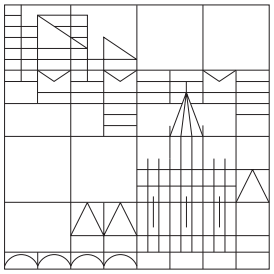
University of Konstanz

Department of Computer and Information Science



Master Thesis

Data Ownership & Privacy Preserving Applications

in fulfillment of the requirements to achieve the degree of

Master of Science (M.Sc.)

Harsh Kedia

Matriculation Number :: 01/752437

E-Mail :: ⟨harsh⟩.⟨kedia⟩@uni-konstanz.de

Field of Study :: Information Engineering
Focus :: *Applied Computer Science*
Topic :: *Distributed Systems*

First Assessor :: *Prof. Dr. M. Waldvogel*
Second Assessor ::
Advisor :: *Prof. Dr. M. Waldvogel*

Any dedications or other fancy stuff???

Abstract

Abstract

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Humans have evolved over thousands of years building systems which deals with land ownership and property rights. With the advent of Internet our lives has become more and more digital, but we have no experience in managing data ownership. It's clear that data is becoming the new currency in today's digital economy. Big tech companies understood this a long time ago and therefore offered their services free of charge in exchange of our data which they then used to generate profits, control our perception about how we see the world and also tamper with public affairs like the election. There's clearly a need to define data ownership and build systems which enable users to own their data.

With data ownership comes the question of digital identity. How can we identify ourselves over the internet? With username and passwords we can uniquely identify ourselves when using a service, but then we have to create an identity for each service we want to use. It has another drawback, i.e. our passwords are stored on a central server which is prone to hacking. There exists systems like *Google Sign-in* or *Facebook Connect* which allows us to carry our identity across multiple services but then again this identity is not owned by the user but by Google or Facebook. Therefore, there is a need for a self-sovereign identity which is owned by the user and can be verified independently by anyone.

To define a model for Data Ownership, lets looks at Land Ownership. In a land ownership model, at any given point in time, a property has a fixed Geo-location while the owner can be anywhere in the Geo-space. Conversely, In a data ownership model, at any given point in time, a user has a fixed identity while the data can be anywhere on the internet.

Blockchain along with Public key cryptography allows us to build a Decentralized Public Key Infrastructure (DPKI) thereby empowering users to create self-sovereign identity. Combining self-sovereign identity with encrypted storage enable us to build systems where users own their identity as well as their data.

Explaining contents of each chapter.

Explores the emerging protocols which enable a decentralize web.

Chapter 2

Background

2.1 Pretty Good Privacy (PGP)

PGP[1] is a encryption program which uses public-key cryptography[1] to provide cryptographic privacy and authentication for data communication. It can be also used to sign messages such that the receiver can verify both the identity of the sender and integrity of the message.

It is built upon a Distributed Web of Trust in which a user's trustworthiness is established by others who can vouch through a digital signature for that user's identity[2].

There are a number of inherent weaknesses which prevented the widespread adoption of PGP. These include the following[2]:

- Trust relationships are built on a subjective honor system.
- Only first degree relationships can be fully trusted.
- Levels of trust are difficult to quantify with actual values.
- Issues with the Web of Trust itself (Certification of Endorsement).

2.2 Public Key Infrastructure (PKI)

PKI is a system for creation, storage and distribution of digital certificates which can be used to verify ownership of a public key[3]. In today's Internet, third parties such as DNS registrars, ICANN, X.509 Certificate Authorities (CAs), and social media companies are responsible for the creation and management of online identities. Thus our online identities lie in the control of third-parties and are borrowed or rented rather than owned. This results in severe usability and security challenges[4].

There is a possible alternate approach called *decentralized public key infrastructure (DPKI)*, which returns control of online identities to the entities they belong to. By doing so, DPKI addresses many usability and security challenges that plague traditional public key infrastructure (PKI)[4].

