Managing Lifetime Healthcare Data on the Blockchain

Mark Hanley
Trinity College Dublin
Dublin, Ireland
mahanley@tcd.ie

Hitesh Tewari
Trinity College Dublin
Dublin, Ireland
htewari@tcd.ie

***Abstract*—The widespread adoption of fax machines in the 1980s revolutionised everyday communications. It was quickly adopted as the standard form of communication across the globe. Since then, the internet has replaced fax as a truly global form of instant communication. However, the fax machine still reigns as the primary form of communication in a number of industries, healthcare being one of them.**

This paper presents a system that uses a blockchain and an off-chain centralised data storage to give patients and medical professionals instant access to their medical records from anywhere. By assigning each medical record a pseudo anonymous identifier, a second layer "blockchain" for each user can be created allowing for the rapid collection and querying of data. The off-chain pseudo anonymous data storage allows for the data to remain unencrypted enabling the rapid generation of anonymous medical datasets which can be used for machine learning and data mining on the data, potentially bringing many benefits to the healthcare industry.

***Index Terms*—blockchain, medical, healthcare, machine learning**

I. Introduction

The current medical systems in Ireland and most of the world, provide no quick, easy method to access a patient's medical history. Medical systems in healthcare institutions are fragmented with no secure, easy way to share medical records between different institutions.

In this paper a system is proposed that allows for the storage and easy sharing of medical records. Records will be kept anonymous in the system to facilitate the use of machine learning on the data by research institutions. Despite this anonymity, the data still needs to be retrievable by the user that owns the data. Medical records will be easily and rapidly shared between medical institutions and users by using a blockchain as a secure, tamper-proof method of communication [1]. A centralised anonymous database that healthcare institutions can store their patient's data would allow for the rapid querying and access to a user's medical history.

The rest of the paper is outlined as follows. The current medical system is examined in section II, how it handles sharing medical records between institutions and the steps required to retrieve a patient's entire medical history. Three blockchain based medical systems currently in development are reviewed and are compared to the proposed system in section III. In section IV, the design of the system is presented, explaining how the user's data can be stored anonymously within the system. Section V gives an overview of the system and how it is used for user purposes. A brief explanation of how the system can interact with machine learning and data analysis is also given. Potential future improvements to the system are outlined in section VI. Finally, a conclusion to the paper is given in section VII.

This work was supported, in part, by Science Foundation Ireland grant 13/RC/2094 and co-funded under the European Regional Development Fund through the Southern & Eastern Regional Operational Programme to Lero - the Irish Software Research Centre (www.lero.ie).

II. Background

The current medical record system is extremely fragmented. If a patient wanted to access their entire medical history from the moment they were born, this would be a long, slow, and potentially costly process. Consider all the different healthcare institutions that a patient would visit in their entire life. They would need to remember where and when they visited each medical institution and then request their medical records from the institution.

A person obtaining a copy of their medical records ends up being a costly endeavour. The Health Service Executive (HSE) of Ireland, has a set price list on accessing a patient's own medical records [2]. The price of photocopying each page of every medical record in an institution is €0.04 per page. Opting for a CD-ROM of data costs €10.16 per CD. Collating records onto a CD would certainly be quicker and more environmentally friendly than photocopying every page of a patient's medical history. There is a legal solution to avoiding these costs somewhat by submitting a Freedom of Information (FOI) request for a fee of €6.35 [2]. However, this still could be costly as a FOI request would need to be submitted to every medical institution visited.

Communication between medical institutions is a slow process. Currently documents are sent between institutions by fax machine [3], [4]. The medical industry is among the last few industries still using fax machines as primary forms of communication. Fax is preferred over quicker, modern forms of communication such as email as it is less prone to remote attackers and documents cannot be manipulated [5], [6]. However, fax is a slow form of communication with the maximum transmission speed being 64Kbits per second and offers no form of authentication which is a security risk [7].

