Systematic analysis of phishing websites

Tanguy Bénard Marie Montégu
Email: {tanbe975,marmo931}@student.liu.se
Supervisor: Niklas Carlsson, {niklas.carlsson@liu.se}
Project Report for Information Security Course
Linköpings universitetet, Sweden

Abstract

Phishing websites impersonate trusted legitimate websites and request sensitive information to the user, usually with the intention of stealing money from them. Today, a systematic detection of these kinds of widespread frauds is therefore a huge challenge for cybersecurity. This paper gathers a dataset of 250 manually filled phishing websites, analyzing different relevant features that could contribute to give deep insights of these types of malicious attacks. We particularly found that passwords were the most requested information, as it gives access to accounts and possibly financial details. Furthermore, banks are the most impersonated websites, and even more pages were asking for bank details. Phishing websites may finally be flagged with different characteristics, such as suspicious URLs or suspicious source codes.

1. Introduction

Phishing is a social engineering attack, a concept based on the victim's credulity. Phishing websites links are often sent by e-mail or by private messages. If the victim clicks on it, he or she will be redirected to a website that looks familiar. Indeed, phishing websites impersonate common verified websites but with fraudulent messages and intentions. Nevertheless, they can't have the same URL as the website impersonated so it is usually changed a bit or totally.

Finally, phishing websites usually ask for very sensitive information such as e-mail, phone number, identity, passwords or even bank details. They can even hide fake executables, and if the user unintendedly downloads it, his machine could be infected by some malware. That is called drive by download.

Unfortunately, not so many people around the world are sufficiently sensitized to the existence of such fraudulent malicious websites. Indeed, phishing is the first cause of compromise : in 2020, a FBI IC3 report recorded the number of phishing's incidents as twice the number of other computer attacks. In 2016, another report estimated that 30% of phishing emails were eventually opened, and, in 12% of cases, the link was clicked on.

Considering that these types of attacks are deployed at a very large scale, victims of phishing actually represent a huge number of people. Furthermore, It also represents a huge economic loss of more than 50 million dollars per year for consumers and companies in the United States.

In this paper, we give some background information, detailing how we could securely access phishing websites and explaining in what ways we collected the information requested within a structured database. We then tried to implement a semi-automatically algorithm to cope with the time-consuming manual tasks of filling this database. We then present our results and give a precise analysis of them. Finally, we compare our findings with some related works we read through during this project, and ultimately jump to conclusions.

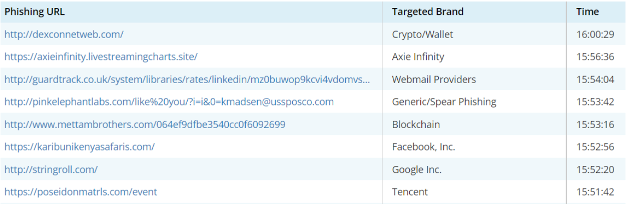
2. Background

It seems important to underline the main issues and challenges coming with analyzing phishing websites. In this paper, we try to give solutions to key questions such as securing our devices, structuring our database or collecting information more efficiently.

2.1 Secured Access to Phishing Websites

Access and security are two essential bullet points when it comes to phishing websites. Indeed, as we wanted to build a whole database on these types of malicious websites to drive analysis on it, we obviously had to feed it with actual phishing websites.

The two major sources we used to find and reach phishing websites are two trusted anti-phishing websites : PhishTank and OpenPhish. Phishtank is a collaborative website where users can vote whether or not it is a true phishing website. On the contrary, OpenPhish does work in a different way. Its administrators actually receive millions of URLs by their partners and they then identify phishing websites by performing intelligence analysis without human intervention. Figure 1 effectively shows a table where phishing websites are recorded on Openphish, giving their URL and the brand they usurp.

