Analysis and Overview of Information Gathering & Tools for Pentesting

Alekya Sai Laxmi Kowta, Karan Bhowmick, Jeev Ratan Kaur, N. Jeyanthi
School of Information Technology and Engineering
Vellore of Institute of Technology, Vellore, Tamilnadu, India
Alekyasai.laxmikowta2018@vitstudent.ac.in; njeyanthi@vit.ac.in

2021 International Conference on Computer Communication and Informatics (ICCCI) | 978-1-7281-5875-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCI50826.2021.9457015

Abstract— **Information Discovery is something that involves identifying and finding sensitive or regulated data to adequately protect or securely remove. It involves auditing sensitive/ remote information. In today's era of remote workers, data spread across several systems, applications and databases necessitates authentication, a challenging task. Data discovery enables us to be aware of our weaknesses and promotes context aware security solutions. Objectives of someone discovering information would be to find out the network data like public or private and associated domain names, whatever UDP and TCP services they're currently running, Information regarding the SSL certificates or open ports and more. Collecting system related information – user enumeration, system groups or OS hostnames, OS system types etc., would prove to be very useful. In this paper, we present how information gathering is very essential for a cyber-analyst in the security point of view and we will look into the tools used for information gathering and discovery. One would be surprised when they know the number of tools that are available online for free that will help uncover such information. While these tools can be very useful if used in the right way, they can also be very dreading because of the vast information gathering properties they possess. We will describe how a simple Google search engine can shed light on a lot of vulnerabilities. Different tools that are used for e-discovery and their workings are explained here. With special emphasis on Google Dorking and how it is a very valuable resource for analysts, we show indexing and the structure of the internet. We will talk about how OSINT is crucial and how information gathering is used in social engineering. We discuss the type of attacks and the defense mechanisms. We will then talk about how intrusion detection systems are coupled with OSINT for protection of data. Then we proceed by discussing how social engineering is very related to information gathering and then discuss defense mechanisms.**

Keywords- OSINT Tools; Final Recon; Photon; Sherlock; Dorking; Whois Database; CVEs; Analyst; Modules; Marketplaces; Docker; Crawling; Indexing;

I. INTRODUCTION

The main goal of this paper is to emphasize on the different ways information gathering can take place. Good information gathering can help in successfully carrying out a pentest and provides maximum benefit to the client. To give an overview of cyber reconnaissance, we will list out the steps a penetration tester would take to gather as much information as possible. This includes trying to get the registration details of the website, or contact details of a certain person, trying to gain information through email harvesting and finding out the IP addresses and determining the network ranges. Some steps include trying to find out the host machine, the DNS record, OS type of the machine and subdomains that they're registered to.. The pen testing process starts with getting acquainted with the client. It's a process where one tries to know whatever they can about the target. This includes trying to know almost all the public information that is available about the target. One can use a lot of tools like port scanners or other tools available online like the ones we describe in this paper. This is the first step of a pen test. After this, one can do a thorough vulnerability analysis. Through this one can manage to either attack the target by using some exploits or try to use this information to make their target/ own system much more secure than it is. The more information you have about your target, the more probable it is for you to perform an attack. Gathering takes place in two forms – passive and active gathering. One must realize how crucial information gathering is and should be well aware of the various ways that your information can be gathered and used. There are a lot of tools, techniques and different sites that do the same for you. Public sites like whois, dns, google hacking helps attackers gather information. This is exactly why we will be showing you how easy it is for one to gather information about anyone through various tools.

II. LITERATURE SURVEY

Clive (2012) [1] has presented OSINT, The Internet and Privacy where they explained about how the web was always intended to be a two-way multi-user system which uses state controlled one-way broadcasting. Kariya and Kher [2] (2016) have social engineering targets the emotional parts of humans to gain access to controlled areas to achieve sensitive/protected information from the users for different purposes. Open-Source Intelligence Base Cyber Threat Inspection Framework for Critical Infrastructures are using OSINT process Lee and Shon [3] (2016).

Vacas et al., [4] (2018) presented Detecting Network Threats using OSINT Knowledge- based IDS, a fully automated approach to update the IDS knowledge. They implemented the same as the IDSOSINT system by accessing 49 OSINT feeds and production traffic. Real-time malicious activities could be identified. However, with continuous sophisticated attacks out there, it would be very difficult to keep the IDS updated.

978-1-7281-5875-4/21/$31.00 ©2021 IEEE

Osamah et al., [5] (2018), Enterprises are obligated to use a variety of tools to detect and try to cover up most of the cybersecurity aspects that they wish to protect. Social engineering, Kevin et al., [6] (2017), is the attack that aimed to manipulate the user to get sensitive information or take actions to help the adversary bypass the privileges so as to complete the attacking goals. Stage theory demonstrates Alain [7] (2018) how information security analysts can utilize the various stages in phishing. It is suggested that recommendations against phishing should target individuals based on their resident stages. Moreover, the processes of change should be applied to the correct stage for the recommendations to be effective.

The human computer interaction involves a digital system wherein it acts as a team member in the discovery of information and augmenting the user‗s cognitive processes, Kerne and Smith [8] (2004). Social engineering can be classified into two types – technology-based deception and human based deception, Comia et al., [9] (2017). Spear phishing emails are usually victim-specific and targeted, Prateek et al., [10] (2014). Because it is targeted, it looks much more realistic, which makes it harder to detect. Additional AI and ML techniques could be used to make the analysis better.

III. Architecture and Design

Ever wonder why you include meta tags in the html pages you build? Metadata web indexing involves giving a particular keyword or phrase to web pages or web sites within a meta-tag. This then helps retrieval of a website from the search engine as shown in Fig.1.

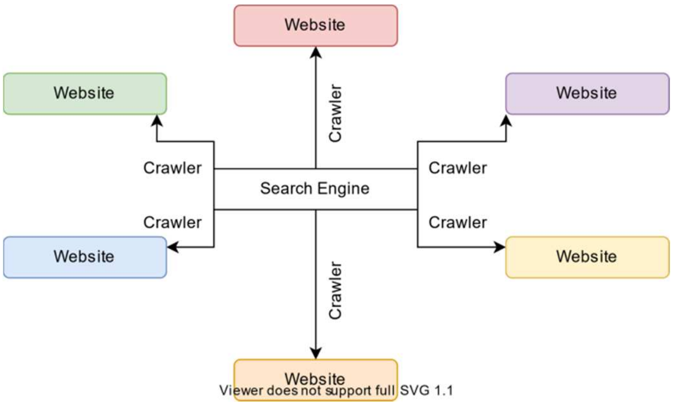


Figure 1. How Search Engines Crawl your Website

Usually search engine indexing works this way. The engine is customized to search the keywords list. Before one performs a search, web crawlers already gather all this information from across billions of sites and pages and organize it in the search index. So once google finds a page, Google tries to understand the content of the page by analyzing it. It catalogs images, video files in a page. This information is stored in a Google-index which is controlled by many computers. Search engines usually have three functions- crawling, indexing and ranking as in Fig.2. At the end of the day, each crawler‗s job is to learn as much as possible about your website.

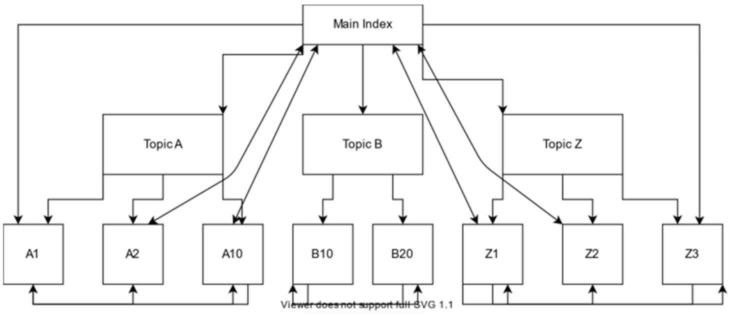


Figure 2. Internal Linking

IV. Overview of Reconnaissance

As already stated, the goal of a hacker or a security analyst would try to gather as much information as possible and this would start with scouring all the exposed ports and running network services. They could collect the DNS names, and IP addresses, remote access capabilities. From this, one could get an idea of all the unpatched vulnerabilities and CVE‗s in operating systems. For websites, usually the first technical information one would gather is the tracking services or their hosting service/ provider. One can try to get the programming language of the website or various widgets of the website. One can also gain insight to the JavaScript functions and libraries and try to get access to the types of web servers that are being run. For the very same one can use tools like Recon-ng and the built-with module in it. Apart from this there are a large variety of tools that perform the initial stage of pentesting.

After one is armed with this information, we can go to the vulnerabilities databases that are available online. These have all the security vulnerabilities listed. Most of the times various exploits associated with each technology are also available online. If the hacker is a professional, they can also make their own exploits. After this is done, one could also use various web crawling techniques to find out what other sites are linked to the main site. Most of the times linked sites which are put out of use can prove to be very useful. Google Reverse Analytics could also be used for the same.

Different kinds of vendor risk assessments can be performed to reveal all the subdomains of the target domain. This would lead us to forgotten domains that could be left unprotected. Because of this, a lot of information regarding services that are run by the subdomain such as VPN, FTP, Webmail etc., could be known. For the very same one can use tools like Recon-ng and the built-with module in it. Apart from this there are a large variety of tools that perform the initial stage of pentesting. Google also can be used to search for sub domains. The ―site‖ operator does exactly this. Advanced Google search operators can be used to find out a lot of sensitive and restricted files online.

Google Hacking: Google ―Dorking‖ is the practice of using Google to find vulnerable web applications and servers by using native Google search engine capabilities. We will discuss about how to do the same with various operators and then give a detailed explanation of how to prevent your websites/ information from being dorked.

Domain Names: Who Is Database Information gathering. These are registered by organizations, governments, public and private agencies and people. DNS or Domain Name System is a system which connects URL with their IP Addresses. Without going into details can think of DNS like a smartphone contact list, which matches the peoples name and their phones numbers and email addresses.

Internet Servers: Authoritative DNS servers are a great source of information, as they often include every single surface point exposed to the Internet—which means a direct link to related services such as HTTP, email, etc. Some tools for information gathering include nmap, sherlock, search4, th3inspector, sublist3r etc. We will give a detailed analysis of how some of these tools work and analyze their functionalities.

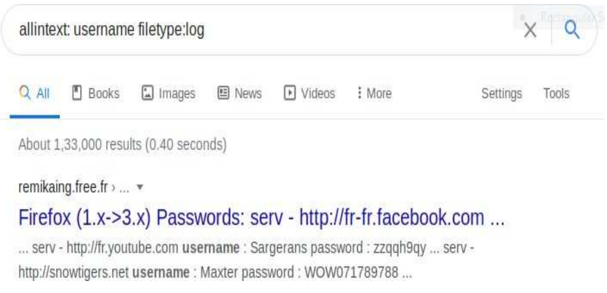
V. OSINT TOOLS

A. Google Dorking:

Google dorks are very simple to use and the amount of information they can get you is very amusing. Search engines index a lot of information – almost anything on the internet, including individual companies and their data. Filetype is used for finding any kind of filetypes. Intext helps you perform queries to search for specific text. Intitle searches for certain keywords in the title. Logs aren‘t supposed to be indexed by search engines; however, all of this is done and there‘s nothing that you can‘t find on the internet. It is a technique used by investigators, organizations to query for various hidden information in public websites and servers. Dorking/ Google-Hacking is a way of using search engines to their full capacity to penetrate web-based services that are not really visible right away.