As a result of the fragmentation in the system and the lack of speedy communication between institutions, there has been cases of delayed diagnosis of patients due to a delay in the reporting of results [8].

III. Related Work

Research And Markets' report on the Electronic Health Records market valued the market at $23 billion in 2016 and expects this to rise to $33 billion by 2023 [9]. As a result there are many companies worldwide aiming to get into this very lucrative business.

Medicalchain [10] aims to use a blockchain to make every interaction with a patient's medical record secure and transparent. User's files are uploaded to the system and are encrypted with a symmetric key. The symmetric key is then encrypted with the user's public key and the file is uploaded to a database. A pointer to the file is then returned from the database and the pointer and hash of the file is written to the blockchain.

MediChain [11] aims to be a big data platform allowing users electronic health records be stored on the system and allow access to these data by specialists, whether that be medical professionals, researchers, or insurance companies. The system uses numerous smart contracts stored on the Ethereum blockchain. Smart contracts are code stored in transactions published to the blockchain [12]. The downsides of using smart contracts in such a system that they must be produced to a high quality because once smart contracts are published onto the blockchain it becomes almost impossible to change the code in the contract [13].

MediBloc [14] is a system that wants to disrupt the current way in which medical data is managed and make it more consumer centered. Different user accounts will exist in the system, users, healthcare professionals, and researchers, which will allow for access control to the user's data. MediBloc uses the Qtum [15] blockchain which is a a blockchain based off Bitcoin and runs the Ethereum Virtual Machine. The user's personal device, e.g. their mobile phone, will be used as the primary data storage in the system which can be a problem if someone lost their phone for example.

Table I compares these systems in development to the proposed system.

TABLE I
Comparing systems in development to proposed system

	Blockchain Used	Tokenised	Machine Learning
Medicalchain	Hyperledger Fabric & Ethereum	MedTokens	Difficult due to key management
MediChain	Ethereum	MediChain Utility tokens	No method outlined
MediBloc	Qtum	MediTokens Medi Points	No easy way
Proposed System	Multichain	None	Easy scraping of anonymous data

IV. Design

Handling user medical data is a very sensitive operation. There are many laws and restrictions around the world protecting user's data making sure it is stored and handled in a secure manner [16], [17]. Despite these regulations, there are numerous instances of patient's data being mishandled [18]. A solution to keeping the user's data anonymous but yet retrievable is to design a *pseudo anonymous identifier* and make use of a suitable *data store*.

A. Pseudo Anonymous Identifier

Stripping a medical record of any identifying information and assigning a pseudo anonymous identifier to the record, allows for the development of a system that keeps user's medical data anonymous yet remains accessible to those with the correct credentials. The construction and assignment of the pseudo anonymous identifiers is given below.

1) Base Record: When a user registers with the system for the first time (ideally when a newborn visits their local GP for vaccinations at two months) a *base record* will be generated. This base record will be the credentials that the person will use to interact with the system. To generate the base record, the patient's date of birth, social security number (SSN), and a secret seed (PIN) are taken and the concatenation of the data is passed through a hashing algorithm, e.g. SHA-256, producing a unique base record for a user. The base record would then be published and stored on the blockchain.

Ideally the PIN chosen to seed the hash would be something unique to the user and would not be left up to random assignment, requiring the user to remember their PIN. The solution to this could be using the last four digits on a user's birth certificate registration. A birth cert is something that is not easily lost and is usually close to hand when visiting a GP for the first time with a newborn.

The base record is a 64 character long string that uniquely identifies the user. It is not expected that the user remembers this unique sequence of characters and so it is possible to generate this record again at any time by passing the correct details through the same hashing algorithm. The user would also be able to generate a QR code [19] containing their base record through an online web portal to present to their doctor or GP to grant them access to their medical records. This is discussed further in section V-B.

2) Linking Subsequent Records: By creating a method that securely links every subsequent record back to the a user's previous record, it is possible to create a second layer "blockchain" on top of the main chain. This second level blockchain is specific to the user that it belongs to and allows for the rapid collection and querying of the user's records. Figure 1 shows this second level chain.

