pre-established protocol.

The data received must be validated and then made available to health officials to help manage their response to public health demands. Due to the sensitivity of the information, These large databases are subject to central control which is usually where all chronic problems of disease surveillance systems come from. Among the main problems brought about by central control we can list: a single point of failure; delays in the aggregation of information; lack of transparency; lack of responsiveness to failures; to cite but a few.

Instead of going through the long list of shortcomings of current surveillance information systems, we can instead try to elicit the desiderata for an ideal surveillance system [3]. As we are focusing only on the case reporting stage, the key features we envision are the following.

Provenance: origin of a case report must always be known. This means effective geographical location but also the id of the health establishment and/or professional collecting the data.

Timing: The date and time of reporting must be accurately known. This time stamp is not the same as the date of the onset of symptoms, but knowing it allows for the assessment of the readiness of of health services.

Uniqueness: Each case report must be uniquely identifiable, to allow for effective data management.

Selective privacy: Some informations on the case report must remain private, since they consist of PII[1]. but other fields can be publicized. Privacy issues should not stand in the way of accessibility of the public information.

Queryability: Data must be queryable in an expressive way, i. e., allowing for filtering, grouping, ordering, etc.

Coverage: The network of health professionals reporting on cases, must maximize capillarity and be cheaply expandable.

Incentives: Health professionals must be incentivized to report promptly and accurately.

Consensus: Data must be validated so there is no disagreement about a report. The validation mechanism must also guarantee the veracity of the data as much as possible.

Versioning: Case reports are usually revised. There must be an easy way to update reports without losing original data.

Durability: The possibilities of data loss must be minimal. Long term retention of data must be guaranteed.

In this paper we are going to discuss how the adoption of a blockchain-based surveillance system can achieve all the desiderata listed above, while being a robust, transparent and cheap solution.

Architecture

Distributed ledger systems come in two basic types of architectures, blockchains, of which a good example is Ethereum[10], and DAGs. In the former, transactions are grouped into blocks which are then linked to form a linear chain. In the latter, the transactions are connected to other individual transactions forming a Directed Acyclical Graph (DAG)[4].

[1] Personally identifiable information

In the proposed surveillance system, we believe DAG-based ledgers are the ideal solution but it is not a requirement since both architectures of distributed ledger provide the required features to achieve what is needed. In the following text we will assume a DAG architecture, for simplicity.

Let's start with some definitions. Let us call a *node*, each reporter in the surveillance network, it can be a health professional, a clinic or a hospital.

Each *node* will have an *address* in the DAG address space, and they are connected amongst themselves forming a *network*.

Let a *transaction* mean a case report broadcast to the network of nodes.

A *private key* is associated to each address, and can be used to *sign* a *transaction*. Each transaction is signed and timestamped before being broadcast to the network.

Before a transaction can be added to the DAG, it must be validated by a minimum number of nodes The validation algorithm is described below. The *DAG* is the permanent record of transactions or reports it is a Directed Acyclical Graph. Every node maintains a complete copy of the DAG, thus giving robustness to the permanent record.

It is important to note that there are two networks at play here: The network of reporting nodes and the DAG wich is a network formed through the interconnections of transactions.

Case Reports Data Structure and Content

Case report data is stored within its transaction data structure (figure 2) in encrypted form in a data structure know as a Merkle tree (figure 3). In a Merkle tree the data is a tree where each node is a hash of the sum of of their children node's hashes. The root of the tree serves as an unique id for the entire record, since any other transaction differing on at least one node will yield a different root hash.

Its interesting to notice that if we group the leaves (terminal nodes) by category we can have branch roots which act as category tags or IDs. In the example of figure 3, we select all reports from the same patient by filtering on the patient root. Besides the patient and case roots shown in the example tree, we can have a location root, a clinical symptoms root etc. defined according to the reporting protocol. These roots can be used to filter or group cases based on any of the categories they represent.

The transaction hash also include the hash(es) of parent transactions (figure 2) thus serving as confirmations for the parent transactions and their grand parents all the way to the *genesis* transaction (the first transaction in the DAG). The hash of the report ID. If someone tries to change a record, thus changing is hash root, it will break compatibility with all the descendant transactions. As the number of descendants to a transaction grows, a larger number of transaction would have to rewritten, which would required the coordination of a large number of users or stealing all of their private keys. Transactions are signed by the reporter's private keys, which is derived from their public address.

Now we are going to show how the features of a DAG-based ledger, can help us achieve the desiderata elicited before.

Provenance

As stated before, transactions originate from nodes, nodes are associated to an address on the chain, and the address to a private key. This private key is used to sign a transaction (case report), making it impossible for anyone to impersonate a node and create fake reports, without first stealing the the private keys of a node. It is possible for anyone to verify a transaction from its hash (ID) and the public key (address) of the node. Another important feature derived from the signing of transactions is "non-repudiation", that is, the node cannot deny having reported a case.

