

# Attack Of The Clones!

## Research Document

Conor Lewis – C00246763

Supervisor – Chris Staff

## Abstract

The ability to clone webpages has been a tool utilized by developers for some time as it eases the workload of creating multiple websites and can lead to innovation and learning opportunities when utilized in a safe manner, but issues can arise when website cloning is being used with malicious intent.

Individuals who wish to cause harm with website cloning can do so by many means. A malicious user may use a cloned website for spreading misinformation, credential harvesting or malware distribution to name some of the possible uses of a malicious cloned website. This can be a great cause of concern for companies who choose to host websites on the internet or individuals who are browsing the internet.

Taking the risks into account, creating a tool that can identify the difference between a malicious clone, and a legitimate website hosted by the original company would be a great asset for every internet would be a great benefit to ensure security for internet users.

## Table of Contents

Abstract.....	1
2. Legality of website cloning.....	3
3. How to clone a website.....	5
4. Identifying cloned websites .....	6
5. High Profile Cases of Cloned Websites .....	7
6. How users may visit cloned websites .....	8
Spear Phishing.....	9
Smishing.....	10
Pharming.....	11
DNS Cache Poisoning .....	12
7. What to do in the case of a cloned website .....	13
Contact website clone’s owner.....	13
Contact the domain registrar.....	14
Submitting a Report to Google .....	14
8. Existing Tools.....	15
Phish Tank.....	15
CheckPhish .....	17
PhishDetector .....	18
9. Coding Languages for Developing the Application .....	19
10. Natural Language Processing .....	20
spaCy.....	21
Natural Language Toolkit.....	21
11. Flask .....	21
12. Summary .....	22
13. References .....	22

## 1. Introduction

A website is a digital file that can contain several interconnected webpages that can be accessed under a single domain name. The website is then stored on a device that has full-time access to the internet so other devices with internet access can view the contents of the website. Websites are created using HTML. These websites can be used for both corporate and personal reasons. Individuals may want a website to show off personal collections or to discuss their hobbies with other people over the internet, and corporations may conduct their business online and engage in e-commerce over the internet allowing their customer to purchase goods over the internet. According to Siteefy [1] the number of websites both personal and corporate being stored on the internet is over 1.14 billion and this number is increasing daily. According to research conducted by Statistica [2] it was found that in the first quarter of 2021 alone over 611,000 of these websites were identified as phishing websites.

Website cloning is the act of copying code, whether that be sections of code or the entirety of code from one website onto another website. Website cloning can be a helpful tool when used with good intentions used, web developers may utilize website cloning in order to create multiple webpages at once with similar styles or content, but website cloning may also have malicious users such as creating a phishing website in order to trick users into believing they are visiting a website from an already established organization.

## 2. Legality of website cloning

The legislation that surrounds the act of website cloning is convoluted, as the act of website cloning is legal, but many websites are protected by Copyright Law.[3] Creating the cloned webpage would be legal but publishing anything covered with copyright protection without the permission of the original copyright holder would lead to legal ramifications. Any piece of original work that is published by an individual is protected by copyright law, but the argument of fair use is a difficult one to argue.

To circumvent the copyright laws that would protect webpages from being protected by copyright laws, individuals who are cloning webpages would have to argue fair use. Fair use allows an individual to use copyrighted materials if it is being used for criticism, teaching, scholarship, news reporting or research. Fair use does not require an individual to notify the original copyright holder of their usage of the original copyrighted material. Fair use considers four separate cases to judge if a copyright infringement case is protected by fair use. The four cases considered are, Is the purpose and character of the use, including whether it is commercial or for non-profit education purposes? What is the nature of copyrighted work? What is the amount and substantiality of the portion used in relation to the copyrighted work? Is the effect of the use the potential market for or value of the copyrighted work?

The first case considered, what is the purpose and character of the use, including whether it is commercial or for non-profit education purposes? This case would specify whether the case of copyright infringement is for a commercial purpose or if this case is for personal use. Fair use cases where the reason of copyright infringement is personal use, the ruling is more favorable in terms of fair use, which in terms of malicious users they would intend to make financial gain and therefore they would be seen as commercial use.

The second case considered is copyrighted work. This case takes into consideration if the copyrighted work is being used as factual in reason, or if there is creative reasoning for the copyright infringement. Factual reasons are more restrictive in content created on specific topic.

The third case that is brought into consideration is the amount of substantiality of the portion used in relation to the copyrighted work, this would take how much of the original copywritten work has been copied into the cloned webpage or has there been enough modifications made to the copywritten website that the new cloned webpage can be considered a new piece of work. This would ensure that cloned webpages that are published are not complete copies of the original website and that alterations are made so users can distinguish between the original and cloned website.

The fourth and final case that is considered to distinguish between normal copyright infringement and fair use is the effect of the use upon the potential market for or value of the copyrighted work. This would be important to be aware of when it comes to website cloning as

these websites tend to have a negative impact on the origin website as the website would drive traffic away from the original site or leave users weary of the security implications that are involved with the original website upon future usage.

### 3. How to clone a website

There are many different methods that can be used to successfully clone a website that is already made. A person may choose to manually copy the code using their browser's own development tools that are prebuilt into the browser itself and require no additional tools or services to be used or a user may choose to use online website cloning services and tools to successfully clone a webpage without much effort on the cloner's behalf.

When using the browser's development tools to clone a webpage a user must take care to ensure that all required files are being cloned. A single website may take advantage of multiple CSS, and JavaScript files to successfully display their webpage to their online users.

There are many online tools readily available for cloners to take advantage of to clone a website with ease, some of these online services are sitepuller<sup>1</sup>, saveweb2zip<sup>2</sup>, HTTrack<sup>3</sup> and many more.

The Social-Engineer Toolkit[4] is a toolkit that can be found on the Kali operating system, which can be used in order to clone a website. One component on this toolkit is a built-in site cloner. In this site cloner an attacker must only provide the URL that they are in control of and the URL of the website that they wish to clone.

---

<sup>1</sup> Sharp, M. (n.d.) *Website Downloader | Website Copier Online | Website Cloner*. [online] Website Downloader | Download any website. Available at: <https://sitepuller.com/> [Accessed 17 Apr. 2023]

<sup>2</sup> Saveweb2zip.com. (n.d.). *SaveWeb2ZIP – Website Copier Online Tool*. [online] Available at: <https://saveweb2zip.com/en> [Accessed 17 Apr. 2023]

<sup>3</sup> Httrack.com. (2017). *HTTrack Website Copier – Free Software Offline Browser (GNU GPL)*. [online] Available at: <https://www.httrack.com/>.

## 4. Identifying cloned websites

Cloned webpages will try to remain as close to the website they are cloning as possible whilst still making changes where necessary so the malicious user can achieve their goals, but there are still weaknesses in these cloned websites that the diligent user can use to identify the discrepancies between a fraudulent cloned website and legitimate website.