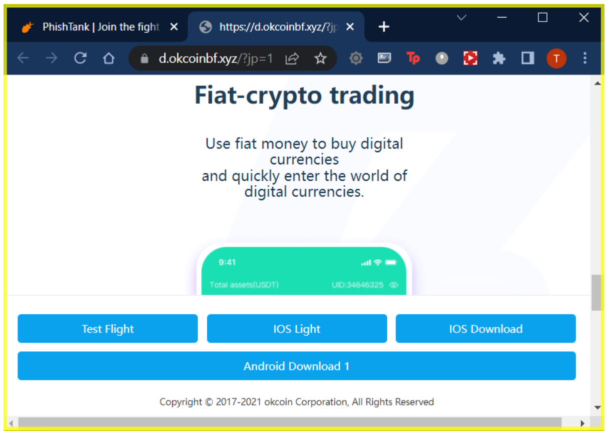


Phishing URL	Targeted Brand	Time
http://dexconnetweb.com/	Crypto/Wallet	16:00:29
https://axieinfinity.livestreamingcharts.site/	Axie Infinity	15:56:36
http://guardtrack.co.uk/system/libraries/rates/linkedin/mz0buwop9kcvi4vdomvs...	Webmail Providers	15:54:04
http://pinkelephantlabs.com/like%20you/?i=i&0=kmadsen@ussposco.com	Generic/Spear Phishing	15:53:42
http://www.mettambrothers.com/064ef9dfbe3540cc0f6092699	Blockchain	15:53:16
https://karibunikenyasafaris.com/	Facebook, Inc.	15:52:56
http://stringroll.com/	Google Inc.	15:52:20
https://poseidonmatrls.com/event	Tencent	15:51:42

Figure 1. Screenshot of Openphish

Moreover, when accessing phishing websites, we have to be very careful to not enter sensitive information or click on other links that could download malware. Even if our only awareness should be sufficient, we want to make sure that our computers are truly safe when navigating through malicious websites. Therefore, we chose to add extra protection by using a sandbox.

We execute our browser in a sandbox with an application called Sandboxie, as the yellow borders around the chrome browser show on figure 2.



Fiat-crypto trading

Use fiat money to buy digital currencies
and quickly enter the world of digital currencies.

Copyright © 2017-2021 okcoin Corporation, All Rights Reserved

Figure 2. Browser used within a Sandbox (with Sandboxie)

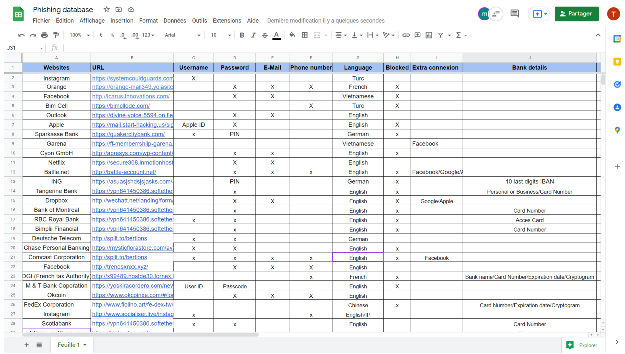
When navigating on phishing websites flagged by Openphish, our browser (both Google Chrome and Mozilla FireFox) blocked it and warned us that we were trying to access a malicious website. Nevertheless, it is the opposite for Phishtank : most of the phishing websites were not blocked by our browser and could access them without any warning.

Obviously, this dissimilarity is due to a difference in the way these two anti-phishing websites collect and record their database, whether they are collaborative or not. It actually seems that OpenPhish mostly records verified malicious websites, with many checks and very few errors whereas PhishTank could sometimes record true websites.

2.2 Collection of Information Requested by Phishing Websites

Now that we are able to access phishing websites securely, the information we collect has to be stored in a database that will finally be analyzed.

We chose to use a collaborative excel sheet to store the data. In the excel table, we fill in the websites impersonated, the URL of the phishing websites and all the information asked on the page : email, password, username, bank details, etc. We then draw an X or put a comment when the field is required, otherwise we leave the cell blank. If we had more than one information requested (such as bank details with card number, IBAN…), we wrote the different information separated with a slash. A clear and precise syntax is important to analyze the data later.