To create the link from one record to the other, we seed a hashing function with the pseudo anonymous identifier of the

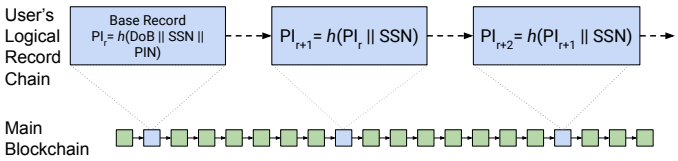

Fig. 1. The logical chain produced by linking records

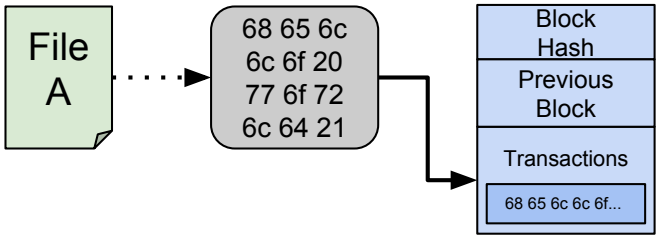

Fig. 2. Embedding the hexadecimal data of File A in the transaction data

previous record, e.g. the user's base record, and a salt, e.g. the user's SSN. The inclusion of the salt into the hash function is important as it provides and extra piece of security because if the link from one record to another is just the hash of the previous record it would be easy for an attacker to access the user's entire medical history if they gained access to one of their records.

Without the second layer "blockchain", assembling a user's entire medical history would be a more difficult challenge as they are not stored consecutively on the blockchain, requiring the entire blockchain be traversed to find their next record. By using every record and a salt to link one record to another, the user only needs to use their base record, which is encoded in their QR code, and their salt, e.g. SSN, to retrieve their entire medical history.

B. Data Storage

User records are now kept anonymously in the system, so a suitable data storage needs to be chosen. There are numerous options that could be used which are explained below.

1) Embedded in Transaction Data: The first approach was to simply embed the file data in the transaction data in a block. Instead of a transaction containing the addresses of the wallets where the transaction is being sent from and to, data can be encoded in these fields, storing it in the blockchain forever [20]. By taking the raw binary data in hexadecimal form from the file as it is being saved, this data could then be embedded in a transaction and published to the blockchain. The advantage of this method of saving medical records on the blockchain is that the data would be distributed across the network by default.

The downsides of writing files in transactions is that anyone that has access to the blockchain would also have access to the data. There would be no possibility of adding in any access control to restrict unauthorised access to the data once someone has access to the blockchain. Another issue is that if someone was to leak their details or get their details phished, there would be no way to recover from this situation and prevent the attacker from accessing the user's medical data. This is further discussed under section IV-C.

2) Emerging Blockchain Based Solutions: The use of blockchain as a secure, decentralised form of communication is on the rise [21]. Storing data directly on the blockchain can lead to the blockchain growing massively in size as well as increasing in cost to store the data on the chain. This has lead to a number of projects being actively developed to solve this problem of storing data for use in a blockchain based system. Two very promising projects are Storj [22] and InterPlanetary File System (IPFS) [23]

Storj encrypts user data using AES256-CTR mode [24] and splits the data into pieces (shards) and spread across the network of nodes. As the network is used more and more and the number of shards in the system increases, locating shards without prior knowledge of their location becomes exponentially harder. Only the owner of the data possesses the key containing the location of the shards across the network as well as the key to decrypt the files.

IPFS aims to replace HTTP as the main protocol for transferring files across the internet. Instead of the current system of requesting files from a single server, with IPFS files would be requested from the network and those that have the files available serve them to the user in a peer-to-peer fashion similar to BitTorrent. This allows for the uncensorable hosting of data. As long as someone has a copy of the file and shares it over the IPFS network, the content cannot be removed or censored.

