Google dorks: Use cases and Adaption study

UNIVERSITY OF TURKU
Department of Future Technologies
Master of Science in Technology Thesis
Networked Systems Security
October 2020
Reza Abasi

Supervisors:
Dr. Ali Farooq
Dr. Antti Hakkala

UNIVERSITY OF TURKU

Department of Future Technologies

Reza Abasi: **Google dorks: Use cases and adaption study**

Master of Science in Technology Thesis, 93 pages.

Networked Systems Security

October 2020

The information age brought about radical changes in our lives. More and more assets are getting connected to the Internet. On the one hand, the connectivity to this ever-growing network of connected devices and assets (the Internet) precipitates more convenience and access to various resources. However, on the downside, the Internet could be the hotbed for malicious actors like hackers, attackers, and cybercriminals' communities. Continuous Penetration testing and monitoring of the sites, and forums providing illicit digital products and services is a must-do task nowadays. Advanced searching techniques could be employed for discovering such forums and sites. Google dorks that are utilizing Google's advanced searching techniques could be applied for such purpose. Google dorks could be used for other areas that we will explain during this thesis in more detail like information gathering, vulnerability detection, etc.

The purpose of this thesis is to propose advanced searching techniques that will help cybersecurity professionals in information gathering, reconnaissance, vulnerability detection as well as cyber criminal investigative tasks. Further, a usability study has been conducted to examine the acceptance of these techniques among a group of cybersecurity professionals. In this usability study, we will measure the significance of 5 variables in the innovation diffusion model (IDT) namely Complexity, Compatibility, Relative advantage, Trialability, and observability in the adoption of Google dorks for search-related tasks for cybersecurity professionals.

Keywords: Google dorks, Cybercriminal forums, Information gathering, Dark web, Defaced sites, Innovation diffusion theory

Table of Contents

List of tables

List of figures

1 Introduction

Thanks to the information age our life has been highly affected by the tremendous growth of access to 24/7 resources. By the time of writing this thesis, there are almost more than 4.5 billion users connected to the Internet throughout the world and the number is constantly increasing [1]. Using social media sites and applications like Facebook, Twitter, WhatsApp, and Telegram has become more and more common. Simultaneously the number of websites and other resources connected to the internet growing rapidly. However, locating the most useful resources for our day in day out search-related activities through this huge amount of resources not always easily possible. Search engines like Google, Yahoo, Bing, Yandex, and Duckduckgo try their best to index more and more of the resources connected to the Internet so the users' searches are more efficient. By far the most widely used search engine is Google [2]. The number of searches for Google search engines is above 5.8 billion daily by the time of writing this thesis [1]. To facilitate the search, Google provided syntaxes that will limit the search results, which will increase the effectiveness of the search [3].

It could be highly beneficial for actors in the cybersecurity community, both white hat, and black hat hackers, to adopt using such syntaxes and even combining them to produce customized advanced Google searching techniques, also known as Google dorks, that facilitate their search-related tasks. Google Hacking Database (GHDB) [4] that is part of the exploit-db.com bestows a good set of such Google dorks. Advanced google searching techniques or Google dorks can be beneficial for various activities for cybersecurity actors including information gathering, vulnerability detection, discovering files containing sensitive information, credentials finding through paste services and file sharing sites, finding criminal forums, locating defaced sites, and even searching in the dark web.

This thesis is organized as follows. Chapter two covers the background and already existent literature that examines Google dorks for various areas, as well as practical issues and areas we believe that applying Google dorks will be beneficial.

In chapter three we cover searching syntaxes of not only Google but other famous search engines like Yahoo, Bing, and Duckduckgo.

In chapter four we discuss how Google dorks can be applied for information gathering about the targets both by white hat hackers and black hat hackers. We briefly explain some of the categories of GHDB such as files containing interesting information, usernames or passwords, login portals, and online devices discovery. Additionally, we review Google dorks that are useful for finding credentials. Another area that Google dorks could be beneficial is finding vulnerabilities or potential points for starting attacks like SQL injection that is going to be covered in this chapter. Eventually, in the chapter‘s last section, we will briefly examine some of the well-known tools for automation of Google dorks like Bingoo, xgdork, Zeus, and pagodo. The pagodo automation tool is discussed in more detail.

In chapter five we shed light on some of the areas where Google dorks could be highly beneficial but are less studied. Detection of cybercriminal forums like carding sites and forums, defaced websites detection, the social media channels and groups cybercriminal and defacers actively participate in, Google dorks for locating cybercrime sites in the dark web, discovering the Clearnet site of a dark web site (if exist) and simple still useful dorks for deanonymization of dark web cybercriminal sites will be covered in this chapter.

Last but not least, in chapter six we analyze results from two surveys that have been designed based on the Innovation Diffusion Theory (IDT), together with a workshop about Google dorks based on quasi-experimental time-series design to measure how cybersecurity actors in our sample group adopt Google dorks for their search-related tasks.

Finally, the thesis ends with concluding remarks in chapter seven that includes a brief overview of the overall process of this thesis.

2 Problem statement and literature review

In this chapter, we will cover some practical issues from the author of this thesis‘s point of view applying Google dorks could be beneficial for solving these issues. Additionally the next part of this chapter includes the literature review of using Google dorks between the cyber security actors.

2.1 Practical Problems

In both hacking and penetration testing the reconnaissance and information gathering is a lengthy process. Google dorks make it easier time-wise. For instance, checking the companies' addresses, emails, the staff is not always easily found but with dorks more easily especially time-wise they can be accessed.

Another practical problem that Google dorks can be utilized is database verification by security experts, of the data breaches, which is possible by different methods. One typical approach is to discover registration links, login pages, or reset password links and checking whether the registration is possible with the credentials that exist in the breach. Finding login, resetting the password, registering pages of sites is not always straightforward. The category “pages containing login portal” helps and facilitates these tasks. This category already exists in exploit-db/GHDB and some of the existing dorks registered by the author of this thesis. Besides, from the hackers and penetration testers' side, the login pages can also be a pivot point for “Brute Force Attacks”. Similarly, login pages of online devices could be targeted for Brute force attacks.

The next practical area that Google dorks can be applied is to finding communication channels for cybercriminals, like the Telegram channels that are somewhat commonly used by cybercriminals, is not convenient. Telegram channels sharing hacked accounts, carding channels, hacking channels as well as Facebook groups and vk.com, and other social media applications are a practical problem for the organizations monitoring these channels and groups. Thanks to Google dorks these are more conveniently accomplishable.

Besides there are companies that inform their clients about their cyber exposure and they need to gather publicly available shared accounts from any possible resources, like paste sites, telegram channels, and so on. They also require to be informed about cybercriminal actors' malicious activities in terms of their business and monitor the cybercriminal site and forums. Finding such activities is a real challenge especially if not using Google dorks.

Moreover, from the cyber defensive perspective as well as the law enforcement side, being informed about the defacer groups, and the already defaced sites that potentially could be compromised is essential. Google dorks as we will examine in this thesis are quite useful. We will provide examples of sites that have been defaced without being noticed. Furthermore, with these dorks, it is possible to find all the defaced sites by a single defacer or a defacer group. So if the security researchers try to identify as much as possible information about the activities of hacking and defacing groups these dorks could facilitate their tasks.

Another practical challenge especially for law enforcement authorities is to find the dark websites and forums and deanonymize them, find if the same dark websites and forums are active, and exist on the surface web, and generally speaking gather as much as possible information about such sites and forums. Using Google dorks and web2tor sites [5] it is made possible to search for dark web content on the surface web.

Last but not least, passive reconnaissance has priority over active reconnaissance in this sense that there is no need to clear the track for the hackers and penetration testers so though there are multiple tools for doing these phases of penetration testing and hacking, the priority is to do it passively. Google dorks facilitate this phase as well as even some more steps like vulnerability detection, WAF detection, Web server detection.

Generally speaking, many of the above-mentioned tasks may be possible without Google dorks but the point is that especially in the industry that lots of tasks need to be done within a short time frame, disregarding Google dorks makes the tasks much lengthier as well as less precise and with more false-positive. Almost all of the topics that will be covered in this thesis are based on real tasks that needed to be done in a limited time-frame.

2.2 Literature Review

In this section, we will try to shed light on the existing literature regarding the areas Google dorks could be beneficial for search-related tasks in the cybersecurity community. However, at the same time, other areas exist that Google dorks can be applied and facilitate the tasks of the cybersecurity actors but less work has been done in those areas. Considering the existing literature, phases like vulnerability detection, files containing sensitive information, information gathering are scrutinized while other areas like discovering cybercriminal forums and their communication channels or searching for darknet sites with Google dorks are not covered in the existing literature but have been covered in the other sections of this thesis. The diagram is a schematic of these two areas.

Figure 1: Google dorks usage

(Blue color indicates areas covered in existing literature and orange color indicates areas not covered in literature but we covered in this thesis)

The existing literature covered the four below areas:

1. Definition of Google dorks from various resources
2. How the attackers used Google dorks
3. How the security researchers used Google dorks
4. Tools that utilized Google dorks as their core for activities

Though in some context there is a definition for Google dorks in some other context there is not a specific definition of Google dorks. For instance, Google dorking is a concept that lacks definition but simply they are queries that use advanced operators offered by Google search engine to retrieve sensitive information or vulnerable systems [5]. Additionally, others elucidated Google dorking is a hacking technique that uses Google's capabilities to locate specific files and vulnerabilities in web applications(using Google dorks for vulnerability detection) [6]. As an example of the usage of Google dorks for vulnerability detection, Google dorks were used to locating the monitor and control of the sluice gate [7]. There are studies that considered the Google dorks capabilities to locate the login portals and passwords, that is part of information gathering about targets, so according to them what is known as google dorking originates from the "google hacking" community and was used for locating login information/passwords or vulnerable systems [8]. Detection of vulnerable sites made pretty straightforward using Google dorks, so Google dorks could be scrutinized as one of the simplest methods to locate vulnerable sites, to put it simply it is a specific search request that uncovers websites that match the parameters in the request [9].

The so-called term "Google dorks" was coined by Johnny Long on his site johnny.ihackstuff.com first time and he explained the search techniques for finding information left on websites that revealing information about their assets that could be exploited by hackers [10].

Gathering the information about the target is one of the reasons attackers and penetration testers use Google dorks. Employing the indexing power of Google as a powerful search engine, Google dorks capable of pinpointing the sensitive information about targets while in some of the cases it is unknown by the site owners that those files containing the

sensitive information exist(another example of Google dorks for information gathering about targets and vulnerability detection) [11]. Additionally, locating the information from a site that is not possible or barely possible applying the typical searching by security practitioners is another usage of Google dorks [12]. Locating the error messages, documents and files, Network devices, open directories as well as numerous other interesting sources of information are made possible using Google dorks [13]. Malicious attackers, as well as cybersecurity actors like penetration testers, security researchers, and cyber investigators, utilize Google dorks also known as Google hacking techniques in some contexts, to acquire interesting information about their target. This worthy information in the majority of cases left unintentionally and mistakenly by organizations or companies on publicly available servers so they can be accessed by Google dorks that are advanced Google searching techniques.

In a study in 2014,1000 various Google dorks were used on numerous targets and found that around 300,000 sites that potentially were vulnerable[14]. This was part of the study work for proving that vulnerability detection of the websites is more conveniently possible utilizing well-crafted Google dorks. However, the questions that whether the attackers using Google dorks for spotting vulnerabilities or not or even which Google dorks checked by them and how they will apply Google dorks against their targets remained unanswered in that academic work [15].

Prominent hacking communities and groups benefit Google dorks for their objectives. For instance two famous hacking groups namely Lulzsec and Anonymous enjoyed the benefits of using Google dorks and used it as their main method of locating vulnerabilities of their targets [16]. Especially in the last decade, the attackers employed Google dorks to locate vulnerable computers all over the US [17].

Attackers benefit from Google dorks to locate potentially vulnerable sites to a web attack known as SQLi (Structured Query language Injection). Additionally the attackers able to search for terms like DOT SQL or DOTPWD and having the list of sites that contains the information about the site [18]. As an example of how the attackers can utilize the benefits of Google dorks in their objective, they can find the credentials of the websites that

mistakenly left in the site directories or files with a well-crafted Google dork. One of the benefits of such cases is that the victim is unaware that the attacker obtained the credentials illegitimately [19]. Attackers also employ Google dorks for their objectives especially to figure out where to launch their attack against their targets, though not always this could assist them in revealing their desired information. For instance, Google dorks were used in 2004 to an e-commerce package named Comersus that was ASP-based and had flaws in one of its files namely cemersus_message.asp. One of the categories that Google dorks can be beneficial is to discover the default pages, especially the login pages or pages containing configuration settings [20].

Cybersecurity researchers and practitioners also benefited from Google dorks to do their objectives. As an example, Google dorks was used for the detection of governmental infrastructure in Estonia [21]. Another group of security researchers is the malware researchers that Google dorks facilitate their activities as they can utilize Google dorks to detect vulnerable websites and malware. For instance, they can apply specific searches as well as Google dorks in google alerts and get notified and alert as soon as a matching result is found by google. It can be used by malware researchers for discovering the newly compromised or rogue sites [10].

Another area that Google dorks facilitate the activities of both security researchers and attackers is the tools that are based on Google dorks and Google dorks operate as their core. Some of the bots using Google dorks as their core for automation of exploitation [22]. Numerous Tools are utilized by Google dorks for discovering vulnerable sites to XSS or SQL injection or even performing XSS or SQL injection attacks [23]. Additionally, there is a countless number of tools that people who lacking enough knowledge of exploitation or cyber-attacks can employ them to spot the information that can pave the way for hacking their targets. These tools are not limited just to utilizing the Google dorks, but other search engines like Baidu, Bing, Shodan advanced search commands. SearchDiggity is an example of such tools that utilized Google dorks, as well as other search engines, advanced searching techniques [19]. In addition to the tools that utilizing Google dorks for locating sites vulnerable to SQL injection or XSS attacks, there are online resources like pentest-tools.com[24] that provide a specific URL to be tested

with nine various Google dork types and the outcome of google search results will be shown [25].

Some studies have been conducted on the observation of Google dorks' attacks. In a case of a web-server honeypot named Dorkpot that is dynamic, low-interaction was employed to detect requests related to Google dorks for analyzing such attacks [15].

When it comes to the resources for finding Google dorks, there are numerous sites on both surface web and darknet for obtaining such dorks. The publicly available, famous site exploit-db.com has a section that is a collection of various categories of Google dorks [26]. Also, it is common to find the Google dorks in paste sites or even the documents that are shared in sites like Scribd. Blogs, News websites, social media sites like twitter or telegram are other examples of sites Google dorks can be found. All the aforementioned resources are fine for finding Google dorks but the most prominent resource by far is the GHDB in exploit-db [27]. Till the time of writing this thesis more than 5500 registered Google dorks can be easily accessed using the GHDB [15].

3 Search engine hacking

The term search engine hacking is the process of using advanced operators and keywords of search engines like Google, Bing, duckduckgo, and even the popular search engine for internet-connected devices known as shodan [28]. We use every then and now use these operators and keywords for finding information, for detecting and even exploiting the vulnerabilities of the targets [29]. To reach this goal the attacker or penetration tester needs to be familiar with the basic search engine operators and keywords and their functionality. As the topic of this thesis is about google search hacking, the concentration will be on google as the search engine for hacking used by both White hat and Black hat hackers. Anonymous hacking groups like Lulzsec [30] consider google hacking as their leading means for diagnosing the vulnerabilities of their targets [16].

3.1 Description of the process

As we know every IT system has a logging system, so any access to a device connected to the internet is logged which makes the process of clearing track that is one of the threats for the attackers to be detected tough, so they prefer to do the whole steps of hacking in a passive way which means without direct communication with the target that logs the attackers' activities and sensitive information that in the end may lead to the disclosure of their identity. So google hacking is a great method as doing steps of hacking with google there will be no footprint of the hackers on the target. It means if the attack happens on a website and server logs, for instance, are investigated in case the webserver is apache the log file is access.log or error.log then the requests of the result of Google dorks coming from google and no trace of the attacker exist in the log file. Another positive point of using search engines, especially Google as the most widely used search engine [2] is that for the reconnaissance, information gathering phase of penetration testing or hacking, the more information about the target the attackers have the simpler the other steps and the more chance for the success of the hacking process.

3.2 Most commonly used search queries

Following the description of google hacking and the process of google hacking the importance of being familiar with the basics and more advanced searching techniques is undeniable. Almost all of the famous search engines ranging from google, yahoo, Bing, and even shodan that is the IOT devices search engine have their operators and keywords that simplifies the hackers' and attackers' activities for detecting their target's vulnerabilities more accurately. In what follows we can see how most of the search engines we studied here briefly have a more or less similar common manner for advanced search techniques and how to develop more advanced search dorks by mixing the basic search keywords and operators.

3.3 Basic search queries: intitle, intext, inurl, site, ext, intitle:" index of"

Typical Google search is the way that every person no matter what level of IT knowledge he/she has can accomplish. But when it comes to the more advanced google search for specific reasons for instance penetration testing, finding vulnerabilities, threat intelligence, finding cybercriminal sites and services, and so on that is some of the varied topics that will be covered in this thesis requires some advanced techniques. The key point that needs to be kept in mind is that Google indexes not only the titles and descriptions of the sites' pages but the entire content of pages [31] that allows searching based on the search_term the whole content of the page including text, title, and so on. For this goal, there is the concept of google advanced operators that help to refine searches to optimize the results. The general syntax of google advanced operators is:

Operator: keyword_for _the_topic

The Boolean operators AND, OR can be used in Google searches to refine the results. The default logical operator for multiple search terms in google is AND [31] which means google searches all of the search terms in case no modifier is used in the search. The OR operator clarifies for the google that you are googling for either of the words. For instance the search term below limits the results to either car or bike:

Car OR bike

Google borrowed the "|" operator for OR operator from the programming domain so for the above example the alternative would be car | bike. If you decide to exclude a term from your results the "-" is the operator that can be used. Another category of google advanced searching is google syntax words to narrow the user's searches [32].

intitle:search_term

Limits the search to the pages having the search_term in the page's title, allintitle is a variation of this syntax but better to avoid using as it's not well mixing other syntaxes.

inurl:search_term

Targets the URL of the pages that include the search_term for searching. Same as the above syntax it has a variation allinurl that does not mix well with other syntaxes.

intext:search_term

Used for searching the body of the pages that have the search_term and it has the allintext variation that does not mix well with other syntaxes.

inanchor: Searches for the description of the hyperlinks

Site: This syntax is used to restrict the search to a specific site or even domains or top-level domains (TLD) [33]. For instance, the site:yahoo.com limits the search just to site yahoo, or site:edu restricts the search just to the educational sites.

Cache: Used for finding the last copy of the page before the page being removed. It is mostly useful in case we face error 404 which means the requested page has been removed but with this syntax, the already removed page can be retrieved.

Both filetype and ext are more or less have the same outcome that is to search for the extension of the file. The extension can be all the files. For instance ext: doc searches among all the indexed doc files by google.

Related: This syntax limits the search for pages that are related to the page specified in the search request.

info: presents more information for the specified page in the search request, that includes a link to the URL's cache, pages that are linked or related to the requested page, info page if exist as well as pages containing the requested URL.

Last but not the least useful syntax for this thesis is intitle: "index of" "search term" that is quite useful for finding the "search term" for sites that directory browsing are active in those sites. This syntax is considerably useful for finding the sensitive files and directories of a website.

For more clarification of the above-mentioned operators and syntaxes here is an example. Suppose that we are going to find an article in pdf format from an academic site like university about malware and wannacry in the text of that file. The below dork gives us the desired result:

site:edu intitle:"malware" intext:"wannacry" ext:pdf

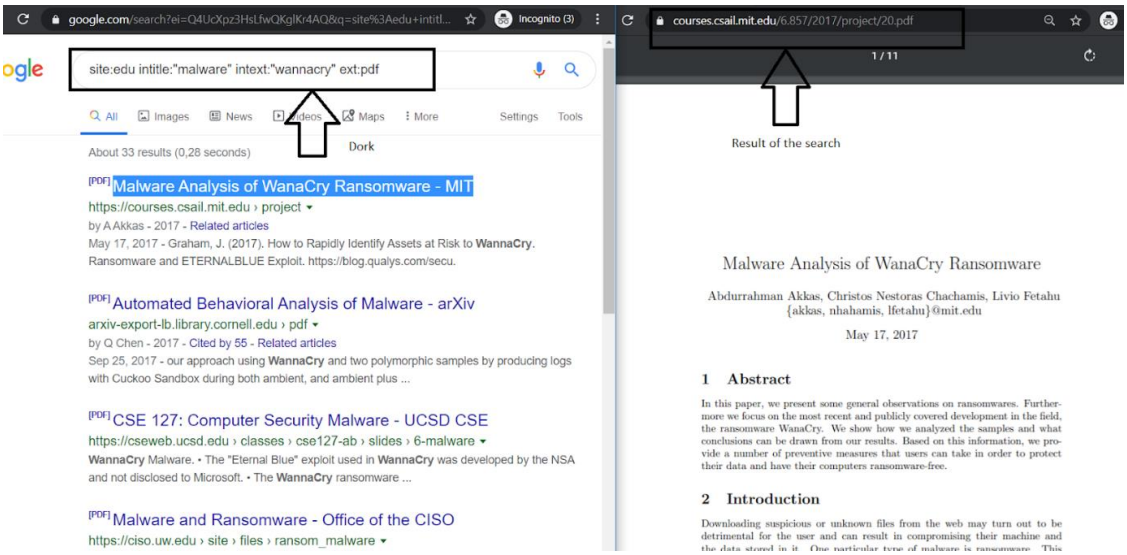

Figure 2: Google dorks example

The last point for this section is that by combining the syntaxes and operators in this part, we can create more effective dorks.

3.4 Creating more advanced queries

Armed with the basic syntax for google dorks now we are capable of writing more advanced and specific dorks. For instance, like a pen tester or attacker if we plan to target the governmental sites designed with php,asp,asp.net or python that are common server-side programming languages the queries below will be conductive, just replacing the ext:php with ext:desired_extension:

site:gov. ext:php*

This search will be advantageous in this sense that one of the tasks during the reconnaissance phase of penetration testing is to find the programming language technology behind the web application.

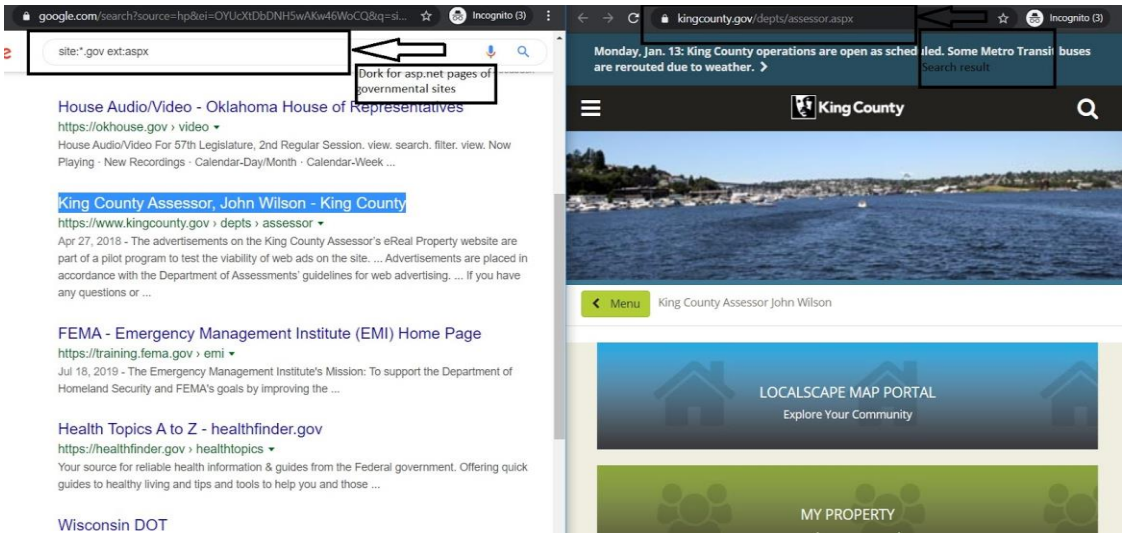

Figure3: Detecting site's web technology

3.4.1 Use Case: Finding mail subdomain of governmental sites with google dorks

From the attacker and penetration testers' perspective finding the subdomains of a website is an issue of high importance as it will lead them to more details about the structure of their target and in phases, after the reconnaissance, they can focus on a specific subdomain of the target site. Regarding this importance, Google dorks can be advantageous for finding the subdomains. Each of the below dorks is beneficial in terms

of finding the mail subdomain of the governmental sites and the last one targets the login page of the mail subdomain that can be targeted by the attacker and penetration tester will use for instance Brute Force attacks to find credentials.

site:smtp..gov.**

site:mail..gov.* intitle:"login"*

3.5 Duckduckgo and Bing search queries

Google.com is the most popular site in the ranking of the top 500 sites on the web [2]. However, other search engines like Yahoo.com, Bing.com, and Duckduckgo.com to name a few of them are commonly used for searching and have their syntaxes and operators that both Black hat and White hat communities can benefit them for their objectives. In this section, we described some syntaxes and operators of duckduckgo.com and Bing.com .However, the concentration in this thesis is on Google.com as a search engine.