Phishing database
Fichier Édition Affichage Insertion Format Données Outils Extensions Aide Dernière modification il y a quelques secondes
Partager
100%
Arial
10
J31
Feuille 1
Explorer

Websites	URL	Username	Password	E-Mail	Phone number	Language	Blocked	Extra connexion	Bank details
Instagram	https://systemcouldguards.com	X				Turc			
Orange	https://orange-mail349.yolasite		X	X	X	French	X		
Facebook	http://icarus-innovations.com/		X	X		Vietnamese	X		
Bim Cell	https://bimcllode.com/				X	Turc	X		
Outlook	https://divine-voice-5594.on.fle		X	X		English			
Apple	https://mail.start-hacking.us/sig	Apple ID	X			English	X		
Sparkasse Bank	https://quakercitybank.com/	x	PIN			German	x		
Garena	https://ff-memberrship-garena.					Vietnamese		Facebook	
Cyon GmbH	http://apresys.com/wp-content/		x	x		English	x		
Netflix	https://secure308.inmotionhost		X	X		English			
Batlle.net	http://battle-account.net/		x	x	x	English	x	Facebook/Google/A	
ING	https://asuasjshdsjsjasks.com/		PIN			German	x		10 last digits IBAN
Tangerine Bank	https://vpn641450386.softether		x			English	x		Personal or Business/Card Number
Dropbox	http://wechatt.net/landing/form/		X	X		English	X	Google/Apple	
Bank of Montreal	https://vpn641450386.softether		x			English	x		Card Number
RBC Royal Bank	https://vpn641450386.softether	x	x			English	x		Acces Card
Simplii Financial	https://vpn641450386.softether		x			English	x		Card Number
Deutsche Telecom	http://split.to/bertions	x	x			German			
Chase Personal Banking	https://mysticflorastore.com/av	X	X			English	x		
Comcast Corporation	http://split.to/bertions	x	x	x	x	English	x	Facebook	
Facebook	http://trendsxnxx.xyz/		X	X	X	English			
DGI (French tax Authority)	http://x99489.hostde30.fornex.				x	French	x		Bank name/Card Number/Expiration date/Cryptogram
M & T Bank Coporation	https://yoskiracordero.com/new	User ID	Passcode			English	X		
Okcoin	https://www.okcoinxe.com/#/log		X	X	X	English			
FedEx Corporation	http://www.fiolino.art/fe-dex-tw/					Chinese	x		Card Number/Expiration date/Cryptogram
Instagram	http://www.socialiser.live/instag	x			x	English/IP			
Scotiabank	https://vpn641450386.softether	x	x			English			Card Number

Figure 3. Organization of our database

In our database, we add extra information that seems to be relevant. For instance, phishing websites are obviously coded in different languages as they do not target the same persons. Besides, we noticed that these websites were often blocked by our browser, but not all the time. We thus reported this information in our table.

As Figure 3 illustrates, most of the existing phishing websites URLs are very far from the impersonated website's one. Therefore, it might be a very first warning to detect that the website is probably malicious. Nevertheless, some phishing websites execute a very accurate impersonification, with a very close URL, and a webpage that really looks familiar for users. This is dangerous because it is likely to reinforce the trust people might have when giving their information, without raising any doubt or mistrust.

Figure 4 is a screenshot taken from a phishing website impersonating Apple. The URL, apprre.com is actually really close to Apple's URL, apple.com. Again, this is a way for the attacker to not raise any distrust from the user that might not pay attention to a URL that looks familiar. The web design is also really close to Apple Website so that users might give their information easily.