Timing

Reports are timestamped. So the exact time of the report is registered automatically and cannot be manipulated by the node. Timestamps are used in the validation algorithm. Accurate timing does not imply synchronous operation, on the contrary, since the timestamp is generated locally in the node it allows for asynchronous update of the ledger without loss of the proper temporal ordering of events.

Uniqueness

Reports are born unique due to the combination of the timestamp, node ID and other data from the report the combined hash of these fields plus the hashes of the parent nodes make up a unique transaction ID (see figure 2). Denoting the hash function used by H the ID of the transaction can be defined as

$$ID_t = H(MTR \parallel PH \parallel NA)$$

where MTR is the Merkle tree root (see fig. 3), PH is the parent transaction hash, and NA is the node address.

Selective Privacy

The case report data payload, described in detail below, contains all PII in encrypted form. Information such as timestamp, reporting establishment's ID and other pieces of information deemed public in compliance with local legislation, can remain unencrypted and thus readable by anyone.

Queryability

Due to the mixed open/encrypted composition of the data. The records allow for extensive SQL-like queries. Open fields can be fully queried, and encrypted fields although not prone the filtering, can be used for groupings, for example one can find the number of cases per household without knowing where the households are located.

Coverage

Traditional information systems rely of fixed (trusted) computer terminals and local network connections. The blockchain client nodes can be packed in a simple mobile App, it can be deployed at next to zero cost. Since it does not require centralized supervision, it can scale without limits even to places where connectivity is deficient, since constant connectivity is not required.

Incentives

All collaborative systems require some form of incentive to work. In this case, a smart-contract associated with the distributed-ledger will emit tokens in exchange for the validation of reports(transactions). the basic validation algorithm for a report to be included in the DAG is fixed. But other higher level validation tasks can be defined and rewarded via the smart-contract.

Consensus

Reports of the same disease for the same patient at different places at roughly the same time can be detected and automatically pruned according to some predetermined criterion. The veracity of a report can be guaranteed by a special type of validation operation, done by health professionals, and rewarded by tokens.

Versioning

Updates to reports which have already been reported can be registered into the DAG as a new report with a reference to the unique ID of the original report thus the full history of the case gets preserved. When querying for cases, updated reports are clearly identifiable and can be used instead of the original reports.

Durability

The records stored on the block chain are permanent, and replicated on a large number of nodes making them robust to data corruption or loss.

Validation algorithm

The validation of a report is done by its child nodes, as a new report is created it is required to validate at least two other reports. the validation consists in the following steps:

1. Checking if the signature on the transaction corresponds to the address of the reporter.
2. Check that the referenced parent transactions exist, and have a timestamp earlier than the current transaction.
3. Validate the parent hashes, by re-hashing the ancestors chain.
4. Check that data root hash is valid.
5. Check that the transaction root hash is valid.

If any of the checks above fail, the verification fails.

Deployment

The implementation and deployment of this distributed-ledger-based surveillance system can rely on different ledger platforms. Even though we presented the solution assuming a ledger with a DAG topology, the system can be deployed on a classical linear blockchain, such as Ethereum's. The main requirement is that the platform provides a way to store the arbitrary data connected to a transaction.

In order to deploy on Ethereum, we can deploy a smart contract that registers the open data in a transaction input data or in contract public variables. The drawback, of this approach is that it would impose a variable cost to registering each report on the blockchain. The alternative would be to store the report data on a peer-to-peer distributed file system, such as IPFS [1], and store only the location hash in the Ethereum transaction or contract.

For A DAG-based ledger, we have a few options: Byteball [4] is such a distributed ledger that is designed from start to be a tamper proof registry of data. Another platform in this category is Iota [6] which also meets the necessary requirements.

We should not rule out the possibility of developing a custom distributed ledger from scratch for data storage, but leveraging more established blockchains such as Ethereum or even Bitcoin simply to save the state of the ledger periodically.

The incentive layer must be integrated with the report workflow. ERC-20 tokens can be used, but in order to give value to the tokens we propose the creation of marketplace for data-analysys. Individuals buy tokens with which to pay for analytical services. Nodes then can also be rewarded for report quality, which is determined by health professionals wanting to earn tokens. Other services can gradually be built on top of this token economy.

Discussion and Conclusion

Adopting a distributed ledger to record disease case reports can bring several advantages over the current information systems backing disease surveillance. Among the main advantages is the immediate validation and availability of data, which can lead to faster responses of health systems during public health emergencies.

The distributed storage of the database offers greater transparency, through data locality, and also eliminates the problem of having central servers as a single point of failure. This means that reporting will never be delayed by system failures and is always accessible for reads.

The existence of a blockchain-based surveillance system such as the one described here is not incompatible with the maintenance of the old system, since it possible to have the reports sent simultaneously to the two databases. But we believe that gradually the distributed surveillance system will make the traditional information system obsolete.