Storj and IPFS have great potential but due to the fact that the system would be handling medical data, control over who has access to this data and where it is stored needs to be maintained. Storj and IPFS both suffer the same problem of storing data on untrusted nodes across the network. A potential idea that can be taken from both systems is that when data is stored on the system, the system returns a pointer to the data that can be used as an index into the storage.

3) Use a Centralised Database: Using a centralised database has plenty of advantages compared to a decentralised one. Firstly, centralised databases are much easier to setup, manage, and operate compared to their decentralised counterparts due to them being less complex. Secondly, providing access control on a centralised database is very easy to implement, securing access to the user data. A central database under the control of a government health agency would also allow the easy process of gaining authorisation and connecting to the database by research institutions for machine learning. Another advantage of using a centralised database is that files are able to be altered or removed which is an important consideration and is addressed in section IV-C.

Medical institutions would continue to maintain their own existing databases of identifiable patient data. Operating in tandem with the existing systems is important as it has been

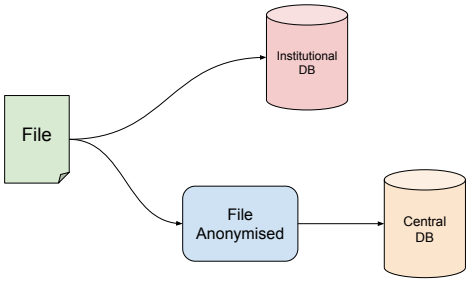

Fig. 3. The central database working in tandem with existing databases

troublesome in the past getting medical institutions worldwide to switch to newer technology [25].

4) Storing Files: As the central database will be working in tandem with existing healthcare institutional databases the system would take a copy of the records as they are being saved and strip away any identifying information from the file, making it anonymous. The file is then uploaded to the central database. The central database saves the file and returns a hashed pointer to the index of the file in the database. In this implementation the hashed pointer is the result of hashing the file data and the timestamp that the file was uploaded to the database. The pseudo anonymous identifier, hash pointer, and the hash of the file are then written to the blockchain. By publishing the hash of the record to the blockchain, we can protect the file from being tampered with or manipulated. If a file is tampered with the system can compare the recorded hash of the file stored in the blockchain against the hash of the file retrieved from the database. If there is a mismatch in the hashes, the user is alerted and can then inform the governmental agency controlling the system that their files have been tampered with and they can investigate the case.

C. Leaking of User Details

In a perfect world, users would never accidentally leak or have their private information phished by attackers. Unfortunately there is no way to overcome this happening and because in this implementation a user's entire medical history is linked to their date of birth, social security number, and PIN there needs to be a way of protecting a user's data. If a user was to divulge this information to an attacker, there must be a way that would allow the user to get their data back under their control.

This can be solved by assigning a new base record to the user. If the attacker has access to the user's date of birth, social security number, and PIN, the only detail that it would be possible to change is the PIN for their base record. By changing the user's PIN to some other sequence, an entirely different base record for the user would be generated creating a new chain for the user. The attacker would no longer be able to identify the new base record of the user unless they managed to retrieve the user's new PIN.

As the attacker has compromised the user's chain on the blockchain, the data needs to be moved in the database and linked to in the new chain. Iterating over a user's medical history, a copy of each of their records is taken and they are reuploaded to the central database. This results in a completely new hash pointer being returned. The new hashed pointer is then written to the blockchain as discussed in section IV-A2. The original file at the index pointed to by the "compromised" hash pointer is then erased, removing the ability of the attacker to access the file again.