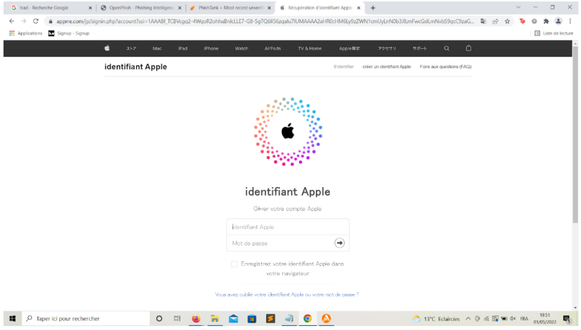

trad - Recherche Google
OpenPhish - Phishing Intelligenc
PhishTank > Most recent unverifi
Récupération d'identifiant Apple
apprre.com/jp/signin.php?account?ssi=1AAABf_TCBVcgq2-4WqoR2ohhaBnlcLLE7-G8-5gTQ6lI56zqalu7IUMAAAA2aHR0cHM6Ly9zZWN1cmUyLnN0b3JlLmFwcGxlLmNvbS9qcC9zaG...
Applications
Signup - Signup
Liste de lecture
ストア
Mac
iPad
iPhone
Watch
AirPods
TV & Home
Apple限定
アクセサリ
サポート
identifiant Apple
S'identifier
créer un identifiant Apple
Foire aux questions (FAQ)
identifiant Apple
Gérer votre compte Apple
Identifiant Apple
Mot de passe
Enregistrez votre identifiant Apple dans votre navigateur
Vous avez oublié votre identifiant Apple ou votre mot de passe ?
Taper ici pour rechercher
13°C Eclaircies
FRA
19:51
01/05/2022

Figure 4. A phishing website impersonating Apple

2.3 Implementation of a Semi-Automatic Algorithm

To reach a bigger database, we wrote a semi-automatic algorithm on python that downloaded the source code of multiple urls of phishing websites and saved the label, class, id of the fields that the website requests.

After that, another python program asked us for each field what was the main field if it cannot figure it out. For example, <input type="password" id="password" name="password" placeholder="Password"> would be automatically turned into “password”, but <input type="text" class="form-control" id="chocolate" placeholder="Mot de passe"> won’t, so we had to write by ourselves “password”.

However, after several tries, we saw that the results were pretty poor. Firstly, lots of urls were not accessible because they were blocked by chrome and we could not bypass it using the python library selenium. The python library requests are allowed to bypass it since we do not use any browser but some pages have to be fully loaded or redirected in order to display all the fields. Globally, less than one third of the websites on openphish contained fields and the half was exploitable. 30 websites are displayed in openphish at the same time and are totally updated hourly so this algorithm was not very useful on openphish. It was more useful on PhishTank.

But even if we managed to get a lot of websites, there were plenty of useless input tags that could be considered as noise. Sometimes, we could not figure out what type of information the field was requested because of the inclarity of class, id or label names. And most importantly, it was impossible to know what website the page was impersonated so we had to connect to the website eventually.

We concluded that it may take a little less time to complete the database but with very poor results. So we decided to enter all the information manually in order to have a more accurate database.

3. Analysis of the results

Having a dense database composed of 250 websites allowed us to analyze it and look for patterns. Indeed, it would be interesting to know what is the part of phishing pages targeting bank websites, social networks… For that, we coded a python program that took in input a csv file of our database and printed relevant statistics of our choice.

3.1 Types of websites impersonated

Practically all kinds of websites that ask for username/password/card numbers can be targeted by phishing attackers. The table 1 indicates the main types of websites that were recorded by our research.

As we expected, the motive is mostly financial. Banks represent 25.2% of our database but other websites can provide money to the attacker. Indeed, if we add the websites asking for cryptocurrencies wallet (14.4%) and shopping websites (6.0%), it is nearly half of the database websites (**45.6%**) that brings direct financial profit to the attacker.

This percentage can even be higher if we add websites that provide software. These sorts of phishing websites can have multiple goals such as : collect banking details while paying the software, have access to the user’s collection of software (Steam), make the user download malicious software.

Other websites that are very often targeted by phishing attackers are communication websites such as social networks, e-mails, clouds... They represent **33.6%** of the database and could have multiple goals. An attacker who has access to the Facebook account of a person can thus share malicious (phishing) links to all of his contacts while being less suspicious than a fake account.