Some examples include:

- ➢ allintext:username filetype:log



allintext: username filetype:log

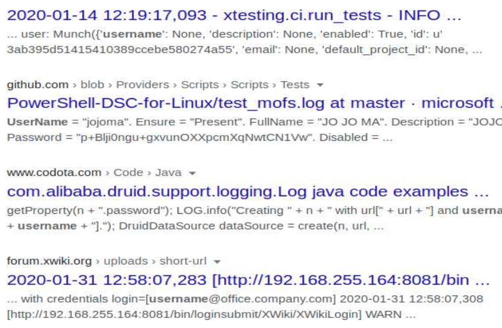
About 1,33,000 results (0.40 seconds)

remikaing.free.fr › ...

Firefox (1.x->3.x) Passwords: serv - http://fr-fr.facebook.com ...

... serv - http://fr.youtube.com **username** : Sargerans password : zzqqh9qy ... serv - http://snowtigers.net **username** : Maxter password : WOW071789788 ...

Figure 3. Filetype command



2020-01-14 12:19:17,093 - xtesting.ci.run_tests - INFO ...

... user: Munch({'**username**': None, 'description': None, 'enabled': True, 'id': u' 3ab395d51415410389ccebe580274a55', 'email': None, 'default_project_id': None, ...

github.com › blob › Providers › Scripts › Scripts › Tests

PowerShell-DSC-for-Linux/test_mofs.log at master · microsoft ...

UserName = "jojoma". Ensure = "Present". FullName = "JO JO MA". Description = "JOJOMA". Password = "p+Blji0ngu+gxvunOXXpcmXqNwtCN1Vw". Disabled = ...

www.codota.com › Code › Java

com.alibaba.druid.support.logging.Log java code examples ...

getProperty(n + ".password"); LOG.info("Creating " + n + " with url[" + url + "] and **username**[" + **username** + "]."); DruidDataSource dataSource = create(n, url, ...

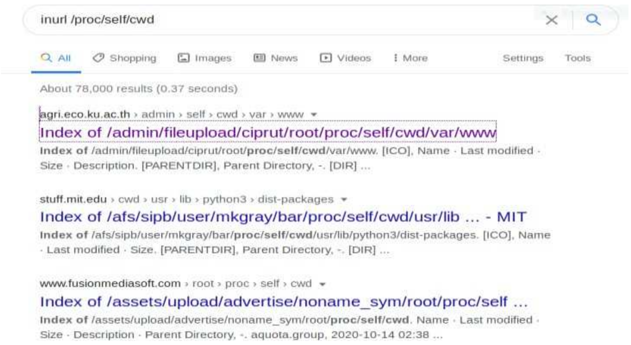
forum.xwiki.org › uploads › short-url

2020-01-31 12:58:07,283 [http://192.168.255.164:8081/bin ...

... with credentials login=[**username**@office.company.com] 2020-01-31 12:58:07,308 [http://192.168.255.164:8081/bin/loginsubmit/XWiki/XWikiLogin] WARN ...

Figure 4. Filetype command results

- ➢ inurl: /proc/self/cwd



inurl /proc/self/cwd

About 78,000 results (0.37 seconds)

agri.eco.ku.ac.th › admin › self › cwd › var › www

Index of /admin/fileupload/ciprut/root/proc/self/cwd/var/www

Index of /admin/fileupload/ciprut/root/**proc/self/cwd**/var/www. [ICO], Name · Last modified · Size · Description. [PARENTDIR], Parent Directory, -. [DIR] ...

stuff.mit.edu › cwd › usr › lib › python3 › dist-packages

Index of /afs/sipb/user/mkgray/bar/proc/self/cwd/usr/lib ... - MIT

Index of /afs/sipb/user/mkgray/bar/**proc/self/cwd**/usr/lib/python3/dist-packages. [ICO], Name · Last modified · Size. [PARENTDIR], Parent Directory, -. [DIR] ...

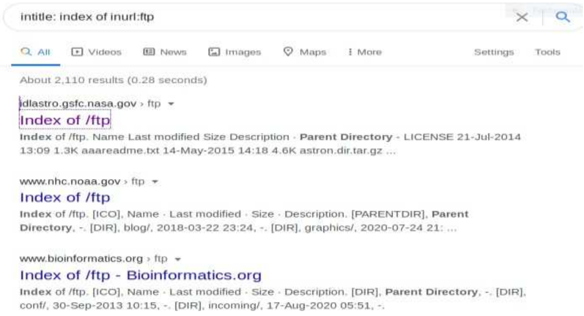
www.fusionmediasoft.com › root › proc › self › cwd

Index of /assets/upload/advertise/noname_sym/root/proc/self ...

Index of /assets/upload/advertise/noname_sym/root/**proc/self/cwd**. Name · Last modified · Size · Description · Parent Directory, -. aquota.group, 2020-10-14 02:38 ...

Figure 5. proc command

- ➢ FTP Servers: intitle: “index of” inurl:ftp



intitle: index of inurl:ftp

About 2,110 results (0.28 seconds)

idlastro.gsfc.nasa.gov › ftp

Index of /ftp

Index of /ftp. Name Last modified Size Description · **Parent Directory** - LICENSE 21-Jul-2014 13:09 1.3K aaareadme.txt 14-May-2015 14:18 4.6K astron.dir.tar.gz ...

www.nhc.noaa.gov › ftp

Index of /ftp

Index of /ftp. [ICO], Name · Last modified · Size · Description. [PARENTDIR], **Parent Directory**, -. [DIR], blog/, 2018-03-22 23:24, -. [DIR], graphics/, 2020-07-24 21: ...

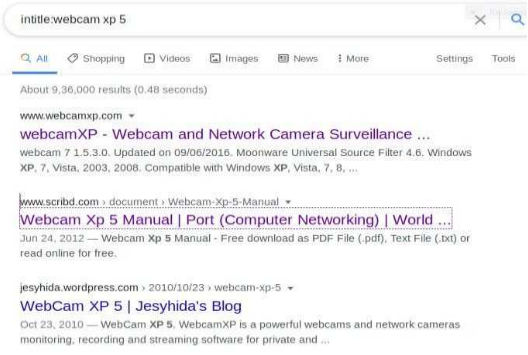
www.bioinformatics.org › ftp

Index of /ftp - Bioinformatics.org

Index of /ftp. [ICO], Name · Last modified · Size · Description. [DIR], **Parent Directory**, -. [DIR], conf/, 30-Sep-2013 10:15, -. [DIR], incoming/, 17-Aug-2020 05:51, -.

Figure 6. ftp command

- ➢ Intitle: webcams XP 5



intitle:webcam xp 5

About 9,36,000 results (0.48 seconds)

www.webcamxp.com

webcamXP - Webcam and Network Camera Surveillance ...

webcam 7 1.5.3.0. Updated on 09/06/2016. Moonware Universal Source Filter 4.6. Windows **XP**, 7, Vista, 2003, 2008. Compatible with Windows **XP**, Vista, 7, 8, ...

www.scribd.com › document › Webcam-Xp-5-Manual

Webcam Xp 5 Manual | Port (Computer Networking) | World ...

Jun 24, 2012 — Webcam **Xp 5** Manual - Free download as PDF File (.pdf), Text File (.txt) or read online for free.

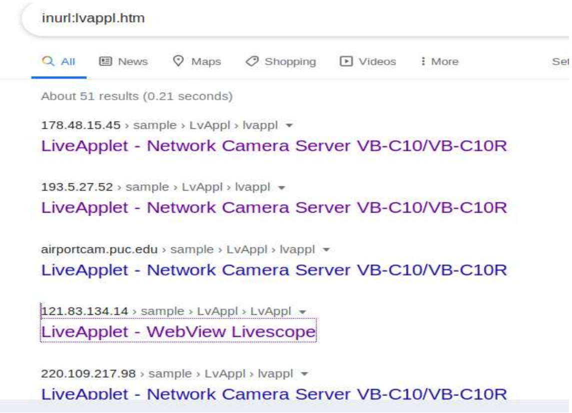
jesyhida.wordpress.com › 2010/10/23 › webcam-xp-5

WebCam XP 5 | Jesyhida's Blog

Oct 23, 2010 — WebCam **XP 5**. WebcamXP is a powerful webcams and network cameras monitoring, recording and streaming software for private and ...

Figure 7. webcam command

- Inurl: lvappl.htm



About 51 results (0.21 seconds)

178.48.15.45 › sample › LvAppl › lvappl
LiveApplet - Network Camera Server VB-C10/VB-C10R

193.5.27.52 › sample › LvAppl › lvappl
LiveApplet - Network Camera Server VB-C10/VB-C10R

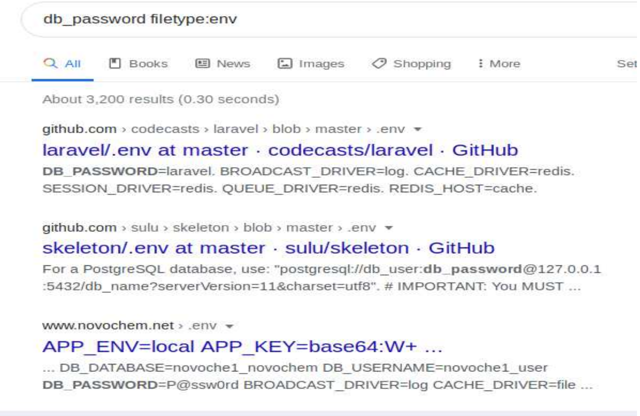
airportcam.puc.edu › sample › LvAppl › lvappl
LiveApplet - Network Camera Server VB-C10/VB-C10R

121.83.134.14 › sample › LvAppl › LvAppl
LiveApplet - WebView Livescope

220.109.217.98 › sample › LvAppl › lvappl
LiveApplet - Network Camera Server VB-C10/VB-C10R

Figure 8. live applet command

- Db_password filetype:env



About 3,200 results (0.30 seconds)

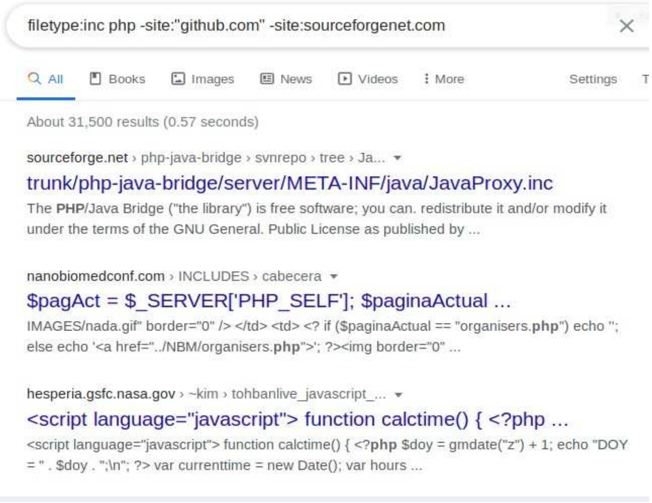
github.com › codecasts › laravel › blob › master › .env
laravel/.env at master · codecasts/laravel · GitHub
DB_PASSWORD=laravel. BROADCAST_DRIVER=log. CACHE_DRIVER=redis. SESSION_DRIVER=redis. QUEUE_DRIVER=redis. REDIS_HOST=cache.

github.com › sulu › skeleton › blob › master › .env
skeleton/.env at master · sulu/skeleton · GitHub
For a PostgreSQL database, use: "postgresql://db_user:**db_password**@127.0.0.1:5432/db_name?serverVersion=11&charset=utf8". # IMPORTANT: You MUST ...

www.novochem.net › .env
APP_ENV=local APP_KEY=base64:W+ ...
... DB_DATABASE=novoche1_novochem DB_USERNAME=novoche1_user DB_PASSWORD=P@ssw0rd BROADCAST_DRIVER=log CACHE_DRIVER=file ...