Duckduckgo.com

This search engine claims to respect highly to the user's privacy in their personal information and gained more popularity nowadays though they have been found in 2008 and according to them their mission is to provide more trust for the users in terms of their privacy [35].

With two keywords "cats" and "Dogs" there is a valuable description of the basic operators and syntaxes used for searching by duckduckgo.com [36].

Table 1: Duckduckgo.com search syntax

Example	**Description**
Cats Dogs	The outcome will be either cats or dogs
“Cats and dogs”	Searches for exact terms that include both “cats and dogs” with the same order. If nothing found shows related results.
Cats –dogs	Fewer dogs in the outcome
Cats+dogs	More dogs in results
Cats filetype:pdf	PDF files containing Cats, the filetype accepts doc(x),html ,pdf ,ppt(x) file types
Cats site:example_site	Pages about cats from example_site
Dogs site:example_site	Pages about dogs except for example_site website
intitle: cats	Pages that their title includes the word “cats”
inurl:dogs	Pages URL that includes the word “dogs”

Bing.com is another search engine that is owned by Microsoft. It has operators and keywords for doing more efficient searches and spending less time through irrelevant results. The table below presents an overview of the keywords for using with Bing search engine [37].

Table 2: Bing.com search syntax

Keyword	**Description**	**Example**
contains:	Focuses on websites having links to the specified file types after contains	contains:wma Searches through websites containing links to wma files
ext:	Only webpages having filename extension specified here are returned	ext:docx ,filetype:docx Returns reports created in DOCX format
inanchor: inbody: intitle:	Webpages containing the specified term in metadata, like anchor, body, or title of the site, respectively	inanchor:msn inbody:spaces Retrieves web pages that contain "MSN" in the anchor, and spaces in the body
IP:	Returns sites hosted by a specified IP address. The IP address should be in the correct format.	IP:207.46.46.231
language:	Finds webpages for the language specified after :	language:fr Returns the results only in French
location: or loc:	Retrieves web pages from a specific region or country.	loc:fi Returns only the results from Finland
prefer:	Returns the results by adding emphasis to an operator or search term	Basketball prefer: organization Will return results about basketball that mainly belong to an organization
site:	Searches in a specified website, domain, or TLD	News site:yle.fi Returns web pages about news from the yle.fi web site

feed:	Retrieves Atom feeds or RSS feeds on a website related to the term searched for	feed: basketball To retrieve Atom or RSS feeds about basketball
hasfeed:	Retrieves web pages containing RSS or Atom feed on a website for the terms searched for	Site:washingtonpost.com hasfeed: basketball Finds webpages in washingtonpost.com containing Atom feeds or RSS feeds
url:	Used for checking if the web address or domain indexed by Bing	Url:nytimes.com Checks if Nytimes domain is in the index
filetype:	Finds web pages created in the specified file type	Subject filetype:docx Returns the reports with docx extension about the specified subject

There are other search engines but as the topic in this thesis is google hacking so we don't go through them.

3.6 Stop words for search engines

“Google stop words” is a topic that is worth considering when studying Google as a search engine though it mainly affects Search Engine Optimization (SEO).To put it more simply they are common words that are ignored by search engines to speed up the search. A comprehensive list can be found by simply searching the “google stop words” [38]. But to give you an overview of the words like:” a”,” about”,” above”,” able”,” her”,” hi” .are to name a few of them. Noticing the concept of google stop words is beneficial as it helps us to write more efficient google search terms and dorks.

The last word for this chapter is the definition of Google dorks, dorks, or google hacking. To put it simply what is known as Google dorks also known as dorks is to use advanced Google search operators and keywords to find information using google that is finding it without these keywords and operators is hard or almost impossible [39]. Considering this definition till now in sections 2.2.3 and 2.2.4 we already have written some simple Google dorks. As we will see in the coming chapters for the InfoSec world Google dorks is a valuable hacking tool [40]. Google as a search engine tries to index as much as it can from whatever tool, software, devices connected to the internet. Then if we always consider this fact that if we are aware of how to ask from google then we can find whatever google indexed ranging from sensitive files, directories, login portals, and so on.

4 Information gathering

Multiple phases for the modern attacks against IT systems exist. Information gathering is the phase for an attacker to gather adequate information about the vulnerability of the target [15]. Utilizing Google dorks for the information gathering is one of the best options as the attacker's identity will remain undercover and the target receives the requests from google rather than the attacker. Google dorks can be obtained through various resources on the Internet both from the clear web and deep web [15]. These resources include paste sites, hacking forums, or legit sites like exploit-db, social media, and many more resources.

4.1 Introduction to information gathering

Information gathering, better known as "Reconnaissance", in the IT world is the process of gathering information about the target [41]. This phase intends to gather as much as possible information about the target. Server names, IP addresses, sensitive directories, log files, contact information, addresses are to name a few useful information found in this phase and could be a pivot point for the rest of the activities of the attacker or penetration tester. Methods for gathering this information are either passive or active and it can be gathered using tools or online sources like social media, websites, or search engines. Passive information gathering is when there is no direct communication with the target so using search engines to obtain this information is considered as passive information gathering. On the other hand, active information gathering is the case when the attacker or penetration tester interacts with the organization's infrastructure. Almost in all cases, the first and easiest way to find information about the target is by googling about the target. It worth mentioning that we will opt to primarily use the Google dorks registered by the author of this thesis in the GHDB throughout this chapter and the rest of this thesis [4].

Having knowledge about the addresses, in charge people in a firm or organization, phone numbers, and any other typical information about an organization that exists normally in a website but can be exploited with the attackers for the next steps is highly beneficial for

the attackers. The more attackers invest in this stage the easier it is to exploit the target [42]. Regarding this fact, and thanks to knowing the typical structure of many sites we found that “contactus” and “aboutus” pages are typically the pages that we can spot contact information, location, and organization structure, and so on. The below dork limits the search to the contactus.asp pages of the governmental sites:

site:gov. inurl:contactus ext:asp*

The site can be modified to any other Top Level Domain (TLD), also the extension could be altered to any other possible back-end programming language like Asp.net, PHP, and Python.

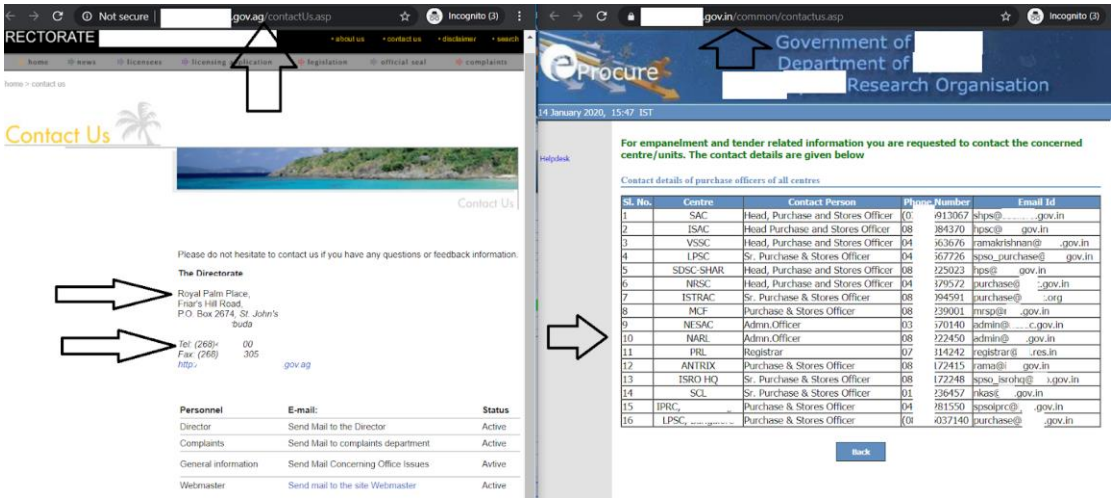

Figure 4: Finding the contactus page

Acquiring the possible subdomains of the target site also possible by straightforward Google dorks. Suppose the target site is www.example.com, so for this case simply typing site:*.example.com will limit the google search to the example.com subdirectories. Discovering the subdirectories of a website gives the attacker a chance to spot the vulnerabilities of the systems in the subdomains.

Being aware of the back-end technology of a website is another stage in information gathering. It is possible by a simple dork if we consider the imaginary target www.example.com then the site:example.com mixed with ext: with common server-side

development technologies file extension [34] will help the attacker to detect the back-end technology which leads him to concentrate on the vulnerabilities related to that programming language.

site:example.com ext:(file extension for server-side development technology)

4.2 Sensitive information disclosure, utilizing exploit-db

For this section, we tried to concentrate on GHDB from the exploit-db website that is the most comprehensive and biggest source of Google dorks [15]. The exploit-db website is maintained by an information security training company named offensive security and is a non-profit project that provides public service. It has two major sections namely Exploits and GHDB. The GHDB is an interesting source of search engine queries to uncover sensitive, and interesting information that is publicly available on the Internet. These queries are categorized into 14 various categories. First time Johny Long, an expert in hacking, in 2000 popularized the process" Google Hacking" by cataloging these queries in GHDB. He also has a book concentrating on Google dorks named "Google Hacking for Penetration Testers". He was also the person that first coined the term "Google Dork" that emphasized the uncovered information by Google dorks, which is not a google problem and is mostly unintentional misconfiguration of the programs installed by users. In November 2010 Johnny turned the GHDB to Offensive Security [4].

The chart below shows the approximate percentage of each of the 14 categories of Google dorks in GHDB. The statistics are based on the numbers till 15 of January 2020 that the overall number of registered dorks were 5196. Not all dorks in various categories useful for our purpose of this thesis. The first registered dork in this database has been registered on the 24th of June 2003. Next, we are explaining some more interesting categories of GHDB.

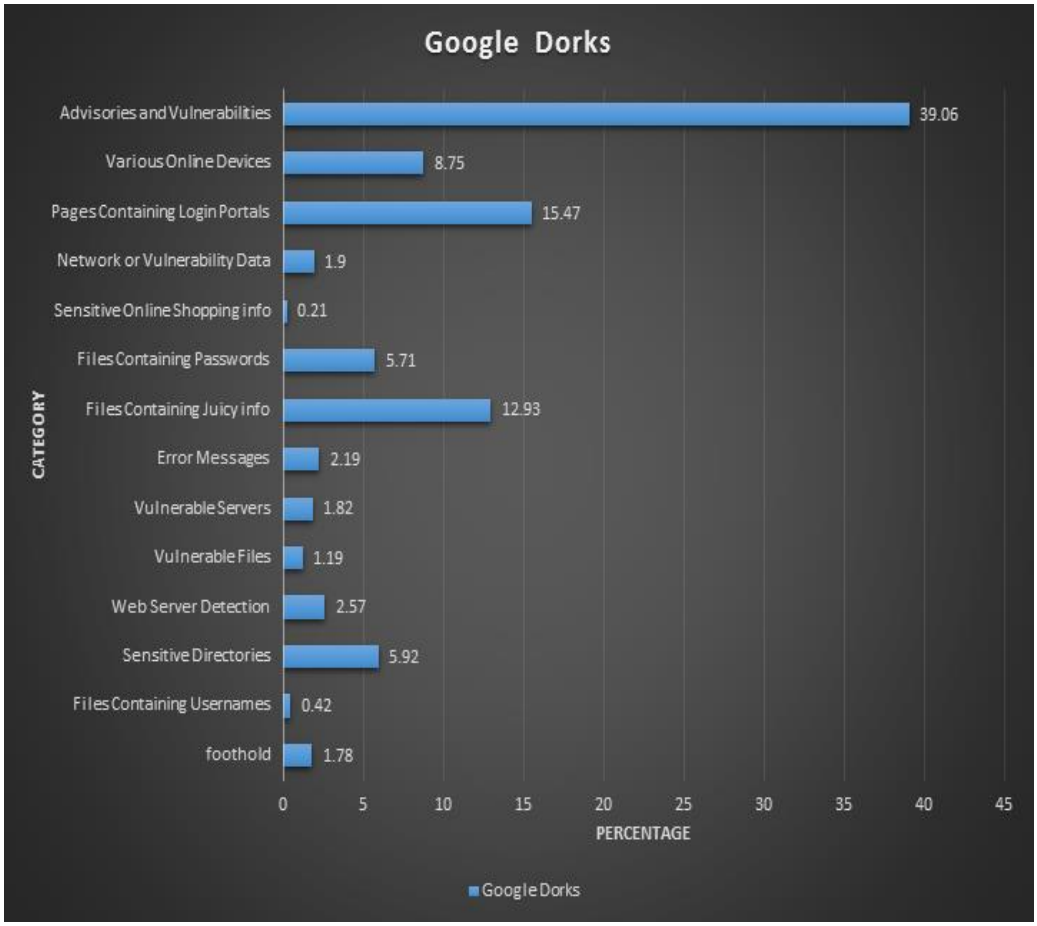

Figure 5: GHDB statistics

4.2.1 Files containing interesting information

The objective for using the category named is to focus on the information publicly available and possible to be found employing Google dorks that provides both hackers and penetration testers what is necessary for their aim.

The first dork in this category was written 2003-06-24 by a user named “anonymous” and the last one till the time of writing this document was 2020-01-09 by the author of this thesis.672 dorks out of 5196 dorks belonging to this category.

To put it simply dorks in this category facilitate the information gathering stage by giving beneficial information to the attacker. Inadvertently left credentials is another possibility in the file contains interesting information.

For clarification on how these files could be beneficial for attackers take a look at the below dork that acquires sensitive data including username and password for connecting to the database. The file databases.yml is a configuration file located by default in the config folder that allows the configuration for the database connection. Malicious users can accomplish attacks by utilizing the sensitive information exposed in this file. Restricting access or even removing this file is highly recommended to reduce the risk [43].

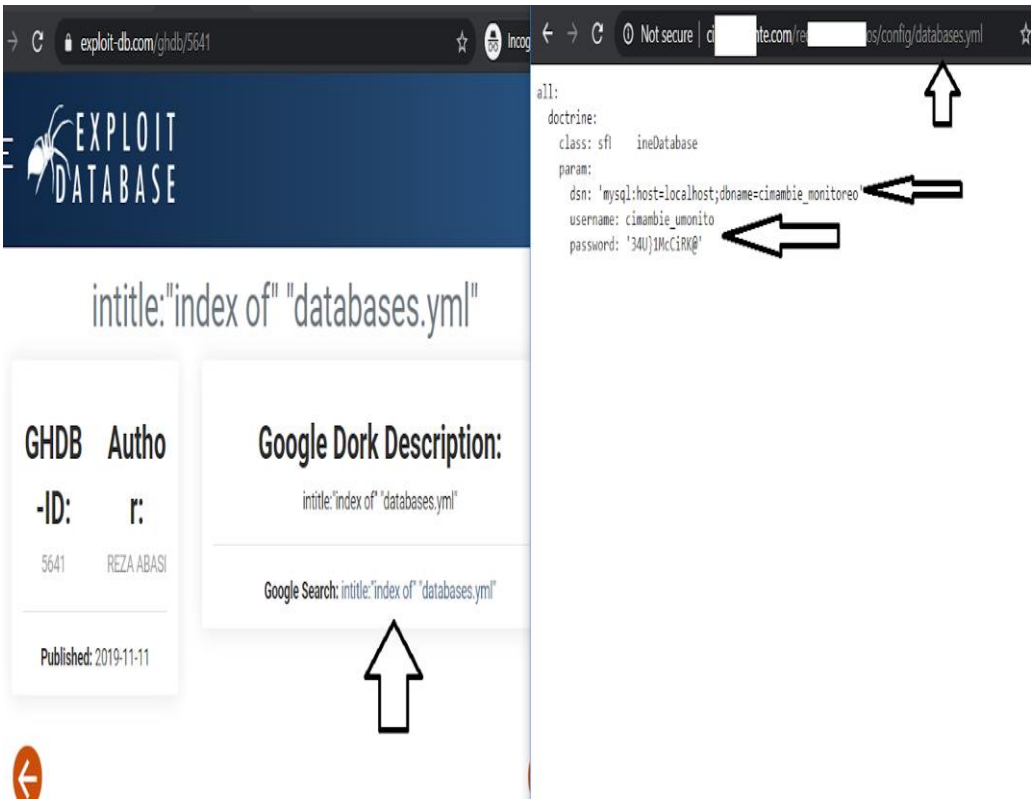

Figure 6: Finding datababases.yml file

As a general rule that can be used with all of the above-mentioned dorks and the coming dorks if the attacker concentrates on a single target he/she can either mix the dork with the “site:example_site” or if there is the “site:” keyword, write the desired example_site in front of the keyword “site:”. In this case, the last-mentioned dork will be:

intitle:"index of" "databases.yml" site:example_site

4.2.2 Files containing usernames

Another category of the exploit-db that can be advantageous in terms of both penetration testing and hacking is the files containing usernames. Typical user authentication is currently done with a combination of a username and password. Tools like Hydra [44] that is by default exist in Kali Linux [45] that is known as the most advanced penetration testing Distribution ever, perform the process of cracking for having access to remote systems. The Hydra covers a vast range of protocols and authentication mechanisms [44]. For instance, if the attacker tries to use the Hydra for cracking the system's credentials, a list of possible Usernames and passwords needed for this process.

Another situation that attackers can benefit the files containing usernames is that some user's passwords are related to their usernames or combination of usernames and other factors related to them like age, social security number, address, the field of study, pet name, and so on. So the attacker can utilize the files containing usernames to create a customized password list using all these features by a technique known as "password profiling". In fact, in password profiling, a customized dictionary is generated for launching an attack to cracking the system credential that is known as "dictionary Attack".

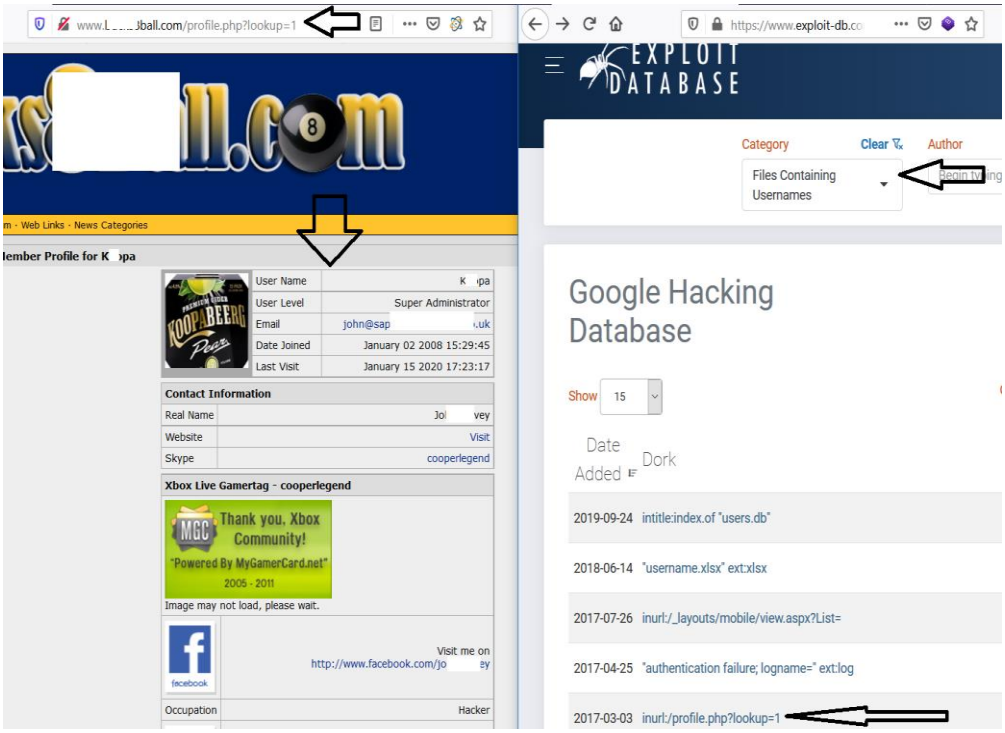

Figure 7: Finding a user's profile

As the above Figure illustrates the presence of "/profile.php?lookup=1" in the site's URL indicates that the page is the profile of a user in the site.

Another beneficial dork when it comes to finding the user lists, especially in forums built with VBulletin [46] that prepares this possibility to know about the forum members even when you are not registered or logged in to the forums.

site:/memberlist.php*

site:/memberlist.php intitle:"carding"*

Here we combined the dork for finding member lists of a carding forum [47]. As we can see in the below illustration it displays a member list of one of the famous carding sites.

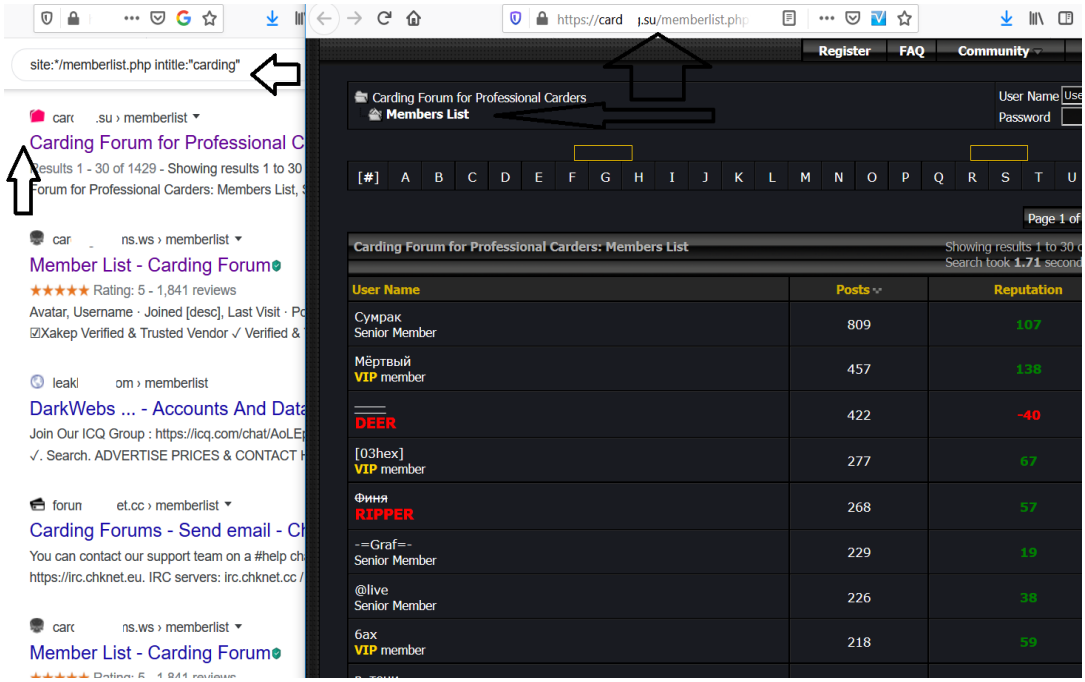

Figure 8: Carding forum's member page

4.2.3 Files containing passwords

Leaving the password in plain text format or even hashed or encoded in configuration files or any other file type to gain access to critical assets is a common mistake.

The Figure below shows how utilizing the dork: intitle: "index of" "db.conf" the attacker spots the credentials including the password in plain text in the file db.conf that is a configuration file for MapR Metrics databases [48].