[1] https://en.wikipedia.org/wiki/Pretty_Good_Privacy

2.3 Distributed Hash Table (DHT)

2.4 Cloud Computing

2.5 IPv6

2.6 Segment Routing

2.7 Blockchain

The current Internet Protocol stack consists of four layers: the *Link Layer* puts data onto a wire; the *Internet Layer* routes the data; the *Transport Layer* persists the data; and the *Application Layer* provides data abstraction and delivers it to the end user in the form of applications. All four layers work seamlessly for exchanging of data, but not value. Bitcoin[5] and other cryptocurrencies help define the fifth Internet Protocol layer which enables the exchange of value as fast and efficiently as data[6].

Exchanging value across the Internet presents two challenges. First, every participant in the network must agree upon a shared state and Second, the asset being exchanged should have a clearly defined owner. These challenges are commonly referred as the *Byzantine General's* problem[7] and *Double-spending* problem[8] respectively. Blockchain, the technology underlying Bitcoin and most cryptocurrencies solved the above problems by means of decentralized consensus[2].

At a higher level, blockchains are append-only, totally-ordered, replicated logs of transactions[9]. A transaction is a signed statement that transfers the ownership of an asset from one cryptographic keypair to another. Peers[3] in the network, append new transactions by packaging them into a block and then executing a leader election protocol which determines who gets to append the next block[10]. This election protocol is determined by the underlying consensus algorithm of the blockchain. Each block contains the cryptographic hash of the previous block along with some transactional data.

[2] https://en.wikipedia.org/wiki/Consensus_(computer_science)

[3] A Node having the full copy of the blockchain.

Chapter 3

Application Design

3.1 Introduction

An Application software or *app* is a computer program designed to perform a specific set of tasks or actions for the end user. There are countless number of applications in use today and the majority of them are web applications following a centralized client-server model[6].

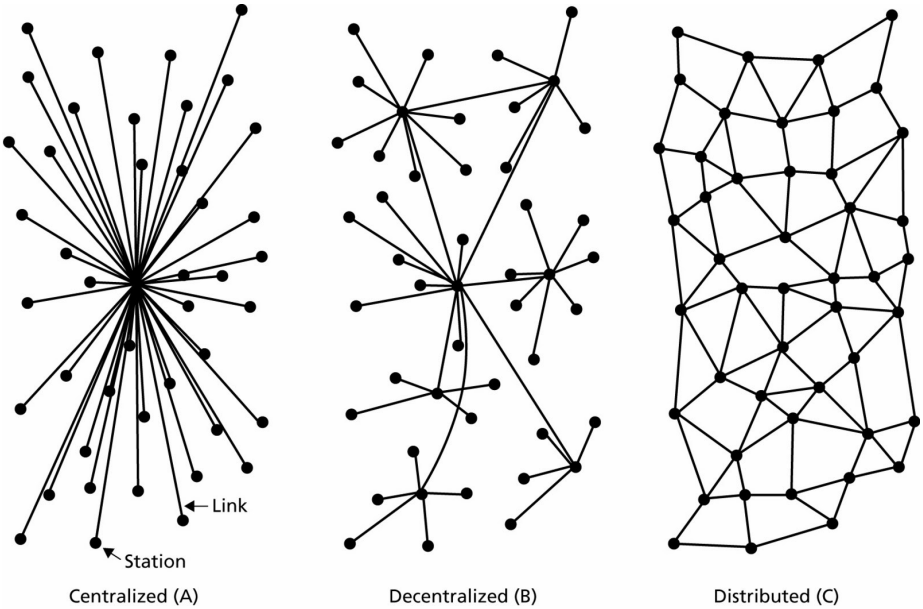


Figure 3.1: The three way of modeling web applications

Figure 3.1 shows a visual representation of three different ways of modeling web applications[11]. Here, *Centralized* and *Decentralized* refers to level of control, while *Distributed* refers to differences of location. Both centralized and decentralized systems can be distributed as well.

3.1.1 Centralized

It's currently the widespread way of building software applications. In this model a central server control the flow of information and governs the operation of individual units. Since the control is centralized, these types of systems suffer from single point of failure risk.

3.1.2 Distributed

In a Distributed model, the control still resides with a central server, however, the computation is spread across multiple nodes or servers.