Figure 9. Db_pass command

- Filetype:inc php -site:" github.com" -site: sourceforge.net


filetype:inc php -site:"github.com" -site:sourceforgenet.com
All
Books
Images
News
Videos
More
Settings
T

About 31,500 results (0.57 seconds)

sourceforge.net › php-java-bridge › svnrepo › tree › Ja...
trunk/php-java-bridge/server/META-INF/java/JavaProxy.inc
The **PHP**/Java Bridge ("the library") is free software; you can. redistribute it and/or modify it under the terms of the GNU General. Public License as published by ...

nanobiomedconf.com › INCLUDES › cabecera
$pagAct = $_SERVER['PHP_SELF']; $paginaActual ...
IMAGES/nada.gif" border="0" /> </td> <td> <? if ($paginaActual == "organisers.**php**") echo "; else echo '<a href="../NBM/organisers.**php**">'; ?><img border="0" ...

hesperia.gsfc.nasa.gov › ~kim › tohbanlive_javascript_...
<script language="javascript"> function calctime() { <?php ...
<script language="javascript"> function calctime() { <?**php** $doy = gmdate("z") + 1; echo "DOY = " . $doy . ";\n"; ?> var currenttime = new Date(); var hours ...

Figure 10. php command

Other commands include:

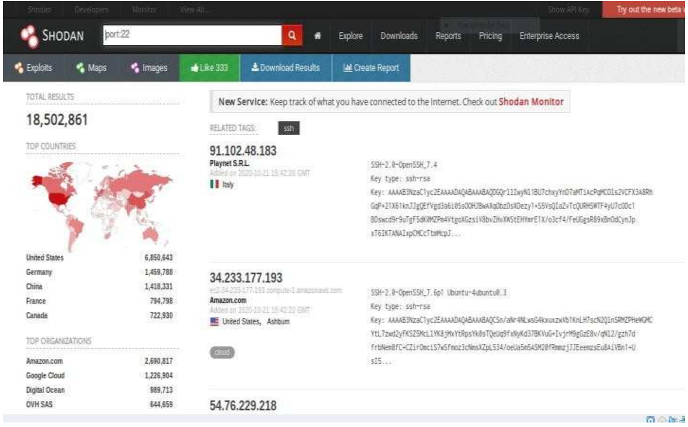
Filetype:php ―notice: undefined variable:data in‖ -forum

Intitle:WAMP server homepage server configuration apache version

Intitle:‖report‖(qualys|acunetic|nessus|netsparker|nmap|) filetype:pdf

B. Shodan:

Shodan is known as the search engine for everything. What does Shodan have more than Google? It can index everything on the internet and not just web-pages. It indexes IoT, devices like web cams, plate readers, smart TV‘s etc. It queries a supported port and checks the IPv4 address on this port, then grabs a service banner. This service banner consists of all the metadata of a specific device. This entire process is re-iterated. It is a scary tool, but a very useful tool when used for defense of one‘s own networks. Some examples of Shodan‘s interface are given below:



New Service: Keep track of what you have connected to the Internet. Check out Shodan Monitor

TOTAL RESULTS
18,502,861

TOP COUNTRIES

United States	6,850,643
Germany	1,459,788
China	1,418,331
France	794,798
Canada	722,930

TOP ORGANIZATIONS

Amazon.com	2,690,817
Google Cloud	1,226,904
Digital Ocean	989,713
OVH SAS	644,659

RELATED TAGS: ssh

91.102.48.183
Playnet S.R.L.
Added on 2020-10-21 15:42:20 GMT
Italy

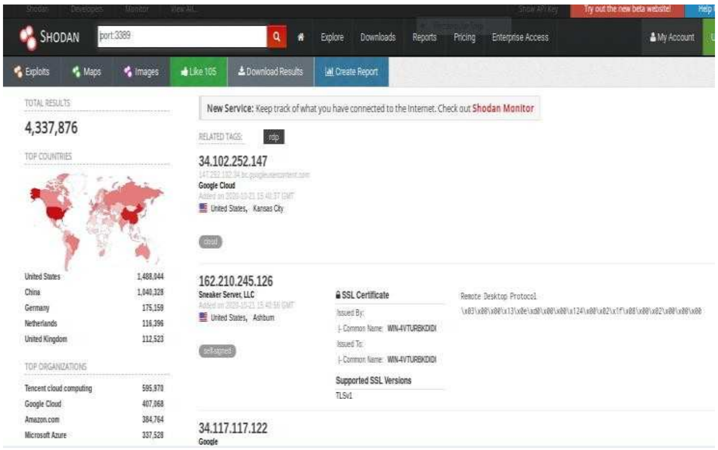
```
SSH-2.0-OpenSSH_7.4
Key type: ssh-rsa
Key: AAAAB3NzaC1yc2EAAAADAQABAAABAQDGQr11IwyN11BU7chxyYnD7aMT1AcPqMC01s2VCFX3A8Rh
GqP+21X61knJJgQEfVgd3a6i05s0OHJBwAXqDbzDsXDezy1+S5VsQIaZvTcQURHSWTF4yU7c0Dc1
BDswcd9r9uTgF5dK0MZPm4VtgoXGzsiV8bvZHvXWStEHYmrE1X/o3cf4/FeUGgsR89xBnOdCynJp
xT6IKTANAIxpCMCcTtmMcpJ...
```

34.233.177.193
ec2-34-233-177-193.compute-1.amazonaws.com
Amazon.com
Added on 2020-10-21 15:42:22 GMT
United States, Ashburn

```
SSH-2.0-OpenSSH_7.6p1 Ubuntu-4ubuntu0.3
Key type: ssh-rsa
Key: AAAAB3NzaC1yc2EAAAADAQABAAABAQCSn/aNr4NLwsG4kxuxzwVb1KnLH7scN2Q1nSRMZPHeWGMC
YtL7zwd2yFKSZ5McL1YK8jMxYtRpsYk0sTQeUq9fxNyKd37BKVuG+IvjrM9gGzE8v/qN12/gzh7d
frbNem8fC+CZirOmciS7wSfmoz3cNmsXZpL534/oeUa5m5ASM20fRmmzjJJEeemzsEu8AiVBn1+U
sI5...
```

54.76.229.218

Figure 11. Shodan port



New Service: Keep track of what you have connected to the Internet. Check out Shodan Monitor

TOTAL RESULTS
4,337,876

TOP COUNTRIES

United States	1,488,044
China	1,040,328
Germany	175,159
Netherlands	116,396
United Kingdom	112,523

TOP ORGANIZATIONS

Tencent cloud computing	595,970
Google Cloud	407,068
Amazon.com	384,764
Microsoft Azure	337,528

RELATED TAGS: rdp

34.102.252.147
147.252.102.34.bc.googleusercontent.com
Google Cloud
Added on 2020-10-21 15:40:37 GMT
United States, Kansas City

162.210.245.126
Sneaker Server, LLC
Added on 2020-10-21 15:40:56 GMT
United States, Ashburn

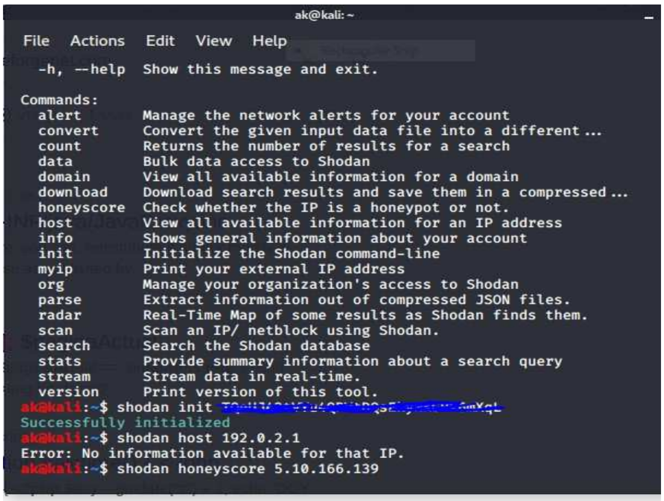
SSL Certificate
Issued By:
|- Common Name: WIN-4VTURBKDIDI
Issued To:
|- Common Name: WIN-4VTURBKDIDI
Supported SSL Versions
TLSv1

```
Remote Desktop Protocol
\x03\x00\x00\x13\x0e\xd0\x00\x00\x124\x00\x02\x1f\x08\x00\x02\x00\x00\x00
```

34.117.117.122
Google

Figure 12. Shodan ports

Shodan also has a command line interface. Different commands that are supported are given below in the picture.



```
  -h, --help  Show this message and exit.

Commands:
  alert       Manage the network alerts for your account
  convert     Convert the given input data file into a different ...
  count       Returns the number of results for a search
  data        Bulk data access to Shodan
  domain      View all available information for a domain
  download    Download search results and save them in a compressed ...
  honeyscore  Check whether the IP is a honeypot or not.
  host        View all available information for an IP address
  info        Shows general information about your account
  init        Initialize the Shodan command-line
  myip        Print your external IP address
  org         Manage your organization's access to Shodan
  parse       Extract information out of compressed JSON files.
  radar       Real-Time Map of some results as Shodan finds them.
  scan        Scan an IP/ netblock using Shodan.
  search      Search the Shodan database
  stats       Provide summary information about a search query
  stream      Stream data in real-time.
  version     Print version of this tool.
ak@kali:~$ shodan init [redacted]
Successfully initialized
ak@kali:~$ shodan host 192.0.2.1
Error: No information available for that IP.
ak@kali:~$ shodan honeyscore 5.10.166.139
```

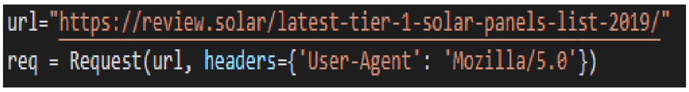
Figure 13 – Shodan CLI

C. Web Crawler:

A web crawler, also known as a spider bot can download and index content from all over the internet. Search Engines like Google are massive web crawlers. The goal of such a tool would be to learn what every webpage on the web will be about so as to retrieve this information when it is needed. They are called web crawls because that is quite literally what they do to get information. Technically, they access a website and obtain the data. Crawler bots start from a list of known URLs and they crawl the pages in these URLs first. As they crawl these webpages, they find out hyperlinks to new webpages and they add these URLs to a list of pages that they can crawl next. Moreover, there are certain policies that the web crawlers follow to make the indexing more selective.

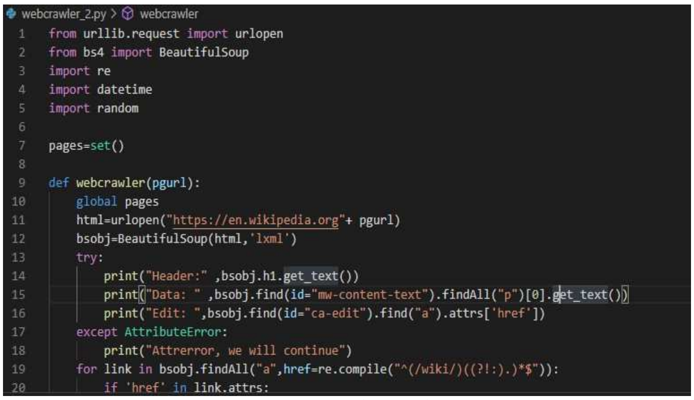
The Web Crawler we made: It is used for crawling Wikipedia Pages, which is a spyder-friendly website. Some websites employ spyder-blockers, in which case we can use a proxy to bypass detection.

