VIETNAM NATIONAL UNIVERSITY - HO CHI MINH CITY
HO CHI MINH CITY UNIVERSITY OF TECHNOLOGY
FACULTY OF COMPUTER SCIENCE AND ENGINEERING

BACHELOR OF ENGINEERING THESIS

AN APPROACH OF HD Wallet ON CURVE Ed25519

COMPUTER SCIENCE COMMITTEE

Supervisors:	Dr. Nguyen An Khuong, Ph.D.	
Examiner:	Dr. No Name, Ph.D.	
Students:	Nguyen Nguyen Phuong	1712726
	Nguyen Dinh Thang	1652318

Ho Chi Minh city, January 2021

Commitment

We commit that the work in this dissertation was carried out following the requirements of the Universitys Regulations and has not been submitted for any other academic organizations. Except where indicated by specific reference in the text, the works are our own.

Ho Chi Minh city, January 2021

Preface

There are two approaches to defense a website in this field: active defense and passive defense. With the expansion of data, it is harder to manually identify and counter cyber threats. However, the rise of using machine learning-based approaches to counter cybersecurity is emerging, with vast applications of machine learning in cybersecurity providers.

In this thesis, we are proposing two machine learning-based approaches to active defense and passive defense. To protect websites actively, we are trying to counter the most common damages to websites: phishing. We have developed a machine learning-based module to actively search for potential phishing websites, verify it and raise alarms about phishing threats online. To the passive defense, traditionally, rule-based WAFs are commonly used. However, they have a high false-positive rate. We are building a machine learning-based validator to validate the requests to support WAFs.

Technically, the phishing website detector will generate a suspicious domain from the legitimate website domain. It will check whether the suspicious website is a phishing one, by comparing the screenshots and textual contents of the suspicious and original one, using a specialized screenshot similarity extraction and GloVe embedding respectively. Then the similarities will be fed to a machine learning model to detect the phishing website.

On the other hand, the malicious request validator is based an observation that legitimate requests to a website usually belong to the same category. The module uses a Convolutional Neural Network to category the suspicious request and checks if that request is in the same category as the normal requests observed or not. This result and the classification of WAF is combined into the final decision.

The dataset for phishing detection is collected from Phishtank. The dataset includes 100 phishing websites and their 64 original websites, forming a dataset with 16300 pairs of phishing-original websites. For malicious request validator, we crawl code snippets from GitHub, as most request payloads are structured languages. We have 127686 records of data in three categories: plain text, JavaScript, and PHP. Also, CSIC 2020 end ECML/PKDD 2007 datasets have been used for verification.

The best experimental results for phishing website detector is a Recall of 96%. The request classifiers achieved almost zero false positive rate and average precision of 95.86%. The malicious request validating module has achieved 96% True Positive Rate and 37% detection rate for CSIC 2020 dataset, while the figure for ECML/PKDD 2007 is 91% and 51%. The module, though not yet completed and cover all the cases to be applicable, show good results in detection XSS attacks and SSI attacks, in which involve the use of JavaScript and PHP in the request classifier.

Contents

List of Figures

List of Tables

1

Introduction

Contents

1.1 Overview

1.1.1 Problem statements

1.1.1.1 HDWallet architect

1.1.1.2 Protocols

1.1.1.3 Algorithm

1.1.2 Explain why this thesis is chosen

1.2 Objectives

1.2.1 Aims

1.2.2 Practical benefits/application

1.3 Scope of the study

1.4 Tentative structure of the study

1.5 Tentative schedule

The target of this project is to build a system that is able to convert hand drawn flowchart image into digital version and allows user to modify the result into the final product before sharing or converting it into other form.

2

Background

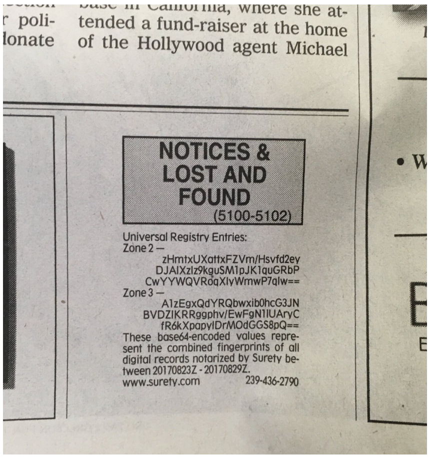
In this chapter, we introduce the foundation knowledge of the thesis, including the history and definition of Blockchain Technology, Cryptocurrency, Hierarchical Deterministic Wallet (HD Wallet) and Cryptography