3.1.3 Decentralized

In a Decentralized model, there is no central point of control as it's spread across all the servers running the application. Applications built using this model don't have a single point of failure and are inherently fault tolerant.

3.2 Enabling Technologies

The document *Information Management: A Proposal[12]* written by *Sir Tim Berners-Lee*[1] conceived the ideas for what would become the WorldWideWeb. It's main goal was to enable information exchange between computers in an accessible way at CERN[2].

HTML[3], URI[4] and HTTP[5] were the fundamental technologies that defined the foundation of the Web. HTTP connected every computer on the planet with a common protocol. The HTTP protocol guidelines defined a set of trusted servers that translated a web address into a server address. Furthermore, HTTPS[6] added another layer of trusted servers and certificate authorities. People would host personal servers for others to connect to, and everyone owned their data[6]. As the Web evolved, applications servers[7] became the common way of interactive with the Web and the centralized model of data ownership as we know it today was born[6]. It was conceptually and programmatically easier to maintain an application server and profit from user's data that utilize it.

Blockchain is the primary technology that enables the creation of applications with a decentralized model of data ownership. It puts the users of an application in control of their data thereby enabling a more open Web, as it was originally intended[8].

The blockchain helped solve the Byzantine Generals Problem[7]. This problem describes a situation where all participating nodes in a distributed network must agree upon every message that is being transmitted between nodes, but where some of the nodes are corrupt and disseminating false information or

[1] https://www.w3.org/People/Berners-Lee/
[2] https://home.cern/
[3] https://developer.mozilla.org/en-US/docs/Web/HTML
[4] https://tools.ietf.org/html/rfc3986
[5] https://tools.ietf.org/html/rfc2616
[6] https://tools.ietf.org/html/rfc2818
[7] https://en.wikipedia.org/wiki/Application_server
[8] https://webfoundation.org/about/vision/history-of-the-web/

are unreliable. This agreement is called as **consensus**. With Bitcoin[5], decentralized consensus became possible. Agreement is achieved in the Bitcoin network by way of *proof-of-work*[9] consensus mechanism which is resistant to Sybil Attack[13]. Proof-of work is both computationally and energy expensive; other consensus mechanism such as *proof-of-stake*[10] relies on stake in the system instead of computational power.

3.3 Concepts

There are five concepts in a web application that have traditionally been implemented in way that puts control with a centralized entity: data, identity, value, computing and bandwidth[6]. Each of these require trust in a 3rd party - a trust which can be betrayed. Recent advancements in distributed-system technology can put users in control of these things. Below sections describes each concept in detail and shows how one can build applications in a way such that centralized control is not required.

3.3.1 Data

Data is the most important concept in any web application. First, let's look at how traditional web applications interact with data. Whenever, a user logs into an application, the application connects to a remote server and sends the authentication details. These details lets the server know which user is interacting with the application. Once authenticated, the user data is fetched from the remote storage and displayed to the user. All complex computations and data storage occurs on dedicated servers maintained in the cloud.

1: Storing Data Directly in a Blockchain

2: Storing Data in a Distributed Hash Table

3: Storing Data in a Cloud in Encrypted Containers

3.3.2 Identity

3.3.3 Value

3.3.4 Computing

3.3.5 Bandwidth

[9] https://en.bitcoin.it/wiki/Proof_of_work
[10] https://en.bitcoin.it/wiki/Proof_of_Stake

Chapter 4

Example Application 1

4.1 Introduction

Existing applications for sharing files are central solutions and therefore suffer from single point of failure risk. Moreover, using central services for securing data means that we have to trust a 3rd party with our data thus exposing it to manipulation risks. Hence, a decentralized application is required to overcome the problems posed by a central application. With the recent developments in Blockchain technology and P2P storage, it is possible to securely store and share data without using any central server.

This chapter describes the workings of the application *dShare*[14] built using P2P technologies enabling a secure way of storing and sharing data between two individuals or entities. The latest version of the application is deployed at https://file-share-dapp.herokuapp.com/

4.2 Technologies Used

4.2.1 Ethereum

Ethereum[15] is a blockchain platform for building decentralized applications. It allows the creation of *Smart Contracts*. Solidity[1] is the primary language for writing smart contracts on Ethereum.