Proxy Detection:



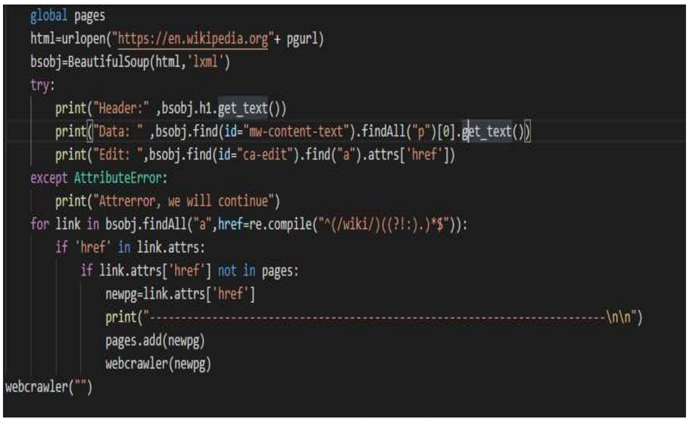
```
url="https://review.solar/latest-tier-1-solar-panels-list-2019/"
req = Request(url, headers={'User-Agent': 'Mozilla/5.0'})
```

Figure 14. Proxy Detection



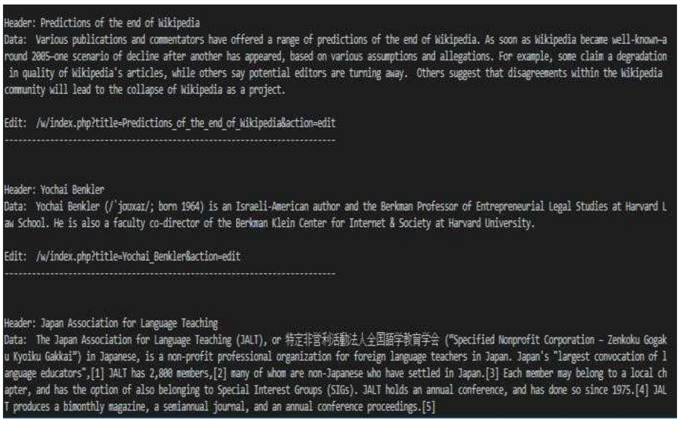
```
from urllib.request import urlopen
from bs4 import BeautifulSoup
import re
import datetime
import random

pages=set()

def webcrawler(pgurl):
    global pages
    html=urlopen("https://en.wikipedia.org"+ pgurl)
    bsobj=BeautifulSoup(html,'lxml')
    try:
        print("Header:" ,bsobj.h1.get_text())
        print("Data: " ,bsobj.find(id="mw-content-text").findAll("p")[0].get_text())
        print("Edit: ",bsobj.find(id="ca-edit").find("a").attrs['href'])
    except AttributeError:
        print("Attrerror, we will continue")
    for link in bsobj.findAll("a",href=re.compile("^(/wiki/)((?!:).)*$")):
        if 'href' in link.attrs:
```

Figure 15. Web Crawler



```
    global pages
    html=urlopen("https://en.wikipedia.org"+ pgurl)
    bsobj=BeautifulSoup(html,'lxml')
    try:
        print("Header:" ,bsobj.h1.get_text())
        print("Data: " ,bsobj.find(id="mw-content-text").findAll("p")[0].get_text())
        print("Edit: ",bsobj.find(id="ca-edit").find("a").attrs['href'])
    except AttributeError:
        print("Attrerror, we will continue")
    for link in bsobj.findAll("a",href=re.compile("^(/wiki/)((?!:).)*$")):
        if 'href' in link.attrs:
            if link.attrs['href'] not in pages:
                newpg=link.attrs['href']
                print("----------------------------------------------------------------------\n\n")
                pages.add(newpg)
                webcrawler(newpg)
webcrawler("")
```

Figure 16. Web Crawler



```
Header: Predictions of the end of Wikipedia
Data:  Various publications and commentators have offered a range of predictions of the end of Wikipedia. As soon as Wikipedia became well-known-a
round 2005-one scenario of decline after another has appeared, based on various assumptions and allegations. For example, some claim a degradation
 in quality of Wikipedia's articles, while others say potential editors are turning away.  Others suggest that disagreements within the Wikipedia
community will lead to the collapse of Wikipedia as a project.

Edit:  /w/index.php?title=Predictions_of_the_end_of_Wikipedia&action=edit
----------------------------------------------------------------------



Header: Yochai Benkler
Data:  Yochai Benkler (/ˈjoʊxaɪ/; born 1964) is an Israeli-American author and the Berkman Professor of Entrepreneurial Legal Studies at Harvard L
aw School. He is also a faculty co-director of the Berkman Klein Center for Internet & Society at Harvard University.

Edit:  /w/index.php?title=Yochai_Benkler&action=edit
----------------------------------------------------------------------



Header: Japan Association for Language Teaching
Data:  The Japan Association for Language Teaching (JALT), or 特定非営利活動法人全国語学教育学会 ("Specified Nonprofit Corporation - Zenkoku Gogak
u Kyoiku Gakkai") in Japanese, is a non-profit professional organization for foreign language teachers in Japan. Japan's "largest convocation of l
anguage educators",[1] JALT has 2,800 members,[2] many of whom are non-Japanese who have settled in Japan.[3] Each member may belong to a local ch
apter, and has the option of also belonging to Special Interest Groups (SIGs). JALT holds an annual conference, and has done so since 1975.[4] JAL
T produces a bimonthly magazine, a semiannual journal, and an annual conference proceedings.[5]
```

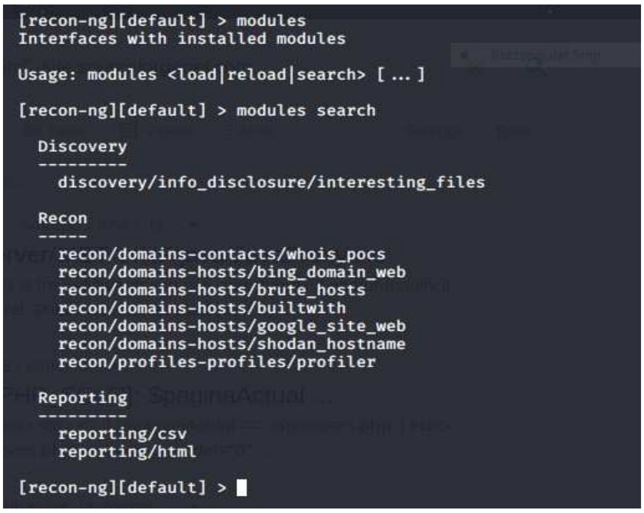
Figure 17. Web Crawler Output

In the pictures above, web crawler used to crawl Wikipedia web pages and extract the Title, Introductory Paragraph, and Links to other pages. We have created a variable _pages' to store the urls of the pages we visit, this is done to avoid revisiting the same page. Essentially this web crawler scours the main div element for Wikipedia URLs. We use Regular Expressions to filter and take only Wikipedia pages for consideration.

This process is iterative and will not stop crawling unless the user stops execution. Web Crawlers show great promise in the world of Vulnerability Testing. It can be used to scour thousands of websites in mere seconds. When used in conjunction with SQL injection it can produce disastrous consequences in terms of web security breaches.

D. Recon-ng [11]:

This is a very popular tool and it is already inbuilt with Kali Linux. This is a reconnaissance framework that is written in python language. Using various modules one can even gather the mobile numbers of the users/ owners of the website. Its interface is very interactive and you'll be guided through the entire process. There are many passive modules. But many modules can be used to directly attack the host. It is designed exclusively for web based open source intelligence. It is a very good OSINT tool.



```
[recon-ng][default] > modules
Interfaces with installed modules

Usage: modules <load|reload|search> [ ... ]

[recon-ng][default] > modules search

  Discovery
  ---------
    discovery/info_disclosure/interesting_files

  Recon
  -----
    recon/domains-contacts/whois_pocs
    recon/domains-hosts/bing_domain_web
    recon/domains-hosts/brute_hosts
    recon/domains-hosts/builtwith
    recon/domains-hosts/google_site_web
    recon/domains-hosts/shodan_hostname
    recon/profiles-profiles/profiler

  Reporting
  ---------
    reporting/csv
    reporting/html

[recon-ng][default] > █
```

Figure 18. Recon-ng

E. Photon [12]

Photon is a real fast web crawler specifically designed for automating OSINT by using a very simple interface and a lot of customizing options. It is also written in python. It acts as a web crawler and helps in extracting URLs with parameters. It can also obtain keys and DNS names. Photon can extract the following data while crawling:

- URLs
- URLs with parameters (forexample.com/id=2
- Different emails, social media accounts
- Files
- Strings matching any custom regex pattern
- JavaScript Files and Endpoints
- Secret keys – API keys/ auth keys
- Subdomains and DNS related data

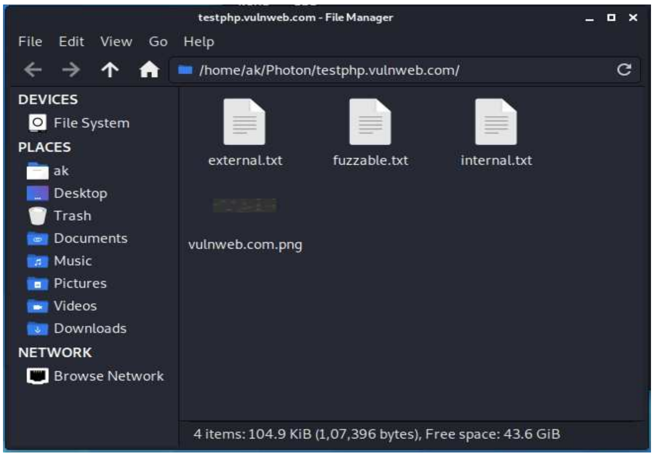
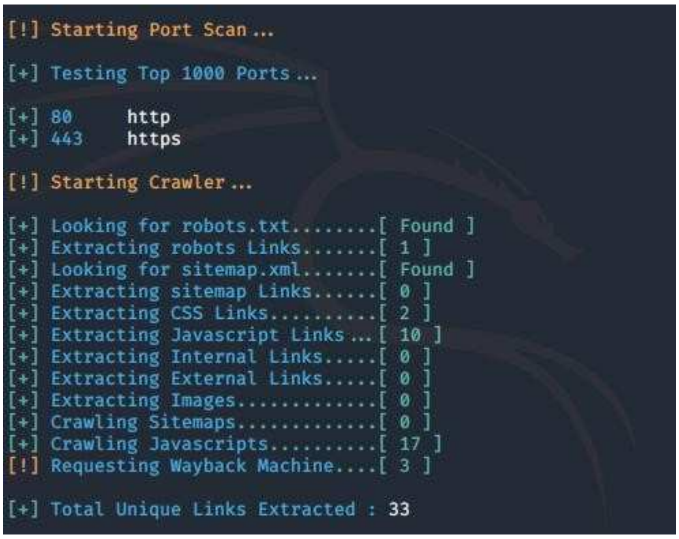


Figure 19. Photon Results

F. Final Recon [13]:

Final Recon is a very simple and intuitive interface also used for web reconnaissance. It also is built with python. It checks SSL certificates, WhoIs information, header analysis and crawling.