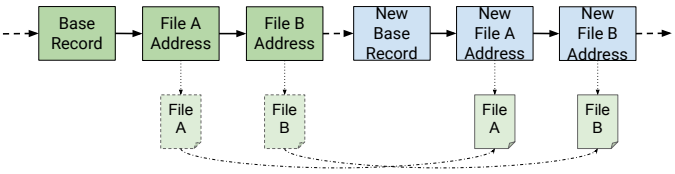

Fig. 4. Copying files from a compromised base record to a new base record

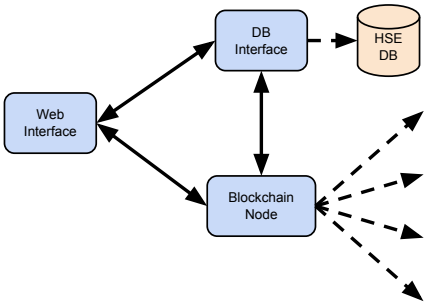

Fig. 5. System Overview

V. System Overview

A. Overview

A simple web based interface would be the easiest way for users to interact with the system. A web based interface would be preferable over a stand alone program as it would save medical institutions from deploying and installing the software on individual computers instead simply hosting it on a server. The system would require a front end for allowing a user to interact and a back end to interact with the database and the blockchain. Figure 5 shows the overview of the system. The "Web Interface" is what allows users to interact with the system and contains both the front and back ends of the system. The "DB Interface" is to allow communication with the central database, allowing data to be read and stored. "Blockchain Node" is the Multichain node running in the system. This handles the requests from the back end to read and write data onto the blockchain.

Multichain was chosen as the blockchain of choice for the system to use. A huge attraction to using Multichain as a blockchain is that it is a fully permissioned blockchain as well as being a fully customisable blockchain, allowing easy configuration of settings such as block size and block timings. It is very easy to publish data to the blockchain using Multichain's stream feature which allows data to be published in a key/value fashion. This method of publishing data makes it very easy for its retrieval off the blockchain.

B. System Usage

Once the system is deployed, users can navigate to the system using the correct URL. Here users are able to generate their base record and view their medical history. Medical professionals would have access to registering new users with the system as well as uploading medical records to the system. This access management would be implemented through a simple username (e.g. medical ID), password combination.

Users can generate their base record by filling out their date of birth, social security number, and their PIN. A QR code containing the user's base record is then generated [19]. It is then possible for the user to save this QR code for offline access. When a user wants to access their medical records the interface opens a QR code scanner through the camera on the user's device. The QR code is presented loading the users base record. The user then needs to input their social security number to retrieve their medical records. The system queries the blockchain for the user's base record and retrieves their records by hashing the current record with their SSN to retrieve the next file on their chain. URLs to the files are then presented to the user for them to open and view their medical records.

Medical professionals are able to register new users, view a user's medical history, and upload files to the system. Medical professionals would access these features through a simple log in process. To register a new user with the system, the medical professional requests the user to input their date of birth, social security number, and their PIN. The process of accessing a patient's medical history is the same process as accessing them as a user. For uploading new records to the system, the user presents their QR code and it is scanned by the camera on the device. The file to be uploaded is selected and the user inputs their SSN to authorise the upload. The file is then uploaded to the central database and recorded on the blockchain.

C. Performing Machine Learning

One of the main desires from using a pseudo anonymous medical storage is the application of machine learning on the data stored there. Multichain's easy to use API allows for the rapid development of applications that can interact with the blockchain, retrieve the anonymous data, and collate it into a massive anonymous dataset for their research.

As a government agency controls access to the blockchain and the central database, research institutions would apply to access the system and its anonymous medical data. The agency would give such an institution temporary access to the system to allow them to perform their research. The institution would be able to iterate over the data on the blockchain, scraping the data that is relevant to their research, e.g. influenza rates, and be able to perform their research. Multichain's permission based blockchain allows researcher's access to the blockchain to be revoked once their research is complete or if the agency in control of the system detects inappropriate use of the system by researchers.

VI. Future Work

There are a few points in the system that could be the focus of further work, especially if this system was to be deployed to the real world.

In this implementation user records were linked using the record identifier and the user's social security number as a salt in a hash function. The social security number would not be the most secure method of linking one record to another as there has been many issues with using SSN as an identifier, especially in the United States [26], [27].