4.2.2 InterPlanetary File System (IPFS)

IPFS[16] is a peer-to-peer file transfer protocol which enables a shared file system between all its connected peers. It achieves this by combining previous peer-to-peer systems such as Distributed Hash Tables (DHT), BitTorrent[17], and Git[18]. The data in the IPFS network are modeled as a Merkle DAG[2] thus providing a throughput storage system with content-addressed hyperlinks.

[1] https://github.com/ethereum/solidity

[2] Merkle directed acyclic graph - similiar to a Merkle tree data structure however they do not need to be balanced and its non-leaf nodes can contain some data.

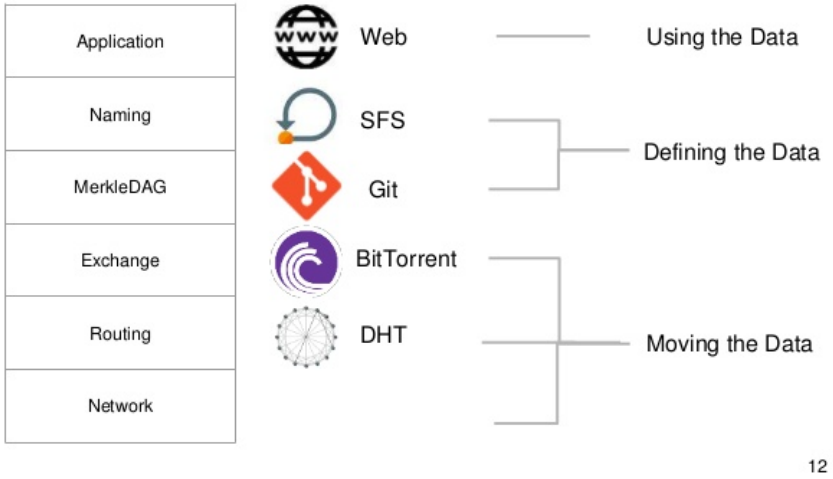


Figure 4.1: IPFS Stack

Figure 4.1[3] shows the IPFS Stack. It consists of sub-protocols, each providing a different functionality.

- Identities - node identification and verification.
- Network - connection management among peers.
- Routing - peer lookup using DHT.
- Exchange - data exchange strategies among peers.
- Objects - a content-addressed Merkle DAG.
- Files - versioned file system.
- Naming - A self-certifying mutable name system.

4.2.3 OriginStamp

OriginStamp[19] is a blockchain based system for decentralized timestamping. It uses the Bitcoin blockchain for the creation of trusted and immutable timestamps for any piece of data. Timestamps created by OriginStamp can be verified independently by anyone.

Figure 4.2 visualizes the timestamping process as implemented in OriginStamp. When a user submits a file, the hash of the data is recorded. It combines all the hashes submitted over a period of time and generates an aggregated hash. After some additional hashing and encoding operations, a Bitcoin address is created to which the smallest possible transactional amount of Bitcoins is transferred. Performing this transaction embeds the hash and the timestamp permanently to the Bitcoin blockchain. Each transaction is part of a block and

[3]Adopted from: https://image.slidesharecdn.com/ipfs-171229085327/95/ipfs-12-638.jpg?cb=1514537643