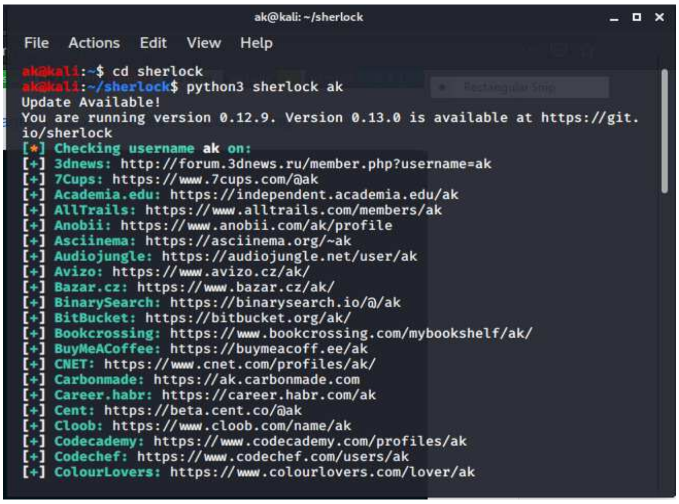


```
[!] Starting Port Scan ...

[+] Testing Top 1000 Ports ...

[+] 80     http
[+] 443    https

[!] Starting Crawler ...

[+] Looking for robots.txt........[ Found ]
[+] Extracting robots Links.......[ 1 ]
[+] Looking for sitemap.xml.......[ Found ]
[+] Extracting sitemap Links......[ 0 ]
[+] Extracting CSS Links..........[ 2 ]
[+] Extracting Javascript Links ...[ 10 ]
[+] Extracting Internal Links.....[ 0 ]
[+] Extracting External Links.....[ 0 ]
[+] Extracting Images.............[ 0 ]
[+] Crawling Sitemaps.............[ 0 ]
[+] Crawling Javascripts..........[ 17 ]
[!] Requesting Wayback Machine....[ 3 ]

[+] Total Unique Links Extracted : 33
```

Figure 20. Final Recon Results

G. Sherlock [14]

Sherlock can be used for finding your username across many different platforms. Written in python, it makes it very easy for one to find your username across any platform. Sites like pipl.com also exist wherein your phone number can be found.



```
ak@kali:~$ cd sherlock
ak@kali:~/sherlock$ python3 sherlock ak
Update Available!
You are running version 0.12.9. Version 0.13.0 is available at https://git.
io/sherlock
[*] Checking username ak on:
[+] 3dnews: http://forum.3dnews.ru/member.php?username=ak
[+] 7Cups: https://www.7cups.com/@ak
[+] Academia.edu: https://independent.academia.edu/ak
[+] AllTrails: https://www.alltrails.com/members/ak
[+] Anobii: https://www.anobii.com/ak/profile
[+] Asciinema: https://asciinema.org/~ak
[+] Audiojungle: https://audiojungle.net/user/ak
[+] Avizo: https://www.avizo.cz/ak/
[+] Bazar.cz: https://www.bazar.cz/ak/
[+] BinarySearch: https://binarysearch.io/@/ak
[+] BitBucket: https://bitbucket.org/ak/
[+] Bookcrossing: https://www.bookcrossing.com/mybookshelf/ak/
[+] BuyMeACoffee: https://buymeacoff.ee/ak
[+] CNET: https://www.cnet.com/profiles/ak/
[+] Carbonmade: https://ak.carbonmade.com
[+] Career.habr: https://career.habr.com/ak
[+] Cent: https://beta.cent.co/@ak
[+] Cloob: https://www.cloob.com/name/ak
[+] Codecademy: https://www.codecademy.com/profiles/ak
[+] Codechef: https://www.codechef.com/users/ak
[+] ColourLovers: https://www.colourlovers.com/lover/ak
```

Figure 21. Sherlock

H. BurpSuite:

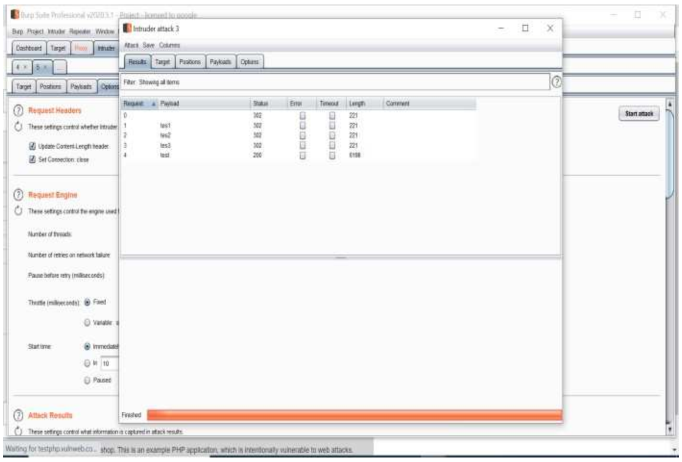
Burp Suite is commonly known as Burp. It is a proxy-based interface that is used to evaluate the security of web-based applications and hand on testing. It is very widely used as a vulnerability scanner.

Below are some of the attacks that can be done using Burp.

Brute Force Attack:

A brute force attack can happen in a million number of ways. Primarily it happens in a way where the attacker has some predefined values configured, and make requests to a server using these values and then analyze the responses.

For efficiency, an attacker can use a dictionary attack – this can be with or without mutations, or a traditional brute force attack. Considering the methods in which the attack is carried out and the number of tries, efficiency of the system used for attacking the attacker can calculate how much time itll take to brute force and get the values that he needs.



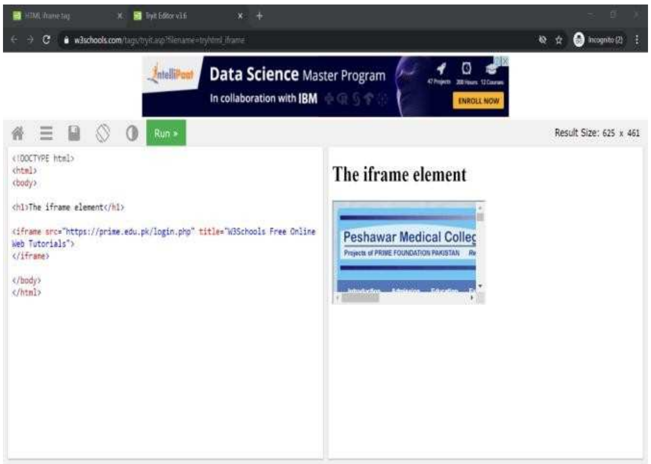
Burp Suite Professional v2020.3.1 - Project - licensed to google
Intruder attack 3
Attack Save Columns
Results Target Positions Payloads Options
Filter: Showing all items
Request Headers
These settings control whether Intruder
☑ Update Content-Length header
☑ Set Connection: close
Request Engine
These settings control the engine used
Number of threads
Number of retries on network failure
Pause before retry (milliseconds)
Throttle (milliseconds): ● Fixed ○ Variable
Start time: ● Immediately ○ In 10 ○ Paused
Attack Results
These settings control what information is captured in attack results.
Start attack
Finished
Waiting for testphp.vulnweb.co... shop. This is an example PHP application, which is intentionally vulnerable to web attacks.

Request	Payload	Status	Error	Timeout	Length	Comment
0		302	□	□	221	
1	tes1	302	□	□	221	
2	tes2	302	□	□	221	
3	tes3	302	□	□	221	
4	test	200	□	□	6198	

Figure 22. Brute Force Attack

Clickjacking:

This is classified as a User Interface redressing attack/ UI redress attack. This is a malicious technique wherein the attacker tricks the user into clicking something different from what the user sees. Therefore, the user may reveal confidential information or allows other to take control of their computer unknowingly while seeing completely normal web pages or UI.



HTML iframe tag
Tryit Editor v3.6
w3schools.com/tags/tryit.asp?filename=tryhtml_iframe
Incognito (2)
Data Science Master Program
In collaboration with IBM
ENROLL NOW
Run »
Result Size: 625 x 461
Peshawar Medical Colleg
Projects of PRIME FOUNDATION PAKISTAN

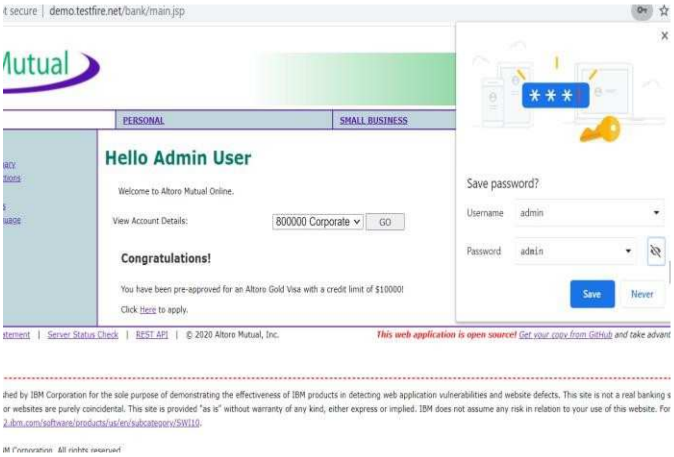
```
<!DOCTYPE html>
<html>
<body>

<h1>The iframe element</h1>

<iframe src="https://prime.edu.pk/login.php" title="W3Schools Free Online
Web Tutorials">
</iframe>

</body>
</html>
```

The iframe element

Figure 23. Click Jacking

Default Credentials:

This vulnerability is very commonly found in many devices like modems, routers and digital cameras, to name a few. Most devices have some pre-set administrative values and controls to access all of the configuration of the devices. Vendors and manufacturers of such devices use a pre-defined set of admin credentials so as to access their configurations and any attacker would be able to misuse this fact to hack these devices and gain access to them.



demo.testfire.net/bank/main.jsp
Mutual
PERSONAL
SMALL BUSINESS
GO
Save
Never
Server Status Check | REST API | © 2020 Altoro Mutual, Inc.
This web application is open source! Get your copy from GitHub and take advant

Hello Admin User

Welcome to Altoro Mutual Online.

View Account Details: 800000 Corporate

Congratulations!

You have been pre-approved for an Altoro Gold Visa with a credit limit of $10000!

Click Here to apply.

Save password?

Username: admin

Password: admin

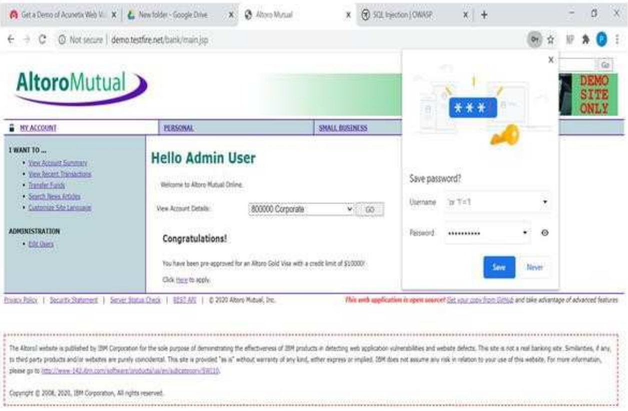
...hed by IBM Corporation for the sole purpose of demonstrating the effectiveness of IBM products in detecting web application vulnerabilities and website defects. This site is not a real banking s... or websites are purely coincidental. This site is provided "as is" without warranty of any kind, either express or implied. IBM does not assume any risk in relation to your use of this website. For ... 2.ibm.com/software/products/us/en/subcategory/SWI10.

...M Corporation. All rights reserved.

Figure 24. Default Credentials

SQL-Injection on Login Pages:

SQL injection is a web security vulnerability that allows an attacker to make changes to the queries that are made by a user through an application to its database. It usually allows the attacker to view data that they cannot be able to retrieve.



Get a Demo of Acunetix Web V...
New folder - Google Drive
Altoro Mutual
SQL Injection | OWASP
Not secure | demo.testfire.net/bank/main.jsp
AltoroMutual
DEMO SITE ONLY
MY ACCOUNT
PERSONAL
SMALL BUSINESS
I WANT TO ...
View Account Summary
View Recent Transactions
Transfer Funds
Search News Articles
Customize Site Language
ADMINISTRATION
Edit Users
GO
Save
Never
Privacy Policy | Security Statement | Server Status Check | REST API | © 2020 Altoro Mutual, Inc.
This web application is open source! Get your copy from GitHub and take advantage of advanced features

Hello Admin User

Welcome to Altoro Mutual Online.

