May-June 2020
ISSN: 0193-4120 Page No. 3469-3473

Recon-Automation using OSINT

[1]Sorna Shanthi D, [2]Jai Krishna B, [3]Jagan RM, [4]Harish M

[1,2,3,4]Department of Computer Science and Engineering
Rajalakshmi Engineering College
Chennai, India

Article Info
Volume 83
Page Number: 3469-3473
Publication Issue:
May-June 2020

Article History
Article Received: *19 August 2019*
Revised: *27 November 2019*
Accepted: *29 January 2020*
Publication: *12 May 2020*

Abstract

In Today's technology driven world, there are many tools and sources to gather publicly available data for the purpose of information security investigation, but the major shortcoming is that it still needs intervention of a person to infer whether the data obtained is relative and genuine. In Open Source Intelligence (OSNIT) the name “open source” clearly depicts that the information obtained is available to anyone and present publicly, but OSINT mainly works on how the information is gathered. OSINT involves information that is not only gathered from search engines but also from various other sources that search engines cannot deliver. Main application of OSINT in cyber-security is threat, predictive intelligence. The system that is proposed not only has the ability to perform link analysis where relationships between endpoints are evaluated but will also try to find correlation between the data fetched from different open source endpoints.

Keywords: *OSNIT , cyber-security, Recon-Automation*

1. Introduction

The more we are connected to the internet, the quantum of information about each and every user keeps scaling up. This information can be put to good and bad use by either penetration testers or blackhats respectively. From an organisation’s perspective this actually means, substantial information has been exposed in public data sources. This is an area that is often flouted by information security consultants and auditors when assessing the security posture of an organisation. Open Source Intelligence (OSINT) is an intrinsic part of penetration testing and anticipating any security threats beforehand. OSINT can be simply described as information that can be gathered on any specific topic, trend or even a person from the web and publicly available resources in it.OSINT searches can be pivotal in terms of conducting investigations when time and cost parameters are taken into consideration. Network attacks on firms are mostly instantiated from such leaked information. OSINT searches not only help in unravelling one’s own vulnerabilities but can also be used to gain valuable information about their competitors.

This paper presents an efficient way for analysing publicly available information on a specific target by crawling upon various sources of data using OSINT gathering. The main goal is to retrieve information available on any particular organization, a person that is made available publicly by intention or unintentionally. This gathered data is processed and provided as intelligence to the user.

2. Related Works

Focussing on OSINT, there are a number of closed source tools and open source endpoints where focus is mainly on particular types of data. There exists a tool called ‘Harvester’ which would take an email address or domain name as the target and try to gather information about the specific given email address the result of which can be usernames of social sites or IP address of the domain specified. But the limitation with this approach is that after getting the usernames or IP out of the Harvester, there is a need to rely on another tool to gather information from social sites according to the resultant usernames. Shodan is a search engine unlike any search engines like google and bing, Shodan gives IT infrastructure related information like IP addresses, network devices, webcam and other internet related devices as a result. Maltego is a proprietary software whose main purpose is to mine data from various open sources and map the obtained data thereby gathering information on the target. This information obtained is depicted to the user henceforth presenting, exploring relationships amongst the information gathered on the target.

Every tool is useful in gathering information on specifics but there is a need to find correlation between the data obtained and to recursively continue processing which is not focussed by much of the tool at present. Considering Harvester, information on usernames can be obtained but one individual has to provide usernames as input and this process is repetitive. There is a need to automate the processing of data after gathering it using OSNIT. There's a need for bringing this correlation in, since this eradicates the necessity of repeating the same process for different data.

3. Framework

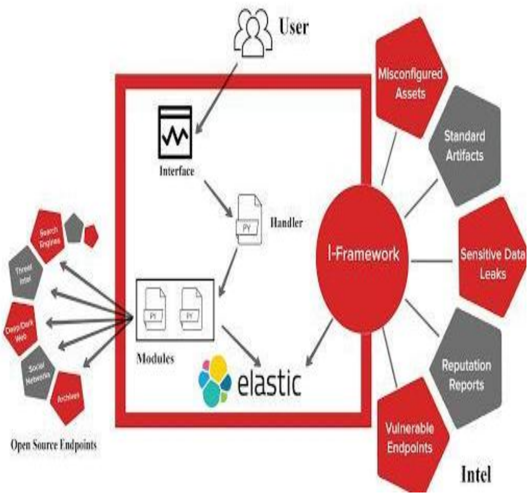


Figure 1: Framework of the Proposed System

The current norms of an individual doing all the tedious investigation of finding valuable data from crucial endpoints and checking for their authenticity can be automated by creating the framework as shown in the figure 1.Since it's impossible for a human individual to go through all the endless data that's available on public resources, optimisation is done by presenting the endless available data to the framework, and try to get more efficient and valuable information. The proposed system has the ability to perform link analysis and find correlation between the data fetched from different source or platform.

The terminologies used in the Framework such as the handler which acts as the control centre and the modules which represent modularly programmed micro services that interact with the handler, extract data from their respective data point, further process the data and report back to the handler are further explained in the methodology section.

4. Methodology

Handler:

Handler does the job of mapping each data point that is to be crawled based on the user input and the type of data extracted from each thread. Handler holds the responsibility of invoking each and every data point module based on the data that is extracted upon execution of each individual thread.