The first step to identify if the user has connected to a malicious clone of their intended website is to check the URL the user has connected to against a database of blacklisted sites. Websites such as OpenPhish<sup>4</sup> contain large databases that contain lists of verified fraudulent sites that users should be wary of when traveling to unknown sites.

The second step to identify if a user has reached their intended website or a malicious clone of the website, they had intended to visit is to check the URL of the website they have accessed, if the URL of the website they are currently on is different to the intended destination the probability of the user being on the wrong website is high. Only one website may have access to a domain URL so attackers must get creative to discover ways about finding a URL that resembles the original websites' domain name but altering ever so slightly as to not raise suspicion from potential victims. One of the ways that attackers can trick users by adding special characters to their URLs so that users may think they have visited the correct URL addresses, so users must ensure that none of the characters are replaced with alternative characters such as 'é' where the letter 'e' should be. The attacker may also change the domain suffix at the end of the URL to deceive victims, they may use '.co.uk' when the original website is '.com.'

The third step for a user to identify whether the site they have accessed is a cloned copy of a website or the original is to read the content that is available on the website. The three countries with the highest level of cybercrime are the China, Russia and India respectively[5], this demonstrates to us that many languages are involved in creation and distribution of these cloned webpages may use a large array of languages and the grammatical accuracy of each webpage and check if the sentences are formed in a manner that makes sense to the reader as the language used may be off as the language the website has been written in may not be the

---

<sup>4</sup> Openphish.com. (n.d.). *OpenPhish – Phishing Intelligence*. [online] Available at: <https://openphish.com/>.

attackers first language and this can be a clear identifier to users that the site they are on might not be entirely legitimate.

The fourth step involved in identifying a legitimate website is checking if the browser displays a padlock in the search bar of the browser beside the URL which Alert Logic[6] claims is an important step in spotting fraudulent clones. The padlock indicates if the browser is transferring information using HTTPS (Hypertext Transfer Protocol Secure) which is an extension of HTTP (Hypertext Transfer Protocol). HTTPS allows for an encrypted transfer protocol that prevents the risk of information being intercepted, read, and altered in transport between the user and the server. Legitimate websites should have this enabled when transferring personal information between the user and the server to ensure security for the end user. When information is being transmitted over HTTP instead, reading a user's personal information is much easier for the attacker. For this reason, users should check that the small padlock icon is being displayed to them when accessing a website as it is another indicator of website legitimacy.

The fifth step involved with verifying legitimacy of a webpage is to perform a Whois lookup on the domain you are visiting. Whois allows a user to search a large record listing to gain more information on the background of the domain and gain information such as contact information of those who are affiliated with domain, the registrant of the domain, the date of registration, contact information for the registrar, name of the servers the domain is hosted on, the date of most recent update and the expiration date. This would provide the user with valuable information in the detection of illegitimate webpages as this information would not match that of a legitimate website.

All the steps required can be run in tandem of one another to detect if a webpage is a clone of another or if they are their own unique webpage, as the answer to one of these steps may not be a complete indicator as to the legitimacy of a webpage, as the authors of a webpage may have had a poor day and misspelled many words, or there may have been a recent update to the website and the domain suffix has been changed. Running all these steps is essential in distinguishing between secure original domains and insecure cloned webpages.

## 5. High Profile Cases of Cloned Websites



In January of 2020, it was discovered that Algebris Investments were a victim of a cloned website attack. A group of cybercriminals cloned the investment firm's website intending to trick unsuspecting victims that they were affiliated with Algebris investments in order to defraud them. The cybercriminals were able to successfully create a convincing clone of the original website as they managed to acquire the domain [www.algebris-investment.com](http://www.algebris-investment.com) which could be easily mistaken as the original domain of Algebris, as the domain that Algebris have ownership of is [www.algebris.com](http://www.algebris.com) [7].

A Portuguese company called Dinheiro Vivo found itself victim to website cloning when an individual replicated the website in its entirety apart from the company's logo. The attacker hosted the cloned website on the domain [vivodiheiro.com](http://vivodiheiro.com) whilst slightly different than the domain [dinheirovivo.pt](http://dinheirovivo.pt), which is website where the original site is hosted but close enough that some victims who are less vigilant may be fooled into thinking that the cloned website was in fact the original [8].

Cryptocurrency websites have been found victim of website cloning, with a financial service provider located in Switzerland having to make an announcement in May of 2022, warning their users to be wary of the cloned website [www.dukacoin.hold-coins.com](http://www.dukacoin.hold-coins.com), as it was in no way affiliated with the original company. Any information that was present on this website was not from the official website and may hold misinformation and this website could be used by the attackers as a launching ground for further cyber-attacks [9].

## 6. How users may visit cloned websites

There are many ways that an attacker may trick the user into visiting their own malicious cloned websites instead of the original website, the methods involved in tricking users can range in sophistication, from simple phishing emails to specific file manipulation in order to achieve their goals of getting users to visit their website over the original website, some methods used by attackers are:

A simplistic way that attackers can trick users into visiting a malicious clone over their intended website is including a link to the malicious website in a phishing email that would be sent directly to the victim. In IBM's X-Force Threat Intelligence Index of 2022 [10] it was reported

that 41% of cyber-attacks use phishing to gain initial access. There are various forms of phishing attacks an attacker could utilize in order to get the victim to click the link to the attacker's cloned website.

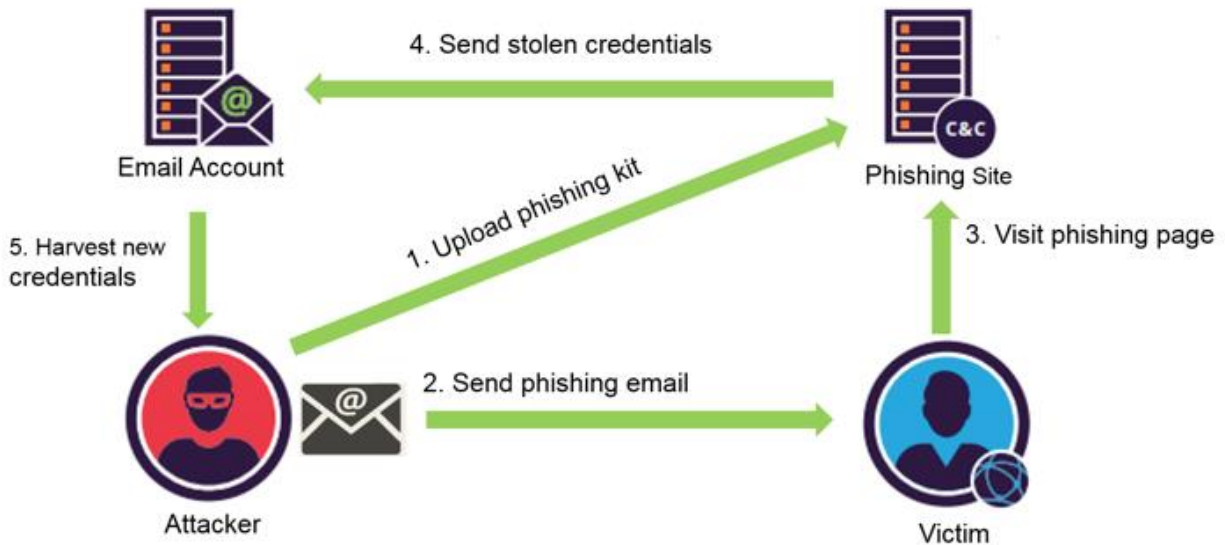


Figure 6.1: Visualization of how a user may be directed to a cloned website through a phishing email [11]