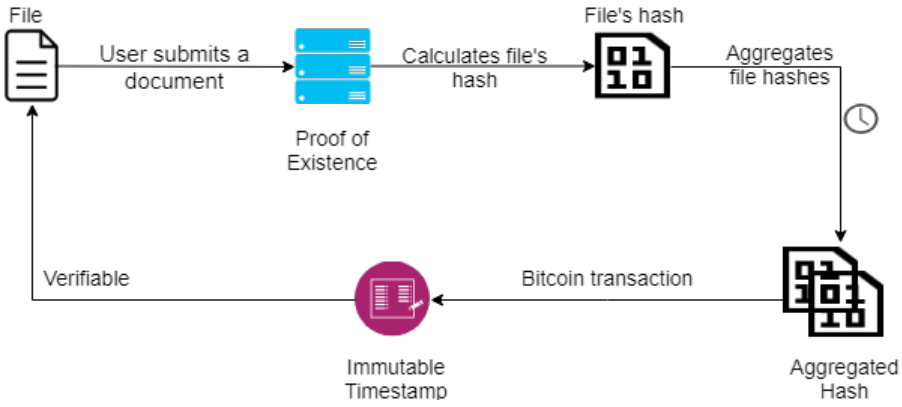


Figure 4.2: Timestamping using OriginStamp

is added to the Bitcoin blockchain by a process called mining. Since each block is linked cryptographically to the previous block, adding a new block confirms the validity of the last block. Changing the timestamp of a transaction becomes impossible once five or six subsequent blocks are mined, which requires an hour on average[5].

4.3 Implementation

For storing of files we used IPFS. Before uploading to the IPFS network, files are encrypted using AES-GCM[4] encryption mechanism. Sharing of encryption keys is facilitated using smart contracts built on Ethereum; thus files can be shared by anyone with an Ethereum address. Finally, OriginStamp is used for immutable timestamping.

The front-end of the application is built using React.js[5], a JavaScript library for building user interfaces. Solidity was used for writing smart contracts and deployed on the Ethereum test network, Rinkeby[6]. Next.js[7] was used for server-side rendering (SSR)[8], and Firebase[9] was used as a database for storing public Ethereum key of the users.

4.4 Working

This section describes the working of the different components of the application.

4.4.1 Smart Contract

The smart contract serves as the bridge between the front-end of the application and the Ethereum Blockchain. Data is read from and written to the blockchain

[4] https://www.aes-gcm.com/
[5] https://reactjs.org/
[6] https://www.rinkeby.io
[7] https://nextjs.org/
[8] https://nextjs.org/features/server-side-rendering/
[9] https://firebase.google.com/

with the help of function calls in the contract. Each function call which modifies some data requires a small fee in the form of gas[10] which defines the cost for a function execution in Ether. Reading from the blockchain does not require any fees.

The application makes use of two contracts, *FileFactory*, which acts as the factory contract for creation of new files and *File*, which represents an individual file.

FileFactory

FileFactory is the contract which is deployed on the Rinkeby test network. It has several mappings which stores the list of file contracts uploaded by a user. Whenever a user uploads a file, a function call is made to the *FileFactory* contract, which in turn deploys the *File* contract and updates the mappings for list of uploaded files and the respective uploader.

File

File contract is deployed to the blockchain whenever a file is successfully uploaded to the IPFS network using the application. Upon deployed, the constructor function is called. It takes the values passed by the *FileFactory* contract and saves the details to its *File* contract.

4.4.2 File Upload

Figure 4.3 visualizes the working of the application when a user uploads a file.

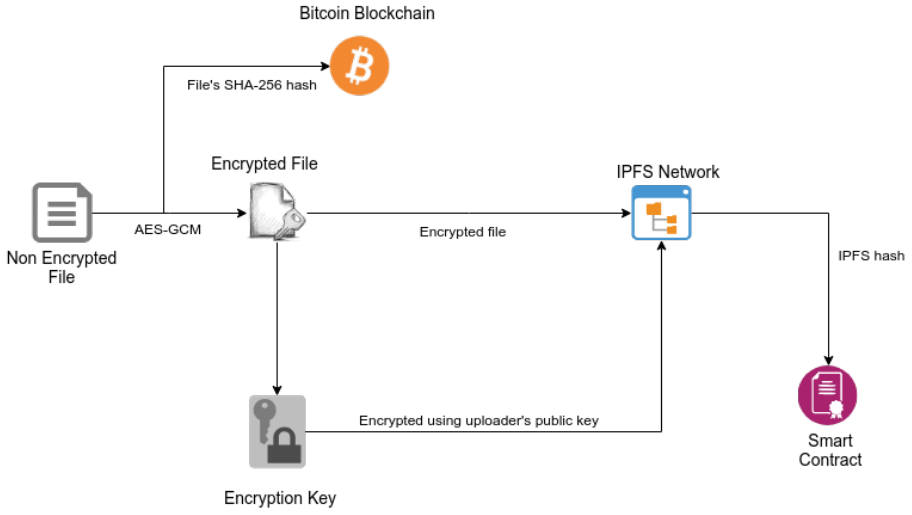


Figure 4.3: File Upload using Ethereum dApp

As soon as a user submits a file to be uploaded, it's SHA-256[11] hash is calculated and a timestamp is created by submitting the hash to the bitcoin