View Account Details: 800000 Corporate

Congratulations!

You have been pre-approved for an Altoro Gold Visa with a credit limit of $10000!

Click Here to apply.

Save password?

Username: 'or '1'=1

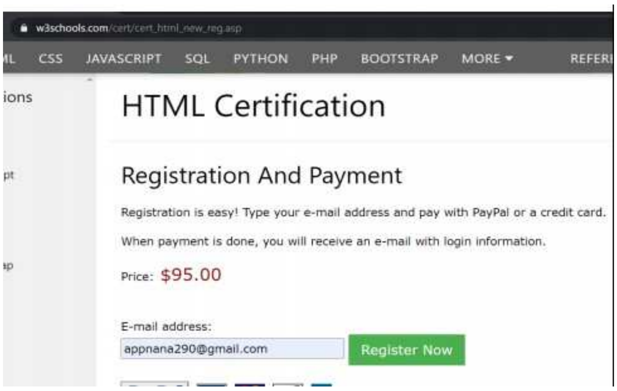
Password: ••••••••••

The Altoro! website is published by IBM Corporation for the sole purpose of demonstrating the effectiveness of IBM products in detecting web application vulnerabilities and website defects. This site is not a real banking site. Similarities, if any, to third party products and/or websites are purely coincidental. This site is provided "as is" without warranty of any kind, either express or implied. IBM does not assume any risk in relation to your use of this website. For more information, please go to http://www-142.ibm.com/software/products/us/en/subcategory/SWI10/.

Copyright © 2008, 2020, IBM Corporation, All rights reserved.

Figure 25. SQL Injection on Login Pages

This includes stuff like data belonging to other users or data that the application itself can access. In most of the cases, the attacker will be able to access or modify or delete this data- this causes persistent changes to the content or behaviour of the application.



w3schools.com/cert/cert_html_new_reg.asp
ML CSS JAVASCRIPT SQL PYTHON PHP BOOTSTRAP MORE ▾ REFERI
ions
pt
ap
Register Now

HTML Certification

Registration And Payment

Registration is easy! Type your e-mail address and pay with PayPal or a credit card.

When payment is done, you will receive an e-mail with login information.

Price: $95.00

E-mail address: appnana290@gmail.com

Figure 26. Web Parameter Tampering

Web Parameter Tampering:

The web parameter tampering attack occurs because of manipulation of parameters that are exchanged between clients and servers so as to modify application data such as user credentials and passwords.

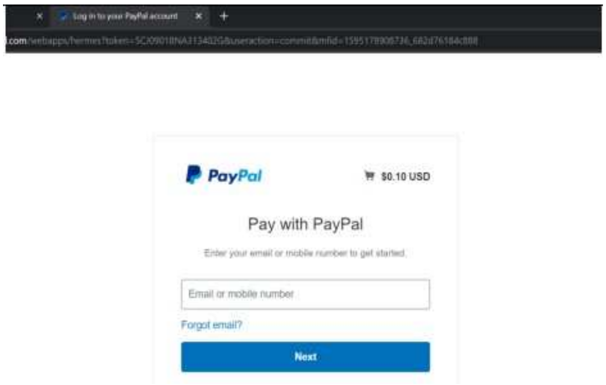
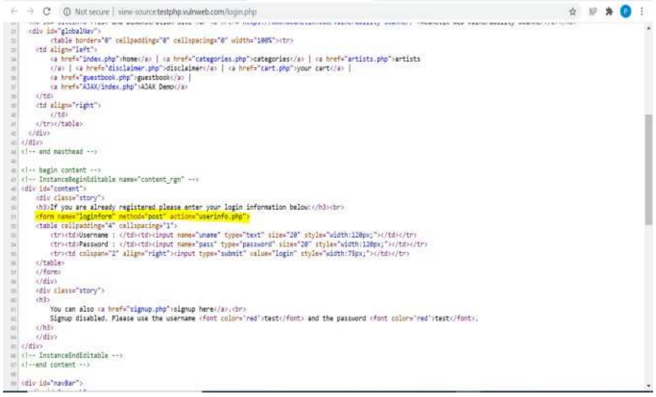

Log in to your PayPal account
.com/webapps/hermes?token=5CJ0901[illegible]&useraction=commit&mfid=1595178908736_682d76184c888
PayPal
$0.10 USD
Pay with PayPal
Enter your email or mobile number to get started.
Email or mobile number
Forgot email?
Next

Figure 27. Web Parameter Tampering

People can also tamper with the prices and quantity of products in transit. Usually this information is all stored in various cookies, hidden form fields or URL query strings. These are used to increase the functionality and control of the application.

Password Field with Autocomplete:

Most browsers offer to remember the user credentials that have been entered into their HTML form fields. This functionality can be configured by the user and also by various applications that employ the user credentials.



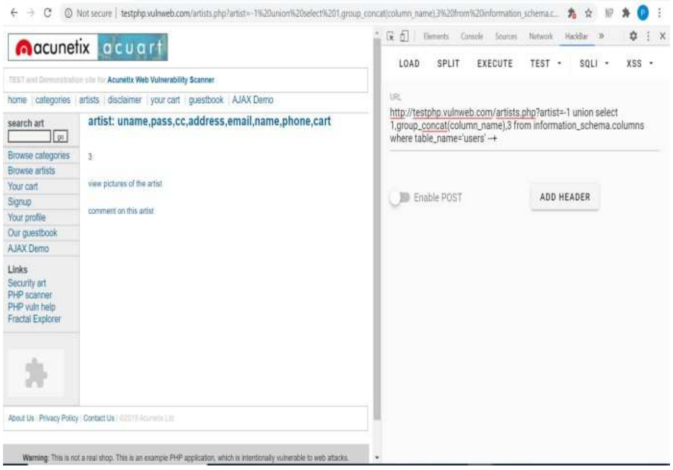
```
  <div id="globalNav">
        <table border="0" cellpadding="0" cellspacing="0" width="100%"><tr>
    <td align="left">
        <a href="index.php">home</a> | <a href="categories.php">categories</a> | <a href="artists.php">artists
        </a> | <a href="disclaimer.php">disclaimer</a> | <a href="cart.php">your cart</a> |
        <a href="guestbook.php">guestbook</a> |
        <a href="AJAX/index.php">AJAX Demo</a>
    </td>
    <td align="right">
        </td>
    </tr></table>
  </div>
</div>
<!-- end masthead -->

<!-- begin content -->
<!-- InstanceBeginEditable name="content_rgn" -->
<div id="content">
    <div class="story">
    <h3>If you are already registered please enter your login information below:</h3><br>
    <form name="loginform" method="post" action="userinfo.php">
    <table cellpadding="4" cellspacing="1">
        <tr><td>Username : </td><td><input name="uname" type="text" size="20" style="width:120px;"></td></tr>
        <tr><td>Password : </td><td><input name="pass" type="password" size="20" style="width:120px;"></td></tr>
        <tr><td colspan="2" align="right"><input type="submit" value="login" style="width:75px;"></td></tr>
    </table>
    </form>
    </div>
    <div class="story">
    <h3>
        You can also <a href="signup.php">signup here</a>.<br>
        Signup disabled. Please use the username <font color='red'>test</font> and the password <font color='red'>test</font>.
    </h3>
    </div>
</div>
<!-- InstanceEndEditable -->
<!--end content -->

<div id="navBar">
```

Figure 28. Password Field with Autocomplete

If this functionality is enabled, then all the credentials entered by the user can get stored in their local computer and retrieved by the browser whenever there‘s a visit to the application. These stored credentials can be captured by an attacker who gains control over this particular user‘s computer. Any attacker who finds any other application vulnerability like cross-side scripting or something can exploit this to get the user‘s browser stored credentials.

SQL Injection:

An SQL injection is essentially an insertion or an injection of an SQL query through the input fields of the application. This is done from the client-side. A successful SQL injection exploit can read any kind of sensitive data from the database or modify the database data.


Not secure | testphp.vulnweb.com/artists.php?artist=-1%20union%20select%201,group_concat(column_name),3%20from%20information_schema.c...
acunetix acuart
TEST and Demonstration site for Acunetix Web Vulnerability Scanner
home | categories | artists | disclaimer | your cart | guestbook | AJAX Demo
search art
go
Browse categories
Browse artists
Your cart
Signup
Your profile
Our guestbook
AJAX Demo
Links
Security art
PHP scanner
PHP vuln help
Fractal Explorer
About Us | Privacy Policy | Contact Us | ©2019 Acunetix Ltd
Warning: This is not a real shop. This is an example PHP application, which is intentionally vulnerable to web attacks.
Elements
Console
Sources
Network
HackBar
LOAD
SPLIT
EXECUTE
TEST
SQLI
XSS
URL
Enable POST
ADD HEADER

artist: uname,pass,cc,address,email,name,phone,cart

3

view pictures of the artist

comment on this artist

http://testphp.vulnweb.com/artists.php?artist=-1 union select 1,group_concat(column_name),3 from information_schema.columns where table_name='users' --+

Figure 29. SQL Injection

It can perform insert, update, delete queries on the database. It can also execute administrational operations like shutting down the DBMS or can recover the content of a given file that‘s present on the DBMS file system. In some cases, one can even issue commands to the operating system that is running. SQL injection attacks are a type of injection attack in which SQL commands are injected into data-plane input in order to affect the execution of already defined SQL commands.

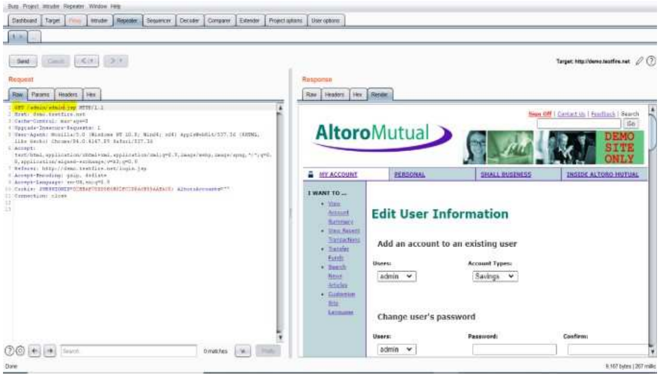
Forced Browsing:

Forced browsing is a type of attack where one tries to enumerate and access resources in the browser that are not really referenced by the application or site, but are accessible nonetheless.

An attacker can use different Brute force techniques and search for unlinked contents in the domain directory- this can include temporary directories and files and very old backup and configuration and auth files. These resources may store sensitive information about the web applications and operating systems- like source code, network addressing and so on. Therefore, it is considered a valuable resource for attackers and intruders.

Privilege Escalation:

Privilege Escalation is the act of exploiting a design flaw, a bug, configuration oversight in an operating system, software application that helps you gain access to the applications, resources that normally would be restricted.



```
GET /admin/admin.jsp HTTP/1.1
Host: demo.testfire.net
Cache-Control: max-age=0
Upgrade-Insecure-Requests: 1
User-Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML,
like Gecko) Chrome/84.0.4147.89 Safari/537.36
Accept:
text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/apng,*/*;q=0.
8,application/signed-exchange;v=b3;q=0.9
Referer: http://demo.testfire.net/login.jsp
Accept-Encoding: gzip, deflate
Accept-Language: en-US,en;q=0.9
Cookie: JSESSIONID=[illegible]; AltoroAccounts="[illegible]"
Connection: close
```

Edit User Information

Add an account to an existing user

Users: admin

Account Types: Savings

Change user's password

Users: admin

Password:

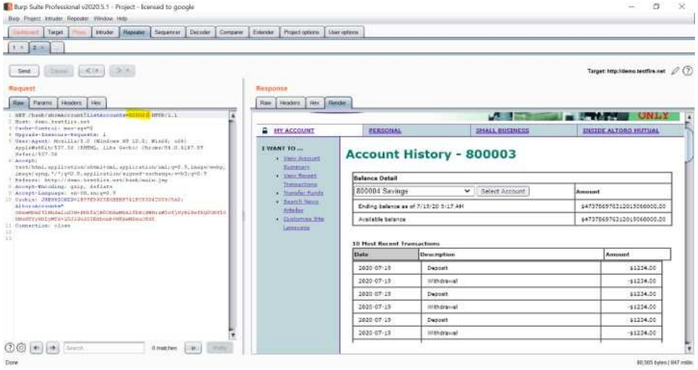
Confirm:

Figure 30. Privilege Escalation

The result is that the attacker can access the application with higher privileges than actually intended by the developer – so the attacker can perform unauthorised actions.