### Spear Phishing

If the attack is targeting a specific individual or specific group of individuals the attacker could use a spear phishing attack, this is a form of phishing attack where an attacker gathers information on a target and uses this information within their phishing email in order to trick the victim into believing the email is legitimate as the sender had prior knowledge on the victim. This form of phishing attack has seen growth in its popularity especially since information gathering on individuals has become increasingly easy especially with the prevalence of social media as an information gathering tool. According to Internet Security Threat Report produced by Symantec [12] 65% of hacker groups used spear phishing as a

primary attack vector.

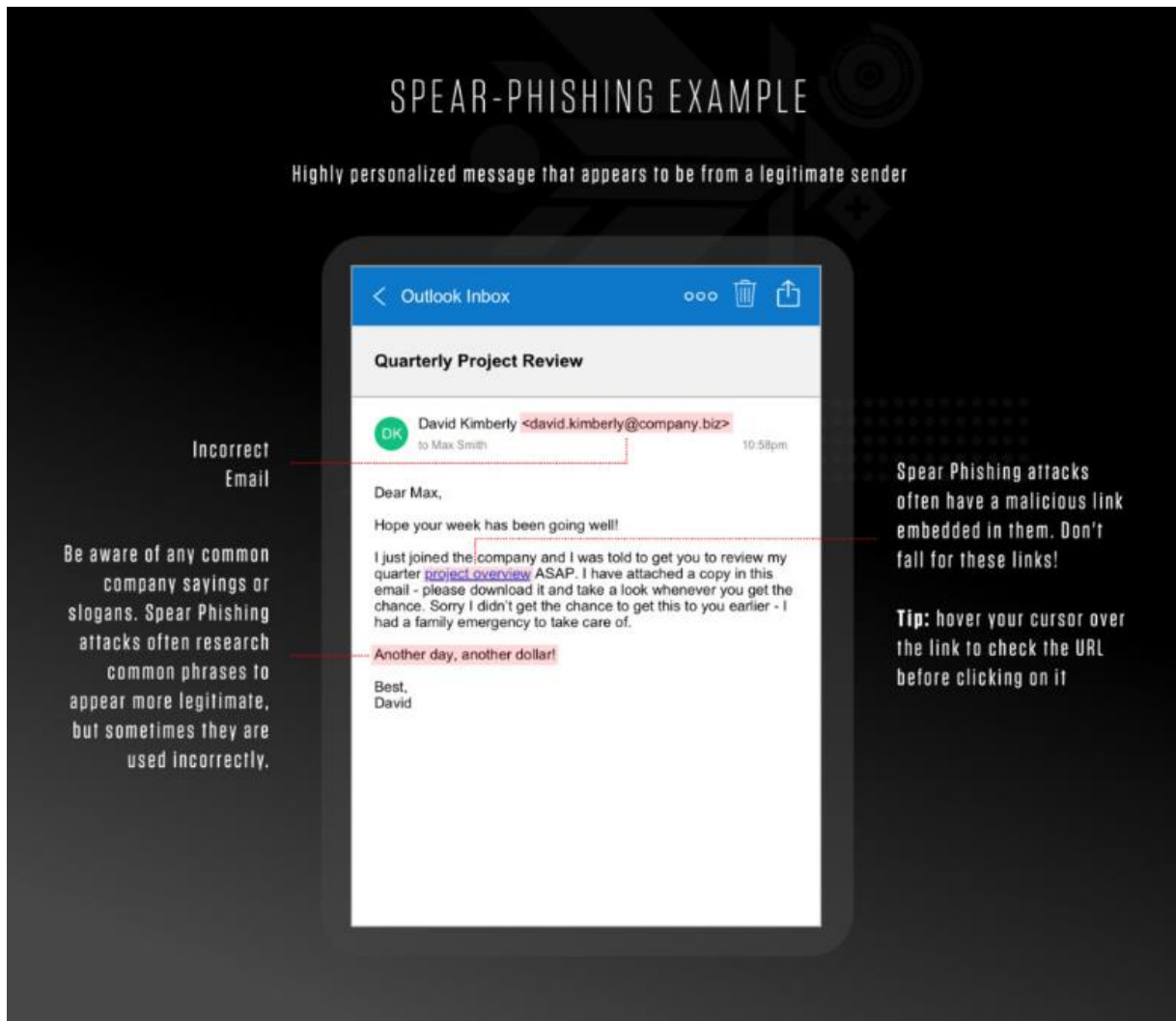


Figure 6.2: Example spear phishing email [13]

## Smishing

Smishing is a form of phishing attack where the attacker contacts the victim through SMS messaging. The attacker will try to impersonate a person or organization with the main goal of the attacker being to trick the victim into clicking on a link to a malicious clone that the attacker has full control of. Smishing attacks can be a particularly powerful form of phishing attack as the attacker can spoof the identity of others, forcing the message to appear as though it has come from a contact the user may already have saved within their phones. Smishing attacks have seen a dramatic rise in recent years with it being reported by Proofpoint [14] that there has been a 328% increase in a single quarter of 2022.