One important drawback of DAG-based ledgers currently available is the lack of support for smart-contracts which would limit the possibilities to add a more complex system of incentives for data maintenance.

We have only looked at the optimization of the first stage of a disease surveillance system through the adoption of distributed ledger technology. However once this stage is on a blockchain, other possibilities open-up, for example a marketplace for analytical models could be built on top of these open records. Any predictive model in this marketplace would be automatically comparable since they are using identical data. If the source code for the models is stored in public repositories they can be easily updated whenever the data changes. Moreover, Public Health agencies could fund the development of analytical models, directly through smart-contracts which would control the validation of the results releasing payments as the research project achieves pre-determined milestones, which can be validated automatically.

In summary, surveillance systems can reap great benefit from a secure system for continuous release of data. The benefits of open public data are well established, but we believe that distributed ledger technology is the key to open sensitive datasets in an effective and secure way.

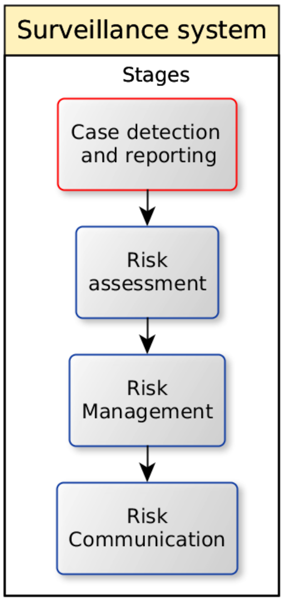

Figure 1 Structure of a surveillance system. We are concerned only with the red block.

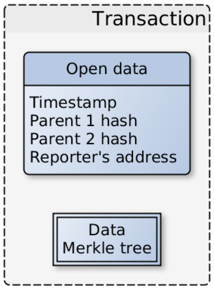

Figure 2 Contents of a transaction. Emphasizing open elements which goes into the determination of the transaction ID.

References

1. Juan Benet. Ipfs-content addressed, versioned, p2p file system. *arXiv preprint arXiv:1407.3561*, 2014.
2. C Catallini. How blockchain applications will move beyond finance. *Harvard Business Rev*, 2, 2017.
3. Centers for Disease Control CDC. Guidelines for evaluating surveillance systems. *MMWR supplements*, 37(5):1, 1988.
4. Anton Churyumov. *Byteball: a decentralized system for storage and transfer of value*. 2016.
5. Ariel Ekblaw, Asaph Azaria, John D. Halamka, and Andrew Lippman. *A Case Study for Blockchain in Healthcare:MedRec prototype for electronic health records and medical research data*, volume 13, page 13. 2016.
6. B Kusmierz. The first glance at the simulation of the tangle: discrete model. *http:/iota.org*, 2017.
7. Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. *https://bitcoin.org/bitcoin.pdf*, 2008.
8. C. Paquet, D. Coulombier, R. Kaiser, and M. Ciotti. Epidemic intelligence: a new framework for strengthening disease surveillance in europe. *Euro Surveill*, 11(12):212–214, 2006.
9. Kevin Peterson, Rammohan Deeduvanu, Pradip Kanjamala, and Kelly Boles. A blockchain-based approach to health information exchange networks. In *Proc. NIST Workshop Blockchain Healthcare*, volume 1, pages 1–10, 2016.
10. G. Wood. Ethereum: A secure decentralised generalised transaction ledger, ethereum project yellow paper. *https://ethereum.github.io/yellowpaper/paper.pdf*, 2014.

Figures

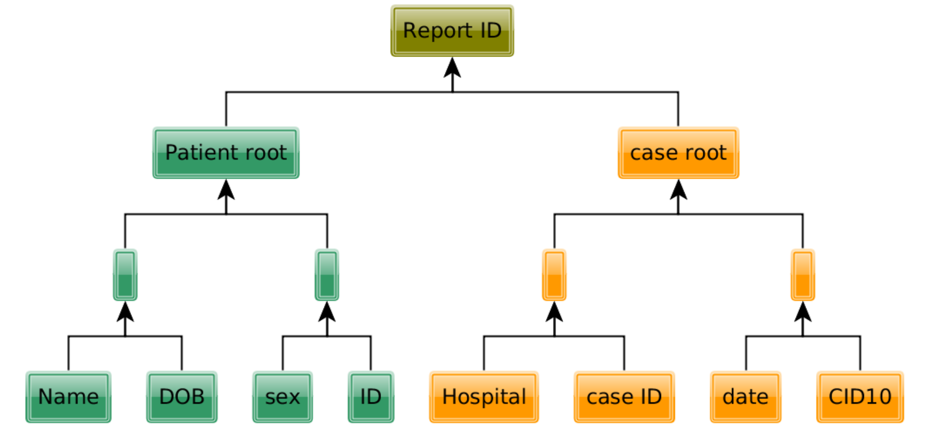

Figure 3 Merkle tree containing the case report data. This is a toy data model, real case reports may have as many fields as required.