Contents

2.1 Blockchain Technology

2.1.1 History and Definition

Blockchains are immutable digital ledger systems implemented in a distributed fashion (i.e., without a central repository) and usually without a central authority. The definition of blockchain was introduced to the world by a person (or a group of people) under the name Satoshi Nakamoto on October 31, 2008. It was applied to enable the emergence of a ”purely peer-to-peer (no financial institution or third party) electronic cash” named Bitcoin where transactions take place in a distributed system. In fact, Satoshi did not invent blockchain, and Bitcoin blockchain is not the first chain that ever created. Back in 1991, cryptographers Stuart Haber and Scott Stornetta published a whitepaper ”How to Time-Stamp a Digital Document” in the Journal of Cryptography. Their goal is to digital time-stamping of documents so that it is infeasible for a user either to back-date or to a forward-date digital document, even with the collusion of a time-stamping service. The technology is called a blockchain because the distributed electronic ledger stores items of data in time-stamped digital groups called blocks. Each block includes an alphanumeric code called a ”hash” summing up its data. The hash of each completed block also appears in the next one in the chain, which means that to alter one block you would have to alter all the ones connected to it. These cryptographic dominos function together to protect against tampering or fraud. Base on this theory, the longest-running blockchain, started in 1995, also by Haber and Stornetta, publishes the weekly summary hash value every week in the New York Times (Figure 2.1) and still running strong today.



tended a fund-raiser at the home of the Hollywood agent Michael

NOTICES &
LOST AND
FOUND
(5100-5102)

Universal Registry Entries:
Zone 2 —
zHmtxUXattxFZVm/Hsvfd2ey
DJAIXzlz9kguSM1pJK1quGRbP
CwYYWQVRoqXlyWmwP7qlw==
Zone 3 —
A1zEgxQdYRQbwxib0hcG3JN
BVDZIKRRggphv/EwFgN1lUAryC
fR6kXpapyIDrMOdGGS8pQ==

These base64-encoded values represent the combined fingerprints of all digital records notarized by Surety between 20170823Z - 20170829Z.

www.surety.com 239-436-2790

Figure 2.1: Weekly summary hash value in The New York Times

But the word ”blockchain” or ”block” and ”chain” wasn’t use back then. Only it become known in Satoshi Nakamoto’s Bitcoin paper in the term of ”chain” of ”blocks”. Later people combined the one-word ”blockchain” in mainstream media publications such as Fortune, Forbes, and the Huffington Post as the technology gained greater interest and use. The comunity use that word for Nakamoto’s invention. Bound to emergence of Bitcoin and cryptocurrency, a concise description of blockchain technology is provided by NIST:

> Blockchains are distributed digital ledgers of cryptographically signed trans-

> actions that are grouped into blocks. Each block is cryptographically linked to the previous one (making it tamper evident) after validation and undergoing a consensus decision. As new blocks are added, older blocks become more difficult to modify (creating tamper resistance). New blocks are replicated across copies of the ledger within the network, and any conflicts are resolved automatically using established rules.

Blockchain technology comes handy in a wide range of areas - both financial and non-financial. Non-Financial application opportunities are endless. We can envision putting proof of the existence of all legal documents, health records, and loyalty payments in the music industry, notary, private securities and marriage licenses in the blockchain. By storing the fingerprint of the digital asset instead of storing the digital asset itself, the anonymity or privacy objective can be achieved. For the sake of our thesis, we will mainly focus on the original and surely the most popular application of blockchains - Cryptocurrency.

Cryptocurrencies are digital currencies that use blockchain technology toFigure 2.1) and still running strong today. record and secure every transaction. A cryptocurrency can be used as a digital form of cash that can be used to buy goods and services. It can be bought using one of several digital wallets or trading platforms, then digitally transferred upon purchase of an item, with the blockchain recording the transaction and the new owner. The appeal of cryptocurrencies is that everything is recorded in a public ledger and secured using cryptography, making an irrefutable, timestamped, and secure record of every payment. The ledger displays user account balances and inter-user payments in a currency defined by the ledger itself and not necessarily in one of the traditional currencies. Nevertheless, cryptocurrency may be traded on the stock exchange and exchanged for traditional money, which makes it hard to distinguish between traditional currency and cryptocurrency and as official vs. non-official currency. The most widely recognized cryptocurrency system is Bitcoin.