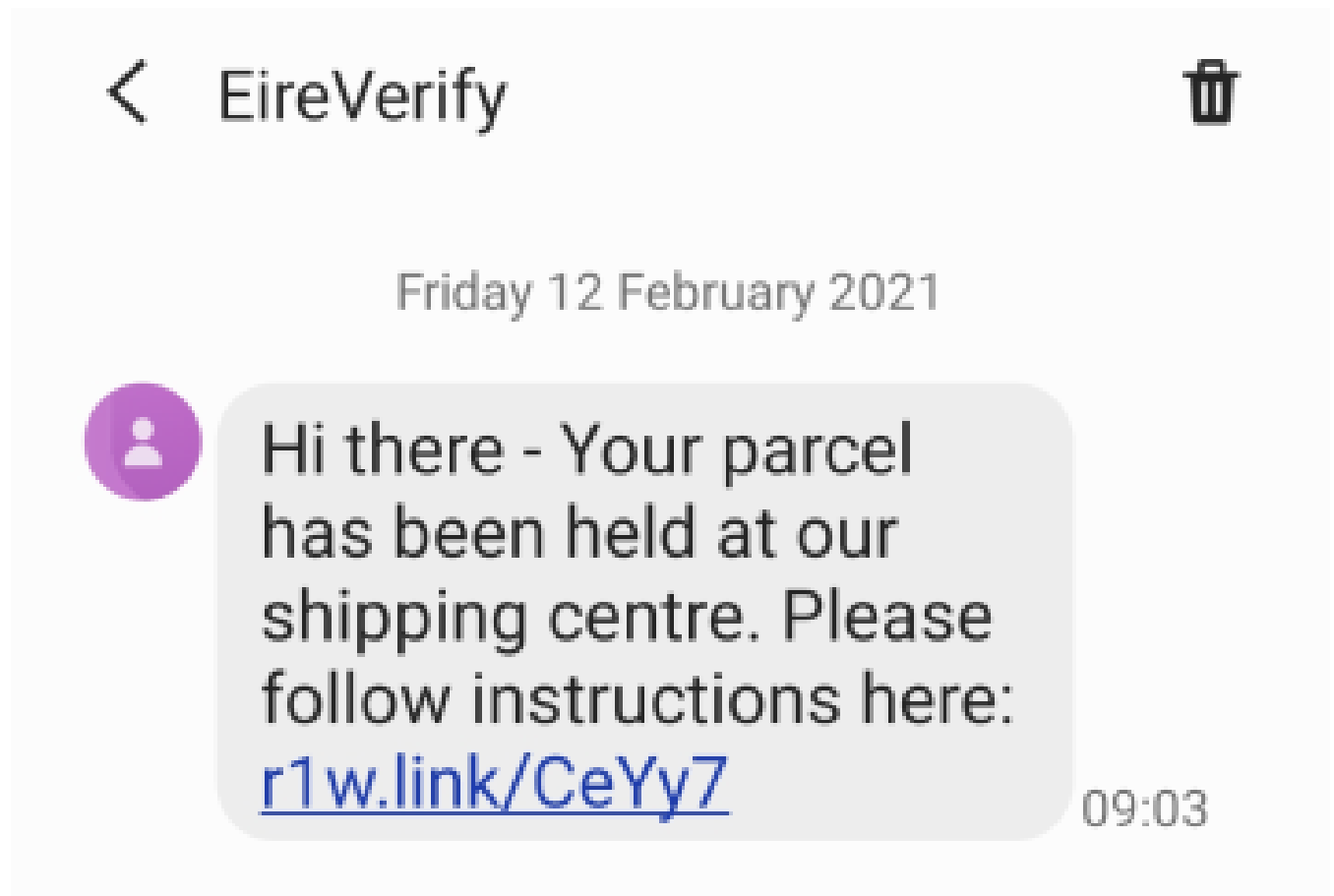


Figure 6.3: An example smishing messages from a spoofed EireVerify number [15]

## Pharming

Pharming is a highly sophisticated form of phishing attack where an attacker can redirect a victim's internet traffic in order to lead the victim different domain than the domain they initially intended on visiting; this is where the victim could be directed to a malicious clone instead of the legitimate website. This type of attack can occur from malware that has altered the host files on a victim's machine which will alter domain name and lead the victim to other websites. These websites would usually be malicious clones of the website the victim intended on traveling to.

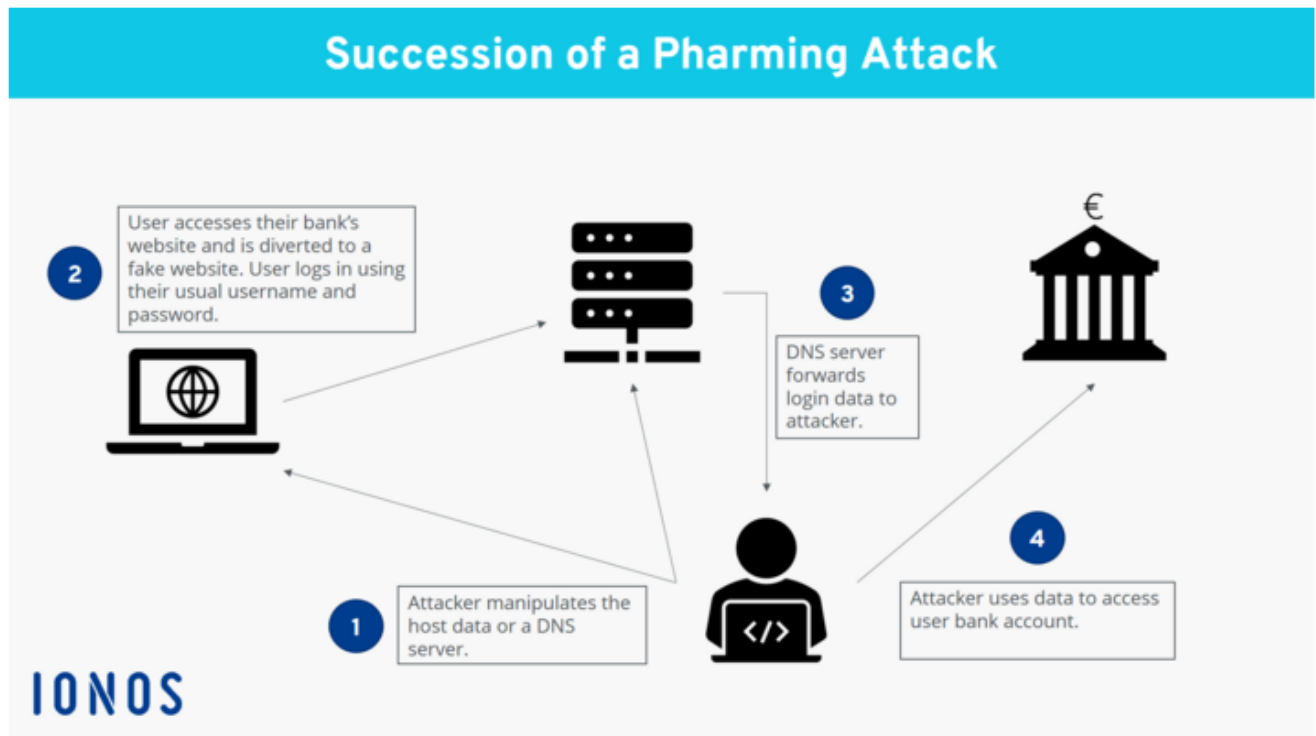


Figure 6.4: A visualization of a successful pharming attack [16]