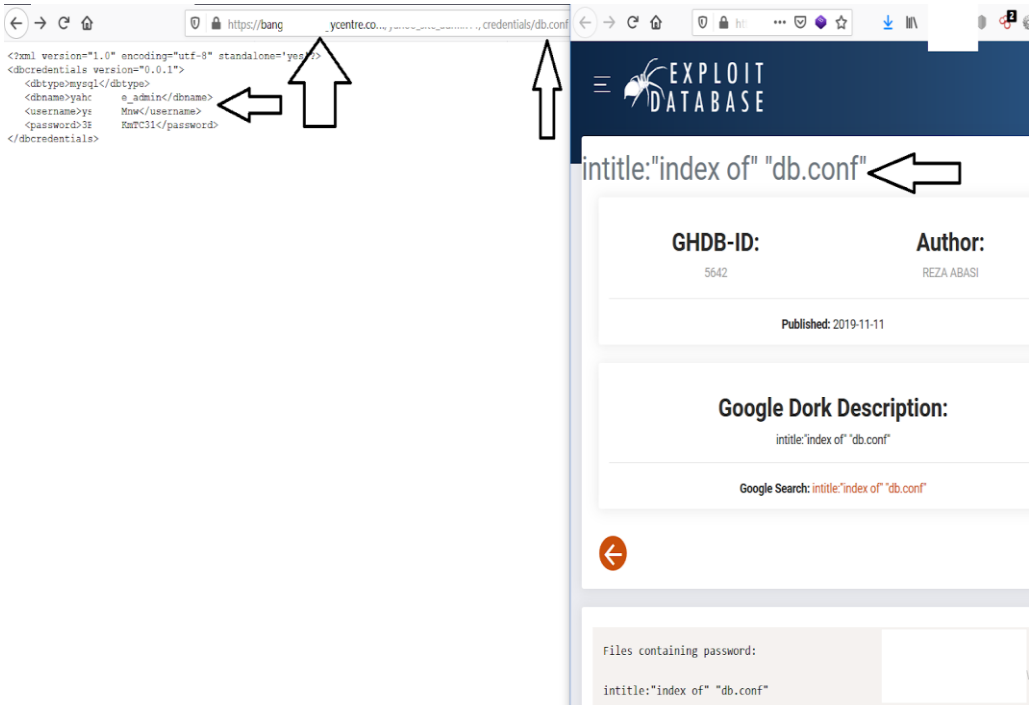

Figure 9: Finding db.conf file

Same as the files containing usernames, this category facilitates the development of combo lists and dictionaries to crack the online services and tools credentials, using dictionary attacks. Also, the attacker is capable of exploiting the user's intrinsic tendency to use single passwords for various services that are better known as "password reuse". This type of password attack is commonly known as a "password reuse attack". Though having a single password for multiple services makes things more convenient, it is perilously insecure. As a real case of such attacks that have been studied by Kaspersky [49] Mark a designer who had an account on multiple social media sites and email that all of them were linked to his email. Attackers gained access to the database of one of the sites he was a member of. The compromised database contained email addresses, passwords, and names. The attackers who had access to this database decided to try if some of the site s' members reused the same password for their email in other sites. Luckily for the hackers, Mark reused naively his password for his Facebook account and many other social media that lead attackers to forward messages from Mark's side to his friends asking for money to be transferred to the attacker's bank account. In the end, losing money and access to many of the portals due to the attacker's changing the

password so that the real owner of the accounts “Mark” can’t change them he decided to never in his life using the same password for different portals and activating the “two-factor authentication”

4.2.4 Enumeration using robots.txt and sitemap.xml files

Search engines like Google utilize web robots also known as Spiders, Crawlers, or Web Wanders that automatically traverse the Web and index the content of the web [50]. The Googlebot is the robot of google that crawls and indexes the web content [51]. A mechanism that is used to avoid sensitive directories and files of a website being indexed by bots is to benefit robots.txt. This is a file that allows the Web Site owners to instruct the robots about their sites. Every robot before visiting a web page of a web site checks the robots.txt of that site. This publicly available text file can be exploited by hackers to detect the structure of the site as well as the critical directories that are “disallowed” in this file.

One possible form that the attacker can exploit such publicly available info through robots.txt is to directly browse to the “disallowed” directory. Sometimes the attacker will be redirected to a 403 error page that means he/she is not authorized to access that resource but still worth trying. The Figure below shows how dork for spotting the robots.txt displays the useful information of a website that can be used by the attacker or pentester.

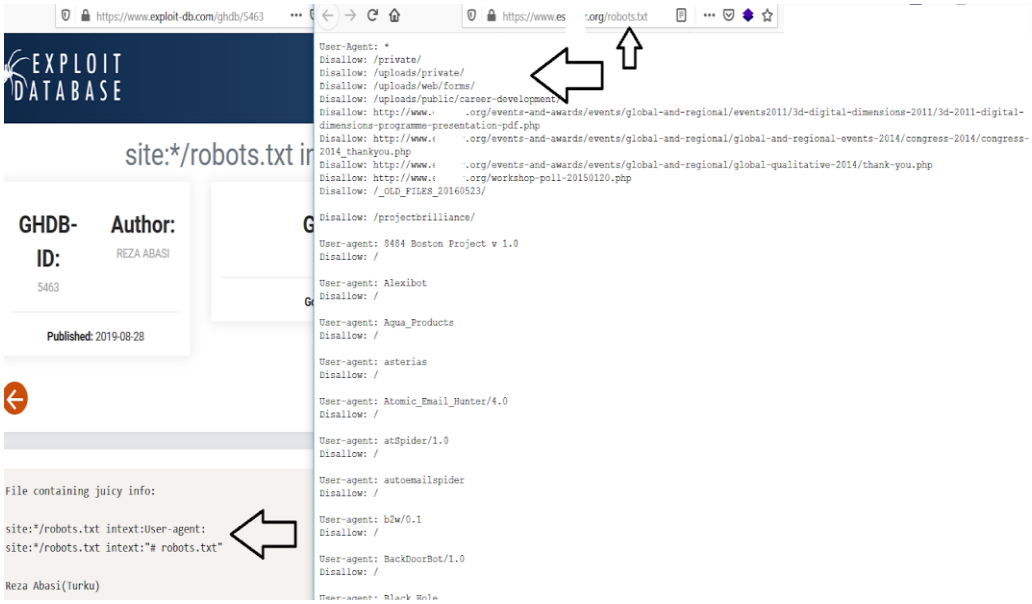

Figure 10: Finding robots.txt

Another dork that is worth describing here is the site:*/sitemap/sitemap.xml. Sitemap.xml is an xml file that provides information about videos, pages, and other files on sites and the relationship between them that can be read by the search engines when the site is being crawled. The valuable information provided to the search engines include the update and changed the time of the file. Sites that are large, or new and have few links to them or rich in media like figures or videos are mostly recommended to have this file [52].

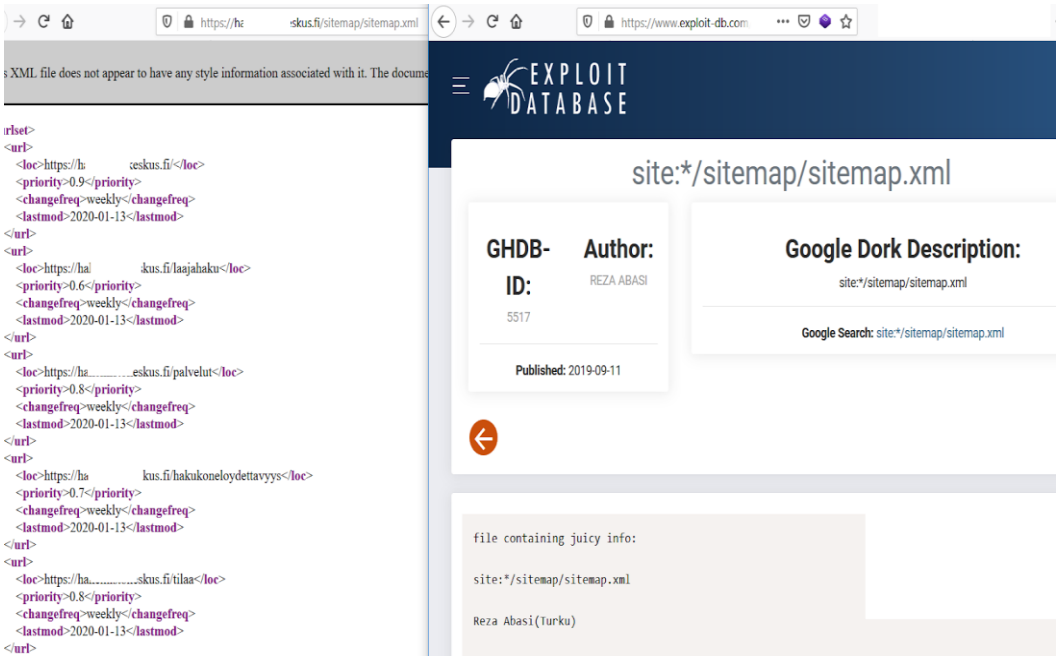

Figure11: Finding the sitemap.xml file

4.2.5 Find emails and passwords from file sharing sites, paste sites

When a data breach happens, the paste sites are one of the most probable places that data breaches are going to be shared. These paste services like pastebin.com are commonplace to find complete dumps or at least a sample of the compromised data [53]. Monitoring such services assist security researchers to find data dumps. To make the process of monitoring automatic, paste monitoring services are available through the net. These paste monitoring sites watch keywords that can be in the regular expression form [54]. Attackers and penetration testers can benefit from this email-password for cracking the authentication of the credentials.

The first step for finding emails and passwords from paste sites is to find probable paste sites.. Pastebin.com is one of the most common paste sites. So the dork below is advantageous for finding paste sites similar to pastebin.com:

site:pastebin. -site:pastebin.com*

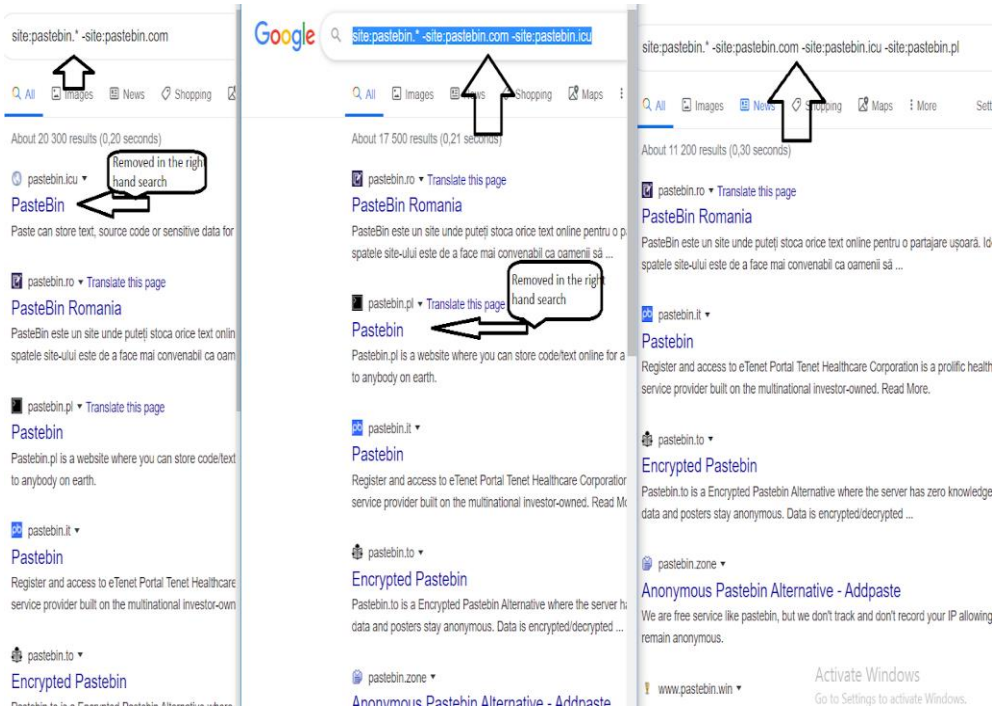

Figure12: Finding some pasting sites

Armed with google search techniques, combined with the found above dorks time to write nice dorks for finding email and password. Here are some examples:

site:paste.in intext:"@gmail.com" intext:"@yahoo.com"

site:pastebin.com intext:"@gmail.com" intext:"@yahoo.com"

As Gmail and Yahoo mail are among the most commonly used emails by users [55] we have used them in the above dorks.

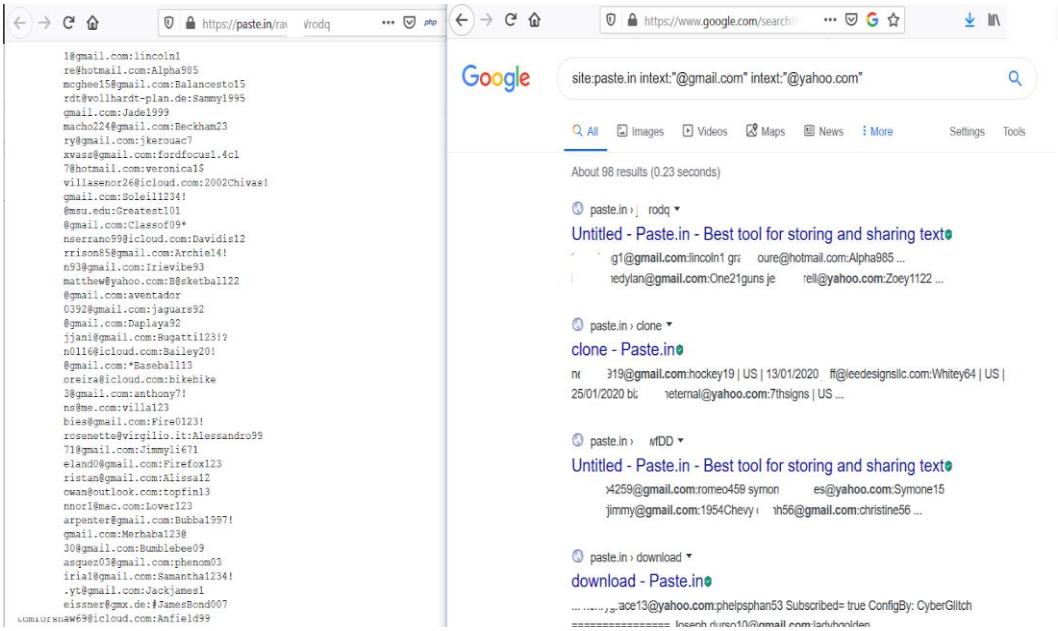

Figure13: Credentials in pasting sites

Besides, to paste sites, anonymous file sharing, and anonymous file upload sites, and similar services are common places for attackers to upload and share breached accounts as well as combo lists. Researching for the current thesis, we came across some of these sites that mixing them with other google advanced search techniques a good list of email, passwords can be acquired. Below are some typical dorks that can be created:

site:anonfile.com intext:"@gmail.com" intext:"@yahoo.com"

site:docdroid.net intext:"@gmail.com" intext:"@yahoo.com"

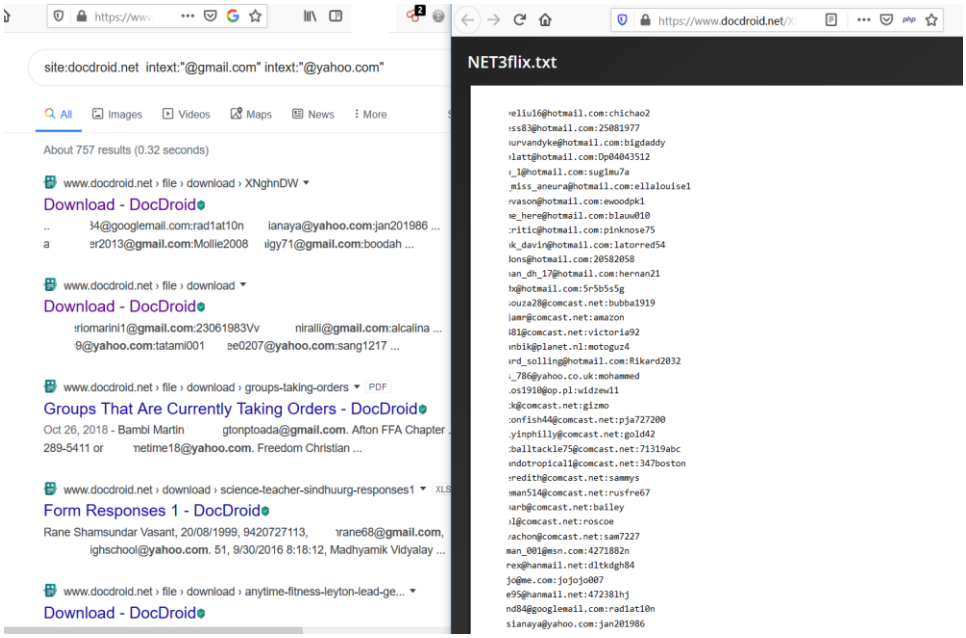

Figure14: Credentials in file-sharing sites

4.2.6 Find files containing emails and passwords in sites designed by WordPress

For this section, we tried to concentrate on acquiring files containing email and passwords from one of the most commonly used Content Management Systems software(CMS) named WordPress [55]. Mixing the keywords like “password”,”@gmail.com” with directories in WordPress that uploading files are possible, could be a good source of emails and passwords. Below are some typical dorks utilizing them the attacker can get access to usernames and passwords. The results can be limited to only doc, docx, or txt files as shown in the examples below:

site:/wp-content/uploads/ intext:"@gmail.com" intext:"password"*

inurl:/wp-content/uploads/ ext:txt intext:"@gmail.com" intext:"password"

inurl:/wp-content/uploads/ ext:txt intext:"username" intext:"password"

4.2.7 Find emails and passwords from random sites

In addition to all resources for finding emails and passwords, some arbitrary web sites share emails and passwords. The dork below alone or mixed with various file extensions like txt, doc, docx, xls, xlsx limits the google search to email, password lists:

intext:"@gmail.com:" intext:"@yahoo.com:" intext:"password"

All in all, files containing usernames or emails and passwords are important to find in this sense that due to broadly password reuse, organized cybercriminals implementing Account Takeover attacks (ATO) to gain access to the credentials of the target. Gaining access to compromised credentials, cybercriminals vending these credentials in underground markets. These stolen or hacked accounts are loaded into brute-forcing tools like Sentry MBA [56] to crack other websites or mobile applications. Lots of successful compromise of credentials as a result of password reuse exist. The success rate of gaining access to other accounts due to password reuse for the attackers is 2% [57]. It can be concluded that being armed with a good repository of email, password the attackers' chances for compromising new accounts will be amplified. Using a password manager in an organization to ensure the uniqueness of passwords as well as activating two-factor authentication (2FA) is highly recommended to reduce the risk of account compromise [58].

4.3 Vulnerability detection and enumeration

The first step for attackers to initiating a SQL injection [59] or XSS attacks [60] regardless of being manual or utilizing some tools for accomplishing it is to detect the vulnerable point. In addition to finding the vulnerable points, enumeration of the target in terms of security solutions that exist to protect the target, and detecting the server type is highly important. Google dorks are beneficial for both hackers and penetration testers for fulfilling these objectives.

4.3.1 Finding sites probably vulnerable to SQL injection, XSS

One of the methods for checking if a site is vulnerable to SQL injection is error-based that the attacker inserts an undefined character for the webserver and then the value will be passed to a variable that will be sent to the server so that using a SQL query the matching row or rows from the database table or tables retrieved and returned to the clients' request [61]. The Figure below shows how an attacker using a simple dork can find vulnerable points in a web site to the SQL injection for testing.

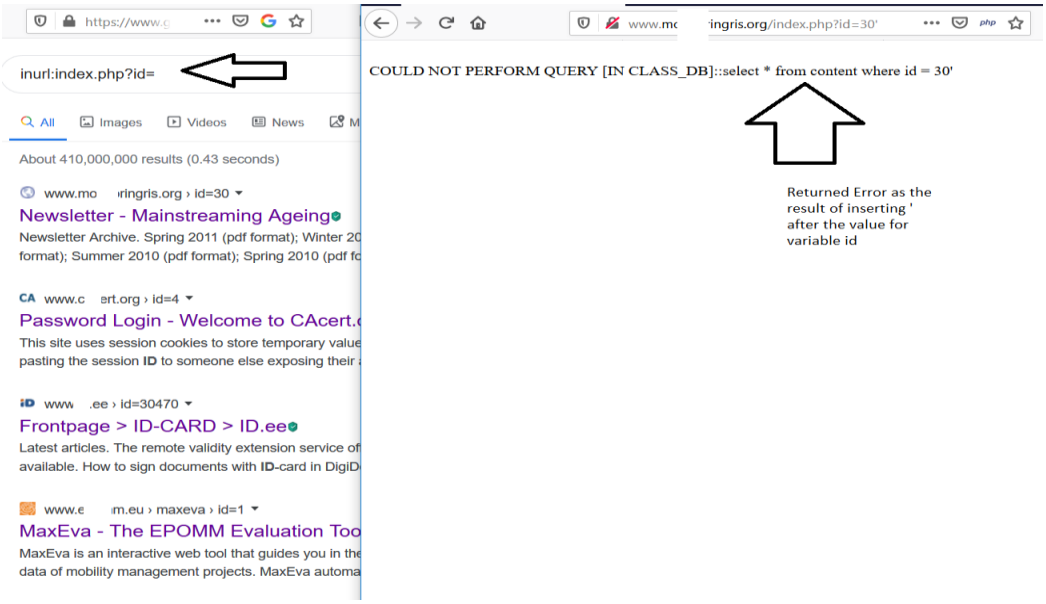

Figure15: SQLi vulnerable sites

In fact, for the above example, the file index.php can be substituted by any other page written by another server-side programming language like asp, aspx, jsp, py, and so on. The same dork works for checking whether a site is vulnerable to XSS attack.

4.3.2 Web server detection

The category named "webserver Detection" in GHDB provides a list of dorks for detecting the webserver. Being armed with these dorks and consequently procuring the

webserver type and version the attacker can simply concentrate on exploits for that web server and version.

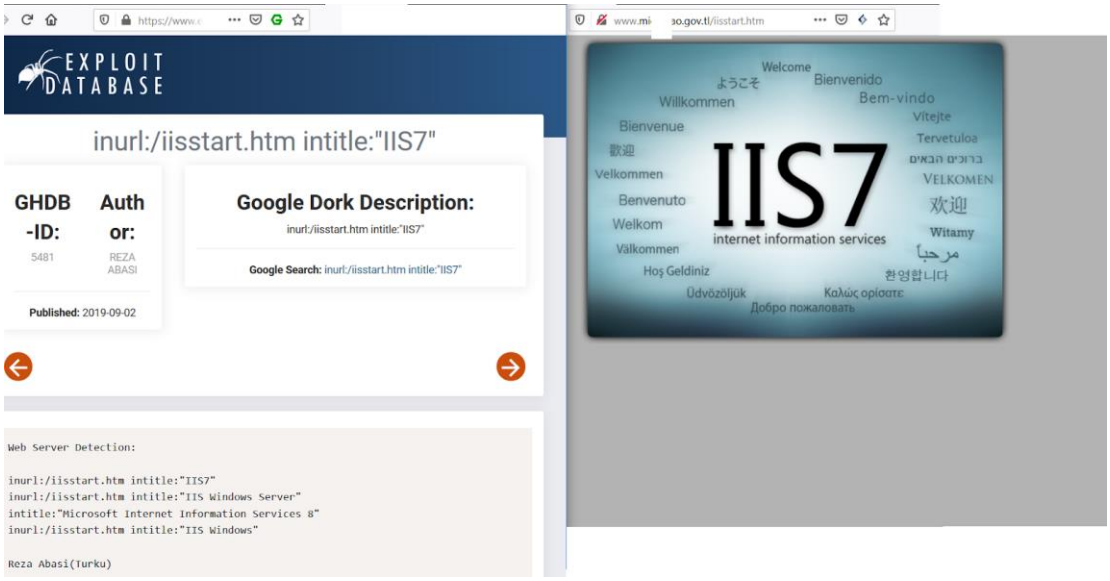

Figure 16: Web server detection

Dorks in other categories are beneficial for uncovering the protective security tools and solutions. For instance, Web Application Firewall (WAF) is a firewall for HTTP applications [62] that using the rules to protect against attacks like Cross-Site Scripting(XSS) and SQL injection. The error 403 could be an indicator of violating file permission or blocking the client by a WAF due to sending malicious requests [63]. The page containing this error displayed when HTTP error code 403 occurs that is when the configuration of the webserver denies the access for some reason to the URL that is requested that is due to the existence of rules in the website ‘s firewall [64]. Even more, it can be an indicator of using Modsecurity WAF [65]. Regarding this dork, the attacker came to this conclusion that the web site is protected by WAF so bypassing WAF techniques need to be applied.

