Interactive Phishing Filter

A Project

Presented to

The Faculty of the Department of Computer Science

San Jose State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Rushikesh Joshi

December 2015

The Designated Project Committee Approves the Project Titled

Interactive Phishing Filter

by

Rushikesh Joshi

APPROVED FOR THE DEPARTMENTS OF COMPUTER SCIENCE

SAN JOSE STATE UNIVERSITY

December 2015

Dr. Thomas Austin	Department of Computer Science
Dr. Christopher Pollett	Department of Computer Science
Teodoro Cipresso	Department of Mathematics

ABSTRACT

Interactive Phishing Filter

by Rushikesh Joshi

With the advancement in technology and the rapid growth of the internet and its awareness, the count of internet users is increasing day by day. Though it is widely accessed from all over the world, it is very tragic that it is not safe. Internet crimes are increasing day by day and hence everyone wants safe and secure browsing. The problem is one can not find a way which guarantee safe and secure browsing. Phishing is one of the prevalent technique used by attackers to breach security and steal private and confidential information. It has compromised millions of users' data. Conventionally there are few methods such as "Blacklist" method which are being used to prevent phishing. None of these techniques or methods which has been implemented till now has given a satisfactory solution. In reference to such big issue, our goal is to come up with a solution and implement it to gain reasonable accuracy for phishing detection. We will be using machine learning algorithm and approximate string matching methods to come up with a filter/add-on that will provide safe and secure browsing to end user.

ACKNOWLEDGMENTS

Final output and results of this project would not have been possible without the kind supervision and help of many individuals. I would like to grab the opportunity to express my sincere thanks to all of them.

I would like to express my gratitude to my project adviser Dr Thomas Austin for his continuous supervision, useful comments and engagement through the learning process of this master project and for suggesting me good topic. Furthermore I would like to thank Dr Christopher Pollett for his valuable suggestions during the execution of the project. Special thanks to Professor Teodoro Cipresso for encouraging me to execute this project and valuable feedback.

TABLE OF CONTENTS

CHAPTER

APPENDIX

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Internet has become another world itself in which you can get everything which is sufficient to live a life. You can watch movie, order food, earn money make friends and find your soul mate even. It has created huge revolution in our life. According to statics, almost 3 billion users have access of internet. Unfortunately, this world has become very vulnerable and unsafe. Internet banking, ecommerce, email services share significant amount of usage of this modern world. Often information and data being transmitted through the internet is very valuable and confidential. Poor security and some vulnerabilities have made easy to gain access of such confidential data for bad programmers.

Among all of these security attacks, Phishing attack is known for stealing private information. According to Kaspersky LabâĂŹs database, 1 million number of phishing attacks has been increased in the first quarter of 2015 compare to previous quarter. World is becoming online which has increased the number of websites so as phishing attacks. Phishing attacks indirectly affect many well know organizationâĂŹs reputation. Many solutions were provided till now to detect phishing websites. None of those solutions provides good accuracy and performance when it comes real time safe browsing.

1.1 What is phishing

Phishing can be defined as activity of collecting unauthorized and confidential data such as username, password, credit card detail, bank account details electronically. First time phishing activity was defined in detail in a paper and presentation delivered to the 1987 International HP Users Group, Interex [2]. First phishing attack was registered by AOHell hacking tool [3].

There are various types of phishing attacks. Phishing can be done in numerous ways. Following is the type of phishing attacks.

- Link manipulation
- Filter evasion
- Website forgery
- Covert redirect
- Phone phishing

Our main focus in on âĂIJLink manipulationâĂİ type of phishing attack. In this attack URLs appear to belong to the valid organization. URLs are obfuscated very smartly that it is difficult sometimes to differentiate by human eye also. LetâĂŹs take an example of such website.

As you can see in the first image this site claims to be genuine Paypal, a worldwide online payments system company. Second image is the real original website of Paypal. If you closely look then you easily see the difference between two websites. There are some visible difference between these two websites like logo of the company, favicon and secured certificate. Different types of techniques is used to redirect such fake websites. Sometimes users do not see this visible difference and becomes the victims of this phishing attacks.