### DNS Cache Poisoning

DNS cache poisoning is a form of cyber vulnerability where an attacker can override a DNS cache and force it to contain unintended DNS information, such as directing traffic to malicious cloned webpages controlled by the attacker. The victim of this form of attack may be completely oblivious to the fact they have visited the attacker's domain over the website they originally intended to visit. This attack occurs when an attacker overwhelms the DNS server's

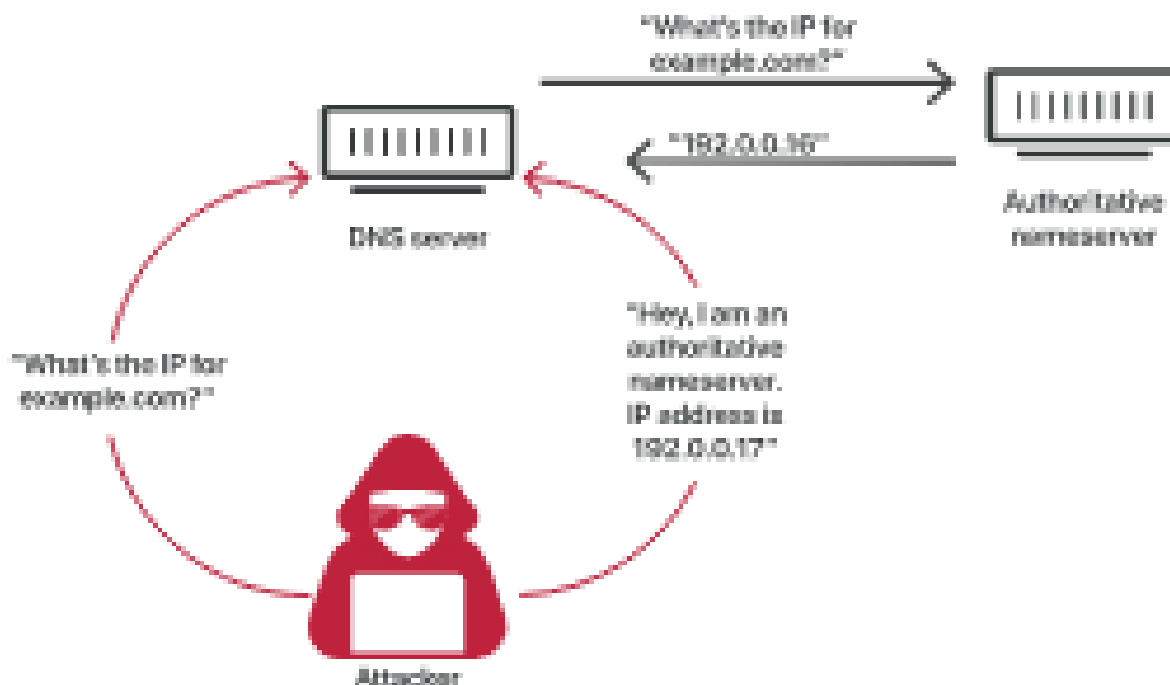



Figure 6.5: Visualization of a successful cache poisoning attack [17]

## 7. What to do in the case of a cloned website

In the case that a cloned website has been accurately identified there are many steps that the website owner can take in order to get the cloned website removed and individuals with malicious intent from benefiting from their cloned website.

### Contact website clone's owner

The owner of the original content can themselves reach out to the individual or corporation that is operating the malicious clone. You can find information on the owner of a domain through a Whois lookup, searching the website for contact information present on the website or reaching out to a domain broker are some of the many ways to gain contact information on individuals running the domain. The person who has had their content cloned can provide the cloner with a cease-and-desist notification as due to copyright law once a website is published the creator owns copyright of the created work.



Enter Domain

DOMAINS WEBSITE CLOUD HOSTING SERVERS EMAIL SECURITY WHOIS

google.com Updated 5 hours ago

Domain Information	
Domain:	google.com
Registrar:	MarkMonitor Inc.
Registered On:	1997-09-15
Expires On:	2028-09-13
Updated On:	2019-09-09
Status:	clientDeleteProhibited clientTransferProhibited clientUpdateProhibited serverDeleteProhibited serverTransferProhibited serverUpdateProhibited
Name Servers:	ns1.google.com ns2.google.com ns3.google.com ns4.google.com

Registrant Contact	
Organization:	Google LLC
State:	CA
Country:	US
Email:	Select Request Email Form at <a href="https://domains.markmonitor.com/whois/google.com">https://domains.markmonitor.com/whois/google.com</a>

Figure 7.1: Example Whois result for google.com [18]

### Contact the domain registrar

If providing the individual who has created the clone a cease-and-desist has proved unsuccessful in removing the malicious clone, the owner of the original content can contact the domain registrar and provide them with the cease-and-desist and request for them to perform a domain takedown. This will cause the domain registrar to take down the website regardless of input from the malicious cloners.

### Submitting a Report to Google

The owner of the original content can also make a report to Google at [https://safebrowsing.google.com/safebrowsing/report\\_phish/?rd=1&hl=en](https://safebrowsing.google.com/safebrowsing/report_phish/?rd=1&hl=en). Here the owner can report the specific cloned website and Google will launch their own investigation into the matter and if they have found the cloned website to be a malicious clone, they will remove the website from the results of Google searches to prevent their users from accessing the cloned website when conducting Google searches.

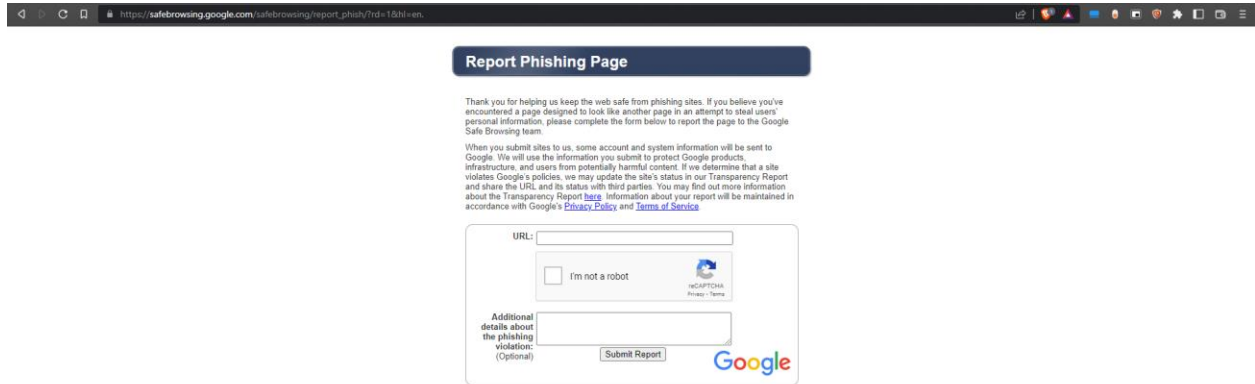


Figure 7.2: Google's report phishing page [19]

## 8. Existing Tools

There are multiple existing tools in this field that cover many areas of website clone detection to detection of phishing websites. The tools use multiple techniques in order to distinguish the difference between a fraudulent clone or a legitimate website. The tools can range from a website that will conduct automated scans against an array of datasets in order to provide its verdict, to a website that allows users to enter a website and many community members provide their own individuals verdict to the legitimacy of the submitted website.