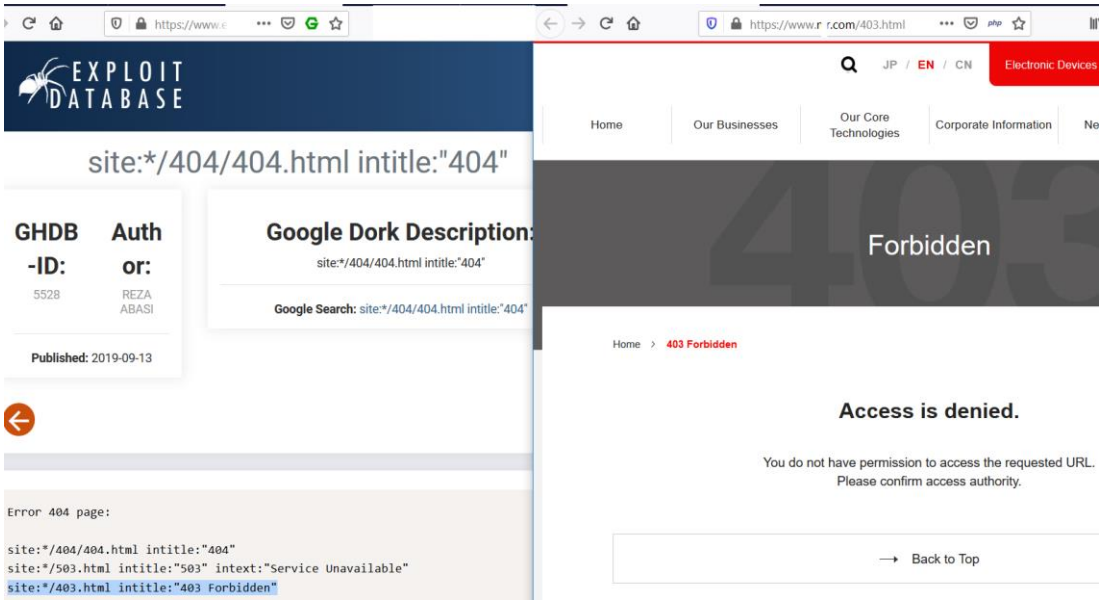

Figure17: Sites protected by WAF

As another example the dork below will assist the attacker to obtain information such as web server information:

site:/phpinfo.php intitle:"phpinfo()"*

This dork searches for finding the phpinfo.php that contains information like operating system, the directory structure of the site, web server technology, PHP configuration settings, enabled functionalities, FTP, LDAP, IMAP protocols enabled or not, MySQL basic configuration, session basic info, environment info, PHP variables, as well as other beneficial information. Accessing information presented in phpinfo.php, the disclosed sensitive information in this file can facilitate the hacking for the attacker. The file phpinfo.php should not be left exposed for random users as it is normally used for debugging purposes [66].

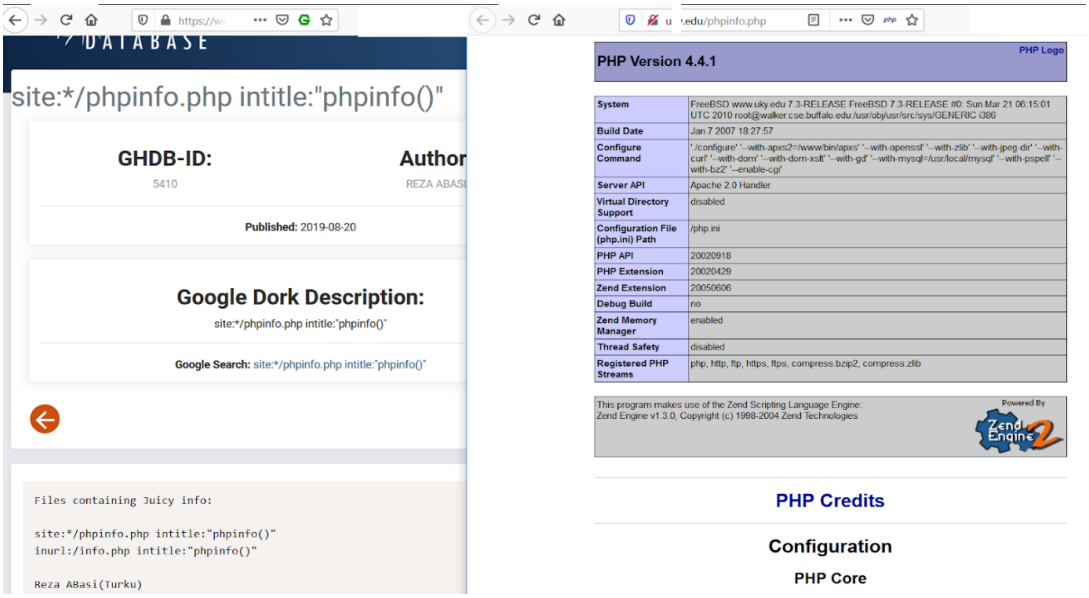

Figure18: Finding the phpinfo.php file

4.3.3 Popular CMS, forum software sensitive file and folder enumeration

Wordpress

When it comes to Content Management Systems software (CMS) according to the latest statistics presented by W3techs.com [67], Wordpress is the most used CMS for developing websites. Detecting the CMS is possible through using some sites [68] as well as even add-ons for browsers like the wappalyzer that exposes technologies used for developing sites [69]. Meanwhile, Google dorks making the CMS detection possible. The detection is primarily conceivable due to not modifying the default configuration and setting. As an example the dork below limits google's result to the sites developed by Wordpress:

intext:"Proudly powered by wordpress" site:/wp-content/*

By default "Proudly powered by wordpress" exists in all the pages of the sites developed by Wordpress, though it is simply feasible to modify or eliminate it [70]. Additionally ,the intitle: "index of" inurl:wp-content/ returns a list of websites with installed WordPress blogging software [15]. Or as another example, the dork below with minor false positive detects the sites built with Wordpress:

site:/wp-*/ -site:wp-*.**

Figuring out the directory structure of Wordpress facilitates writing more efficient dorks for enumerating sites and extracting useful information in terms of hacking and pen-testing. The list below presents a good list of directories and files in Wordpress. Combining them with inurl: intitle:" index of" and site: keywords will uncover interesting information about the structure of the sites built with Wordpress and have the default configuration [71].

/wp-admin/: The name of this folder is self-explanatory and deals with the administrator folder for the website. Connecting to the database is enabled using the admin.php file in the wp-admin directory. The Figure below is an illustration of the Google dork for finding the admin.php page of a site built with Wordpress

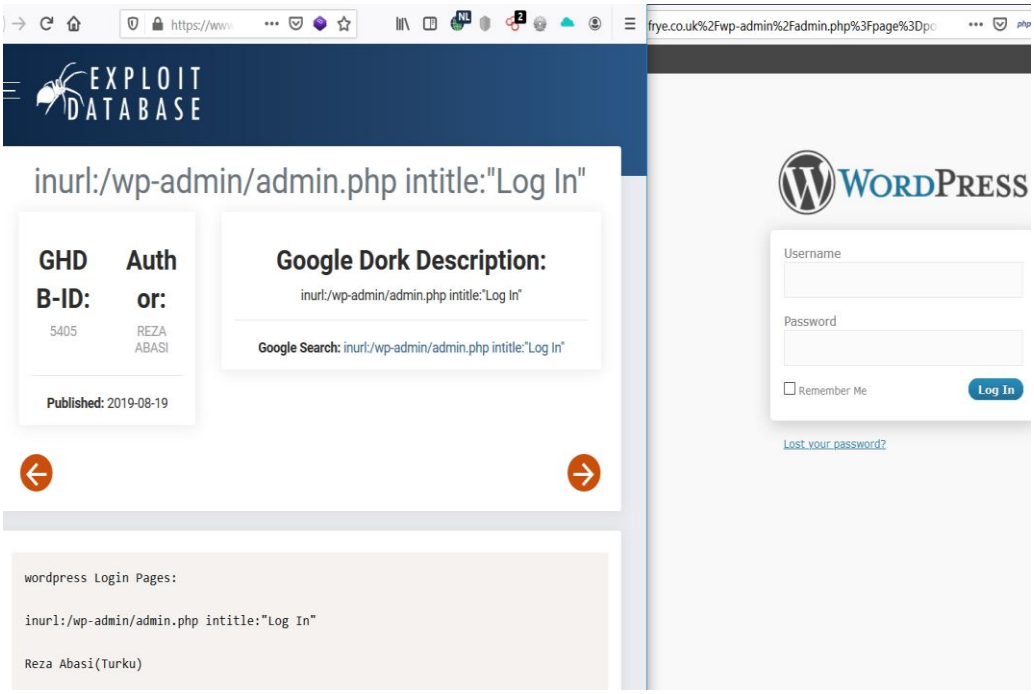

Figure19: Wordpress admin login page

/wp-content/: The themes and plugins are stored in this directory. This directory has a subdirectory /plugin/ that comprise the installed plugins for WordPress. Using the dork:

intitle:"index of" "/wp-content/plugins/"

The attacker can access to the content of the /wp-content/plugins/ or better to say the installed plugins and then searches for the exploits related to these plugins. One commonly installed plugin for WordPress is WordPress-SEO that there is a registered exploit for it in exploit-db [72]. A valuable site concentrating on Wordpress, its themes, and plugins is wpvulndb.com. Another subdirectory in /wp-content/ is /uploads/ that is primarily used in sites that having upload for instance a site that accepts the job applicants' resumes. This is a directory that is important especially for the security researchers and cyber forensic experts as one of the popular methods by attackers to hack a website is to upload their web shells through the submission sections of sites, then accessing their uploaded web shells and for instance performing "privilege escalation" with their uploaded shell and compromise the site with the most powerful privileges. One of the locations probably attackers upload their web shell could be /wp-content/uploads/ [73].

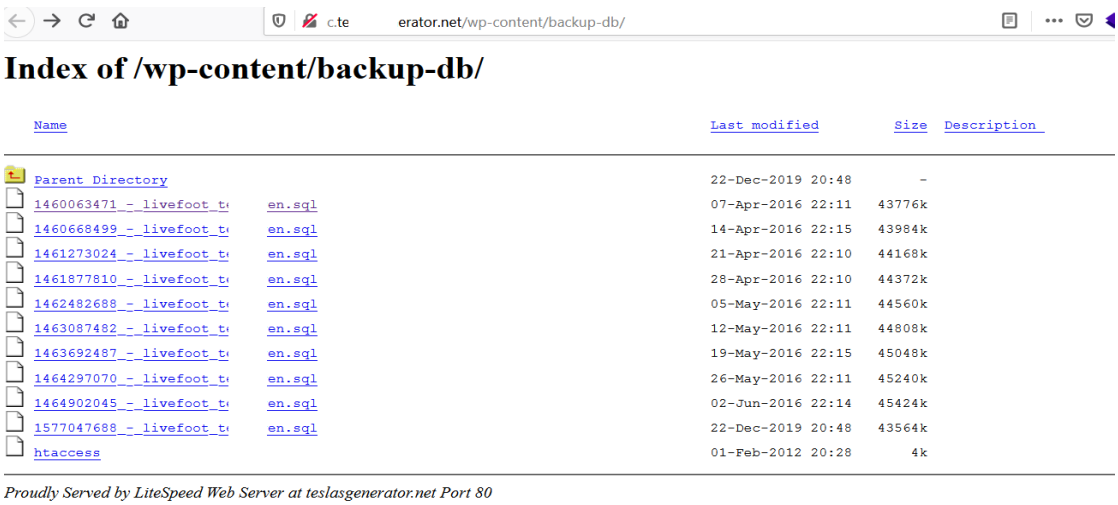

Figure 20: Backup files in wordpress

/wp-includes/: The last top-level directory in Wordpress default configuration is wp-includes that contains whatever permits the site to run and most of Wordpress core files stored here.

By the time of writing this document, searching in the GHDB section of exploit-db for wp- 88 dorks for various categories are displayed.

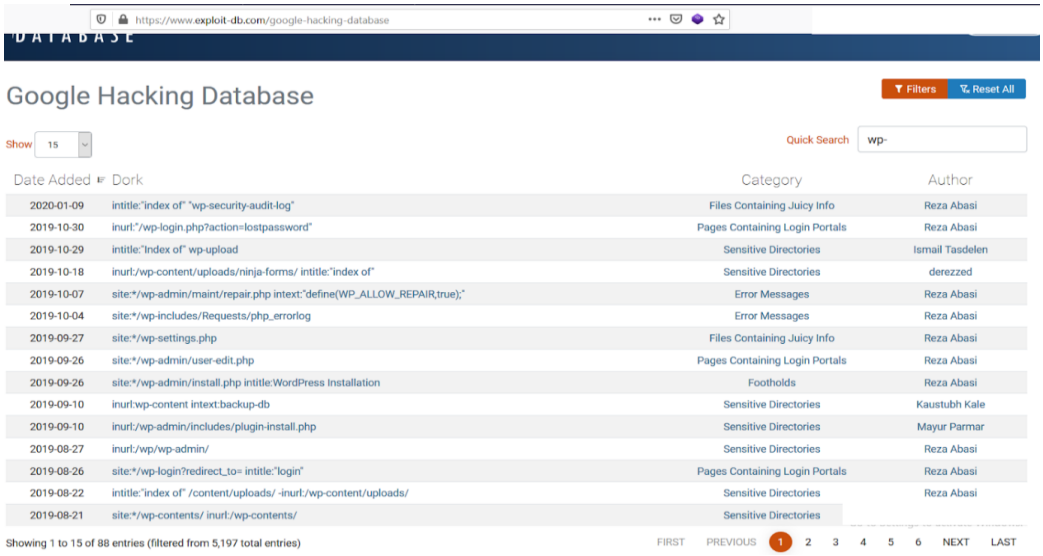

Figure 21: Wordpress dorks in GHDB

VBulletin

As a Platform for building forums, blogs, community publishing, and content**,** VBulletin is based on PHP as the server-side programming language and MySQL as the DBMS [74]. Simply searching for intext: "Powered by VBulletin Copyright" will limit the Google results to sites and mainly forums built with VBulletin. This assists the security researchers to enumerate more in terms of forums that prepare cybercrime services. When it comes to the attackers as a consequence of figuring out a forum that is built with VBulletin, they can deduce that the detected VBulletin site or forum is based on PHP and MySQL so to concentrate on exploiting the vulnerabilities related to them.

MyBB

Another dork that simply used by attackers and penetration testers to detect what forum software used to build a forum is:

intext: "Powered By MyBB, © "

Same as VBulletin, MyBB is based on PHP [75].

4.3.4 Error Messages, log files

Discovering the directory structure of the website as well as even sometimes an operating system of the web server, database tables, and column names can be obtained through the error messages and log files. Error handling needs to be considered with site developers to avoid rendering extra information which may lead to uncovering the site and database structure. For instance, the below Google dork finds error pages that some of them have the SQL query that causes an error, in that query the table name, column name exists that the attacker tries to find them during SQL injection and this error message paves the way for attacker for a more convenient SQL injection attack.

inurl:"/errors/report.php" intext: "There has been an error processing your request”

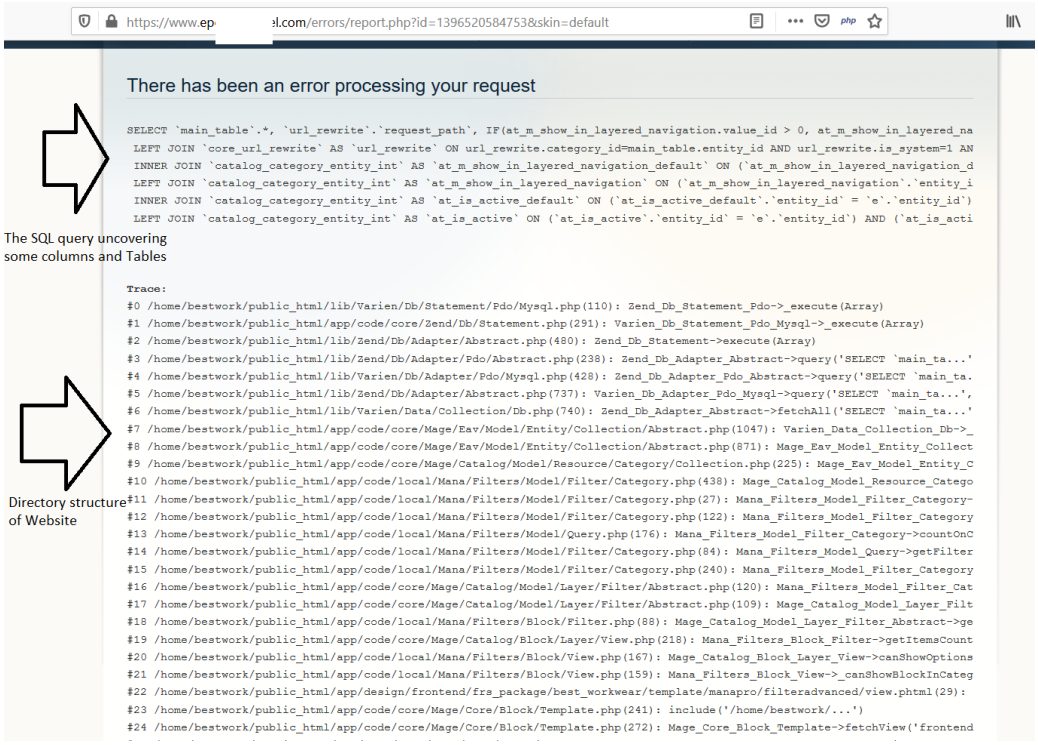

Figure 22: Finding error pages

As another example, the content of the log file discloses some of the emails registered on the site. The Figure below illustrates the emails that tried to subscribe to the site. The attacker can utilize the list for accomplishing a dictionary attack for the login page of the site for instance.

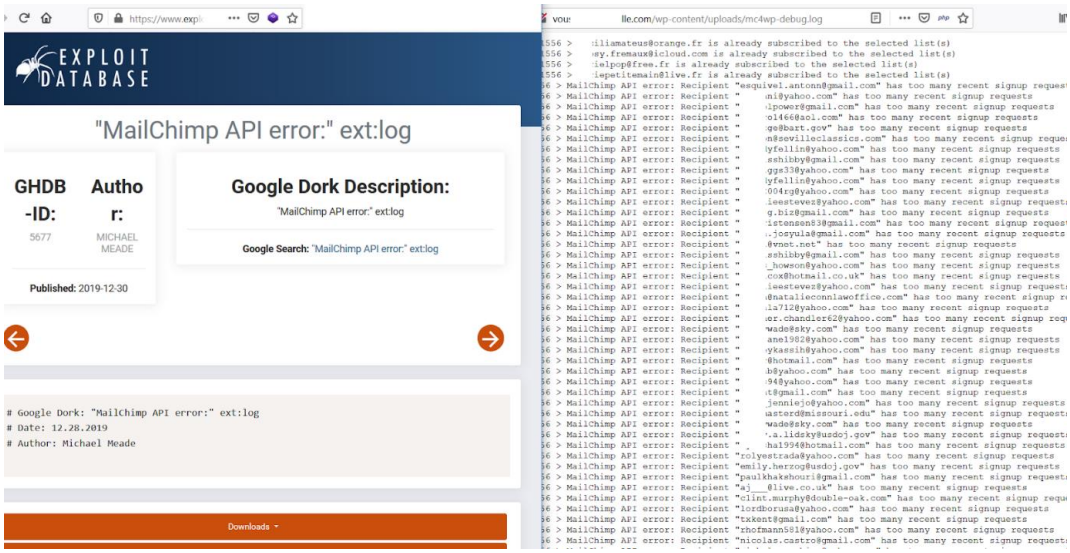

Figure 23: Error pages revealing emails

4.3.5 Vulnerable servers

As stated formerly VBulletin is one of the most famous forum software. One recently found vulnerability for the VBulletin with the CVE code: CVE-2019-16759 with the severity of the level high[76] allows the attacker to successfully run the remote code execution and running the commands on the remote vulnerable machine and the solution for it is to update them to the VBulletin patch according to the version used. Here is a dork that exists for finding sites built with VBulletin:

"Powered by vBulletin Version 5.5.4"

In addition to this vulnerability, there are a couple of other vulnerabilities for the Vbulletin [77] that the attacker can easily detect the version of the VBulletin and test if they are vulnerable to the attacks exist. The attacker can search for dorks related to

VBulletin as shown below in the GHDB section of exploit-db to discover sites built with VBulletin:

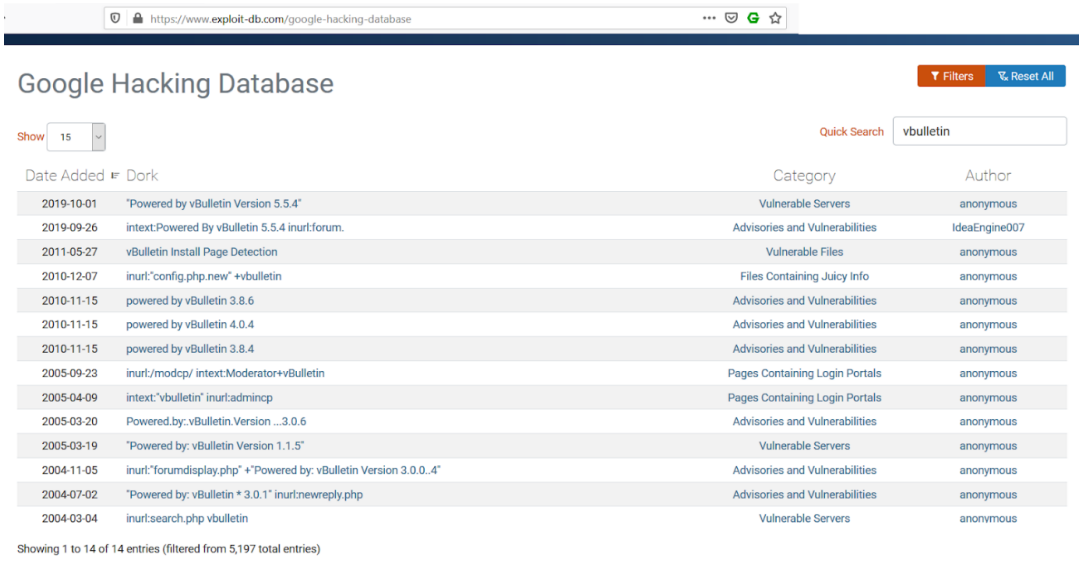

Figure 24: VBulletin dorks in GHDB

4.3.6 Vulnerable files

The attacker can find these vulnerable or "bad" files using a category named as vulnerable files in GHDB. Another method to find these vulnerable files is to utilize CGI scanners that list vulnerable directories and files in a data file. Detecting the vulnerable files could be a clue of the purportedly vulnerable program, but it is not guaranteed the vulnerability of the program [78]. The last dork added to this category dates back to 12 May 2016.