Once the data is collected, different types of forgeries is done by hackers. This scam can be of Millions of money sometimes. Sometimes its all about private and confidential information of celebrity which can be leaked to spoil the image of the same person.

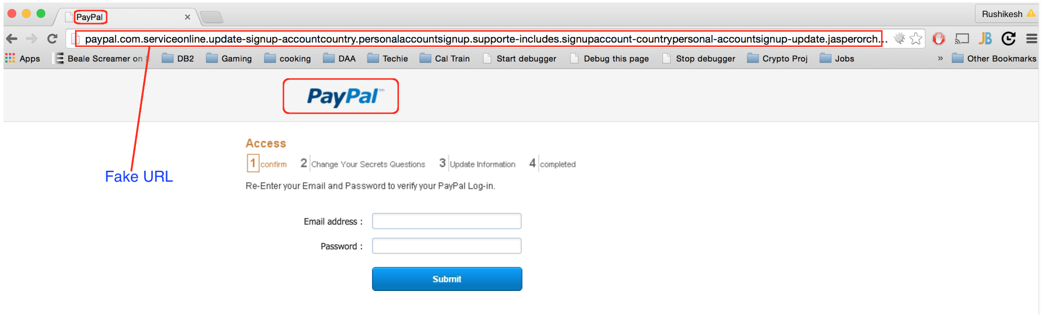
PayPal
paypal.com.serviceonline.update-signup-accountcountry.personalaccountsignup.supporte-includes.signupaccount-countrypersonal-accountsignup-update.jasperorch...
Fake URL
PayPal
Access
1 confirm
2 Change Your Secrets Questions
3 Update Information
4 completed
Re-Enter your Email and Password to verify your PayPal Log-in.
Email address :
Password :
Submit

Figure 1: Fake Paypal website

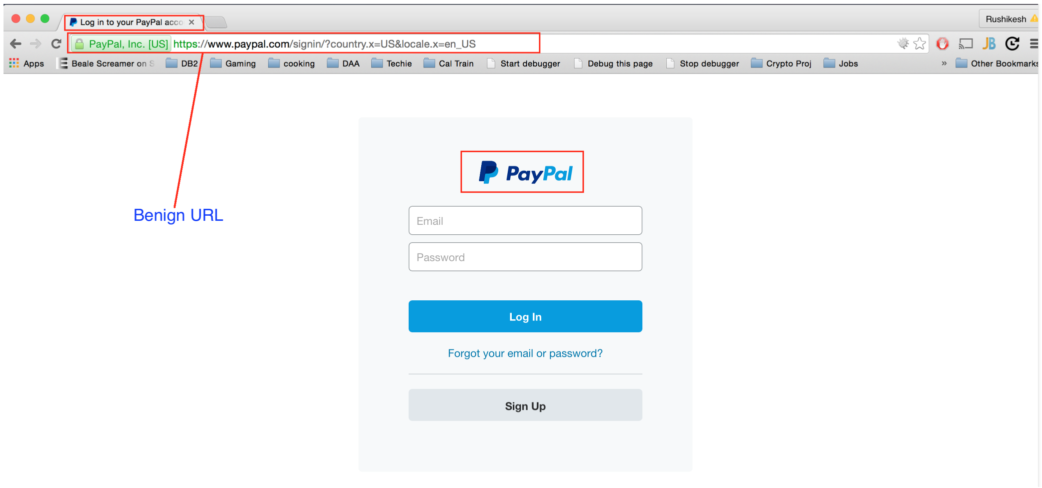
Log in to your PayPal acco
PayPal, Inc. [US] https://www.paypal.com/signin/?country.x=US&locale.x=en_US
Benign URL
PayPal
Email
Password
Log In
Forgot your email or password?
Sign Up

Figure 2: Original Paypal website

1.2 Problems

To address this Phishing problems lots of academic and business research has been done so far. These methods either uses blacklist methods or some type of heuristics or machine learning techniques. Blacklist method is more common and popular for all the browser. All browser contains some manually verified list of phishing websites. This technique contains fairly low positive rate. But when it comes to fresh phishing website, it does fail. It is not much effective to newly developed phishing websites.