Phish Tank



<https://phishtank.org> is an existing website that aims to identify phishing websites but does so with the aid of their community. In Phishtank users may submit a website they believe to be a phishing website and the members of the community then gain the ability to vote on whether they believe the website to be legitimate or if they believe the website is an illegitimate phishing website. If the members vote the website as a phishing website, it will get added to a large database of other phishing websites. Users can search up links that they may believe to be a potential phishing website and can view if there is an existence verdict on the legitimacy of the website, or they can submit the webpage to be reviewed by the community. As of the 19th of November 2022, Phish Tank [20] states that there have been over 7.7 million submissions made to the website with over 87,000 websites being identified as phishing websites.

ID	Phish URL	Submitted	Valid?	Online?
2953698	https://6817643d-c0d-420a-86e8-cc323f69237.id.repl.co/...	by sociatam	VALID PHISH	ONLINE
2953697	http://cliffordchance-attorney.com	by j3eeres	VALID PHISH	ONLINE
2953694	http://cnfmbon51acorevyspt283632443zscriptblmfmskrs.co.vu/verif.p...	by ledfelix	VALID PHISH	ONLINE
2953692	http://9et-5t4nted-565865.cf/confirmid.php	by ledfelix	VALID PHISH	ONLINE
2953691	http://1109-56563595356356.ml/confirmid.php	by ledfelix	VALID PHISH	ONLINE
2953690	http://9et-5t4nted-565865.ga/confirmid.php	by ledfelix	VALID PHISH	ONLINE
2953689	http://9et-5t4nted-565865.kl/confirmid.php	by ledfelix	VALID PHISH	ONLINE
2953688	https://tngparadise.weebly.com/	by bobbyb	VALID PHISH	ONLINE
2953686	http://1109-56563595356356.cf/confirmid.php	by ledfelix	VALID PHISH	ONLINE
2953685	http://9et-5t4nted-565865.ml/confirmid.php	by ledfelix	VALID PHISH	ONLINE
2953681	https://suasasetngaretyddccom.co.vu/dwefdf%3E%3Egdsesedf/...	by ledfelix	VALID PHISH	ONLINE
2953677	https://nataliaiz.ru/derive?est	by KellanMaxwell	VALID PHISH	ONLINE
2953676	https://ad-109795.weeblysite.com/	by eradiyabuse	VALID PHISH	ONLINE
2953675	https://attloginpage-att.square.site/	by eradiyabuse	VALID PHISH	ONLINE
2953674	https://543.sites.google.com/view/lizachvateacct/home...	by eradiyabuse	VALID PHISH	ONLINE
2953673	https://mail-101445.square.site/	by eradiyabuse	VALID PHISH	ONLINE
2953672	https://japan-smbobank.juvcj.top/	by dms	VALID PHISH	ONLINE
2953668	https://attportalonlineff-104934.square.site/...	by eradiyabuse	VALID PHISH	ONLINE

Figure 8.1: Phishtank database of confirmed Phishing websites [21]

This type of solution allows for users to gain extra information on potentially fraudulent webpages but does not offer real time protection to individuals visiting the website. Due to the community component of this solution, gaining information on the website's legitimacy would take some time as if the verdict has not been voted on, the community would have to provide their verdicts on the website. As the verdict of the legitimacy of the website is decided by the community a particularly advanced phishing website may bypass detection if the number of individuals supplying verdicts fail to identify a website as being fraudulent and outnumber individuals who have correctly identified the website as being a phishing website.

## CheckPhish

<https://checkphish.ai> is URL and website scanner, that operates in real-time. The website manages to perform phishing website detection with the aid of deep learning, computer vision and natural language processing. Checkphish claims it conducts these scans by utilizing an automated headless scanner to capture a screenshot of the webpage, then uses this screenshot to gain information regarding the website and feeding all this information to its deep learning module where they can provide a verdict to the likelihood of a website being a fraudulent clone [22]. Checkphish displays its verdict from a scale of 1 to 4. A verdict of 1 means that the scanned website would most likely be a phishing website and an obvious replication of another website. A verdict of 4 would mean that the website is most likely a legitimate website associated with the organization that website is representing.

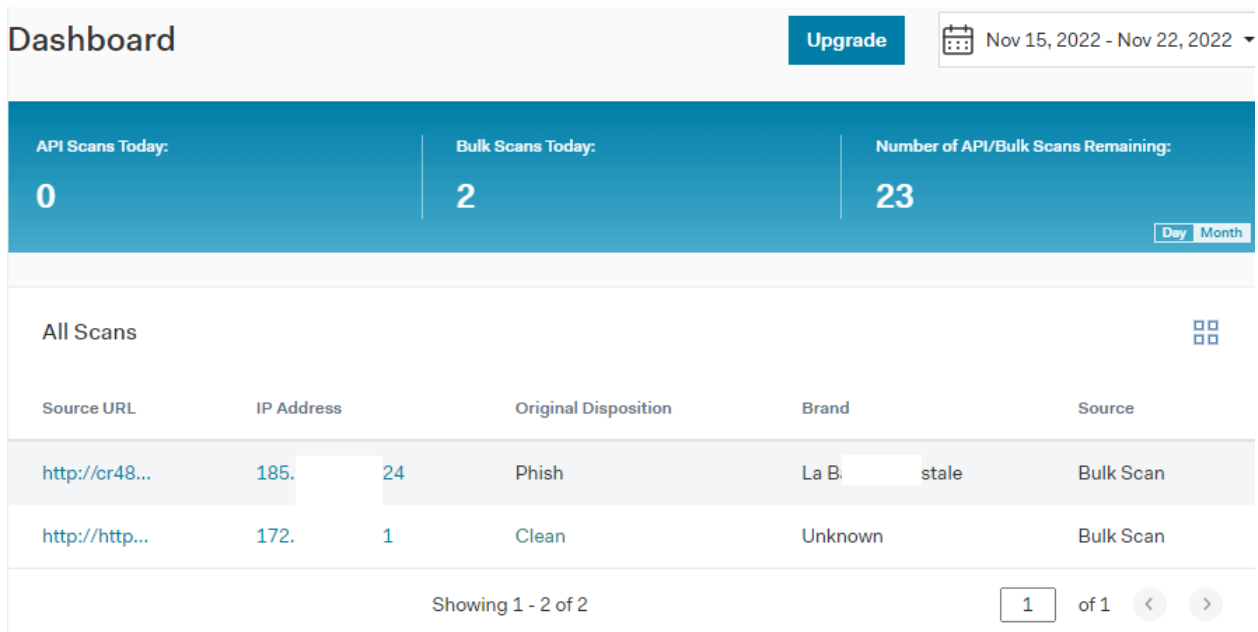


Figure 8.2: Example of a Clean and a Pish verdict from CheckPhish

Checkphish makes use of automated scanning with the aid of deep learning and natural language processing, allows Checkphish to be active around the clock and provide users with near instant feedback, but may not be entirely accurate as it lacks multiple inputs being taken into consideration when coming to a verdict. Providing the user with a verdict that ranges from 1 to 4 allows for indefinite answers since considering the range of datasets that are examined

for when testing whether a cloned website is legitimate or not can range a user can take an educated risk when visiting potentially cloned websites.