4.3.7 Web asset and online device discovery

Two different categories namely "Various Online devices" and "Login Portals" are the subject of this section. Loads of useful dorks come to both categories that assist attackers to do the process of hacking more conveniently. As an example in terms of login portals the dork below finds all the Mikrotik hotspot login portals publicly available. Clients, to access the public networks, need to be authenticated through Mikrotik hotspot gateway [79]. Lack of captcha for this login page makes it more defenseless to the dictionary

attack. So the typical solution to stop attackers from being successful for a dictionary attack is to have a standard captcha to hinder automated dictionary attacks or even account lockout policy to lock the page for a specific time according to the security policy of the organization.

intext: "Please log on to use the mikrotik hotspot service" intitle:"mikrotik hotspot >login" -github -site:mikrotik.com

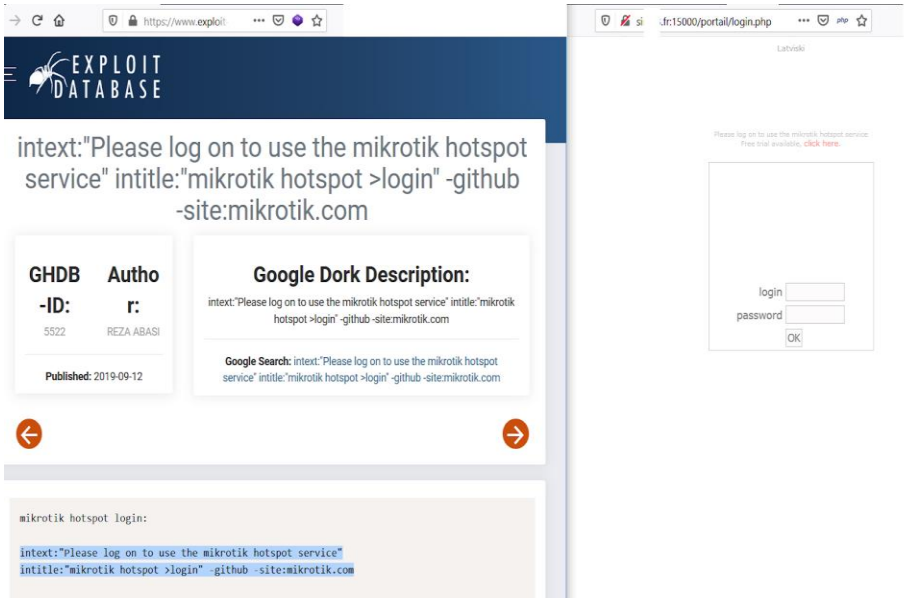

Figure 25: Mikrotik hotspot login page

The next category for this section is various online devices. This category includes a large number of devices both hardware and software. It ranges from network and application monitoring solutions like Zabbix [80], Cisco IP phones [81], Marshal Video Servers [82], and many other online devices. As mentioned earlier the “various online devices “category contains a login portal in most of the cases. So, for instance, the dictionary attack can be used by the attacker for attacking and finding the credentials. Similarly to the login portals category, security solutions to diminish the possibility of success for dictionary attacks is the captcha usage, Account lockout, and 2-factor authentication and that is when login to your accounts like a bank account, Facebook, or any other account or portal needs authentication is not just relied on password but a specific code in the

form of an SMS for instance [58]. The illustration below is an example of utilizing google dorks for finding the login portal of Marshal Video Server login portal:

intitle:"Marshall VS Server"

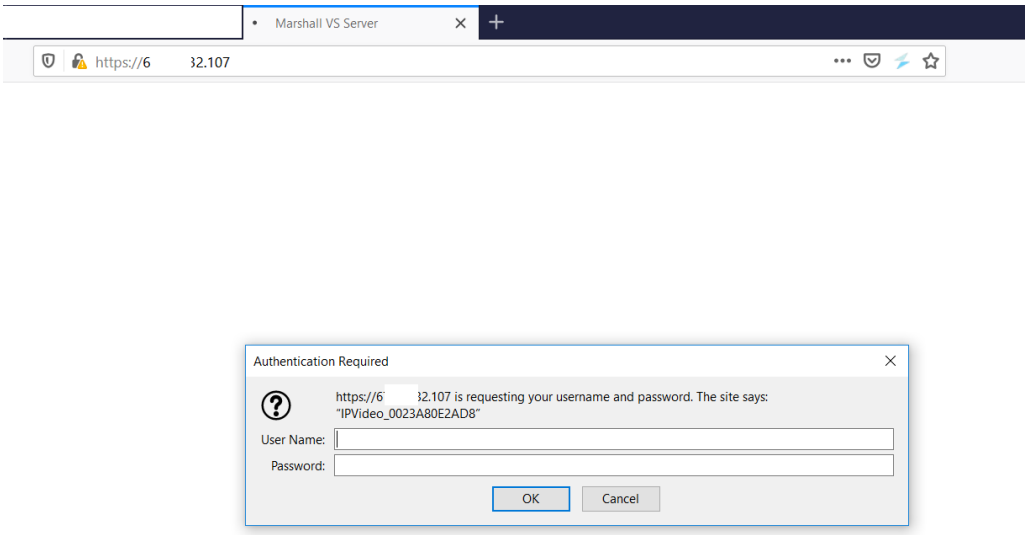

Figure 26: Marshal video login portal

4.4 Automation tools for using dorks

Both attackers and hackers utilize tools that automate the dorking process. Searching darknet sites as well as hacking forums the author of this thesis recognized that the same tools applied by both black and white hat hackers and the only difference is their objective not their tools for this phase. Automating dorks made the process of information gathering and the rest of the phases which Google dorks applied much straightforward, However, there are barriers to automation. For instance, google considers strict limitation on the queries sent to it per day per IP address, and the IP addresses disobeying this, and sending lots of queries are punished by facing captcha, still, attackers remain unaffected as they use proxies to deceive google and bypass this barrier. Searching through the web, we located some automated Google dorking tools. Dorkme,Zeus,xgdork,Bingoo,dork-cli,GoogD0rker,M-dork,gD0rk,fast-recon,snitch, goodork, sitedigger ,gooscan,pagodo are to name a few of them. Here we decide to concentrate on more details on pagodo [83]. The pagodo was developed as a passive Google dork script that collects probably vulnerable web pages and sites on the Internet. The ghdb_scraper.py is in charge of retrieving Google Dorks while the second part pagodo.py is responsible for leveraging the gathered information by ghdb_scraper.py. The pagodo is written by python and to tackle the problem of getting an HTTP 503 error that is due to being detected by Google as a bot and blocking your IP by google for some time the Pagodo uses a bank of proxies and a proxy chain. So the user needs to install proxychains4 also. After installation of proxychains4 the configuration file located in /etc/proxychains4.conf required to be manipulated by setting dynamic socks proxies and ports. To make the searches more like human searches, pagodo uses User-Agent randomization, and randomization of time between search queries is done.

5 Cyber investigation Use Cases

For this chapter the focus is on finding the cybercriminal forums, sites, activities as well as dorks regarding detection of defaced sites, cybercriminal Facebook, telegram, and other social media accounts, searching through the darknet utilizing Google dorks and web2tor proxy sites like onion.ws [84].

5.1 Threat intelligence hunting

Threat intelligence is the knowledge that is based on evidence in the form of context, indicators, implications, mechanisms, and actionable recommendations, considering existing or emerging threats to assets. Threat intelligence is helpful in terms of establishing a more precise picture of the threats and environment [85]. Besides, identifying emerging attacks made easier with threat intelligence so overall, defensive activities to avoid attacks made simpler. For this part, we mainly focused on finding indicators of defacement of web sites as well as finding the most famous hacking forums and sites both in clear net and dark web. The detection of defacements is beneficial in this sense that the web sites, especially the ones built with common CMS's changing the default settings and configurations that made their sites and assets vulnerable to defacements. Finding hacking forums and other resources that cybercriminals share their experiences and products are advantageous in this sense that monitoring such activities make the prediction of upcoming attacks or even ongoing attacks more convenient.

5.2 Defaced Sites

Website defacement includes the changes in the appearance of a website's pages by pictures or words. Attackers or better known as "hacktivists" for this case may have a varied number of motivations for defacing websites. They change the content of a governmental or particular website that they are against with a message or picture of their own choice. In addition to this group of attackers, some other groups' main objective is

to mock site owners by exploiting the website's vulnerabilities and doing defacement. In any case, the reputation of the web site owner was damaged [86]. In addition to that, it is probable that the website that has been defaced is compromised or even a web shell uploaded and the attacker has access to the web site's directories with different levels of privileges that can be escalated if the web site is vulnerable. It is common between the defacers and defacer groups to register their defacements in websites like zone-h.org [87], mirror-h.org [88], or defacer.id [89]. However, sometimes the defacers avoid registering their defacements in the above-mentioned websites. For such cases, Google dorks can be used. Generally, the defaced sites contain pages with "Hacked by", "Defaced by", "Your box owned by", and "Your security is low".

The first three texts are commonly followed by the defacer or defacer's group name. So using the above-mentioned texts in dorks that targeting either intext of the page or intitle of the page or even combining them will retrieve sites that have been defaced. Here are some examples of such dorks:

intext:"your security is low" intitle:"hacked by"

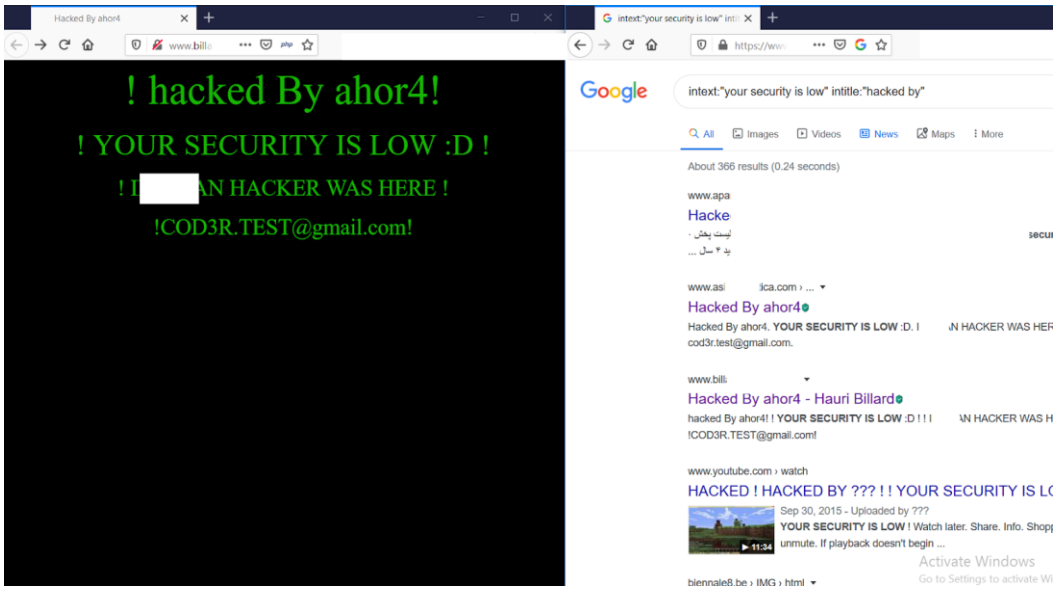

Figure 27: Defaced sites

Similarly the following droks facilitate finding the defaced sites: intitle:"hacked by" ext:htm ,intitle:"hacked by" ext:php , intitle:"touch by" intext:"hacked by".

The other place that defacers upload their pages is the upload directory of the sites. For instance, the dork below returns sites defaced and the defacement page uploaded to /upload/ subdirectory of the site:

site:/upload/ intext:"hacked by"*

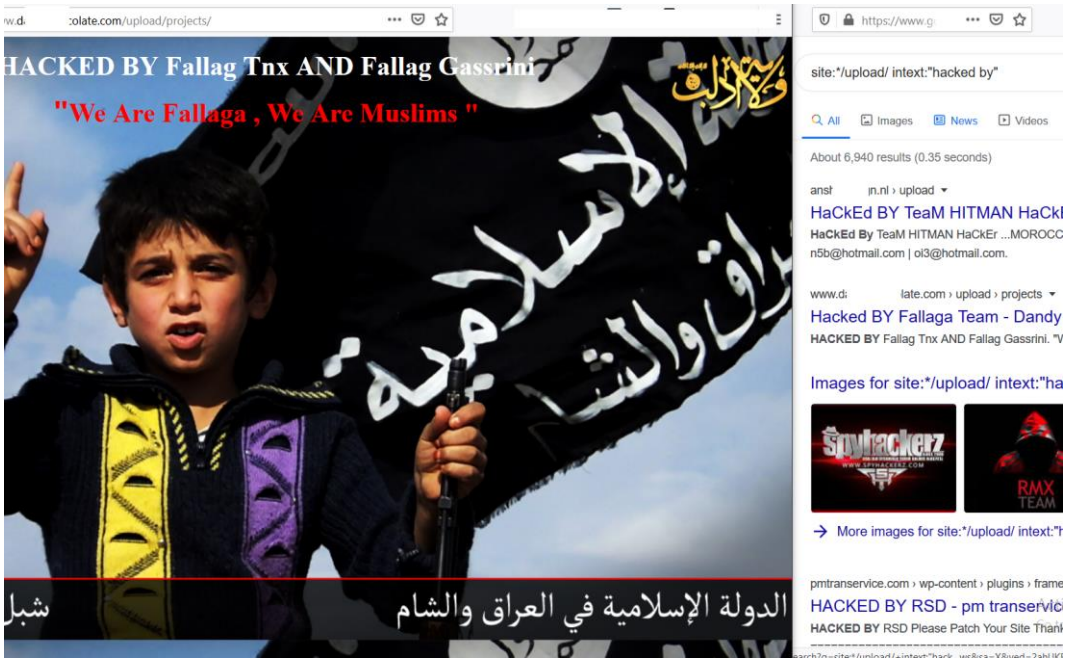

Figure 28: Defaced page in /upload/ directory

Similarly, the /Figures/ subfolder is potentially one of the folders that commonly considered by the defacers to upload their pages:

site:/Figures/ intext:"hacked by"*

It is also common for defacers to upload their page or even sometimes their web shell to the default uploading directory of Content Management Systems (CMS). Dorks below return the defaced sites that are built with Wordpress and their default uploading directory remained unchanged:

site:/wp-content/uploads/ intext:"hacked by"*

site:/wp-content/uploads/*/*/ intext:"hacked by"*

inurl:"/wp-content/uploads/" intext:"hacked by"

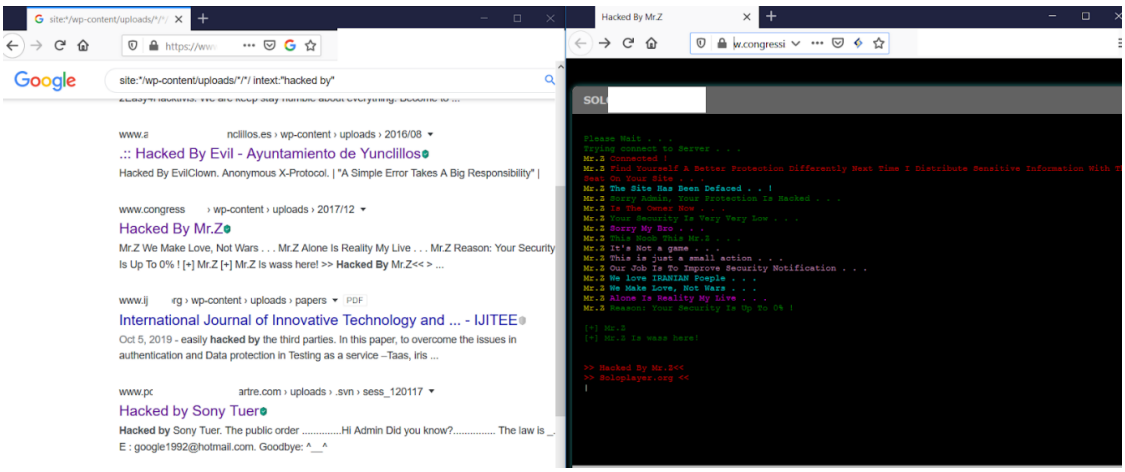

Figure 29: Defaced page found in sites built with wordpress

5.2.1 Finding defaced sites by the same defacer group

Finding sites hacked by a defacer group made it easier by using Google dorks. For this goal combining the intext, inurl and intitle can assist us. The below dorks return all the sites defaced by a group named "indonesian error system" and they mainly upload a php page named “dit14.php”.

intext:"indonesian error system" inurl:dit14.php , intitle: "Hacked By Mr.OXiG3n"

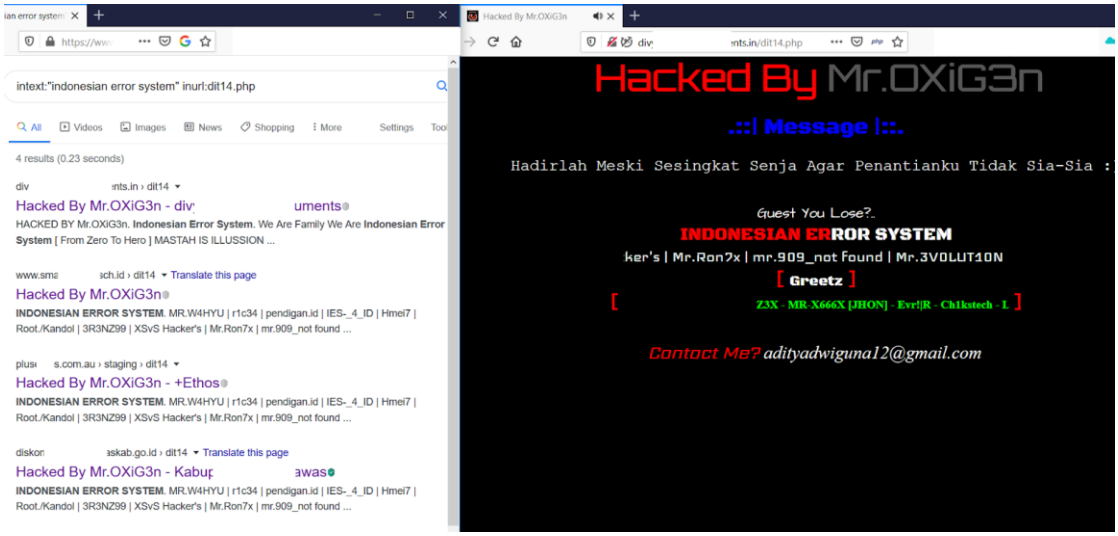

Figure 30: Indonesian error system

5.2.2 Finding defacers' telegram, Facebook, twitter, skype, and other social media accounts

It is common for the defacers to leave their contact information including social media accounts like Facebook account, twitter account, email, telegram account, Instagram account, and other social media accounts in the defaced site. Finding and collecting their social media accounts assist the cybersecurity researchers and cybersecurity investigators to improve the chance of disclosing the real identity of the defacer or defacer groups. Every small piece of information is worthy in terms of putting them together and could be parts of a puzzle of uncovering the hackers' identity. Some typical dorks for finding the defacers' social media accounts can be found below as well as a screenshot of an example of such dorks. The below screenshot is a governmental site defaced by a hacker named "FlyBoy" and his/her Gmail address and Facebook account.

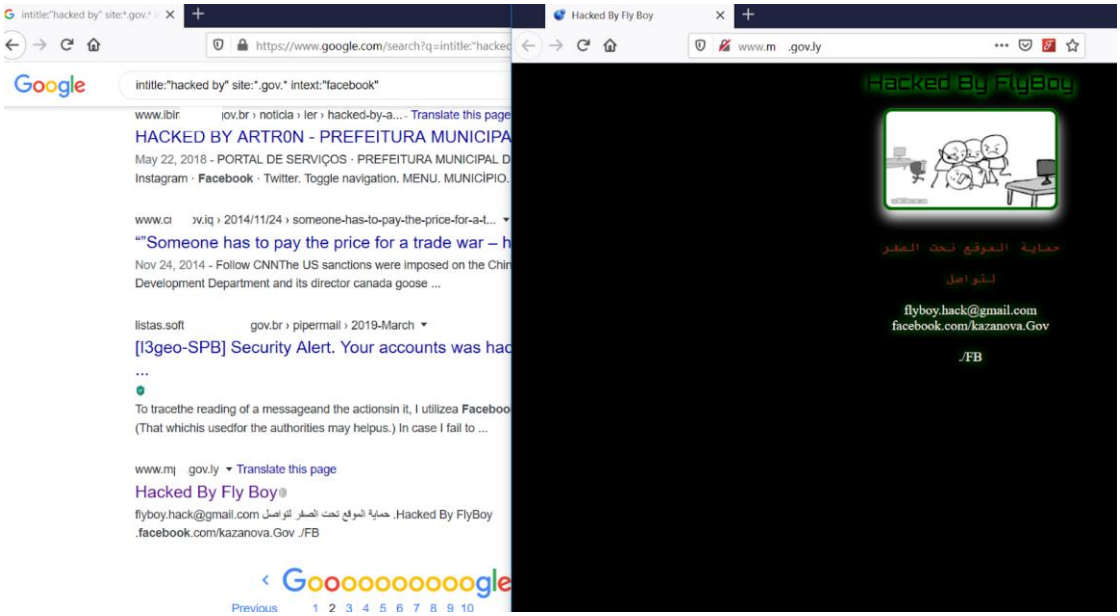

Figure 31: The Facebook account of defacer

intitle:"Hacked By" site:.gov.* intext:"facebook.com" | intext:"fb.com"*

intitle:"Hacked By" intext:"facebook.com/groups/" -site:facebook.com

Manipulating these dorks by substituting Facebook with desired social media can lead to the desired social media account of the defacers.

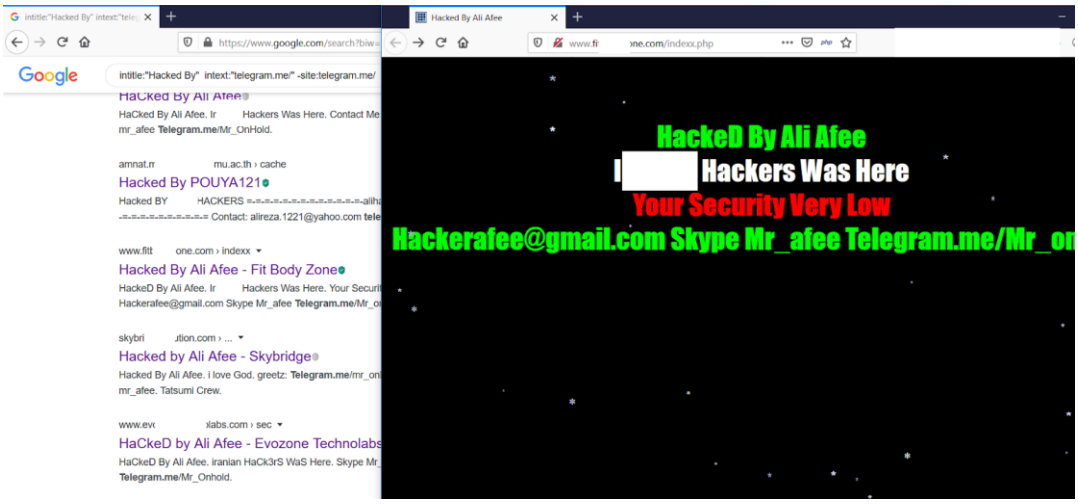

Figure 32: Telegram account of defacer

5.3 Cybercriminal activities

A factor that plays a key role to have optimal search results is to pick the best and most suitable keyword. So, being familiar with cybercrimes, their definition is a key factor for searching. Besides, according to the author of this thesis experience translation of the keywords and using the translated version will be productive in terms of the findings.

5.3.1 Cybercriminal forums

Underground Online forums in both the surface web and dark web are providing platforms for trading stolen goods and illicit services. Financial information trading is the area carding forums focus on[90].In addition to carding forums, hacking forums that mainly focus on trading data dumps as well as other software and hacking tools, pirated operating systems are to name a few of the products offered there. Organizations need better cyber defense as the cyber-attacks are getting more and more complicated. Cyber threats are hard to predict accurately as attackers do their best to conceal their traces yet it is common among the attackers to discuss techniques and methods for hacking in hacking forums [91]. Monitoring hacking and criminal forums seem an essential part of predicting cyber-attacks. Google dorks facilitate finding the cybercrime forums by combining the cybercrimes and queries for finding the forums. The below Figure displays a sample query for finding “carding forums”. Vividly other products and services are traded in such forums.

intitle:"carding forum"

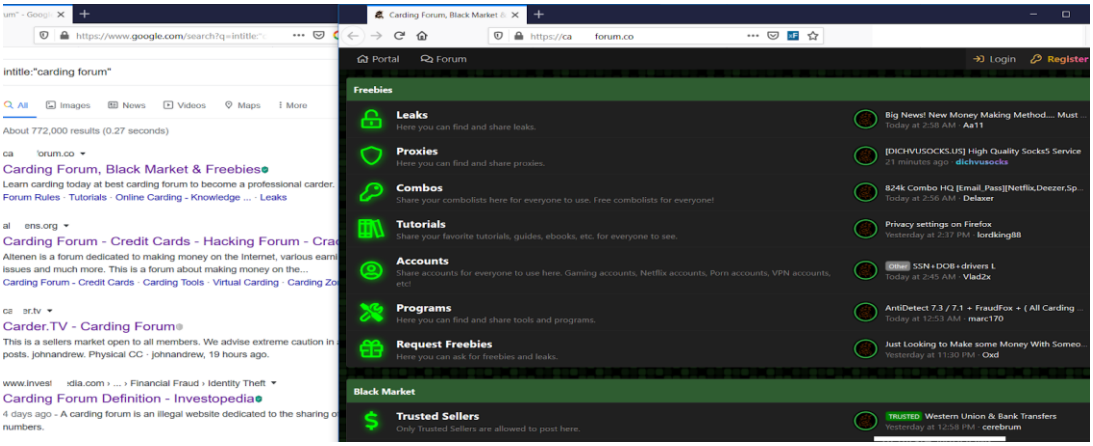

Figure 33: Carding forum

The breached databases are sold in such forums both in the dark web and clear net. The following dorks limit the search results to forums trading data dump:

inurl:forum intext:"leaks" intext:"database download"

intext:"leaked accounts" intitle:"forum", intitle: "Accounts and Database dumps"

intext:"email:pass" intitle:"forum", intitle:"account dumps" intitle:"forum"

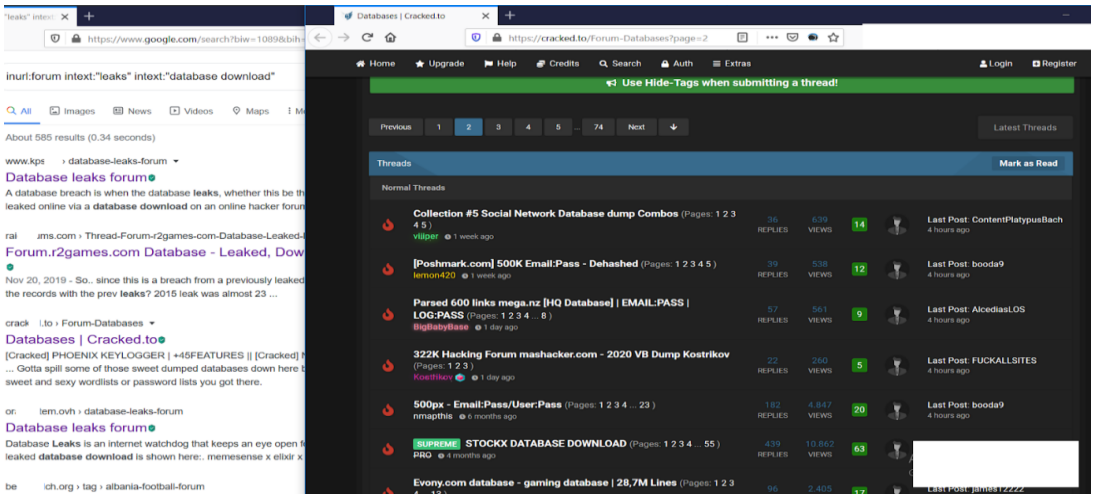

Figure 34: Hacking forum

The abovementioned dorks facilitate finding forums having dumps that the author of this thesis wrote. Registration in such forums can be done free of charge, except for some of the forums that either invite-only membership or need registration fee.