Insecure Direct Object Reference (IDOR):

In a web application whenever a user sends, receives or generates any kind of request from a server, there are different kinds of parameters like id, uid, pid etc., that have unique values that the user has assigned.



Account History - 800003

Balance Detail	
800004 Savings	Amount
Ending balance as of 7/15/20 5:17 AM	$47378697621201506000.00
Available balance	$47378697621201506000.00

10 Most Recent Transactions

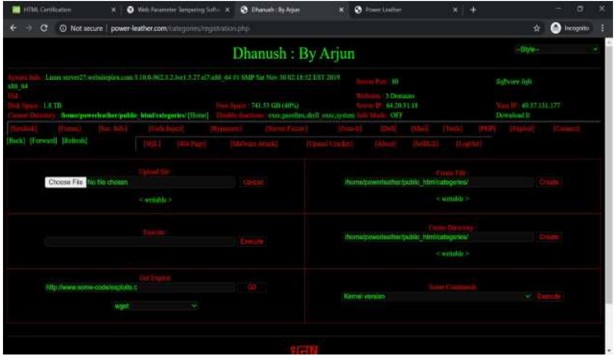
Date	Description	Amount
2020-07-15	Deposit	$1234.00
2020-07-15	Withdrawal	-$1234.00
2020-07-15	Deposit	$1234.00
2020-07-15	Withdrawal	-$1234.00
2020-07-15	Deposit	$1234.00
2020-07-15	Withdrawal	-$1234.00

Figure 31. IDOR

Unrestricted File Upload:

Uploaded files can cause a significant amount of damage to the applications. The first step in all attacks is to get some code to the system to be attached. After one finds the exploit, executing it is the only thing that‘s remaining. File upload form field helps the attacker exploit and execute his code into the application.

The consequences of file uploading can vary from application to application. This can range from a complete system takeover, an overloaded file system or database, to forwarding attacks to back-end systems, client-side attacks or simple defacement. This depends on what the application does and what kind of file is uploaded and where the file is stored.

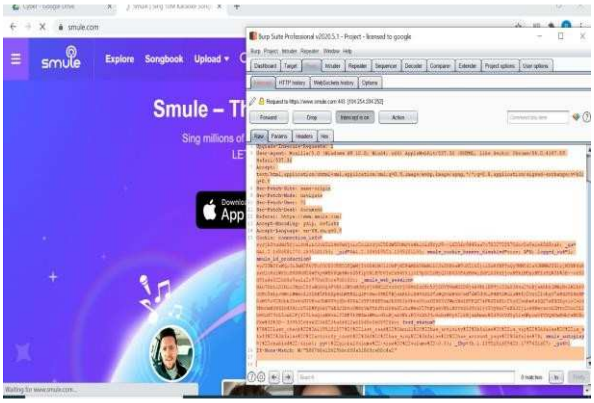


Not secure | power-leather.com/categories/registration.php
Dhanush : By Arjun
Choose File
No file chosen
Upload
/home/powerleather/public_html/categories/
Create
< writable >
Execute
http://www.some-code/exploits.c
wget
Kernel version

Figure 32. Unrestricted File Upload

Session Hijacking:

Session Hijacking attacks comprise of exploiting the web session control mechanism. This includes managing and manipulation session tokens. HTTP communication uses many different TCP connections. This attack compromises the session token by stealing or predicting a valid session token to gain unauthorised access to the web application or server.



smule.com
Explore
Songbook
Upload
Smule – T
Burp Suite Professional v2020.5.1 - Project - licensed to google
Forward
Drop
Intercept is on
Action
Raw
Params
Headers
Hex
Waiting for www.smule.com...

Figure 33. Session Hijacking

HTML Injection:

It is very similar to a cross-site scripting attack (XSS). In XSS vulnerability, the attacker can attack by injecting and executing JavaScript code. In HTML attack, it allows only injection of certain HTML tags.

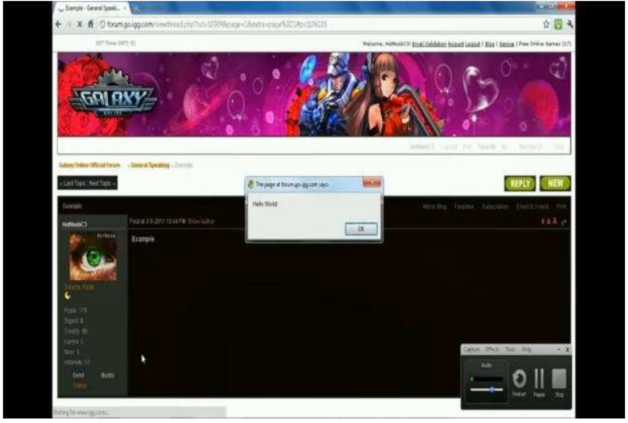


Figure 34. HTML Injection

If an application cannot properly handle user supplied data, the attacker can supply a valid HTML code mostly via parameter value and then they can inject their own content into the page. This is used along with some form of social engineering because the attack exploits code along with user‘s trust.

Command Injection:

This is an attack where the goal is the execution of arbitrary commands on the host operating system along with a vulnerable application.

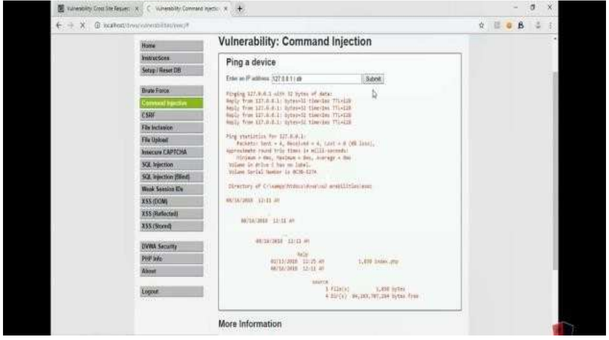


Figure 35. Command Injection

These attacks are possible when the application passes user supplied data like forms, cookies, HTTP parameters, headers etc., to a shell. Command injection attacks are usually possible because of insufficient input validation. The default functionality of the application is extended – this executes system commands without the necessity of injecting code.

VI. ANALYSIS

Sherlock is a tool that helps one find a username amongst a predefined set of platforms. The current list includes 303 sites! Some of these include 9GAG, AskFM, Blogger, BuzzFeed, Codeacademy, Codechef, Facebook, Etsy, GitHub, GitLab, Imgur, Lichess, Munzee, Pastebin, Pinterest, Quora, Reddit, SlideShare, TikTok, Trello, Wattpad, Wikipedia, phpRU, toster etc. [14]

TABLE I. COMPARISON AMONG TOOLS

Tool Name	Update Frequency	Languages Used	Supporting Operating Systems
Shodan	Random	Python, Ruby, Php, C++, C, Crystal, Go, Haskell, NodeJS, Perl, Rust	any OS, comes with a CLI interface
Recon-NG	Frequent	Python, JavaScript, CSS, TML, Dockerfile	Linux OS – Kali, Parrot OS etc
Photon	Has 19 releases – using the v1.3.0	Python, Dockerfile	Linux, Ubuntu, Kali Linux. **Cross-Platform**
Final Recon	two releases – latest is v1.1.2	Python, Dockerfile	Kali Linux, SecBSD, Linux, BlackArch Linux.
Sherlock	Currently using v0.7.1	Python, Dockerfile	Mac, Linux, Kali Linux.

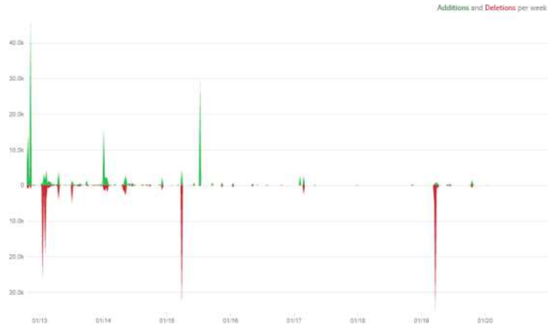


Figure 36. Recon-ng Code Frequency from their GitHub Repository

Recon-Ng also aims at reducing the time spent at gathering information from various sources by integrating a bunch of modules together to make these actions easier for the attacker/ pentester. It is a very widely used tool for information gathering and reconnaissance.

Photon is an incredibly fast crawler designed for OSINT. We can use such tools to do the heavy lifting and maximum work like sifting through URLs and retrieve information. It provides an easy to use command line interface. Apart from looking out for vulnerabilities, it also parses all this data in a user-friendly way for the attacker/pentester to know.

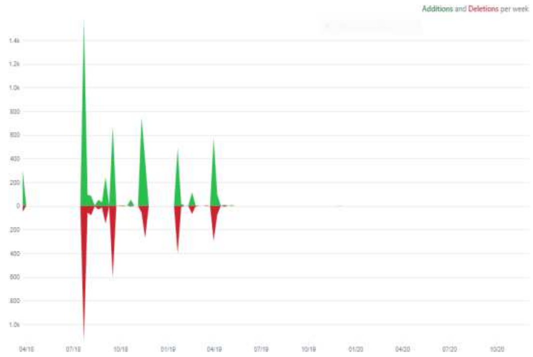


Figure 37: Photon Code Frequency from their GitHub Repository

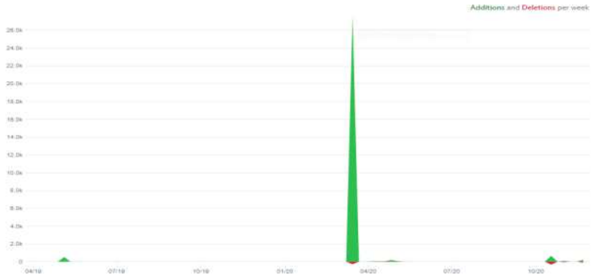


Figure 38: Final Recon Code Frequency from their GitHub Repository

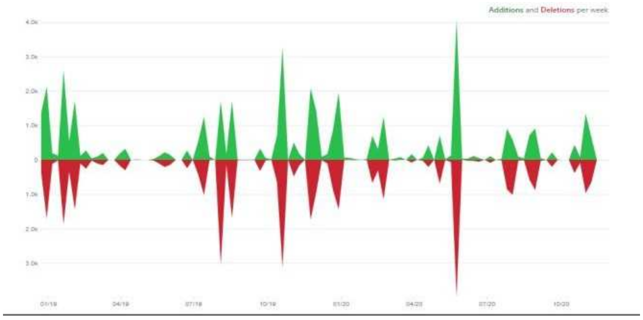


Figure 39: Sherlock Code Frequency graph from their GitHub Repository

Use Cases:

Final Recon provides an overview of the target in a short amount of time. Final Recon is an automatic web reconnaissance tool written in python. Goal of Final Recon is to provide an overview of the target in a short amount of time while maintaining the accuracy of results. Instead of executing several tools one after another it can provide similar results keeping dependencies small and simple. The features of it include, header information, whois, SSL information. The crawler provides information about the JavaScript, CSS, external and internal links, sitemaps etc. It has DNS Enumeration features and Subdomain Enumeration. Other features include – traceroute, Directory Searching, port scanning. All this information can be exported into text, xml or csv formats. [13]

Popularity of these tools based on number of stars on GitHub:

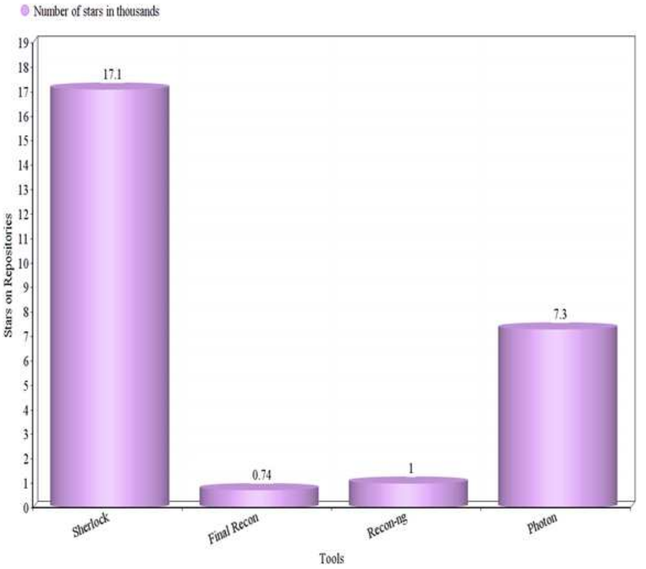


Figure 40. Bar Graph of Popularity

A proper information gathering technique would include trying to get detailed information like all the open ports, OS names, DNS, hidden links etc. A lot of our analysis provides information and content as to how to advance to doing these searches- how to use various tools online and get data like headers, whois, dns information etc. A brief overview of what these are is given below:

Header information – One needs to know what server the application is using. We can get to know various security measures like Http, hsts etc.

Whois – Once a domain is registered, contact information is usually stored. Now, whether this is public or private is the owner‘s wish. But sometimes one can find contact details through this. Also, when domains are not renewed from time to time, attacker can make use of the domain.

DNS Enumeration – It is the process of locating all the servers in a company/ organization. This gives us a lot of information about the company.

Directory Searching and Port Scanning – This can help one find directories that are hidden and not linked in the webpages. Port Scanning is also very important. Nmap is a well-known port scanner.

VII. How to Stay Protected and Safe?

Data is such a vague concept and encompasses a wide range of information that it is worth briefly breaking down different collections before examining how each area is relevant to one‘s privacy and security. Cookies and browser plugins can track your website across multiple websites. Even our email accounts act as a singular hub for devices and a single compromise can snowball into the hijack of many accounts and services, especially when you authorize other apps to be able to use the same mail account throughout.

One can secure their privacy by clearing out one‘s cookie caches and browser histories. Cookies collect a lot of information about the user. Make sure you use a https website rather than a http website. Http uses a layer of encryption to enable secure communication between a browser and a server. One must check if the website is https or not when it comes to online purchasing so that one cannot steal your credit card credentials. Search Engines like Google, Bing and Yahoo index everything that you use. This is done to ensure that people have ‒personalized‖ experiences. To prevent such data from being logged the best practice could be using DuckDuckGo, Qwant and Startpage search engines instead of the traditional ones. One can use browser plugins such as HTTPS Everywhere, NoScript Security Suite, Disconnect, Facebook Container, Blur, Privacy Badger to protect themselves from security breaches. Enabling 2Factor Authentication is a good measure to add an extra layer of security to your accounts and services. Keep patching and updating whatever software‘s or OS‘s that you are using.

Some other wise examples for keeping your devices include:

- Using Security Software
- Avoiding Phishing Emails
- Being Wise about Wi-Fi
- Locking your laptops or devices
- Reading Privacy policies

Recommendations to save oneself from the Burp Suite Attacks:

Brute Force Attack - Use of Captcha, Locking accounts, Unique URL Assignment, Unique and long passwords with mixed characters.

ClickJacking- Using Frame busting code, X-Frame Header, Implement CSP Header

Default Credentials- Do not use default passwords like admin/admin, admin/password, etc.

SQL Injections on Login Pages - Using parameterized queries to process log in request, Not allowing the use of special characters in log in fields.

Web Parameter Tampering - Do not rely on client-side validation

Password Field with autocomplete - Set autocomplete to offSQL Injection - Implementing detection systems, Parameterized statements, etc.

Forced Browsing- URL Masking at application level, Base 64 encoding of URL

Privileged Escalation - Data Execution Prevention, Address Space Layout Randomization, etc.

IDOR - Proper implementation of Parameter validation, verification of all the referenced objects

Session Hijacking - Encryption of data traffic, Use of long random number or string as session key.

HTML Hijacking - The Attacker can inject a malicious html file into the server which can then steal the user details if mistaken for the original site login page.

Command Injection - Validating against a whitelist of permitted values, validating that the input is a number.

VIII. Social Engineering and Information Gathering

Social engineering assaults ordinarily include some type of mental control, tricking, in any case, clueless clients or workers into giving over classified or confidential information. Regularly, social engineering includes email or other correspondence that conjures desperation, dread, or comparative feelings in the person in question, driving the victim to quickly uncover touchy data, click a malevolent link, or open a malicious record. Social engineers are fraudsters or cheats of today's era. It is regular them to depend on the common helpfulness of individuals or to endeavor to abuse their apparent character shortcomings. For instance, they may call with a pretended earnest issue that requires prompt organization access.

Social engineers may proceed with a baiting attack at the point when they leave a malware- contaminated gadget, for example, a USB streak drive or CD, in a spot where somebody probably will discover it. The achievement of such an assault relies on the idea that the individual who finds the gadget will stack it into their PC and unwittingly introduce the malware. Once introduced, the malware permits the attacker to progress into the casualty's system, thereby giving access to all the information and hence leading to information discovery. Another form of attack that could lead to information discovery is the phishing attack where an attacker makes fake correspondences with a victim that are veiled as authentic, regularly guaranteeing or appearing to be from a confined source. In a phishing assault the beneficiary is fooled into introducing malware on their gadget or sharing individual, money related, or business data. Or in another instance, the attacker may send a mail that could be camouflaged to appear as though it originates from somebody inside your association. Yet, on the off chance that you react to that email with your username and password, your PC is handily compromised, thus leading to the revelation of sensitive information. The attackers may also resort to measures such as pretexting or quid pro quo so as to fool the victim into uncovering login accreditations or giving PC access.

Social engineering is a genuine and progressing danger for many associations and individual customers who succumb to these cons. Education is the initial phase in keeping your association from succumbing to shrewd assailants utilizing progressively advanced social engineering strategies to access sensitive information.

The social engineering framework is a good resource of information for people who want to learn about the physical, psychological and historical aspects of social engineering. Usually, information gathering and social engineering go hand in hand. The more an attacker knows about you, the more likely he is to attack you in a way that will completely seem benign. In reality, it is not. We‘ve already shown you how information can be gathered. There‘s multiple number of tools that can help with the very same. A social engineer can combine many small things that they‘ve learnt about you from various sources to obtain useful information. Wherever the information comes from, it‘ll still be useful. As already iterated, information is power and having any amount of this would prove to be very fruitful. According to the definition by the FBI, elicitation is a technique used to discreetly gather information. That is to say, elicitation is the strategic use of casual conversation to extract information from people (targets) without giving them the feeling that they are being interrogated or pressed for the information [15]. Elicitation attacks can be simple or involve complex cover stories, planning, and even co-conspirators. Social engineers use elicitation techniques to gather valuable information.

Techniques to Social Engineering include:

- Flattery
- False Statements
- Artificial Ignorance
- Sounding Board
- Bracketing

In order to keep the research focused, you need to begin with defining your goal for success because a clear objective will determine what information is relevant and what can be ignored as you search. After this, gathering information to support social engineering exercises is much the same as research you do for anything else [16]. This holds true not only for the type of information gathered but also for how it‘s gathered.

IX. Conclusion and Future Enchancements

The volume of information being made, shared, and stored away is developing dramatically with no indication of easing back down. The objective of information disclosure is to comprehend what sort of information is being put away and handled so the business can get an incentive from it. When talking as far as consistency and individual information preparing, information discovery is principal to comprehend and distinguish all information handling exercises. The yield of the information revelation measure is a store of information spaces (for example name, email address, VAT number, racial or ethnic beginning, blood classification, family status, business status, and so on), information classifications (contact data, work data, clinical data, and so forth) and specialized directions of the information (frameworks, information bases, patterns, tables, sections, organizers, records, and so on), and ought to be connected with information preparing stock. There is a likelihood that by finding the information, you will likewise recognize handling that has not been represented, and in any event, preparing managed without a reasonable reason. With more individuals getting access to and storing away records in a huge number of organizations and cloud stores, your sensitive information could be anyplace. Cooperation among workers, accomplices, and clients is critical, however there must be a harmony between information sharing and information assurance.

The review carried out on OSINT shows that there is already a substantial amount of work in the topic. Numerous techniques and tools have been developed up to now. However, there are some gaps and limitations in this field to continue exploiting the offered opportunities. Our review shows how OSINT and Social Engineering go hand-in-hand. Data privacy and security is very important for an individual. There‘s a vast number of tools that could gather any kind of information about you. There‘s a lot of security measures that one can take because of the vast amount of security measures that are available.

REFERENCES

[1] C. Best (2012). OSINT, the Internet and Privacy. *European Intelligence and Security Informatics Conference, Odense*, 4-4.

[2] Kher, Tejasvini & Kariya, Swati. (2016). A Survey on Social Engineering: Techniques and Countermeasures. *International journal of Scientific Research and Development*. 4. 2321-613.

[3] S. Lee and T. Shon (2016). Open-Source Intelligence Base Cyber Threat Inspection Framework for Critical Infrastructures. *Future Technologies Conference (FTC), San Francisco, CA*. 1030-1033.

[4] I. Vacas, I. Medeiros and N. Neves (2018). Detecting Network Threats using OSINT Knowledge-Based IDS. *14th European Dependable Computing Conference (EDCC), Iasi.* 128-135.

[5] O. M. Al-Matari, I. M. A. Helal, S. A. Mazen and S. Elhennawy (2018). Cybersecurity Tools for IS Auditing. *2018 Sixth International Conference on Enterprise Systems (ES), Limassol*. 217-223.

[6] Kevin, Fan, Wenjun & Lwakatare, & Rong, Rong. (2017). Social Engineering: I-E based Model of Human Weakness for Attack and Defense Investigations. *International Journal of Computer Network and Information Security*. 9:1-11.

[7] Alain Claude, Tambe Ebot. (2018). How Stage Theorizing Can Improve Recommendations Against Phishing Attacks. *Information Technology & People*. 32.

[8] Kerne, Andruid & Smith, Steven. (2004). The Information Discovery Framework. 357-360.

[9] Comia, Hazel. (2017). Social Engineering: Exploring Social Engineering Toolkits.

[10] Prateek, Dewan & Kashyap, Anand & Kumaraguru, Ponnurangam. (2014). Analyzing Social and Stylometric Features to Identify Spear phishing Emails.

[11] https://github.com/lanmaster53/recon-ng

[12] https://github.com/s0md3v/Photon

[13] https://github.com/thewhiteh4t/FinalRecon

[14] https://github.com/sherlock-project/sherlock

[15] https://www.redteamsecure.com/blog/5-effective-social-engineering-elicitation-techniques/

[16] https://www.social-engineer.org/framework/information-gathering/