Perhaps for a production system, the data from smart ID cards could be used. The smart ID cards contain details about the person on a chip within the card. Ideally the data stored on the cards would be encrypted to protect the owner's details. The encrypted data could be read off the card using chip readers and then used as a salt into the hash function. By using an entirely separate card's encrypted data, only retrievable through a chip readers, it would decrease the potential for an attacker to gain access to a users medical data through phishing.

As part of the aims, the system is open to the use of machine learning on the data. To allow for this to occur, data needs to be stored unencrypted in the system. The data is kept anonymous so that it cannot be linked back to a person which is acceptable but could be improved upon by encrypting the data. There is active research into homomorphic encryption [29], [30] which is being able to perform machine learning operations on encrypted data. However, currently it is only possible on very simple operations and would not be feasible for use in an encrypted medical scenario [31]. Perhaps in the future when homomorphic encryption has improved, the user data can be encrypted with homomorphic encryption methods allowing for the data to be encrypted without impacting the applications of machine learning.

VII. Conclusion

In this paper a system has been proposed that would allow the easy and quick communication of medical records between institutions, the ability to rapidly collect and view a patient's entire medical history, and to keep the data in the system anonymous for the use of machine learning.

The use of blockchain as a form of communication between different medical institutions in the network allows for the secure transfer of records between institutions and users. Users of the system can be confident that their records are stored anonymously within the system. The system is tamper proof and the files stored in the system cannot be manipulated with.

In comparison to other systems that are being developed, the proposed system is able to operate efficiently without the inclusion of a monetary system where users pay to use the system. By using a permissioned blockchain such as Multichain, there would be more control and governance of the system by the government agency operating the system, selectively choosing who can access the blockchain and database, and in extreme cases, such as professional misconduct by a healthcare professional, revoke access to the system.

If such a system was to be developed and deployed in medical institutions around the world, there could be massive improvements in the healthcare system. Doctors and GPs would have instant access to a person's entire medical history, allowing them to see all of the patient's previous treatments from every medical institution they visited, improving patient care. Users of the system would be able to quickly and freely view their entire record when ever they wanted which was not possible before. Researchers would now have access to a massive anonymous dataset that they could use for research. There is a huge potential for many machine learning applications to use this data to predict trends that could improve the healthcare system, e.g. length of stay in a hospital.

REFERENCES

[1] "Blockchain as an audit-able communication channel." *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), Computer Software and Applications Conference (COMPSAC), 2017 IEEE 41st Annual, COMPSAC*, p. 516, 2017.

[2] Citizensinformation.ie, "Access to medical records." [Online]. Available: http://www.citizensinformation.ie/en/health/legal_matters_and_health/access_to_medical_records.html

[3] S. Kliff, "Why american medicine still runs on fax machines," Jan 2018. [Online]. Available: https://www.vox.com/health-care/2017/10/30/16228054/american-medical-system-fax-machines-why

[4] "Faxing medical records." [Online]. Available: https://provider.ghc.org/open/render.jhtml?item=/open/workingWithGroupHealth/records-faxing.xml

[5] "Security of faxes vs emails," Sep 2015. [Online]. Available: https://security.stackexchange.com/a/100800

[6] "Millions of people still buy and use fax machines," Jul 2015. [Online]. Available: http://www.thejournal.ie/fax-machines-2189600-Jul2015/

[7] T. Recommendation, "Facsimile coding schemes and coding control functions for group 4 facsimile apparatus," *International Telecommunication Union, Geneva*, 1988.

[8] V. Traynor, "2.5m award for terminally-ill woman over missed cancer," Apr 2018. [Online]. Available: https://www.rte.ie/news/courts/2018/0425/957122-vicky-phelan/

[9] "Electronic health records (ehr) market by product, type, application and end user - global opportunity analysis and industry forecast, 2017-2023," *Research And Markets*, January 2018. [Online]. Available: https://www.researchandmarkets.com/research/65n56m

[10] "Blockchain for electronic health records." [Online]. Available: https://medicalchain.com/

[11] "Medichain." [Online]. Available: https://medichain.online/

[12] M. Bacina, "When two worlds collide: Smart contracts and the australian legal system." *Journal of Internet Law*, vol. 21, no. 8, pp. 1 – 27, 2018.