## PhishDetector

PhishDetector is a browser extension that provides the user with the ability to scan websites to gain information on the suspected legitimacy of the website. The extension automatically searches the website that the user has visited and provides the user with its result as to whether the website is in fact legitimate or a fraudulent clone.

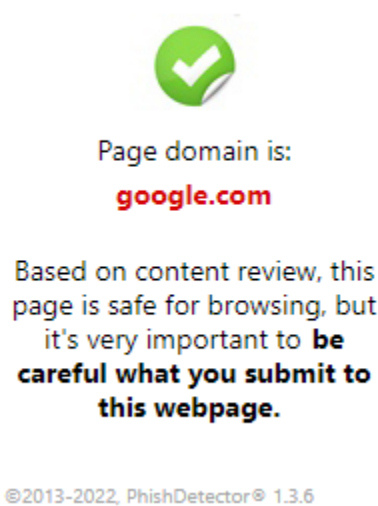


Figure 8.3: An example of a legitimate response from Phish Detector web extension

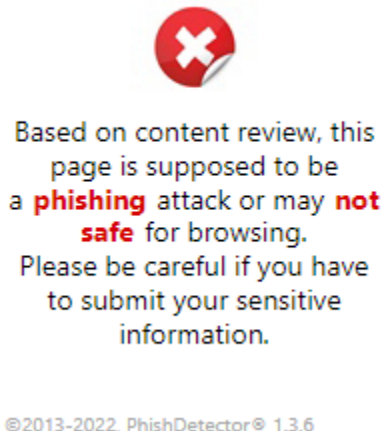


Figure 8.4: An example of phishing website response from Phish Detector web extension

Phishdetector can come to a conclusion on the legitimacy of a website by running a web API that implements a rule-based method. They create this rule-based method using the RESTful API. Phishdetector can analyze the DOM of websites in order to extract information where needed to form their verdict. The extension then shows the result within the extension tab as to whether or not the extensions believe the website to be a phishing site or a legitimate website [23]

## 9. Coding Languages for Developing the Application

As part of this project, I wish to develop an application that would allow me to automatically scan websites to check if a user is on a cloned webpage, or if the user is on a legitimate website, and as part of my research into browser extensions I had to research what language I would use to create this browser extension. In my research I discovered that Chrome extensions are made using a combination of HTML, JavaScript, and CSS.

As part of the browser extension, the browser extension would be required to read website information and make multiple searches in order to check the web contents against a criteria that would provide the user with feedback regarding the legitimacy of the website. This would be outside of the bounds of the capabilities of HTML, JavaScript, and CSS, so research was required in order to find a solution that would allow us to perform the functions required by the extension outside of HTML, CSS, and JavaScript.

Python is a high-level programming language that would aid in the development of a cloned website detection tool. The benefit of using python to develop this application is the natural language toolkit. The natural language toolkit is an open-source library that can be used in python. The natural language toolkit can be used in natural language processing, which would allow the program to be able to utilize the text found on websites to provide the user feedback regarding the website's legitimacy. BeautifulSoup is another python library that would be useful for developing this application as BeautifulSoup allows for the pulling of data from HTML and XML files, which is vital when using information that is present on websites.

C# is an alternative programming language that could be used to develop this cloned website detection application. C# is like python in the sense that it is a high-level programming language. There are C# libraries that would aid in the development of the cloned website detection application, Anglesharp is a library that would enable the parsing of html content within a C# application, since this application is investigating websites and their contents in order to provide a verdict on legitimacy this application would prove extremely useful. SharpNLP is a natural language processing library that is usable with the C# programming language.

Both programming languages both have separate use cases where an application developer may choose one language over the other, before developing this application it was important to identify which programming language was more suitable for this specific application. The python programming language utilizes automatic variable declarations, meaning that the type of variable does not necessarily need to be declared, whilst C# programming language requires manual variable declaration. Having automatic variable declaration in this application would be useful given the wide range of information and datasets that the application must process in order to provide a verdict on the website. Python is a dynamically interpreted language which allows greater flexibility when it comes to the many operating systems and allows easier integration with various browser types and websites during development.

## 10. Natural Language Processing

According to IBM [24], Natural language processing (NLP) refers to the branch of computer science – and more specifically, the branch of artificial intelligence or AI – concerned with giving

computers the ability to understand text and spoken words in much the same way human beings can. For this project being able to understand the information present on a website and having the capability to process the text within the application would be a large benefit for this project.

There are two main processes that would use natural language processing in this application those would be to perform a spell check to see if there are many grammatical errors present within the contents of a webpage and to extract organizations stated on the website and check if there are many other websites with the same organizations mentioned. In order to perform these tests, it is important to first check which natural language processing library to use.

## spaCy

spaCy is a widely used natural language processing library that was developed with use in production being a focus behind spaCy with its fast operation times and the high level of customization [25] that can be performed with the spaCy library; however, this high level of customization can make spaCy rather daunting for developers new to natural language processing and is better tailored to those more familiar to natural language processing.

## Natural Language Toolkit

Natural Language Toolkit is another natural language processing library that offers a wide range of various natural language processing functions. Natural Language Toolkit also offers a lot of educational resources regarding natural language processing with python targeted at a large range of individuals depending on familiarity. One great resource that the team behind the natural language toolkit have made public is a book that the team states “provides a practical introduction to programming for language processing” [26]. The mixture of higher quality documentation and educational resources alongside the wide range of uses for natural language toolkit is the reason I chose this natural language processing library over alternatives.

## 11. Flask

The Flask framework enables the creation of web applications using the python programming language. This enables the development of more sophisticated web applications than just using the already existing web application development languages. In order learn flask I discovered a resource by Miguel Grinberg [26], which has informative labs on multiple sections of Flask development. This gave me the understanding of how Flask can be used in order to create inter-process communication between the browser extension and the underlying python application.