Another risk is the gain of very sensitive or intimate information. This information can be used as leverage to yet again, make profit (e.g. sextortion if this information involves intimate videos or pictures), for personal purposes (e.g. revenge porn)...

Email accounts suffer from the same danger, but another arises : the recovery of tons of passwords linked to the email. After that, all kinds of websites can be targeted.

We identified several logistic companies such as DHL which were victims of phishing attacks. Indeed, people waiting for their delivery are more likely to commit errors and enter their credentials. Attackers could ask for the payment of the delivery before the goods are delivered, fooling the victim into a problem with the delivery and asking to call a premium rate number, etc… Looking for these types of attacks linked to DHL led us to the Brand Phishing Report of 2021 (*Check Point Research*). In the 4th quarter of this year, DHL was in fact the most impersonated website with 23%, even ahead of Microsoft (20%).

The final type identified is entertainment. We did not find a lot of these websites (3.6%) but among them, Netflix was the main target. Phishing attackers rarely use the "entertaining part" of the accounts they hacked. It is often linked to financial goals. After being hacked, these accounts can thus be sold to other users at a lower price than the subscription price of the account.

Even if all the websites are not linked to the collecting of banking details immediately, the immense majority can bring profit to the attacker in a second term.

Table 1. Different types of websites identified: On a database of 250 websites, all were labeled with a different type (16 types in total)

Type of website	**Proportion**
Bank	25.2%
Social network	14.4%
Cryptocurrencies	14.4%
E-mail	14.0%
Shopping	6.0%
Software distribution	5.2%
Telecom	4.8%
Postal service	4.4%
VOD	3.6%
Others	8%

3.2 Types of information requested

Besides the types of websites that phishing attackers target, the analysis of what information is requested by these pages has been done. Table 2 lists all the results of our analysis.

Passwords are obviously very sensitive information and are thus asked 78.4% of the time. This number could have been even higher if we added the websites that request first username and then passwords (we did not enter any information in our search so we would not know if they asked for passwords later on).

The passwords are very often linked to a way of identifying an account (username, email, phone number). 84.4% of the websites want their users to enter one of these three types of information. With practically a half demanding an email and password (48.0%), this is way more than the first analysis that declared only 14.0% of the websites impersonated email websites. If the user does not change passwords between his email box and other websites, he could be in great danger if he enters his authenticating information on a phishing website.

Phone numbers are also often asked by phishing websites: little less than a third (28.8%). Again, it could be a breach to communicate with the victim but not only. Now, telephone lines could be hacked and be used to call overtaxed numbers owned by the attackers, use reverse engineering and so on. Knowing the phone number is not harmless and could be used against the victim.

In total, 19.6% of the websites ask for bank details, such as cards or account numbers. Asking directly for card numbers is straight-forward but seems to work a bit with 6.0% of the websites demanding to enter it.

13.6% of the websites allow a connection through other ways such as Google and Facebook. If the user does not notice all the authorisations given to the phishing website, it could have access to the reading and editing of his Google drive for instance. Too many permissions given through a Facebook connection could also lead to a lack of privacy with date of birth, photos, names, public profile, pages liked transmitted to the attacker.

A small portion of websites asked for personal information such as names, address, birth date (9.2%). Information, even very sensitive, is harder for the attacker to take advantage of.

Table 2. Different information requested by phishing websites : On a database of 250 websites. The proportion is the number of websites where the information was requested, often through an input field.

Information requested	**Proportion found**
Password	78.4%
Email	50.4%
Username	36.8%
Phone number	28.8%
Bank details	19.6%

Extra connection	13.6%
Name	7.6%
Address	6.0%
Birth date	2.8%

3.3 Other results

Through our analysis, we add other columns that information requested and websites impersonated.

Concerning the language of the websites tested, we found more than a half in English (63.2%), followed by French (8.4%), German (6.0%) and Vietnamese (4.0%). Other languages are below 3% and there are a total of 18 languages in the database. It is difficult to draw conclusions from these percentages as it is biased by the methodology used by both OpenPhish and PhishTank to add malicious links on their page.