Second approach is to adopt smart heuristics which is given some training data to train the heuristic. This heuristic collects different types of features of the websites and based on that it decides the authenticity of the website. So far no heuristic has defined which has better accuracy, performance and seamlessly browsing experience. In addition, scope of the detection of phishing website should not be limited. We referred to similar kind of one research paper which talks about all of these features. We tried to collect all the mentioned attributes and experimented with different machine learning classifiers. This method highly depends on approximate string matching algorithms and WHOIS server queries. In the development procedure we used only one third party services which is WHOIS server. Otherwise every algorithm run on native system. Our tests with this method shows some promising results. With the help of this attributes and approach we were able to achieve almost 97

One more research paper has proposed to use image processing techniques. They tried to emphasize on the favicon of the website. They proved that significant difference is found when you compare the favicon of original website and spoofed website. Limitation of this approach is website must have favicon otherwise this method will fail.

1.3 My contribution

My main role was go give practical shape to the idea which they talked in the research paper. Since everything was on the paper only and they did not talk about any implementation part, it required lot of efforts to make it practical.

There are couple of different way to make things happens. We split this entire addons in three different parts. First component resides on client side. Second component is web services which is kind of middle layer between third component, machine learning algorithms and addons.

We opted for making firefox addons. It has got pretty nice documentation to start with. Then there were couple of options available for machine learning libraries. Weka seemed quite distant options for us. It provides api for different languages.

We collected different samples of phishing websites from Phishtank[] which is famous organization for providing database of phishing websites. We used Alexa to get samples of benign websites. Total 1773 URLs collected for training data set. Out of 1773 URLs 829 URLs were phishing and 944 URLs were benign. We tried to apply different machine learning classifiers to get the best accuracy. As per our results, our filter achieves almost 97

CHAPTER 2

Symbolic Cross-References

This chapter has more information on symbolic cross-referencing. In addition, there are examples showing how to typeset a figure and a simple table.

2.1 Figures

Figure 3 has been typeset so that it occupies half the width of text on a page (`width=0.5`). Note the symbolic cross-reference, that is, you simple `\ref{fig:encrypted_virus}` to obtain the appropriate number for Figure 3.

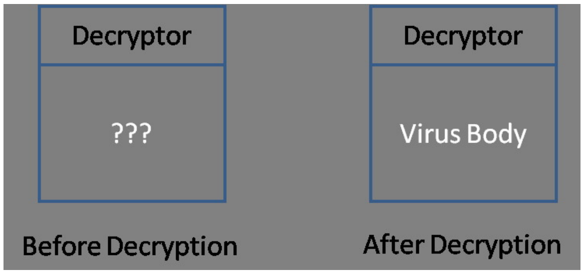

Figure 3: Encrypted virus before and after decryption.

If you move things around, insert or delete text, Figure 3 will still give you the correct number. This is a very handy feature.

2.2 Tables

Table 1 shows some interesting stuff. I hope you enjoy it.

2.3 Citations

References are cited using the `\cite` command. For example, see [1] for information on the Zodiac killer. You can include more than one reference in a single

Table 1: MysteryTwister Zodiac Challenge

Case	Description	Percentage
1F	Fast outer hill climb and English stats	13.00%
1S	Slow outer hill climb and English stats	4.00%
2F	Fast outer hill climb and Zodiac 408 stats	70.00%
2S	Slow outer hill climb and Zodiac 408 stats	84.00%

citation, like this [1, 2, 3, 4, 5, 6, 7, 8, 9] or this [10, 11].

2.4 The Last Word

There are a lot of other uses for symbolic cross-referencing. For example, this chapter title has a label so I can refer to Chapter 2 (as I just did). My recommendation is that you include a label with anything and everything that is numbered. If it has a number, there is a good chance that you will refer to it at some point.

Finally, note that you need to LaTeX a file at least twice (three times to be safe) without changing anything to be sure that the numbering is correct. Why is that? Good question. When you LaTeX your document, various auxiliary files are created. Most of these files contain information related to cross-references that appear in your `.tex` file (or files). The next time you LaTeX your document, the information from the previous run is read from these auxiliary files. Consequently, after making changes to any `.tex` file, you will need to LaTeX your document at least twice to be sure that the cross-reference information is updated in the resulting pdf. In fact, you should run LaTeX three consecutive times to be certain that cross-references are correct.