[13] J. M. SKLAROFF, "Smart contracts and the cost of inflexibility." *University of Pennsylvania Law Review*, vol. 166, no. 1, pp. 263 – 303, 2017.

[14] "Reinventing your healthcare experience!" [Online]. Available: https://medibloc.org/en/

[15] "Qtum the blockchain made ready for business." [Online]. Available: https://qtum.org/en/

[16] "A guide to data protection legislation for irish general practice," 2011. [Online]. Available: https://www.icgp-education.ie/confidentiality/ICGP_Data_Privacy_Doc.pdf

[17] "An as is analysis of information governance in health and social care settings in ireland," Jan 2010. [Online]. Available: https://www.hiqa.ie/sites/default/files/2017-02/Info_Governance_As_Is_Analysis(1).pdf

[18] C. D'Arcy, "Hospital records found strewn across road in drogheda," *Irish Times*, 2015. [Online]. Available: https://www.irishtimes.com/news/crime-and-law/hospital-records-found-strewn-across-road-in-drogheda-1.2371173

[19] "Information automatic identification and data capture techniques qr code barcode symbology specification," International Organization for Standardization, Geneva, CH, Standard, Feb. 2015.

[20] K. Shirriff, "Hidden surprises in the bitcoin blockchain and how they are stored: Nelson mandela, wikileaks, photos, and python software," Feb 2014. [Online]. Available: http://www.righto.com/2014/02/ascii-bernanke-wikileaks-photographs.html

[21] O. Kharif and C. Russo, "Venture capital surges into crypto startups," Mar 2018. [Online]. Available: https://www.bloomberg.com/news/articles/2018-03-26/icos-can-wait-venture-capital-surges-into-crypto-startups

[22] "Storj, a peer-to-peer cloud storage network." [Online]. Available: https://storj.io/

[23] "Ipfs, a peer-to-peer hypermedia protocol to make the web faster, safer, and more open." [Online]. Available: https://ipfs.io/

[24] R. Housley, "Using advanced encryption standard (aes) counter mode with ipsec encapsulating security payload (esp)," Internet Requests for Comments, RFC Editor, RFC 3686, January 2004.

[25] D. of Health, *URGENT ACTION REQUIRED FOR THOSE NHS ORGANISATIONS THAT HAVE NOT MIGRATED AWAY FROM MICROSOFT XP*, Apr 2014. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/314721/DHandCabinetOfficeMicrosoftXPupdate8April14.pdf

[26] A. Acquisti and R. Gross, "Predicting social security numbers from public data," *Proceedings of the National Academy of Sciences*, vol. 106, no. 27, pp. 10 975–10 980, 2009. [Online]. Available: http://www.pnas.org/content/106/27/10975

[27] G. BLOCK, G. M. MATANOSKI, and R. S. SELTSER, "A method for estimating year of birth using social security number," *American Journal of Epidemiology*, vol. 118, no. 3, pp. 377–395, 1983. [Online]. Available: http://dx.doi.org/10.1093/oxfordjournals.aje.a113645

[28] "Public services card." [Online]. Available: https://psc.gov.ie/

[29] C. Gentry, *A fully homomorphic encryption scheme*. Stanford University, 2009.

[30] M. Van Dijk, C. Gentry, S. Halevi, and V. Vaikuntanathan, "Fully homomorphic encryption over the integers," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2010, pp. 24–43.

[31] M. Naehrig, K. Lauter, and V. Vaikuntanathan, "Can homomorphic encryption be practical?" in *Proceedings of the 3rd ACM Workshop on Cloud Computing Security Workshop*, ser. CCSW '11. New York, NY, USA: ACM, 2011, pp. 113–124. [Online]. Available: http://doi.acm.org/10.1145/2046660.2046682