However, as we expected, English websites are predominant, but phishing attacks do not touch only the USA (or other English speaking countries) but target the entire world.

A column of the table was intended to check whether or not the website was blocked by the browser used (Google Chrome in the study). Finally, we found that 70.8% of the websites were previously blocked by Chrome before entering it. Again, this number could be biased knowing that the websites were already on an anti-phishing site, thus certainly already reported as malicious. Nevertheless, it shows the efficiency of browsers to analyze the source code and find patterns specific to phishing sites.

4. Related work

Some research underlined a scope of various features that could characterize phishing websites. Common properties of phishing attacks consist for instance in the use of logos from the legitimate website, suspicious URLs, or some malicious user inputs requesting sensitive information. [2]. Furthermore, phishing websites are most of the time short lived, they copy HTML code from the impersonated website and suffer from lack of familiarity with English, containing grammatical errors and mistakes.

Domain-based features were also pointed out, such as the age of the domain, the DNS record, the website traffic, the PageRank, the Google Index, or the number of links pointed to the webpage [1]. Abnormalities could also concern HTML and JavaScript based-features. Indeed, phishing websites are way more redirected than legitimate ones and may use JavaScript Code to fake the URL shown to the user. Furthermore, these kinds of malicious websites sometimes disable right-clicking for the user or use pop-up windows, which can be seen in the source code.

Analyzing the source code is actually a good way of detecting whether or not a web page is phishing. It is really common to observe that phishing websites' source codes are getting out of W3C standards [2], going against the norms about images, https usage, domain, email or more commonly suspicious URLs. URL detection is by the way an efficient method for systematic detection of phishing websites [3]. The latter might use “basename” URLs, using a very similar URL than the impersonated website, changing a single letter for example. Other techniques are used by phishing websites such as subdomain, path domain and brand name based URLs.

A current challenge concerning phishing is the accurate detection of malicious websites and this is why a good extraction of features is important. Meanwhile URL based detection seems to gain one’s spurs, machine learning detection, which could by the way also be URL based, brings hopes and promises for cybersecurity [4]. Random Forest is an example of an algorithm that got a high precision in the detection process, reaching an accuracy of 98,35%.

5. Conclusion

In the detection process of phishing websites, choosing relevant features is usually essential to get a high accuracy. Deeply looking at phishing websites is an interesting task giving more precise insight on the way they are created and what they aim at. In this paper, we manually analyzed a database filled with 250 malicious websites impersonating legitimate ones. To get access to them, we chose to use Phish Tank and Open Phish which update in real time their storage of fraudulent URLs. Besides, we navigated within a sandbox to ensure complete security for our computer devices.

We filled our database with phishing websites URLs, name of the brand impersonated, and information requested. Banks seem to be the most common target for phishing, concerning 25.2% of the websites we inspected. Password is obviously the most requested information, with 78.4% asking for it. This is easy to understand because it is used by plenty of websites, and gives access to accounts and sensitive information when combined with an ID. Besides, money usually fuels the heart of the process : 19.6% of websites ask for bank details, but most of the time a single access to ID and password allow access to sensitive financial information such as a recorded bank card for instance. Access to emails also pave the way to the findings of more precious details on various other key websites.

References

[1] R.M. Mohammad, F. Thabtah, L. McCluskey, “Phishing Websites Features”, 2015.

[2] M.G. Alkhozae, O.A. atarfi, “Phishing Websites Detection based on Phishing Characteristics in the Webpage Source Code”, International Journal of Information and Communication Technology Research, 2010.

[3] M.E. Maurer, L. Höfer, “Sophisticated Phishers Make More Spelling Mistakes: Using URL Similarity Against Phishing”, 2012.

[4] T.O. Ojewumi, G.O. Ogunleye, B.O. Oguntunde, O. Folorunsho, S.G. Fashoto, N. Ogbu, “Performance evaluation of machine learning tools for detection of phishing attacks on web pages”, Scientific African, 2022.