We believe the ”magic” that brings the above concept of digital currencies to reality, besides blockchain technology, is Nakamoto’s proof-of-work consensus model.

2.1.2 Blockchain Categorization and Generations

Blockchain systems can be:

- *Permissioned blockchain*, where users publishing blocks must be authorized by some authority (be it centralized or decentralized). Users of blockchain have to trust that entity or user who published blocks. Permissioned blockchain networks may thus allow anyone to read the blockchain or they may restrict read access to authorized individuals. This maybe used by organizations that need more control over their blockchain. Some permissioned blockchain networks support the ability to selectively reveal transaction information based on a blockchain network users identity or credentials. Some of famous permissioned blockchain applications are Ripple, which enables interbank transactions, or Sovrin, which is managed by financial institutions and is seeking to build a global decentralized identity system.

- *Permissionless blockchain*, where service providers are not fixed and, in principle, anyone can start operating the service. For example, Bitcoin and the early versions of Ethereum.

Based on the intended audience, three generations of blockchains can be distinguished (Zhao et al., 2016):

- Blockchain 1.0 which includes applications enabling digital cryptocurrency transactions
- Blockchain 2.0 which includes smart contracts and a set of applications extending beyond cryptocurrency transactions
- Blockchain 3.0 which includes applications in areas beyond the previous two versions, such as government, health, science and IoT.

We are now developing blockchain 2.0 but our thesis just focus on cryptocurrency aspect.

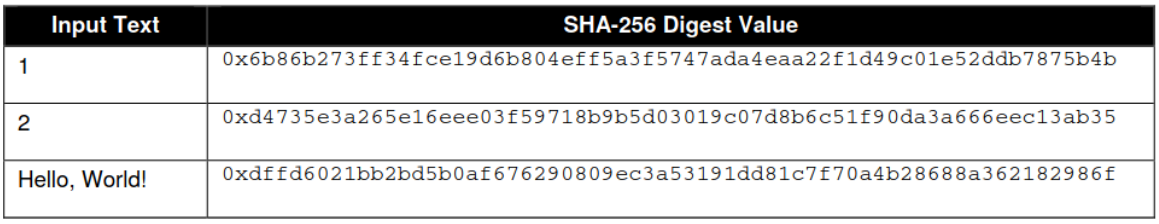
2.1.3 Bitcoin blockchain

Bitcoin is the first application of blockchain and the most famous digital currency ever. As mentioned above, Bitcoin was invented with the publication of a document entitled ”Bitcoin: A peer-to-peer electronic cash system” in 2008 by Satoshi Nakamoto, mentioned as a purely P2P version of electronic cash would allow online payments to be sent directly from one party to another without going through a financial institution. The currency began to use in 2009 when its implementation was released as open-source software. The Bitcoin blockchain is considered to be a world-changing technology because in the first time in human history its solved the biggest problem of distributed system: The Byzantine General’s Problem. We will talk about this in the Bitcoin game of theory and incentives section.

Bitcoin application is one of the permissionless blockchain. It utilize well-known computer science mechanisms (linked lists, distributed networking) as well as cryptographic primitives (hashing, digital signatures, public/private keys) mixed with financial concepts (ledgers, games of theory) in high level. Base on the problems Bitcoin has solved, we examine by dividing it into 3 components:

- Secure and Prevent tempering the data

  Hashes - Cryptographic hash functions (CHF) are used for hashing the content of a block, validating the integrity of data, reduce the size of the message or keys, generating a Bitcoin address. We will show detail at Section 2.3.1. Hashing is a method of calculating a relatively unique fixed-size output (called a message digest, or just digest) for an input of nearly any size (e.g., a file, some text, or an image). Even one single bit change of input will result in a completely different output digest. In Bitcoin and most blockchain technologies, SHA-256 (Secure Hash Algorithm with output size of 256 bits) appear the most. Many computer support hardware level for this algorithm. NIST specified this algorithm for SHA-256 in Federal Information Processing Standard (FIPS) 180-4 as it passed every properties of a cryptographic hashing. Figure 2.2 is an example of SHA-256.