[10] https://ethereum.stackexchange.com/questions/3/what-is-meant-by-the-term-gas
[11] https://www.movable-type.co.uk/scripts/sha256.html

blockchain using the OriginStamp API[12].

Next, the file is encrypted using the SubtleCrypto[13] interface with 'AES-GCM'[14] as the encrypting algorithm. The encrypted data is then combined with the random salt to generate a `Uint8Array` buffer ready to be uploaded to the IPFS network.

The key used to encrypt the file is converted to `JSON` and is encrypted using the uploader's Ethereum public key which is retrieved from the database. This encrypted key and the encrypted data is then uploaded to the IPFS network. Once the file is successfully uploaded, `createFile()` in the `FileFactory` contract is called which deploys a new `File` contract with all details regarding the file saved to the blockchain.

4.4.3 File Sharing

Sharing a file requires the recipient's Ethereum address and uploader's private key. Figure 4.4 visualizes the working of the application when a user shares a file with another user.

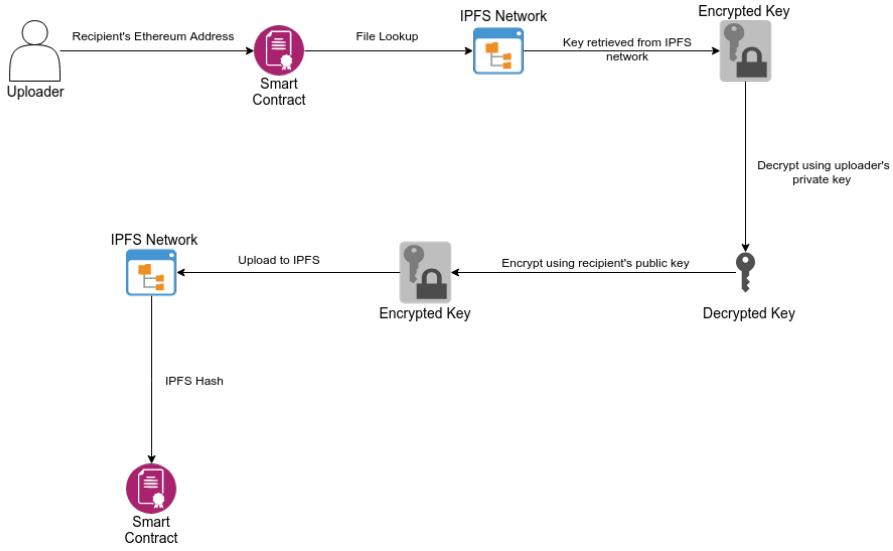


Figure 4.4: File Sharing using Ethereum dApp

Firstly, the file's IPFS location is retrieved from the `File` contract. From this location, the encrypted key is download and decrypted using uploader's private key. Once decrypted, the key is again encrypted using a recipient's public key. The new encrypted key is again uploaded to the IPFS network. Finally, the IPFS location of the key is saved into the `File` contract by calling `shareFile()`.

To stop sharing a file, a function call can be made to the `File` contract with the recipient's address, which deletes the contract reference from the `recipientFiles` array.

[12] http://doc.originstamp.org/
[13] https://developer.mozilla.org/en-US/docs/Web/API/SubtleCrypto
[14] https://en.wikipedia.org/wiki/Galois/Counter_Mode

4.4.4 File Download

Downloading a file requires the user's Ethereum private key. Depending on whether the file is uploaded or shared one, corresponding function from the `File` contract is called to retrieve the file's details. The key is then decrypted using user's private key and is converted to a valid JSON web key (jwk)[15] format. The encrypted file data is then converted to a file buffer, and the original file content and the random salt used for encrypting the file is retrieved. Finally, the file is decrypted and saved to the user's local storage.