5.3.2 Finding cybercrime channels in telegram

Hackers, crackers, and cyber criminals utilizing social media for their targets. Social media applications are getting more and more commonplace for them to achieve their aims. Telegram [92] with its fascinating features like end-to-end-encryption, self-destruct, secret chat messages attract many cybercriminals. Distributing malware using telegram is reported recently by the North Korean "Lazarus" group for stealing cryptocurrency according to Kaspersky [93]. They created channels that the customers can join and communicate with the sellers and deal with their desired products. Almost all the products traded in underground forums and darknet sites are offered in these channels. The dork below limits the search results to channels offering freebies or giveaways credentials.

intext:"@gmail.com" intext:"@yahoo.com" site:telegram.me/ | site:t.me/

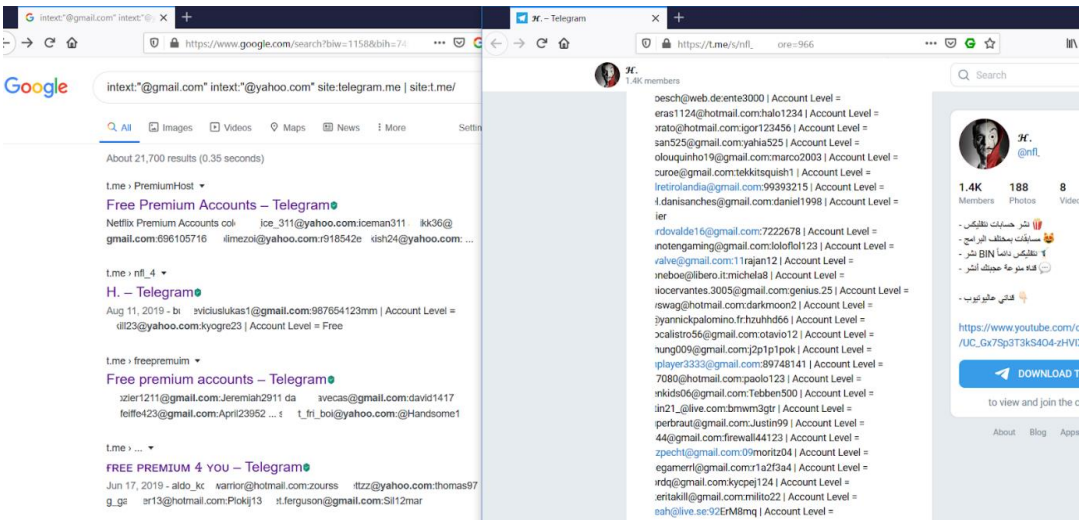

Figure 35: Credentials in telegram channels

The template below assists in finding interesting channels containing cybercrimes materials.

intext: ”cyber_crime_keyword” site:telegram.me/

5.3.3 Finding cybercrimes in paste sites

As explained in the former chapters “paste sites” are popular among the cybercriminals for sharing the breached databases and credentials. In addition to that, it is common to share cybercriminal sites and forums in the paste sites. One good example of such activities is the ISIS using a paste site named JustPaste.it in combination with the telegram application to disseminate the kill-list [94].

The typical template for writing dorks for locating such forums or sites are as below:

intext:”cyber_crime_keyword” site:paste_site

5.3.4 Finding autobuy shops using Google dorks: selly.gg, shoppy.gg

Online stores for selling digital goods facilitate e-commerce. However, cybercriminals benefit from such sites for selling the credit cards’ information, hacked accounts, and other digital products. These sites also are known as autobuy shops integrate with a various number of payment gateways to facilitate the payments by the customers even in the form of cryptocurrency payments.

Sites like shoppy.gg, brutshop.ru, bayacc.store, selly.gg are some of the famous autobuy shops. Hacking forums are the best place for finding more of such “autobuy shops”. As an example searching for intitle:”autobuy shop” will limit the search results to the forums advertising such sites. Moreover, to find exact pages in such sites containing premium accounts that are most for sale product use a query like this:

site:shoppy.gg/ intext:"premium accounts"

5.3.5 Finding ICQ and vk.com channels, and Facebook groups offering cybercriminal products

ICQ is one of the famous social media applications that is used by cybercriminals for advertising their products and channels containing their digital products. Though the main purpose of this application was chatting, sharing files, photos, and videos, hackers use it for selling their products.

site:icq. intext:"carding"*

Another social media application that is worth mentioning and focusing on is vk.com [95] as according to alexa.com [96] at the time of writing this thesis around 67% of the audience for this site are from Russia, it is ranked the 20th top site, that makes this social media application prominent among others for checking.

site:vk.com intext:"@gmail.com" intext:"@yahoo.com"

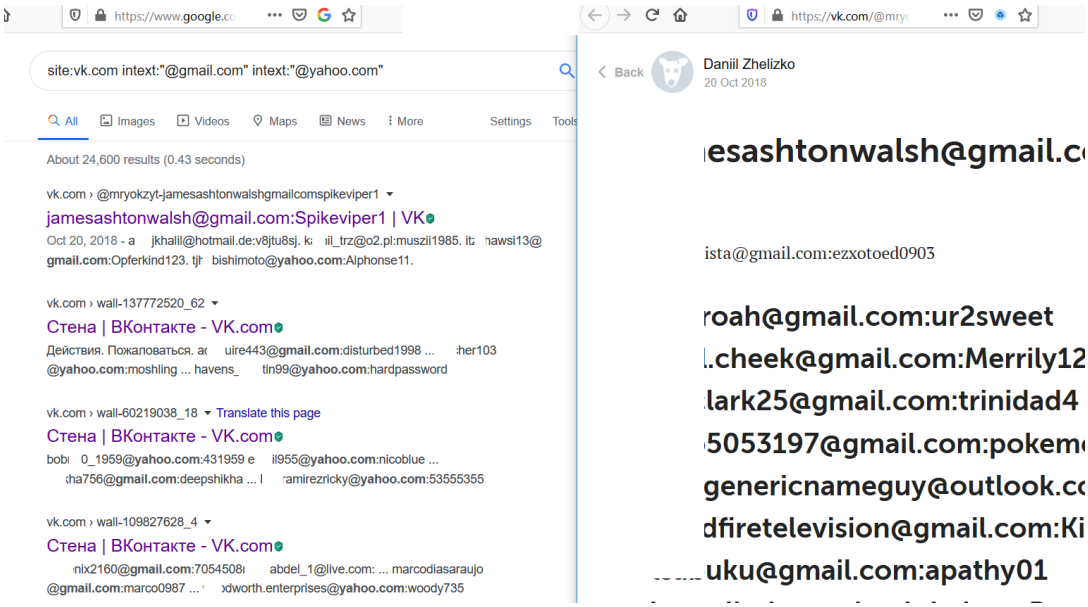

Figure 36: Credentials in vk.com

Ranked the 4th most visited site in alexa.com [96], Facebook.com is another valuable site to check for cybercriminal products and activities. Though such activities are advertised by Facebook personal accounts, Facebook groups are more likely to have more

interesting materials. The screenshot below illustrates groups in Facebook advertising carding sites.

site:facebook.com/groups/ intext:"cc+cvv"

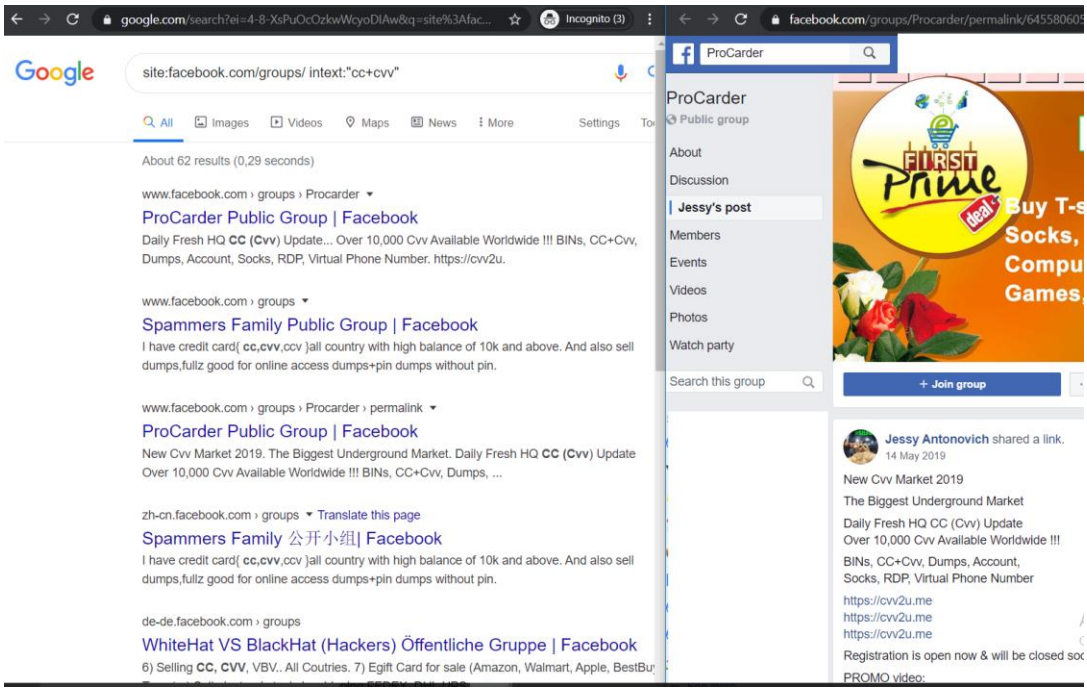

Figure 37: Carding groups on Facebook

5.3.6 Find carding, hacking sites in darknet using dorks and pastebin sites

Though according to the typical definitions of the dark web, it is part of the Internet that is not indexed by typical search engines like google and can't be searched by them [97] and browsing through such sites requires Tor Browser [98], still thanks to google and some sites like "Paste" sites we can create Google dorks for finding sites and forums active in the dark web. As explained in the former sections sharing accounts, famous hacking and cracking sites and forums are commonly done in paste sites. Regarding this the following dorks are assisting us to identify hacking and cracking .onion sites and forums:

*intext:".onion" intext:"carding" site:pastebin.**

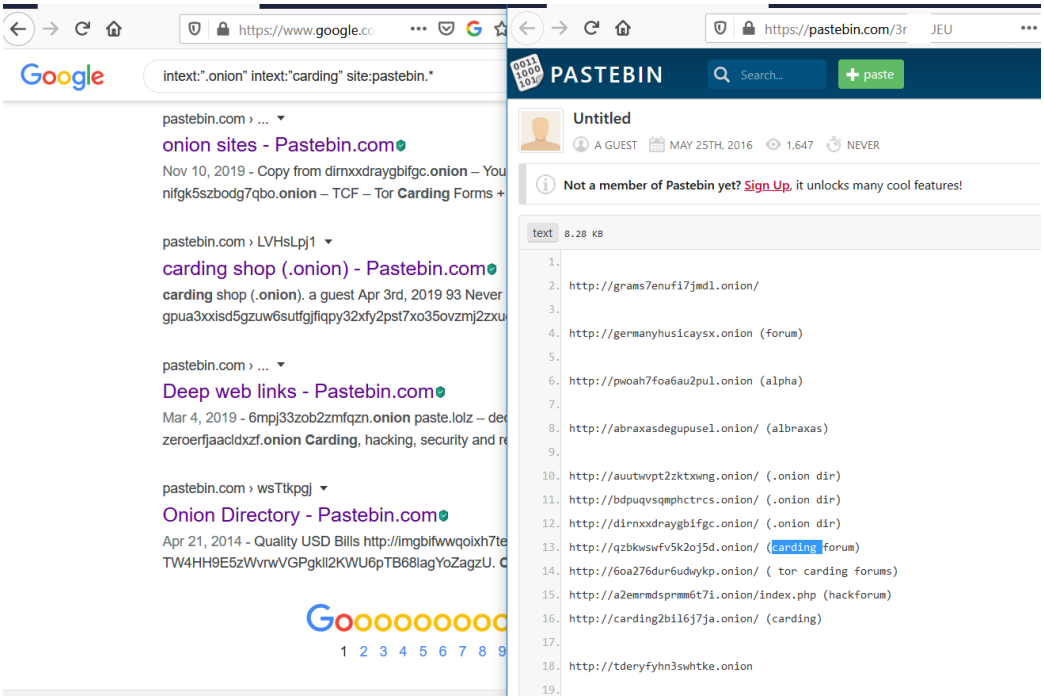

Figure 38: Finding onion sites

5.3.7 Finding darknet criminal sites using Google dork and web2tor proxies

What is known as Tor Project‘s onion services prepares a method of the anonymous running of network service. The anonymity provided for both Tor clients by obfuscating their IP address and the Tor onion services by allowing them obfuscation of their network location and IP address. Developed in 2004, according to Tor Project’s statistics the onion services count more than 100,000 onion services daily, which the serving traffic of nearly 1Gbps. The onion services not limited to web sites, but includes file sharing, instant messaging. In 2016 Facebook reported more than 1 million of its users monthly logged into its onion service. Being accessible through the Tor network, onion services, and typical web services are different from each other. Besides onion services are discoverable by users organically, rather than by search engines [99]. Typically the first and most straightforward definition of the dark web is that part of the Internet that is not indexed by search engines and accessing their content is possible using the Tor browser that is an anonymizing browser [97] and commonly considered as the criminal activities hotbed. A various number of studies support this claim that cybercriminals find the dark

web a safer place for trading their products [100]. Two different solutions are recommended for mitigating the negative effect of cybercriminals especially terrorists of the dark web. One is to block the access to Tor Network that violates the freedom of speech, especially in the totalitarian, non-democratic countries that reporters, whistleblowers, and other NGOs use for good. The other solution is to reduce the number of illicit hidden services [9]. De-anonymizing such illicit hidden services with google dorks are the topic for this part. A good example is the operation Onymous that was a coalition of US and some European countries' law enforcement organizations that led to the shutdown of 400 hidden services [5]. The illicit activities range from selling hacked accounts, credit card information, hacking tools, guns, databases, drugs, counterfeit money, and many more products and services. Our target in this section is not giving information about dark web activities but using dorks and techniques to search darknet through a typical search engine named Google. For this section, the idea behind darknet gateways or proxies is facilitating. These gateways are reverse proxies forwarding HTTP requests to the desired Tor service while no need to install the Tor browser or any other software. Below is a list of such sites and such sites and services would not provide anonymity for the users while browsing the darknet:

https://onion.sh/,http://onion.ws/,http://onion.ly/,http://onion.dog/,http://onion.glass/, http://onion.cab/

*intitle:"HQER" site:onion.**

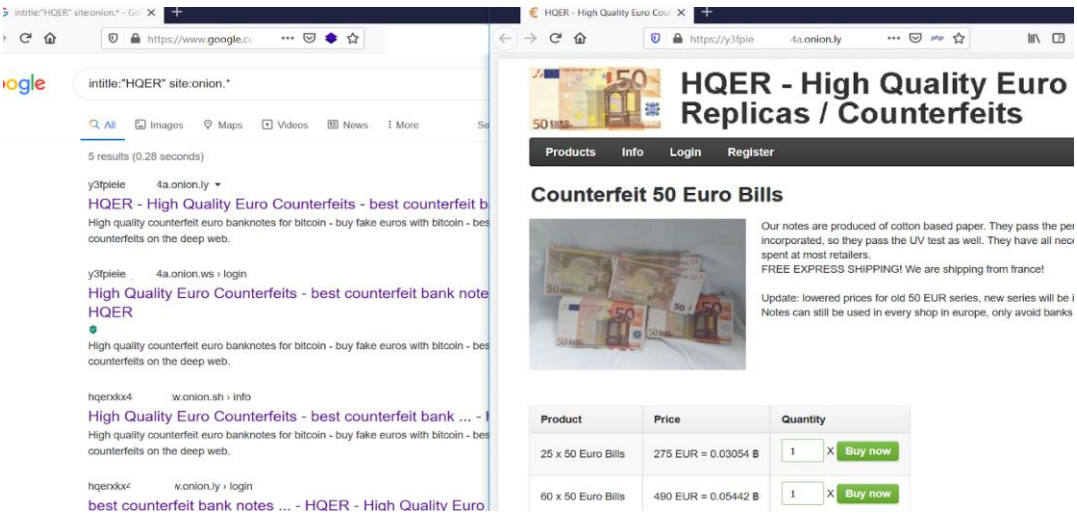

Figure 39: HQER counterfeits

5.3.8 Use Case: Find Clearnet of an onion site with Google dorks

Regarding the former section, we came to this conclusion that thanks to Google dorks that utilize gateway sites that act as reverse proxy searching and discovering more about onion sites made more convenient. Besides, almost all dorks in exploit-db can be used for our aim. Herewith a use case, we will check that conveniently finding a clearnet site of an onion carding site is possible by:

intitle:"club2CRD"

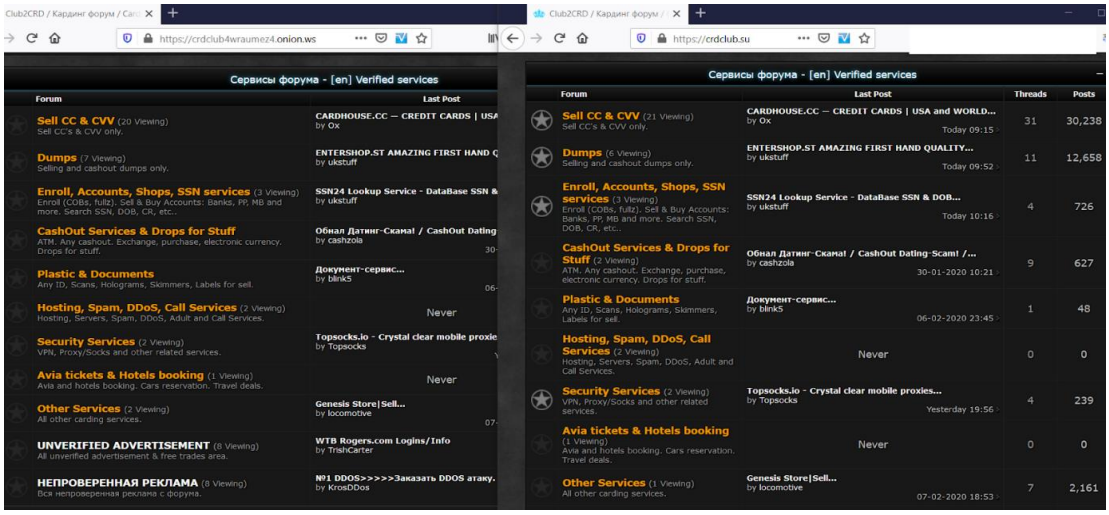

Figure 40: Darknet site mirror in the surface web

In another example this dork that was registered by the author of this thesis in exploit-db.com/GHDB and useful when it comes to web server detection, site:*/server-status intext: "Apache server status for" with a minor change can be converted to:

site:onion./server-status intext:"Apache server status for"*

The screenshot illustrates useful information regarding an onion site providing murder services, torture services, and kidnapping services that could facilitate the seizure procedure of such onion sites for legal authorities.

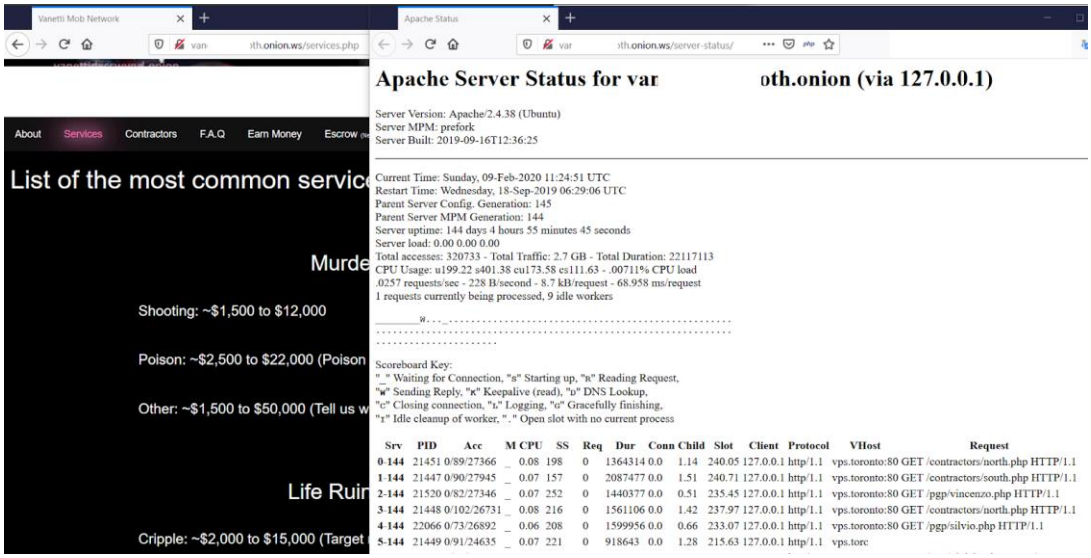

Figure 41: Server info of onion site

6 Usability Study

As the final step regarding all the mentioned above chapters about how using Google dorks would facilitate the cyber security search-related tasks for finding practical solutions for problems that exist in this area, we planned to conduct a study to verify if using Google dorks could be beneficial or not. The purpose of this study is to observe the reaction of a target group of participants mostly students or working in the field of IT, or engrossed in cybersecurity topics attentively, toward the adoption of using Google dorks for their search-related tasks.

6.1 Experimental design

The method we have applied in this study is based on the quasi-experimental design and the model was based on the constructs in the Innovation Diffusion Theory (IDT) model. We believed awareness could affect the adoption of the use of Google dorks. The data was the result of the participation of 32 partakers that actively present in pretest and posttest surveys as the least number of participants in the surveys could be between 30 and 500 participants [101]. The questions in the posttest survey were similar to the questions in the pretest survey. The pretest survey was followed by a treatment that was intended to improve the awareness of participants, a workshop that elucidated some less–known aspects of using Google dorks followed by the posttest survey.

During the workshop, we shed light on some areas Google dorks can be applied for the cybersecurity practitioners. This areas include information gathering, vulnerability detection, credential discovering from paste services, and file-sharing services, locating cybercriminal forums, and services, and communication channels of cybercriminal actors. Nonetheless, not many studies scrutinized elements facilitating the adoption of Google dorks by actors in cybersecurity specifically and regular users of Google search engines generally. This study aims to fill the aforementioned gap and inspect possible factors that have an impact on Google dorks adoption.

However, not many research studies have been conducted for the examination of the factors that have importance in the adoption of Google dorks. We have found that the Innovation Diffusion Theory (IDT), best fits for this study and it includes multiple constructs, the constructs are complexity, trialability, compatibility, observability, and relative advantage [102]. There are numerous former studies conducted on the acceptance of Internet-based technologies and the acceptance of such technologies by end-users [103]. In what follows a concise summary of five attributes proposed by Rogers in the IDT model and their relationship with the acceptance of technology will be explained.

The extent that innovation or technology is relatively hard to use or understand is regarded as complexity [104]. The complexity is distinguished as a factor that has a negative impact on internet usage adoption. The ease of use in the TAM model is the opposite of complexity. We think that users will be reluctant to apply Google dorks for their search-related tasks if they find the Google dorks is time-consuming, exhausting, or needs more mental effort. Consequently, the hypothesis for the complexity construct is that perceived complexity hinders Google dorks adoption.
H1. Perceived complexity has a significant (negative) effect on the Intention to use Google dorks.

The next construct we have used is trialability. Trialability is the degree of feasibility to experiment with the new technology before adoption. It is found that the likelihood of adoption of new technology will soar if the potential adopters have the chance to experiment with new technology [105]. In another study [106], it is found that adopters are more likely to opt for innovation if they are given the chance to try the innovation, as it will mitigate their undetermined fear. Adopters of Google dorks are free to use it on a trial basis, and it is simply possible to compare with typical searching that reduces their fear about using Google dorks that eventually motivates them to apply Google dorks for their search-related tasks.
H2 Perceived trialability has a significant (positive) effect on the Intention to use Google dorks.

Compatibility is the third construct that has been used in our model. Compatibility is considered as the degree that a specific service accordant with the user's existing beliefs, values, habits, and past and current experiences [107]. The key role of compatibility on conformance with the user's lifestyle likely propels a quick rate of adopting innovation [102]. As a result, it can be concluded that in our context there is a significant relationship between adoptions of Google dorks and compatibility.

H3. Perceived compatibility has a significant (positive) effect on the Intention to use Google dorks.

