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

BACHELOR OF ENGINEERING THESIS

MACHINE LEARNING APPROACHES TO CYBER THREATS DETECTION

COMPUTER SCIENCE COMMITTEE

Supervisors: Dr. Nguyen An Khuong, Ph.D.
Dr. Le Lam Son, Ph.D.
Mr. Le Dinh Thuan, M.Sc.
Mr. Nguyen Van Hoa, B.Eng.
Examiner: Dr. Le Thanh Sach, Ph.D.

Students: Nguyen Tri Chan Hung 1552159
Le Nguyen Minh Khoi 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

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. Infact, 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 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, also by Haber and Stornetta, started in 1995, publishes the weekly summary hash value every week in the New York Times and still running strong today.

PICTURE

According to NIST:

Blockchains are distributed digital ledgers of cryptographically signed transactions that are grouped into blocks. Each block is cryptographically linked to the previous one after validation and undergoing a consensus decision. As new blocks are added, older blocks become more difficult to modify. New blocks are replicated across all copies of the ledger within the network, and any conflicts are resolved automatically using established rules.

2.1.2 Bitcoin

2.2 HDWallet

2.2.1 Category

2.2.2 Coins

2.2.3 Wallet structure

2.3 Cryptography

2.3.1 Cryptographic hash

2.3.2 Diffie-Hellman algorithm

2.3.3 RSA

2.3.4 ECC

2.3.4.1 Blockchain: secp256k1

2.3.4.2 Discrete logarithm problem

2.3.5 Twisted-Edward curve and Ed25519

2.3.6 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