## 12. Summary

In conclusion whilst website cloning can be a useful tool for website development it also leaves many security vulnerabilities present for those who intend to use the tools for malicious purposes. Creating an application in order to protect individuals when traversing the web is not only viable by using many test cases in order to come with a conclusion of website legitimacy but also a necessary tool of defense for all internet users. There are many ways attackers can cause users to be tricked into visiting a malicious site so a user would have to take necessary precautions in order to defend themselves from the many attack vectors that an attacker may take advantage of in order to trick a user into visiting a malicious clone. There are legalities defending website owners from having their original content, but as laws regarding digital content have failed to keep up with the constantly evolving digital word, receiving justice in a court of law can be burden for digital content creators and instead reaching out individual to companies is a much more effective method to have a fraudulent clone taken down.

## 13. References

[1] Huss, N. (2020). *How Many Websites Are There Around the World? [2020]*. [online] Siteefy. Available at: <https://siteefy.com/how-many-websites-are-there/>.

[2] Statista. (n.d.). *Number of global phishing sites 2020*. [online] Available at: <https://www.statista.com/statistics/266155/number-of-phishing-domain-names-worldwide/>.

[3] IPOI. (n.d.). *Copyright protection*. [online] Available at: <https://www.ipoi.gov.ie/en/types-of-ip/copyright1/understanding-copyright/copyright-protection/>.

[4] GitHub. (2020). *trustedsec/social-engineer-toolkit*. [online] Available at: <https://github.com/trustedsec/social-engineer-toolkit>.

[5] ACMC, E.S.C. (2022). *Countries With The Highest Rate Of Cybercrime 2023: Top 10*. [online] Bschorly. Available at: <https://bscholarly.com/countries-with-the-highest-rate-of-cybercrime/> [Accessed 17 Apr. 2023].

[6] *Spotting Cloned Websites* (n.d.). [online] Available at: <https://support.alertlogic.com/hc/en-us/articles/360057785872-Spotting-Cloned-Websites> [Accessed 17 Apr. 2023].

[6] FCA. (2020). *Algebris Investment (clone of FCA authorised firm)*. [online] Available at: <https://www.fca.org.uk/news/warnings/algebris-investment-clone-authorised-firm> [Accessed 24 Nov. 2022].

[7] [www.dukascoin.com](http://www.dukascoin.com). (n.d.). *Dukascoin. Swiss Bank Cryptocurrency*. [online] Available at: <https://www.dukascoin.com/?cat=news#ns0> [Accessed 24 Nov. 2022].

[8] Pedro (2019). *Clone websites: a story about fraudulent profiteering • pedrosaurus*. [online] pedrosaurus. Available at: <https://pedrosaurus.com/online-fraud/clone-websites-profiteering/> [Accessed 24 Nov. 2022].

[9] X-Force Threat Intelligence Index 2022 2. (n.d.). [online] Available at: <https://www.ibm.com/downloads/cas/ADLMYLAZ>.



[10] Dikbiyik, F. (2019). *What can be learnt from a phishing domain*. [online] Medium. Available at: <https://medium.com/@fdikbiyik/what-can-be-learnt-from-a-phishing-domain-44397c26a7d0> [Accessed 22 Nov. 2022].

[11] ISTR Internet Security Threat Report Volume 24 | . (2019). [online] Available at: <https://docs.broadcom.com/doc/istr-24-2019-en>.

[12] crowdstrike.com. (n.d.). *What is Spear Phishing? Definition with Examples | CrowdStrike*. [online] Available at: <https://www.crowdstrike.com/cybersecurity-101/phishing/spear-phishing/>.

[13] Proofpoint. (2020). *Security Brief: Mobile Phishing Increases More Than 300% as 2020 Chaos Continues | Proofpoint US*. [online] Available at: <https://www.proofpoint.com/us/blog/threat-protection/mobile-phishing-increases-more-300-2020-chaos-continues> [Accessed 22 Nov. 2022].

[14] <https://www.bankofireland.com/>. (n.d.). *Gallery of phishing and smishing examples - Bank of Ireland Group Website*. [online] Available at: <https://www.bankofireland.com/security-zone/gallery-of-phishing-and-smishing-examples/>.

[15] IONOS Digitalguide. (n.d.). *Pharming: Protecting against Redirections to Fraudulent Websites*. [online] Available at: <https://www.ionos.com/digitalguide/e-mail/e-mail-security/what-is-pharming/>.

[16] *What is DNS cache poisoning? | DNS spoofing | cloudflare* (no date). Available at: <https://www.cloudflare.com/learning/dns/dns-cache-poisoning/> (Accessed: November 22, 2022).

[17] Whois.com. (2018). *Whois google.com*. [online] Available at: <https://www.whois.com/whois/google.com>.

[18] [safebrowsing.google.com](https://safebrowsing.google.com/safebrowsing/report_phish/?rd=1&hl=en). (n.d.). *Report a Phishing Page*. [online] Available at: [https://safebrowsing.google.com/safebrowsing/report\\_phish/?rd=1&hl=en](https://safebrowsing.google.com/safebrowsing/report_phish/?rd=1&hl=en). [Accessed 22 Nov. 2022].

[19] [phishtank.org](https://phishtank.org). (n.d.). *PhishTank > Statistics about phishing activity and PhishTank usage*. [online] Available at: <https://phishtank.org/stats.php> [Accessed 22 Nov. 2022].

[20] [phishtank.org](https://phishtank.org). (n.d.). *PhishTank > Phish Search*. [online] Available at: [https://phishtank.org/phish\\_search.php?valid=y&active=All&Search=Search](https://phishtank.org/phish_search.php?valid=y&active=All&Search=Search) [Accessed 22 Nov. 2022].

[21] *Checkphish ai* (no date). Available at: <https://checkphish.ai/fag/> (Accessed: November 23, 2022).

[22] Varjani, M.M., Ali Yazdian (n.d.). *PhishDetector | A true phishing detection system*. [online] PhishDetector Landing Page. Available at: <https://moghimi.net/works/PhishDetector> [Accessed 24 Nov. 2022].

[23] IBM (n.d.). *What is Natural Language Processing? | IBM*. [online] [www.ibm.com](http://www.ibm.com). Available at: <https://www.ibm.com/topics/natural-language-processing>.

[24] spaCy. (2015). *spaCy · Industrial-strength Natural Language Processing in Python*. [online] Available at: <https://spacy.io/>.

[25] NLTK (2009). *Natural Language Toolkit — NLTK 3.4.4 documentation*. [online] [Nltk.org](http://nltk.org). Available at: <https://www.nltk.org/>.

[26] Grinberg, M. (2017). *The Flask Mega-Tutorial Part I: Hello, World! - miguelgrinberg.com*. [online] [Miguelgrinberg.com](http://miguelgrinberg.com). Available at: <https://blog.miguelgrinberg.com/post/the-flask-mega-tutorial-part-i-hello-world>.