Observability is the next construct of the IDT model. The degree that an innovation is visible to a social systems 'members is described as observability, while simultaneously the benefits are conveniently observable and can be communicated [102]. In the Google dorks context, observability is considered as the degree that users can utilize Google dorks for their search-related tasks at any time, without depending on any specific extra tool or conditions, just a Browser needed, and conveying the advantages to other users. Using this exposure, the users will acquire knowledge about Google dorks benefits for their search-related tasks and more likely adoption to using Google dorks.

H4. Perceived observability has a significant (positive) effect on the Intention to use Google dorks.

The next construct in our model is the relative advantage. Relative advantage introduces the amount of benefit an innovation perceived over its precursor [108]. Increased economic profits, efficiency, and enhanced status are the results of relative advantage [102]. According to numerous former researches adoption of innovation or technology is positively affected by relative advantage [108]. As the user perceives relative advantage or usefulness, the user is more inclined to adopt the new technology or innovation [109]. In Google dorks adoption, the benefits such as more efficient results, less time-consuming, the simple syntax to apply and learn are considered as the advantages. Consequently, we hypothesize that when the users apply the Google dorks for their search-related tasks they enjoy more benefits time-wise and efficiency of the outcomes so it is more likely that they adopt it.

H5. Relative Advantage has a significant (positive) effect on the Intention to use Google dorks.

Google dorks adoption is also considered in our model. Adoption is a decision to apply a new technology or innovation [102]. Innumerable studies describe adoption in terms of usage, utilization, implementation, or satisfaction.

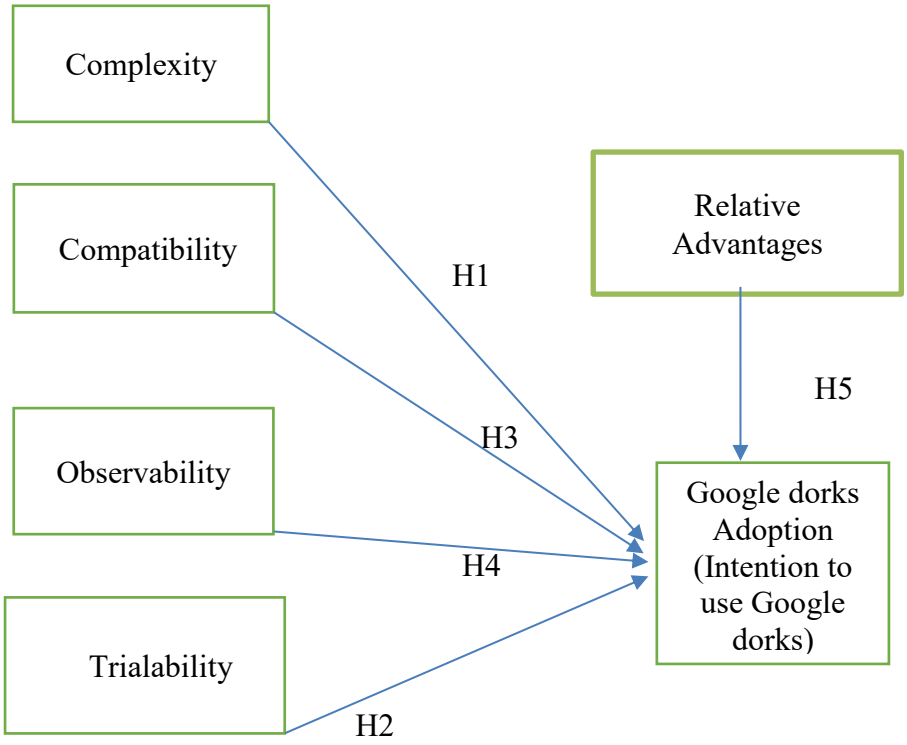

Figure 42: Conceptual model

The proposed conceptual model is based on the Innovation Diffusion Theory model which has revealed its suitability in this study. As it has been explained earlier the Innovation Diffusion Theory (IDT) has these 5 constructs that we will use the acronyms in parenthesis especially in the coming statistical tables: Relative Advantage (RA), Complexity (CX), Compatibility (CO), Observability (OB) and Trialability (TR) [102]. Additionally, we have considered the Workshop as a mechanism that upsurges the awareness of the participants in the posttest survey that is conducted after the workshop. So we consider any improvement in the correlation between the variables and adoption

to the use of Google dorks or Intention to Use (IU) as the result of the increase in the awareness.

6.2 Measures

The whole process of the study included a pretest survey followed by a workshop and a post-test survey after the workshop. We planned to conduct the workshops online due to the exceptional situation that is the result of Covid-19 and gathering people in meetups was forbidden. Overall three online workshops were held for this purpose. Various materials related to the Google dorks were covered in these workshops. The questions were categorized into demographic and general information questions and the other category was the questions for checking hypothesizes. We also had a control question somewhere in the middle of the survey to avoid random responses to the survey.

The distribution method of the questionnaire for the participants was to send a link for the online questionnaire (multiple options exist like Google forms, survey monkey, webropol…). For accomplishing this objective we have used the webropol of our university.

We also used the Likert scale for most of the questions of the survey except for demographic and general questions to measure. The scores were formed according to the Likert 5 point scale, which strongly agree coded as 5 and strongly disagree coded as 1. Overall 19 questions were used for measuring the variables in our IDT model. For the relative advantage construct, 4 questions have been used. Additionally, the 15 other questions were used for compatibility, observability, complexity, trialability, and intention to use with the numbers 4 questions, 2 questions, 4 questions, 3 questions, and 2 questions respectively. The items and the corresponding references can be found in the “appendix”. The items were adapted from literature to enhance content validity [114].

6.3 Research Method

This study is designed according to the quasi-experimental method design [119]. The quasi-experimental for our study includes three major stages. The first stage was a survey that has been provided to measure the constructs of the innovation diffusion theory (IDT) model. For the second stage in this study, we conducted three various online workshops that covered various areas of utilizing Google dorks for search-related tasks and created a YouTube video covering the contents of the workshop and sharing with the participants. Both workshops and videos played the role of treatment in the quasi-experimental design. Eventually, in the last stage, a posttest survey to measure the correlation between the variables of the IDT model with the intention to use Google dorks has been provided and given to the participants of the workshop and YouTube video tutorial. The figure below illustrates these stages.

Figure 43: Research stages

6.4 Result and Discussion

To analyze the gathered data when comparing the pretest and posttest we considered a question for matching the responses of the same respondents in both pretest and posttest. So in case that participants answered the pretest survey but not the posttest survey we disregarded the respondents' pretest by removing their answers from the dataset [111]. Regarding this out of 56 respondents to pretest either familiar or unfamiliar, removing the ones just answered the pretest and not the posttest and the ones that likely answered some of the questions randomly(we had a control question in the surveys for detecting the random responses) only 32 of the surveys were eligible for our analysis.

We have used Statistical Package for the Social Sciences, SPSS version 26 for analysis of the data gleaned for this study. As can be seen from Table 3, out of 32 participants in both pretest and posttest surveys 34.4% of them were female while 56.3% were male and 9.4% preferred not to mention their gender in the surveys. Majority of the participants considered in the age range of 25 to 34 with 65.6%. The other respondents were in the age range of 35 to 44, 18 to 24, and under 18 with the percentages 15.6%, 12.5%, and 6.3% respectively. Regarding the occupation, the majority group of the participants is students with 40.6%. The other groups of the occupation are IT security, Other, Engineering, and IT operation with 25%, 15.6%, 12.5%, and 6.3% respectively. The other factor for demographic statistics that we have considered in this study was the field of study. While 28.1% of the participants were working the rest of the participants were studying in various fields like artificial intelligence, bioscience, social sciences, education, computer engineering. Finally, in terms of the search engine used by the participants, Google is used by 87.5%, and DuckDuckGo is used by 12.5% of the participants.

Table 3: Demographic statistics

Variable		**N**	**Percent**
Gender	Female	11	34.4
	Male	18	56.3
	I prefer not to tell	3	9.4
Age(year)	Under 18	2	6.3
	18 to 24	4	12.5
	25 to 34	21	65.6
	35 to 44	5	15.6
Occupation	IT security	8	25.0
	IT operations	2	6.3
	Engineering	4	12.5
	Student	13	40.6
	Other	5	15.6
	Artificial Intelligence	1	3.1

Field of Study(if student)	Bioscience	1	3.1
	Business Administration	1	3.1
	Chemical engineering	1	3.1
	Chemistry	1	3.1
	Computer engineering	1	3.1
	CS	1	3.1
	Cybersecurity	1	3.1
	Education	2	6.2
	Digitalization	1	3.1
	Information Security	2	6.2
	Information Technology	1	3.1
	IT and Management	1	3.1
	Languages	1	3.1
	Law	1	3.1
	N/A	1	3.1
	NanoMedicine	1	3.1
	Social Sciences	3	6.2
	Software engineering	1	3.1
	Working	9	28.1
Search Engine Used		4	
	DuckDuckGo		12.5
	Google	28	87.5

6.4.1 Scales reliability and validity testing

To check the validity of the scales, assessment of the definition of constructs has been conducted for research findings that are relevant to the constructs in our model as well as reviewing the theories. Items adapted from the existing literature, pilot testing, and pre-testing of the adapted constructs have been other methods for checking the validity of the items. Thus, we believe that the measures have adequate content validity.

The Cronbach‘s alpha was used for testing the reliability of the items of the surveys. The minimum coefficient was 0.325 belongs to the observability construct with 2 items in the surveys while the maximum amount of coefficient for relative advantage. As it is shown

in Table 4 apart from OB (observability) the rest of the variables Cronbach‘s alpha is above 0.6 that indicates the items are reliable.

Table 4: Standard deviation, mean, and Cronbach‘s alpha reliability

Constructs	Mean	Std. Deviation	No of Items(N=20)	Alpha
RA	3.8047	.60487	4	.834
CO	3.3594	.65666	4	.797
OB	3.9063	.54532	2	.325
CX	3.0703	.67571	4	.639
TR	3.7813	.59030	3	.754
IU	3.8333	.57424	2	.808

Note: The scores are formed according to the Likert 5 point scale, that Strongly Agree coded as 5 and Strongly Disagree coded as 1.

6.4.2 Descriptive statistics

Table 5 presents the mean and standard deviation values of constructs in our IDT model for pretest and posttest surveys. A paired sample t-test has been conducted using SPSS to evaluate the statistical significance of the differences between means of the pre-test and post-test. The analysis results indicated a statistically significant difference in terms of mean value between pretest and posttest at the 95% confidence level with p-value=.019 ($p<0.05$).

For the pretest survey, the observability has the highest mean value of 3.9063 while the complexity has the lowest mean value of 3.0703. Other variables namely relative advantage, compatibility, trialability and intention to use have the mean values of 3.8047, 3.3594, 3.7813, and 3.8333 respectively. As we have explained in the research method section the workshop was followed by a posttest survey. The mean value for relative advantage raised from 3.8047 in pretest to 4.0781. Similarly, the compatibility mean value raised from 3.3594 in the pretest survey to 3.5469, the observability mean value raised from 3.9063 to 4.0625, the complexity mean value raised from 3.0703 in pretest to 3.2031 in the posttest, the trialability mean value increased from 3.7813 to 3.9896 and

finally the mean value of intention to use increased from 3.8333 in pretest to 3.9271 in the posttest. Consequently, it can be concluded that treatment (workshop) increased awareness which has a positive effect on other constructs' value and mean values.

Table 5: Comparison between the mean and standard deviation of pretest and posttest surveys (Paired Sample Test)

Posttest-pretest Mean	**Mean**	**Std. Deviation**	**Std. Error Mean**	**t**	**Sig. (2-tailed)**
	.17535	.40080	.07085	2.475	.019

	Pretest survey(N=32)		**Posttest survey(N=32)**				
Constructs(item)	**Mean**	**Std. deviation**	**Mean**	**Std. deviation**	**Mean difference**	**Std. deviation difference**	**Sig. (2-tailed)**
RA(4 items)	3.8047	.60487	4.0781	.45983	.27344	.59010	0.13
CO(3 items)	3.3594	.65666	3.5469	.61381	.18750	.70711	.144
OB(2 items)	3.9063	.54532	4.0625	.47093	.15625	.55992	.125
CX(4 items)	3.0703	.67571	3.2031	.63002	.61232	.61232	.229
TR(3 items)	3.7813	.59030	3.9896	.45187	.57268	.57268	.048
IU(2 items)	3.8333	.57424	3.9271	.69681	.65711	.65711	.426

Note: The scores are formed according to the Likert 5 point scale, that Strongly Agree coded as 5 and Strongly Disagree coded as 1.

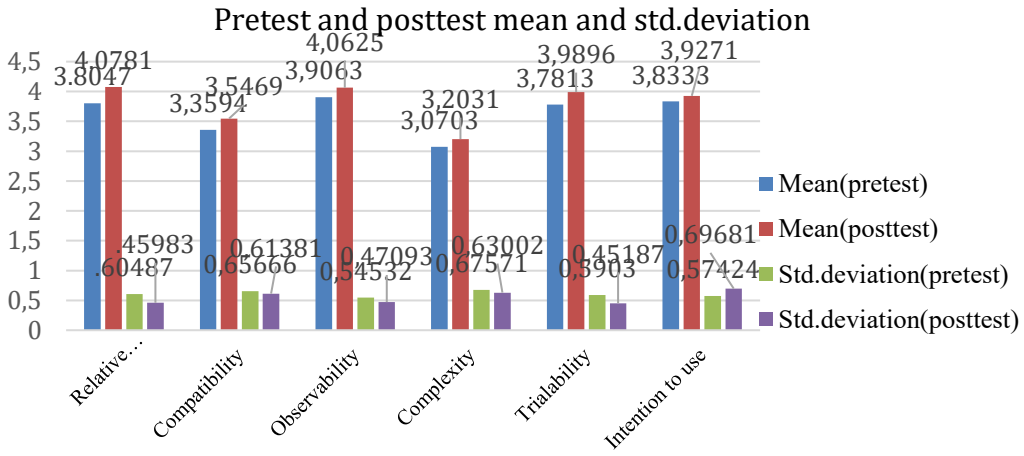

Figure 44: Mean, std.deviation –pretest, posttest comparison

6.4.3 Correlation and Multiple Regression Analysis

We have used the bivariate Pearson Correlation for measuring the direction and strength of the relationship between continuous variables of our model for both pretest and posttest surveys. As we have mentioned formerly we have applied the quasi-experimental design for this study. We have examined that during the pretest phase, 23 of the participants out of 32 that is equal to 71.9% of participants were unfamiliar with Google dorks and the rest 9 participants that is 28.1% were aware of Google dorks. In our study, the role of the workshop of Google dorks, as the treatment, was to increase the awareness so we conclude that if the intention to use Google dorks among the participants surged it was thanks to the elevation of awareness.

As it is shown in Table 6 there is a strong and positive correlation between the relative advantage of using Google dorks and the intention to use Google dorks (r=0.600). The correlation between compatibility and intention to use Google dorks is strong and positive (r=0.592), similarly the correlation between trialability and intention to use is positive and high (r=0.545). Finally, there is a positive and medium correlation between observability and intention to use (r=0.446), and, a positive and weak correlation between complexity and intention to use (r=0.204). Similarly for the posttest survey, as it is shown in Table 7 there is a medium and positive correlation between the relative advantage of using Google dorks and the intention to use Google dorks (r=0.421). The correlation between compatibility and intention to use Google dorks is strong and positive (r=0.511), similarly the correlation between trialability and intention to use is positive and high (r=0.578). Finally, there is a positive and strong correlation between observability and intention to use (r=0.539), and, a positive and weak correlation between complexity and intention to use (r=0.127).

Eventually, we can mention that workshop caused the increase in correlation between intention to use of Google dorks (adoption of Google dorks) and observability and trialability comparing the pretest survey and posttest survey from 0.446 to 0.539 for observability and 0.545 to 0.578 for trialability, while on the contrary relative advantage, compatibility and complexity correlation with the intention to use reduced after the

workshop from 0.600 to 0.592 and from 0.204 to 0.421 and from 0.511 to 0.127 respectively.

Table 6: Inter-correlation of constructs

	RA	CO	OB	CX	TR	IU
RA	1.000					
CO	.751**	1.000				
OB	.737**	.593*	1.000			
CX	.469**	.323	.248	1.000		
TR	.449*	.466**	.686**	.026	1.000	
IU	.600**	.592**	.446*	.204	.545**	1.000

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 7: Inter-correlation of constructs

	RA	CO	OB	CX	TR	IU
RA	1.000					
CO	.537**	1.000				
OB	.666**	.492**	1.000			
CX	.034	.100	.133	1.000		
TR	.327	.419*	.610**	.414*	1.000	
IU	.421*	.511**	.539	.127	.578**	1.000

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

A regression analysis has been conducted to examine the proposed hypotheses in our study using SPSS. We have provided the details of the regression analysis for pretest and post-test separately in Tables 8 and 10 respectively. We took the 5% significance level(2-tailed) and based on that for the pretest survey as it is indicated in Table 8, the p value for the complexity construct is .869 that is higher than .05(p>0.05) and (H1, Beta=-.026,t=-.167, p=.068) so H1 is rejected. The p value for the trialability construct is .022 that is

lower than .05(p<0.05) and (H2, Beta=.457, t=2.493, p=.022) so H2 is accepted. The p value for the compatibility construct is .315 that is higher than .05(p>0.05) and (H3, Beta=.216, t=1.025, p=.315) so H3 is rejected. The p value for the observability construct is .135 that is higher than .05(p>0.05) and (H4, Beta=-.384, t=-1.544, p=.135) so H4 is rejected. The p value for the relative advantage construct is .068 that is higher than .05(p>0.05) and (H5, Beta=.520, t=1.933, p=.068) so H5 is rejected.

Table 8: Multiple Regression (Pretest)

Dependent Variable: IU	**R**	**R Square**	**Adjusted R square**	**F**	**Sig.**
	.722	.521	.429	5.666	.001
	Unstandardized Coefficients(B)	**Std.Error**	**Standardized Coefficients(Beta)**	**t**	**Sig.**
(constant)	3.668	1.925		1.906	.068
RA	.370	.192	.520	1.933	.064
CO	.142	.138	.216	1.025	.315
OB	-.607	.393	-.384	-1.544	.135
CX	-.017	.101	-.026	-.167	.869
TR	.462	.189	.475	2.439	.022

Dependent variable: IU (Intention to Use)

The overall outcome of the multiple regression analysis of the constructs of the pretest can be seen in table 9.

Table 9: Hypothesis test for pretest

Hypothesis		Result
H1	Perceived complexity has a significant (negative) effect on the Intention to use Google dorks.	Not supported
H2	Perceived trialability has a significant (positive) effect on the Intention to use Google dorks.	Supported
H3	Perceived compatibility has a significant (positive) effect on the Intention to use Google dorks.	Not supported
H4	Perceived observability has a significant (positive) effect on the Intention to use Google dorks.	Not supported
H5	Relative Advantage has a significant (positive) effect on the Intention to use Google dorks.	Not supported

Similarly for the posttest survey as it is indicated in Table 10,the p value for the complexity construct is .601 that is higher than .05($p>0.05$) and (H1 , Beta=-.086 ,t=-.530 , p=.601) so H1 is rejected. The p value for the trialability construct is .063 that is higher than .05($p>0.05$) and (H2, Beta=.405, t=1.939, p=.063) so H2 is not accepted. The p value for the compatibility construct is .185 that is higher than .05($p>0.05$) and (H3, Beta=.216, t=1.360, p=.185) so H3 is rejected. The p value for the observability construct is .567 that is higher than .05($p>0.05$) and (H4, Beta=.137, t=.579, p=.567) so H4 is rejected. The p value for the relative advantage construct is .748 that is higher than .05 ($p>0.05$) and (H5, Beta=.068, t=.325, p=.748) so H5 is rejected.

Table 10: Multiple Regression (Post-test)

Dependent Variable: IU	**R**	**R Square**	**Adjusted R square**	**F**	**Sig.**
	.672	.451	.346	4.278	.006
	Unstandardized Coefficients(B)	**Std.Error**	**Standardized Coefficients(Beta)**	**t**	**Sig.**
(constant)	-1.483	3.396		-.437	.666
RA	.077	.239	.068	.325	.748
CO	.210	.246	.216	1.360	.185
OB	.304	.524	.137	.579	.567
CX	-.071	.134	-.086	-.530	.601
TR	.624	.322	.405	1.939	.063

Dependent variable: IU (Intention to Use)

The overall outcome of the multiple regression analysis of the constructs of the posttest can be seen in table 11.

Table 11: Hypothesis test for posttest

Hypothesis		**Result**
H1	Perceived complexity has a significant (negative) effect on the Intention to use Google dorks.	Not supported
H2	Perceived trialability has a significant (positive) effect on the Intention to use Google dorks.	Not supported
H3	Perceived compatibility has a significant (positive) effect on the Intention to use Google dorks.	Not supported
H4	Perceived observability has a significant (positive) effect on the Intention to use Google dorks.	Not supported
H5	Relative Advantage has a significant (positive) effect on the Intention to use Google dorks.	Not supported

6.5 Discussion, Limitation and Future Research

Though considering the already existent literature about the various areas Google dorks have been used by cyber security professionals that we have provided in chapter 2, by conducting the surveys we can assume that apart from the hypothesis H2 that is Perceived trialability has a significant (positive) effect on the Intention to use Google dorks and has been accepted in pretest survey, the rest of the hypotheses in both pretest and post-test have been rejected.

There were limitations for this study like finding the participants for both pretest and posttest surveys as during the time of writing this study the whole world was affected by COVID-19 and many of the meetups (cybersecurity meetups included) was canceled due to the lockdown policies by governments to avoid the spread of the COVID-19, which undoubtedly affected the participants' number. The future works could be lengthier work with more participants in a better time while considering searching for extra variables affecting adoption (intention to use) of Google dorks like facilitating conditions, technical support, or even social influence that needs to be added to the potential future work's research model. As this study was quantitative likely the qualitative study or mixed-method could be other alternatives for future studies and comparing those studies with the current quantitative study would lead us to more accurate results. The experience in the field of cybersecurity, education, and age as moderating variables could also be considered in future studies. Additionally, studies that focus on the black hat hackers or cyber criminals' community and how they find using Google dorks useful could be another area for future studies.

7 Conclusion

The information age brought about access to a large amount of data for us. Connection to the Internet, posting to social media frameworks become an inseparable part of our life. The flip side of this rapidly growing network is locating the desired information. Search engines play a pivotal role in such an object. The most common search engine is Google by far [2] with more than 5.8 billion searches daily [1]. However, the more skilled users of the Internet using advanced searching techniques also known as Google dorks [3]. Cybersecurity professionals either white hat or black hat hackers enjoy the benefits of such techniques for their search-related tasks. By far GHDB is the most popular resource for Google dorks [4]. The GHDB comprises various worthwhile categories that actors in cybersecurity can utilize them for their tasks. We have tried to cover the various areas Google dorks were used and mentioned in the existent literature that can be utilized for numerous purposes ranging from information gathering, finding interesting information about targets by both white hat and black hat actors, finding useful information from pasting and file sharing sites, detecting defaced sites, finding communication channels and forums for cyber criminals. The whole thesis concluded by a usability study of the variables facilitates the adoption of Google dorks for the search-related tasks. The general outcome of the study is that apart from our second hypothesis that is perceived trialability has a significant (positive) effect on the intention to use Google dorks that has been confirmed in our pretest the rest of the hypotheses have been rejected. The constructs for our surveys were based on the innovation diffusion model (IDT), namely compatibility, complexity, relative advantage, observability, trialability, and intention to use.

The general purpose of this thesis was to clarify some areas Google dorks have been already used by cyber security professionals as well as the areas Google dorks can be used by cyber security actors and intensifying its role for them in dealing with their search-related tasks by shedding light on the areas less studied and conducting a study to verify how it has been useful for such actors.

References

[1] “Internet Live Stats - Internet Usage & Social Media Statistics.” https://www.internetlivestats.com/ . [Accessed: 07-Oct-2020].

[2] “The top 500 sites on the web The sites in the top sites lists are ordered by their 1 month Alexa traffic rank. The 1-month rank is calculated using a combination of average daily visitors and pageviews over the past month. The site with the highest combination of visitors and pageviews is ranked #1,” *Alexa*. [Online]. Available: https://www.alexa.com/topsites. [Accessed: 07-Oct-2020].

[3] “Advanced Searching in Google - Resources and Search Strategies,” *Google Sites*. [Online].Available: https://sites.google.com/site/resourcesandsearchstrategies/google/advanced-searching-in-google. [Accessed: 07-Oct-2020].

[4] “Offensive Security's Exploit Database Archive,” *Exploit Database*. [Online]. Available: https://www.exploit-db.com/. [Accessed: 07-Oct-2020].

[5] N. V. Denic, “Government Activities to Detect, Deter and Disrupt Threats Enumerating from the Dark Web,” *DTIC Online*. [Online]. Available: https://apps.dtic.mil/docs/citations/AD1038658. [Accessed: 07-Oct-2020].