4.4.5 File Archiving

Instead of deleting a `File` contract, the application provides a way to achieve files. This is also useful to keep track of archived files and restore them at a later date if required. When a file is archived, the `File` contract address is saved in an array which is later used for filtering the archived files from the UI. Restoring a file removes the `File` contract address from the archived files array.

[15] https://tools.ietf.org/html/rfc7517

Chapter 5

Example Application 2

5.1 Introduction

5.2 Technologies Used

5.3 Implementation

5.4 Working

Chapter 6

Results

Chapter 7

Discussion

Chapter 8

Conclusion

Appendix A

Acknowledgements

Bibliography

[1] W. Stallings, *Cryptography and Network Security: Principles and Practice*, ser. The William Stallings books on computer and data communications technology. Prentice Hall, 1999. [Online]. Available: https://books.google.de/books?id=Dam9zrViJjEC

[2] D. Wilson and G. Ateniese, "From pretty good to great: Enhancing pgp using bitcoin and the blockchain," in *International conference on network and system security*. Springer, 2015, pp. 368–375.

[3] J. Weise, "Public key infrastructure overview," *Sun BluePrints OnLine, August*, pp. 1–27, 2001.

[4] C. Allen, A. Brock, V. Buterin, J. Callas, D. Dorje, C. Lundkvist, P. Kravchenko, J. Nelson, D. Reed, M. Sabadello *et al.*, "Decentralized public key infrastructure. a white paper from rebooting the web of trust," 2015.

[5] S. Nakamoto *et al.*, "Bitcoin: A peer-to-peer electronic cash system," 2008.

[6] S. Raval, *Decentralized applications: harnessing Bitcoin's blockchain technology*. " O'Reilly Media, Inc.", 2016.

[7] L. Lamport, R. Shostak, and M. Pease, "The byzantine generals problem," *ACM Transactions on Programming Languages and Systems (TOPLAS)*, vol. 4, no. 3, pp. 382–401, 1982.

[8] U. W. Chohan, "The double spending problem and cryptocurrencies," *Available at SSRN 3090174*, 2017.

[9] J. Bonneau, A. Miller, J. Clark, A. Narayanan, J. A. Kroll, and E. W. Felten, "Research perspectives and challenges for bitcoin and cryptocurrencies (extended version)," *Cryptology ePrint Archive, Report 2015/452*, 2015.

[10] J. Nelson, M. Ali, R. Shea, and M. J. Freedman, "Extending existing blockchains with virtualchain," in *Workshop on Distributed Cryptocurrencies and Consensus Ledgers*, 2016.

[11] P. Baran, "On distributed communications," 1964.

[12] T. J. Berners-Lee, "Information management: A proposal," Tech. Rep., 1989.

[13] J. R. Douceur, "The sybil attack," in *International workshop on peer-to-peer systems*. Springer, 2002, pp. 251–260.

[14] H. Kedia, "hKedia/dShare: First release of dShare built with Ethereum," Aug. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3359852

[15] V. Buterin *et al.*, "Ethereum: A next-generation smart contract and decentralized application platform," *URL https://github. com/ethereum/wiki/wiki/% 5BEnglish% 5D-White-Paper*, vol. 7, 2014.

[16] J. Benet, "Ipfs-content addressed, versioned, p2p file system," *arXiv preprint arXiv:1407.3561*, 2014.

[17] B. Cohen, "The bittorrent protocol specification," 2008.

[18] J. Loeliger and M. McCullough, *Version Control with Git: Powerful tools and techniques for collaborative software development*. " O'Reilly Media, Inc.", 2012.

[19] T. Hepp, A. Schoenhals, C. Gondek, and B. Gipp, "Originstamp: A blockchain-backed system for decentralized trusted timestamping," *it-Information Technology*, vol. 60, no. 5-6, pp. 273–281, 2018.