CHAPTER 3

The Files

For this thesis format, you LaTeX the file `thesis.tex`. But first, you need to make some modifications to `thesis.tex`. The title of your report, committee members, etc., are specified in `thesis.tex`. All of the things that you need to modify are indicated by comments beginning with five consecutive asterisks, so search for "*****" in `thesis.tex` and make the necessary changes.

You put the actual content of your report in the following files:

- `abs.tex` — abstract
- `ack.tex` — acknowledgements
- `chap1.tex`, `chap2.tex`, and so on — chapters
- `bib.tex` — references
- `appA.tex`, `appB.tex`, and so on — appendices (if any)

LIST OF REFERENCES

[1] Sheng, Steve, *An empirical analysis of phishing blacklists*, V-2:2009,Sixth Conference on Email and Anti-Spam (CEAS)

[2] Announcing the Advanced Encryption Standard (AES), NIST, FIPS 197, November 26, 2001,
`http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf`

[3] P. K. Basavaraju, Heuristic-search cryptanalysis of the Zodiac 340 cipher, Master's report, Department of Computer Science, San Jose State University, 2009,
`http://scholarworks.sjsu.edu/etd_projects/56/`

[4] K. Briggs, English and Latin digram and trigram frequencies,
`http://keithbriggs.info/documents/english_latin.pdf`

[5] G. Claston, 340-cipher — overview and examination,
`http://www.zodiackiller.com/mba/zc/69.html`

[6] Glurk, The final 18 symbols are nothing but filler!,
`http://www.zodiackillerfacts.com/forum/viewtopic.php?f=49&t=423`

[7] T. Jakobsen, A fast method for the cryptanalysis of substitution ciphers, *Cryptologia*, 19:265–274, 1995

[8] D. Kreher and D. Stinson, *Combinatorial Algorithms: Generation, Enumeration and Search*, CRC Press, 1998

[9] G. Shanmugam, R. M. Low, and M. Stamp, Simple substitution distance and metamorphic detection, *Journal of Computer Virology and Hacking Techniques*, 9(3):159–170, 2013

[10] M. Stamp and R. M. Low, *Applied Cryptanalysis: Breaking Ciphers in the Real World*, Wiley, 2006

[11] M. Stamp, *Information Security: Principles and Practice*, second edition, Wiley, 2011

@booksheng2009blacklist, title=An empirical analysis of phishing blacklists, author=Sheng, Steve, volume=2, year=2009, publisher=Sixth Conference on Email and Anti-Spam (CEAS)

@articlejoinsoon2013interactive, title=Interactive Website Filter for Safe Web Browsing, author=Jo, Insoon, Eunjin Jung, and Heon Young Yeom, journal=Information Science and Engineering, volume=29, number=1, pages=115-131, year=2013, publisher=IEEE

@bookwangguang2014favicon, title=Favicon âĂŞ a Clue to Phishing Sites Detection, author=Wei Wang, Guang-Gang Geng Xiao-Dong Lee, and Shian-Shyong Tseng, year=2014, publisher=IEEE

APPENDIX A

Zorak Likes Beans

A.1 Oh Yes He Does

Appendices can have sections and subsections and so on.

A.2 Really

Sections, subsections, or whatever should come in pairs.

APPENDIX B

Everybody Wants to Be Space Ghost

Space Ghost: Everybody wants to be Space Ghost
Everybody near and far
Hey, ma, look at me, I'm on TV
Everybody wants to be a star
I'm Space Ghost, Mr. Space Ghost
I've got big muscles, And I can dance
When Zorak tries to bug me, I zap him with my power bands

Zorak: Uhhhh… I don't think that will be necessary

Space Ghost: I think I'll zap him with my power bands
On Saturn and Jupiter and Neptune, too
I've been hearing it from coast to coast
I'd give anything, If for just one day
I could be a super hero like Space Ghost

Zorak: Ugghh… Why would anyone want to be Space Ghost?

Space Ghost: Everybody wants to be just like me!