[6] A. Hassanzadeh, A. Rasekh, S. Galelli, M. Aghashahi, R. Taormina, A. Ostfeld, and M. K. Banks, “A Review of Cybersecurity Incidents in the Water Sector,” *Journal of Environmental Engineering*, vol. 146, no. 5, p. 03120003, 2020.

[7] J. Runyon and Jeffp, “Google Dorking and Shodan,” *POWERGrid International*, 03-Sep-2019. [Online]. Available: https://www.power-grid.com/2016/11/09/google-dorking-and-shodan/. [Accessed: 07-Oct-2020].

[8] D. Keegan and Braun, “GOVERNMENT AS A PLATFORM: EXPLOITING OPEN GOVERNMENT DATA TO DRIVE PUBLIC SERVICE CO-CREATION,” 2017.

[9] M. S. S. Sadegh, F. Zarafshan, M. Safari, and A. Rahimian, “Optimization of Multi-Agent Security Solution for Prevent Web-Based System of SQL Injection Attack,” p. 16.

[10] H. Martin, “Virus Bulletin :: Search engines in research and vulnerability assessment,” *www.virusbulletin.com*, 01-Nov-2007. [Online]. Available: https://www.virusbulletin.com/virusbulletin/2007/11/search-engines-research-and-vulnerability-assessment. [Accessed: 07-Oct-2020].

[11] K. Foster, “Ten Recommendations for Improving Government (or anyone’s) IT Security,” 2015.

[12] A. Roy, L. Mejia, P. Helling, and A. Olmsted, “Automation of cyber-reconnaissance: A Java-based open-source tool for information gathering,” in *2017 12th International Conference for Internet Technology and Secured Transactions (ICITST)*, Dec. 2017, pp. 424–426, doi: 10.23919/ICITST.2017.8356437.

[13] J. Billig, Y. Danilchenko, and C. E. Frank, “Evaluation of Google hacking,” in *Proceedings of the 5th annual conference on Information security curriculum development*, Kennesaw, Georgia, Sep. 2008, pp. 27–32, doi: 10.1145/1456625.1456634.

[14] J. Zhang, J. Notani, and G. Gu, “Characterizing Google Hacking: A First Large-Scale Quantitative Study,” in *International Conference on Security and Privacy in Communication Networks*, Cham, 2015, pp. 602–622, doi: 10.1007/978-3-319-23829-6_46.

[15] F. Quinkert and R.-U. Bochum, “DorkPot: A Honeypot-based Analysis of Google Dorks,” p. 11.

[16] F. Brown and R. Ragan, “Pulp Google Hacking: The Next Generation Search Engine Hacking Arsenal,” p. 71.

[17] R. Broadhurst *et al.*, “Cyber Terrorism: Research Review: Research Report of the Australian National University Cybercrime Observatory for the Korean Institute of Criminology,” Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2984101, Jun. 2017. doi: 10.2139/ssrn.2984101.

[18] “Mayuranathan, M., Murugan, M., & Dhanakoti, V. A Performance Perspective Analysis: A Detailed Vision on Denial of Service and Distributed Denial of Service on Cloud Computing.”

[19] J. F. P. Disso, K. Jones, P. Williams, and A. Steer, “A distributed attack detection and mitigation framework,” in *2011 IEEE 5th International Conference on Internet Multimedia Systems Architecture and Application*, Dec. 2011, pp. 1–6, doi: 10.1109/IMSAA.2011.6156366.

[20] “Doshi - Notorious ways of exploitation search engine.pdf.” [Accessed: 07-Oct-2020].[Online].Available: https://www.researchgate.net/profile/Sanket_Suthar/publication/319416485_Notorious_ways_of_exploitation_search_engine/links/5b879560a6fdcc5f8b711806/Notorious-ways-of-exploitation-search-engine.pdf.

[21] R. Krimmer, “GOVERNMENT AS A PLATFORM: EXPLOITING OPEN GOVERNMENT DATA TO DRIVE PUBLIC SERVICE CO-CREATION,” p. 72, 2017.

[22] F. Toffalini, M. Abbà, D. Carra, and D. Balzarotti, “Google Dorks: Analysis, Creation, and New Defenses,” in *Detection of Intrusions and Malware, and*

Vulnerability Assessment, Cham, 2016, pp. 255–275, doi: 10.1007/978-3-319-40667-1_13.

[23] R. Pelizzi, T. Tran, and A. Saberi, "Large-Scale, Automatic XSS Detection using Google Dorks," p. 10.

[24] Pentest-Tools.com, "Pentest-Tools.com | Powerful Pentesting Tools, Easy to Use," *Pentest-Tools.com*. https://pentest-tools.com/home. [Accessed: 07-Oct-2020].

[25] "Bae, M. Y., Lim, H. K., & Cho, D. J. (2016). A study on security diagnosis using automated Google hacking tools-focusing on the US government website. Journal of Advances in Information Technology, 7(2), 93-97."

[26] "Swart, I. P. (2015). Pro-active visualization of cybersecurity on a National Level: A South African Case Study (Doctoral dissertation, Rhodes University)."

[27] "Offensive Security's Exploit Database Archive." https://www.exploit-db.com. [Online]. Available: https://www.exploit-db.com. [Accessed: 07-Oct-2020].

[28] "Shodan," *www.shodan.io*. [Online]. Available: https://www.shodan.io.

[Accessed: 07-Oct-2020].

[29] "Search Engine Hacking – Manual and Automation," *Infosec Resources*, May 01, 2013. https://resources.infosecinstitute.com/search-engine-hacking-manual-and-automation/.[Online]. Available: https://resources.infosecinstitute.com/search-engine-hacking-manual-and-automation/. [Accessed: 07-Oct-2020].

[30] C. Arthur, "LulzSec: what they did, who they were and how they were caught," *The Guardian*, May 16, 2013.

[31] "Google Hacking 101." https://www.oakton.edu. [Online]. Available: https://www.oakton.edu/user/2/rjtaylor/cis101/Google%20Hacking%20101.pdf. [Accessed: 07-Oct-2020].

[32] "Google Hacking and Defense Cheat Sheet." https://www.sans.org. [Online]. Available: https://www.sans.org/security-resources/GoogleCheatSheet.pdf. [Accessed: 07-Oct-2020].

[33] "List of Top-Level Domains - ICANN." https://www.icann.org/resources/pages/tlds-2012-02-25-en .[Accessed: 07-Oct-2020].

[34] “Introduction to the server-side - Learn web development | MDN.” https://developer.mozilla.org/en-US/docs/Learn/Server-side/First_steps/Introduction [Accessed: 07-Oct-2020].

[35] “About DuckDuckGo.” https://duckduckgo.com/about . [Accessed: 07-Oct-2020].

[36] DuckDuckGo, “DuckDuckGo Search Syntax,” *DuckDuckGo Help Pages*. https://help.duckduckgo.com/duckduckgo-help-pages/results/syntax/. [Accessed: 07-Oct-2020].

[37] “Advanced search keywords.” https://help.bing.microsoft.com/#apex/18/en-us/10001/0 . [Accessed: 07-Oct-2020].

[38] “Online SEO Guide: Google Stop Words - A Comprehensive List Of Words Google Ignores.” http://www.link-assistant.com/seo-stop-words.html . [Accessed: 07-Oct-2020].

[39] “Google Dorks: An Easy Way of Hacking | Cybrary.” https://www.cybrary.it/blog/0p3n/google-dorks-easy-way-of-hacking/. [Accessed: 07-Oct-2020].

[40] “Security Trails | Exploring Google Hacking Techniques.” https://securitytrails.com/blog/google-hacking-techniques. [Accessed: 07-Oct-2020].

[41] B. K. K. Roopkumar, “Ethical Hacking Using Penetration Testing,” p. 105.

[42] F. Breda, H. Barbosa, and T. Morais, “SOCIAL ENGINEERING AND CYBERSECURITY,” Valencia, Spain, Mar. 2017, pp. 4204–4211, doi: 10.21125/inted.2017.1008.

[43] “Symfony databases.yml configuration file - Vulnerabilities - Acunetix.” https://www.acunetix.com/vulnerabilities/web/symfony-databases-yml-configuration-file/ .[Accessed: 07-Oct-2020].

[44] “THC-Hydra.” https://tools.kali.org/password-attacks/hydra. [Accessed: 07-Oct-2020].

[45] “Our Most Advanced Penetration Testing Distribution, Ever.” https://www.kali.org/. [Accessed: 07-Oct-2020].

[46] “VBulletin 5 Connect, The World’s Leading Community Software.” https://www.vbulletin.com/. [Accessed: 07-Oct-2020].

[47] “All About Carding (For Noobs Only) [Updated 2019].” https://resources.infosecinstitute.com/all-about-carding-for-noobs-only/#gref. [Accessed: 07-Oct-2020].

[48] “MapR 5.0 Documentation : db.conf.” https://mapr.com/docs/archive/mapr50/db.conf_26982888.html. [Accessed: 07-Oct-2020].

[49] “Why you should never reuse passwords | Kaspersky official blog.” https://www.kaspersky.com/blog/never-reuse-passwords-story/24808/. [Accessed: 07-Oct-2020].

[50] “The Web Robots Pages.” https://www.robotstxt.org/. [Accessed: 07-Oct-2020].

[51] “The Web Robots Pages.” https://www.robotstxt.org/db/googlebot.html. [Accessed: 07-Oct-2020].

[52] “Learn about sitemaps - Search Console Help.” https://support.google.com/webmasters/answer/156184?hl=en&ref_topic=4581190 .[Accessed: 07-Oct-2020].

[53] “Have I Been Pwned: Pastes.” https://haveibeenpwned.com/Pastes .[Accessed: 07-Oct-2020].

[54] “PasteMonitor.” https://www.pastemonitor.com/ .[Accessed: 07-Oct-2020].

[55] “W3Techs - extensive and reliable web technology surveys.” https://w3techs.com/. [Accessed: 07-Oct-2020].

[56] “Shape Security Blog : A look at Sentry MBA – the most popular cybercriminal tool for credential stuffing attacks,” *Shape Security Blog*, Mar. 09, 2016. https://blog.shapesecurity.com/2016/03/09/a-look-at-sentry-mba/.[Accessed: 07-Oct-2020].

[57] “Dropbox employee’s password reuse led to the theft of 60M+ user credentials,” *TechCrunch*. https://social.techcrunch.com/2016/08/30/dropbox-employees-password-reuse-led-to-theft-of-60m-user-credentials/. [Accessed: 07-Oct-2020].

[58] “How to Secure Your Accounts With Better Two-Factor Authentication | WIRED.” https://www.wired.com/story/two-factor-authentication-apps-authy-google-authenticator/ .[Accessed: 07-Oct-2020].

[59] “SQL Injection | OWASP.” https://owasp.org/www-community/attacks/SQL_Injection .[Accessed: 07-Oct-2020].

[60] “Cross-Site Scripting (XSS) Software Attack | OWASP Foundation.” https://owasp.org/www-community/attacks/xss/ .[Accessed: 07-Oct-2020].

[61] Ninja Hatori, “Example of an Error-Based SQL Injection,” *Medium*, May 22, 2019. https://medium.com/@hninja049/example-of-a-error-based-sql-injection-dce72530271c. [Accessed: 07-Oct-2020].

[62] “Web Application Firewall | OWASP.” https://owasp.org/www-community/Web_Application_Firewall .[Accessed: 07-Oct-2020].

[63] “403 Forbidden errors when working on your website? Firewalls, firewalls, firewalls,” *Websavers*, Feb. 13, 2017. https://websavers.ca/unexpected-403-errors-working-website .[Accessed: 07-Oct-2020].

[64] “How to fix ‘HTTP Error 403’ in WordPress - Press Customizr Documentation.” https://docs.presscustomizr.com/article/195-http-error-403. [Accessed: 07-Oct-2020].

[65] “Website is not available with enabled ModSecurity: 403 Forbidden,” *Plesk Help Center*. http://support.plesk.com/hc/en-us/articles/213378669. [Accessed: 07-Oct-2020].

[66] “PHP: phpinfo - Manual.” https://www.php.net/manual/en/function.phpinfo.php .[Accessed: 07-Oct-2020].

[67] “Usage Statistics and Market Share of Content Management Systems, June 2020.” https://w3techs.com/technologies/overview/content_management [Accessed: 07-Oct-2020].

[68] “Detect which CMS a site is using - What CMS?” https://whatcms.org/ [Accessed: 07-Oct-2020].

[69] “Wappalyzer – Get this Extension for Firefox (en-US).” https://addons.mozilla.org/en-US/firefox/addon/wappalyzer/ [Accessed: 07-Oct-2020].

[70] “How to Remove the Powered by WordPress Footer Links,” *WPBeginner*, Oct. 06, 2016. https://www.wpbeginner.com/wp-themes/how-to-remove-the-powered-by-wordpress-footer-links/ [Accessed: 07-Oct-2020].

[71] “WordPress Files and Directory Structure,” *Interserver Tips*. https://www.interserver.net/tips/kb/wordpress-files-directory-structure/ [Accessed: 07-Oct-2020].

[72] R. Dewhurst, “WordPress Plugin SEO by Yoast 1.7.3.3 - Blind SQL Injection,” *Exploit Database*, Mar. 16, 2015. https://www.exploit-db.com/exploits/36413 [Accessed: 07-Oct-2020].

[73] “Compromised Web Servers and Web Shells - Threat Awareness and Guidance | CISA.” https://www.us-cert.gov/ncas/alerts/TA15-314A [Accessed: 07-Oct-2020].

[74] “Usage Statistics and Market Share of VBulletin, June 2020.” https://w3techs.com/technologies/details/cm-vbulletin [Accessed: 07-Oct-2020].

[75] “Usage Statistics and Market Share of MyBB, June 2020.” https://w3techs.com/technologies/details/cm-mybb [Accessed: 07-Oct-2020].

[76] “vBulletin 5.x < 5.5.4 Patch Level 1 Remote Code Execution Vulnerability.” https://www.tenable.com/plugins/was/98764 [Accessed: 07-Oct-2020].

[77] “Vbulletin : Security vulnerabilities.” https://www.cvedetails.com/vulnerability-list/vendor_id-8142/Vbulletin.html [Accessed: 07-Oct-2020].

[78] J. Long, B. Gardner, and J. Brown, *Google Hacking for Penetration Testers*. Elsevier, 2011.

[79] “Manual: IP/Hotspot - MikroTik Wiki.” https://wiki.mikrotik.com/wiki/Manual:IP/Hotspot [Accessed: 07-Oct-2020].

[80] “Zabbix: The Enterprise-Class Open Source Network Monitoring Solution.” https://www.zabbix.com/ [Accessed: 07-Oct-2020].

[81] “IP Phones, VOIP Phones,” *Cisco*. https://www.cisco.com/c/en/us/products/collaboration-endpoints/ip-phones/index.html [Accessed: 07-Oct-2020].

[82] “Marshall Electronics - VS-103E-3GSDI, 1080p60 Full HD Video Encoder with Embedded Audio. Encodes 2 channels of Embedded Audio from HDMI or HDSDI input.” http://www.marshall-usa.com/hardware-accessories/encoders-decoders/VS-103E-HDSDI.php [Accessed: 07-Oct-2020].

[83] Opsdisk, *opsdisk/pagodo*. 2020.

[84] “Onion.Ws Tor2Web Gateway.” https://onion.ws/ [Accessed: 07-Oct-2020].

[85] L. Marinos, “ENISA Threat Landscape 2016,” p. 11.

[86] “Website Defacement - Definition - Trend Micro USA.” https://www.trendmicro.com/vinfo/us/security/definition/website-defacement [Accessed: 07-Oct-2020].

[87] “Zone-H.org - Unrestricted information.” http://zone-h.org/?hz=1 [Accessed: 07-Oct-2020].

[88] “Mirror-h.org - Hack World Analyse and Attack Shadow System.” https://mirror-h.org/ [Accessed: 07-Oct-2020].

[89] “Defacer.ID | Global Defacements & Cyber Vandalism Mirror Database,” *Defacer.ID*. https://www.defacer.id [Accessed: 07-Oct-2020].

[90] A. Haslebacher, J. Onaolapo, and G. Stringhini, “All your cards are belong to us: Understanding online carding forums,” in *2017 APWG Symposium on Electronic Crime Research (eCrime)*, Apr. 2017, pp. 41–51, doi: 10.1109/ECRIME.2017.7945053.

[91] A. Deb, K. Lerman, and E. Ferrara, “Predicting Cyber Events by Leveraging Hacker Sentiment,” *Information*, vol. 9, no. 11, p. 280, Nov. 2018, doi: 10.3390/info9110280.

[92] “Telegram – a new era of messaging,” *Telegram*. https://telegram.org/ [Accessed: 07-Oct-2020].

[93] “North Korean hackers use Telegram to steal cryptocurrencies,” *TechBriefly*, Jan. 09, 2020. https://techbriefly.com/2020/01/09/north-korean-hackers-use-telegram-to-steal-cryptocurrencies/ [Accessed: 07-Oct-2020].

[94] C. K. P. Romano, “Government Activities to Detect, Deter and Disrupt Threats Enumerating from the Dark Web,” p. 90.

[95] “Vk.com Competitive Analysis, Marketing Mix, and Traffic - Alexa.” https://www.alexa.com/siteinfo/vk.com [Accessed: 07-Oct-2020].

[96] “Facebook.com Competitive Analysis, Marketing Mix, and Traffic - Alexa.” https://www.alexa.com/siteinfo/facebook.com [Accessed: 07-Oct-2020].

[97] D. Guccione, “What is the dark web? How to access it and what you’ll find,” *CSO Online*, Mar. 05, 2020. https://www.csoonline.com/article/3249765/what-is-the-dark-web-how-to-access-it-and-what-youll-find.html [Accessed: 07-Oct-2020].

[98] “The Tor Project | Privacy & Freedom Online.” https://torproject.org [Accessed: 07-Oct-2020].

[99] P. Winter, A. Edmundson, L. M. Roberts, A. Dutkowska-Żuk, M. Chetty, and N. Feamster, “How Do Tor Users Interact With Onion Services?,” 2018, pp. 411–428, [Accessed: 07-Oct-2020]. [Online]. Available: https://www.usenix.org/conference/usenixsecurity18/presentation/winter.

[100] “Bromium-WoP-Behind-the-Dark-Net-Black-Mirror.pdf.” [Accessed: 07-Oct-2020]. [Online]. Available: https://www.bromium.com/wp-content/uploads/2019/06/Bromium-WoP-Behind-the-Dark-Net-Black-Mirror.pdf.

[101] “Sekaran, U., & Bougie, R." Research methods for business: A skill-building approach." John Wiley & Sons, 2016.”

[102] E. M. Rogers, *Diffusion of Innovations, 4th Edition*. Simon and Schuster, 2010.

[103] L. R. Vijayasarathy, “Predicting consumer intentions to use on-line shopping: the case for an augmented technology acceptance model,” *Information & Management*, vol. 41, no. 6, pp. 747–762, Jul. 2004, doi: 10.1016/j.im.2003.08.011.

[104] W. Cheung, M. K. Chang, and V. S. Lai, “Prediction of the Internet and World Wide Web usage at work: a test of an extended Triandis model,” *Decision Support Systems*, vol. 30, no. 1, pp. 83–100, Dec. 2000, doi: 10.1016/S0167-9236(00)00125-1.

[105] R. Agarwal and J. Prasad, “A Conceptual and Operational Definition of Personal Innovativeness in the Domain of Information Technology,” *Information Systems Research*, vol. 9, no. 2, pp. 204–215, Jun. 1998, doi: 10.1287/isre.9.2.204.

[106] M. Tan and T. S. H. Teo, “Factors Influencing the Adoption of Internet Banking,” *Journal of the Association for Information Systems*, vol. 1, no. 1, Jul. 2000, doi: 10.17705/1jais.00005.

[107] L. Chen, “A model of consumer acceptance of mobile payment,” *International Journal of Mobile Communications*, vol. 6, no. 1, pp. 32–52, Jan. 2008, doi: 10.1504/IJMC.2008.015997.

[108] G. C. Moore and I. Benbasat, “Development of an Instrument to Measure the Perceptions of Adopting an Information Technology Innovation,” *Information Systems Research*, vol. 2, no. 3, pp. 192–222, Sep. 1991, doi: 10.1287/isre.2.3.192.

[109] D. W. McCloskey, “The Importance of Ease of Use, Usefulness, and Trust to Online Consumers: An Examination of the Technology Acceptance Model with Older Customers,” *Journal of Organizational and End User Computing (JOEUC)*, Jul. 01, 2006. www.igi-global.com/article/importance-ease-use-usefulness-trust/3814 [Accessed: 07-Oct-2020].

[110] B. Howcroft, R. Hamilton, and P. Hewer, “Consumer attitude and the usage and adoption of home-based banking in the United Kingdom,” *International Journal of Bank Marketing*, vol. 20, no. 3, pp. 111–121, Jan. 2002, doi: 10.1108/02652320210424205.

[111] “Comparison of intent-to-treat analysis strategies for pre-post studies with the loss to follow-up | Elsevier Enhanced Reader.” https://reader.elsevier.com/reader/sd/pii/S2451865417301941?token=894833E15D33A2771967B3B50307170104E94EB8DCF89DD1F959A87A91929E32ADF8A85E3FA4977CB1A1A7798626BCD8 [Accessed: 07-Oct-2020].

[112] Straub D, Boudreau M-C, Gefen D (2004) Validation guidelines for IS positivist research. Communications of the Association for Information Systems 13:380–427

[113] Tan, M. and Teo, T., 2000. Factors Influencing the Adoption of Internet Banking. *Journal of the Association for Information Systems*, 1(1), pp.1-44.

[114] Atkinson, N., 2007. Developing a Questionnaire to Measure Perceived Attributes of eHealth Innovations. *American Journal of Health Behavior*, 31(6), pp.612-621.

[115] H.-F. Lin, “An empirical investigation of mobile banking adoption: The effect of innovation attributes and knowledge-based trust,” *International Journal of Information Management*, vol. 31, no. 3, pp. 252–260, 2011.

[116] T. Laukkanen and P. Cruz, “Comparing Consumer Resistance to Mobile Banking in Finland and Portugal,” *Communications in Computer and Information Science e-Business and Telecommunications*, pp. 89–98, 2009.

[117] D. Jeffrey, “Testing the Technology Acceptance Model 3 (TAM 3) with the Inclusion of Change Fatigue and Overload, in the Context of Faculty from Seventh-day Adventist Universities: A Revised Model.”

[118] Y.-Y. Wang, Y.-S. Wang, and S.-E. Jian, “Investigating the Determinants of Students’ Intention to Use Business Simulation Games,” *Journal of Educational Computing Research*, vol. 58, no. 2, pp. 433–458, 2019.

[119] B. A. Thyer, “Interrupted Time Series Designs,” *Quasi-Experimental Research Designs*, pp. 107–125, 2012.

Appendix: Measurement items

(1) Relative Advantage :

- RA1: Google dorks might be a convenient way to find desired search information. (adapted from Tan & Teo 2000[113])
- RA2: Google dorks may improve the efficiency of searching. (adapted from Tan & Teo 2000[113])
- RA3: Google dorks are likely to give greater control over searching. (adapted from Tan & Teo 2000[113])
- RA4: It is better to use Google Dorks rather than other methods for searching for information. (adapted from Atkinson, 2007 [114])

(2) Compatibility :

- CO1: Google dorks highly fit with the way I do internet searching. (adapted from Lin 2011 [115])
- CO2: Google dorks seems to fit with the way I do internet searching. (adapted from Lin 2011 [115])
- CO3: I think everyone should use Google dorks for searching. (adapted from Atkinson 2007 [114])
- CO4: My colleagues might be interested in Google dorks when they see me using it. (adapted from Atkinson 2007 [114])

(3) Observability (adapted from Atkinson, 2007 [114]):

- OB1: I might tell my friends about the benefits of using Google dorks.
- OB2: Skilled cyber security practitioners probably like using Google dorks.

(4) Complexity :

- CX1: I think using Google dorks requires a lot of mental effort. (adapted from Atkinson, 2007 [114])
- CX2: Use of Google dorks may require technical skills. (adapted from Laukkanen & Cruz 2009 [116])

- CX3: Using Google dorks might be frustrating. (adapted from Tan & Teo 2000 [113])
- CX4: Google dorks look clear and understandable for me. (adapted from Jeffrey 2015 [117])

(5) Trialability :

- TR1: I might use Google dorks on a trial basis to see what it can do for me. (adapted from Tan & Teo 2000[113])
- TR2: I might use Google dorks to try them out for my task. (adapted from Atkinson, 2007 [114])
- TR3: I won't lose much by trying Google dorks, even if I don't like it. (adapted from Atkinson, 2007 [114])

(6) Intention to use (adapted from Wang et al., 2020 [118]):

- IU1: I will likely use Google dorks in the future.
- IU2: I might use Google dorks for my searching tasks.