Input Text	SHA-256 Digest Value
1	0x6b86b273ff34fce19d6b804eff5a3f5747ada4eaa22f1d49c01e52ddb7875b4b
2	0xd4735e3a265e16eee03f59718b9b5d03019c07d8b6c51f90da3a666eec13ab35
Hello, World!	0xdffd6021bb2bd5b0af676290809ec3a53191dd81c7f70a4b28688a362182986f

Figure 2.2: Example I/O of SHA-256 Digest Value

Public/Private Key - Asymmetric-key cryptography (or public-key cryptography) uses a pair of keys: a public key and a private key that are mathematically related. It could be infeasible to generate one key from the other. The private key is kept secret while the public key can be to everyone, both keys are hold inside user's Wallet, which we present in Section 2.2. One can encrypt with a private key and then decrypt with the public key. Alternately, one can encrypt with a public key and then decrypt with a private key. Bitcoin uses asymmetric-key cryptography to digitally sign transactions, verify signatures or in some cases, exchange the key. Asymmetric-key cryptography is discussed in Section 2.3.2. Figure 2.3 briefly show message exchange usage of the asymmetric protocol.

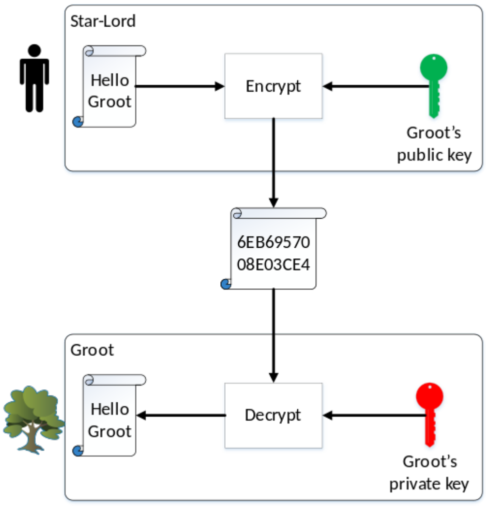


Figure 2.3: Sending private message using Asymmetric-key cryptography

Transactions - Transactions represent transfers of the cryptocurrencies between wallets in the system. A transaction contains input and output. The inputs are usually a list of the digital assets to be transferred. Outputs are the accounts that will be the recipients of the digital assets along with how much digital asset they will receive. All values of in and out cannot be tampered.

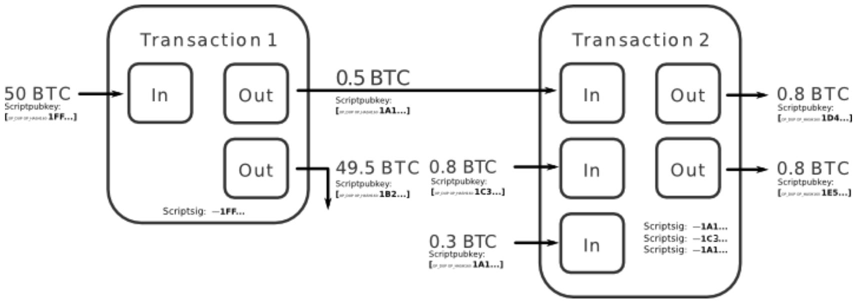


Figure 2.4: An example of bitcoin transaction

All transactions are broadcast to the network and usually begin to be confirmed within 10-20 minutes, through a process called *mining*. Transactions are typically digitally signed by the senders associated private key and can be verified using the associated public key.

Ledgers - A ledger is a collection of cryptographic transactions. Bitcoin ledgers are distributed, the blockchain holds all accepted transactions within its ledgers. Every user can maintain their own copy of the ledger. Whenever new full nodes join the blockchain network, they reach out to discover other full nodes and request a full copy of the blockchain networks ledger, making loss or destruction of the ledger difficult.

The network utilizes cryptographic mechanisms such as digital signatures and cryptographic hash functions to provide tamper-evident and tamper-resistant ledgers. Due to the public distributed network, the Bitcoin blockchain is harder to attack. There is nothing to steal because everything is distributed. If one individual node got taken down, the network will still be running. If targeting the blockchain itself, the attackers will face resistance from the honest nodes present in the system.

Blocks - Transactions, after sent to the network (by wallets, web applications, etc.), will be, if accepted, added to a block that is published by a chosen node. Bitcoin blocks include block header and block data. Figure 2.5 show basic component of a block. Block header contains version, previous block headers hash value (prevBlockHash), a hash representation of the block data (usual Merkle tree* hash), a timestamp, size of the block (bits), a *nonce*. The *nonce* value is manipulated by the publishing node to solve the hash puzzle (see Section 2.1.3) Block data contains a list of transactions and ledger events. Some include other data.

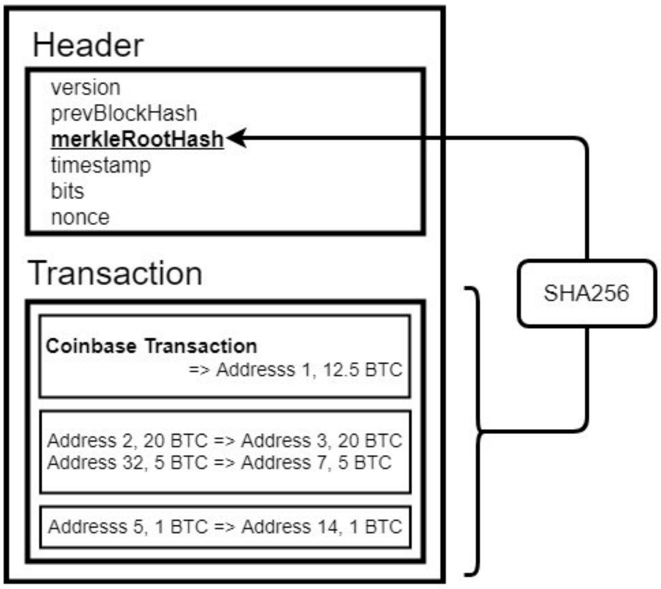


Figure 2.5: Components of Bitcoin block

Chain of Blocks - Blocks are chained together through each block containing the hash digest of the previous blocks header, thus forming the blockchain. If one of the previous blocks were changed, it would result in a different hash. This makes it possible to easily detect and reject altered blocks

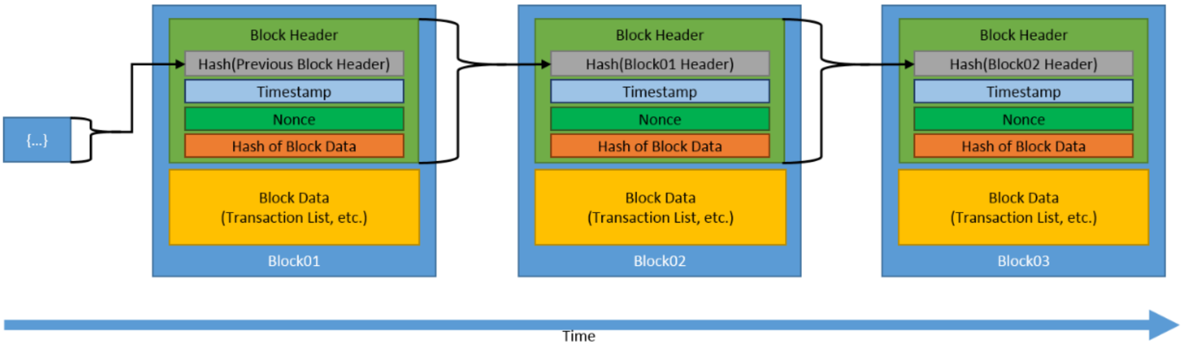


Figure 2.6: Components of Bitcoin block

- Game of theory and Incentives

  Hashes - Cryptographic hash functions (CHF) are used for hashing the content of a block, validating the integrity of data, reduce the size of the message or keys, generating a Bitcoin address. We will show detail at Section 2.3.1. Hashing is a method of calculating a relatively unique fixed-size output (called a message digest, or just digest) for an input of nearly any size (e.g., a file, some text, or an image). Even one single bit change of input will result in a completely different output digest.

- Communication Network

Hashes - Cryptographic hash functions (CHF) are used for hashing the content of a block, validating the integrity of data, reduce the size of the message or keys, generating a Bitcoin address. We will show detail at Section 2.3.1. Hashing is a method of calculating a relatively unique fixed-size output (called a message digest, or just digest) for an input of nearly any size (e.g., a file, some text, or an image). Even one single bit change of input will result in a completely different output digest.

2.2 HD Wallet

2.2.1 Category

2.2.2 Coins

2.2.3 Wallet structure

2.3 Cryptography

2.3.1 Cryptographic hash

2.3.2 Asymmetric-key cryptography

2.3.2.1 Diffie-Hellman algorithm

2.3.2.2 RSA Cryptography

2.3.2.3 EC Cryptography

2.3.3 Twisted-Edward curve and Ed25519

2.3.4 Key derivation function

3

Methodologies and Approaches

3.1 Challenges

3.1.1 HDWallet architecture for Ed25519

3.1.2 Key managements

3.1.3 Attacks on HDWallet

3.2 Approaches

4

Goals

5

Related works

6

Conclusion and Future Work