Modules:

1.*Fetching the target data from Social Media:*

The main goal of this module is to find if any Social Media profiles exist with a particular username or its combination. The username can be specified as the target by the user or upon crawling a website, individual names can be extracted. By using the combinations of such extracted names can also be passed to this module to check the existence of Social media profiles. The profile returned may not be entirely accurate but the ratio of getting legitimate results is greater than that of failures.

Working:

The username that is to be used by this module can be given by the user as input or it may be obtained from processing of other modules. This module analyses if any matching profile exists by appending this username to various social media URLs and by analysing the response, outcome can be determined. For example, upon receiving a username it can be determined if any matching profile in Instagram exists by using {www.instagram.com/username}. Response from this request helps us determine whether such profile exists or not

This goal can also be achieved by passing this username to the search engine module which determines an account exists by using dork.

2. *Fetching the target data from Search Engines:*

There are a variety of search engines out there to provide specific search responses. Using these search engines would result in vast volumes of valuable information that can be gathered about a particular target, which may assist us in better understanding our target. Content that has been retrieved from these search queries may expose different types of valuable data like some website address, or phone numbers, or any user related information. These different types of results can be further processed and put into other modules to find even more valuable findings.

Working:

Search Engine based data gathering involves using advanced search techniques which can help to uncover data that may be gathered with usual web search with keywords,

Dorking involves using search engines to their full potential to unearth results that are not visible with a regular search. It allows to refine the searches and dive deeper, and with greater precision, into web pages and

documents that are available online.
Example:
Dork: intext:(password | passcode) intext:(username| userid| user) insite:www.target.com This dork provides results with pages that contain keywords password username etc within the target domain indexed by the search Engine.

3. *Email Reconnaissance:*
This module does the job of detecting if an account exists with this Email-Id across various forums. It also verifies if the specified Email-ID has been part of any data breach, if so it also notifies the user of the source site of the breach. User names can also be extracted from email-id's by using regular expressions to split data based on certain special characters. Domain names within which the email falls in can also be determined. Domain names can be used in who is lookup to extract valuable information about it.
Working:
1.**Email-verification**:Initially this modules verifies if the mail address exists or if it is invalid, this operation is performed in stages by first connecting to the target mail server and then verifying if the target mailbox exists in the mail server.
2.**Breach data Analysis :** Breach data analysis involves checking if the particular email address has been present in data breaches,this is performed by querying the Breach data monitoring websites such as have i been pwned ,Dehashed etc.

4. *Fetching information on SSL Certificates:*
Since most of the domains which are hosted on the internet nowadays have a higher probability of having SSL certificate to enable secure data sharing between the server hosted and client programs. This module focuses on these SSL certificates to gather valuable information associated with owners of these certificates.
Working:
1.**CT-logs**- Google's Certificate Transparency Program Provides organizations an option to verify their subdomain and certificates further since it is open source it can be leveraged to gather the subdomains of an organization.
2.**crt.sh** - Certificate Search based on crt.sh provides on the organization's certificates issued along with the respective subdomain

5. *Fetching Location Information:*
Location based reconnaissance can be done just by making use of an IP address. This IP address can be specified by the user or maybe obtained upon resolving DNS of the host. This IP address can be used to extract its current physical location.

Working:
Upon resolving the IP address of the host, its exact physical location can be obtained by passing it to the freegeoIP API. Upon passing the IP address, this API returns the physical location of the host pertaining to the IP address in json format.
Physical location can also be obtained by using GeoCoding API. The response/output provided by this API is a set of latitude and longitude coordinates for the address provided.
For example:
Request:
"formatted_address" : "56/69,BR Gardens, Gandhi Nagar, Delhi-110019".

Response:

```
 "geometry" : {
 "location" : {
"lat" : 37.4224764,
          "lng" : -122.0842499
        }}
```

6. *Identifying Interesting Artifacts:*
Interesting artifacts refers to all readable files that may be available on websites, servers. These files hold sensitive data and may or may not be leftover on purpose. In some negligent cases, this has resulted in serious security issues to organisations since the data on the file may be of high significance.

Interesting information can also be obtained by using files like robots.txt, sitemap.xml files. Robots.txt file contains a list of directories that should not be indexed by web crawlers. This helps in extracting interesting locations on the website.

7. *DNS:*
DNS holds major information about a website. This can be used to extract the IP address of the server. In turn, this IP address can be passed to various modules for processing based on the IP address.

Working:
DNSSEC Zone walking:
Zonewalking is a technique where it unveils Internal records if the zone is not configured properly.The information that can be obtained can help us to map network hosts by enumerating the contents of a zone.

Configuration Enumeration:
Gather and analyse NS, MX, AXFR and A records, as well as remote BIND version from the DNS server which may provide data about the target if the target has a dedicated dns server either in On-prem or cloud.

8. *Tor Based Recon:*
Tor uses a relay network to mask the original point of data. When the data or packet reaches its destination server, it appears to have originated from the Tor exit point i.e. the last node in the relay. The main focus of this module is to check whether the target IP has been listed

as a Tor exit point, this provides valuable Intel on whether the target is behind a tor or relay network.

Working:

In order to detect whether the IP is using Tor, we provide this IP as a parameter to Exonerator-Tor Metrics. Upon passing, it checks it against a database of IP's enlisted that have been a part of Tor network. If it matches an IP in the list, it is affirmative that the IP is a Tor node.

Example:
"https://metrics.torproject.org/exonerator.html?ip="IP"&timestamp="Timestamp"&lang=en"

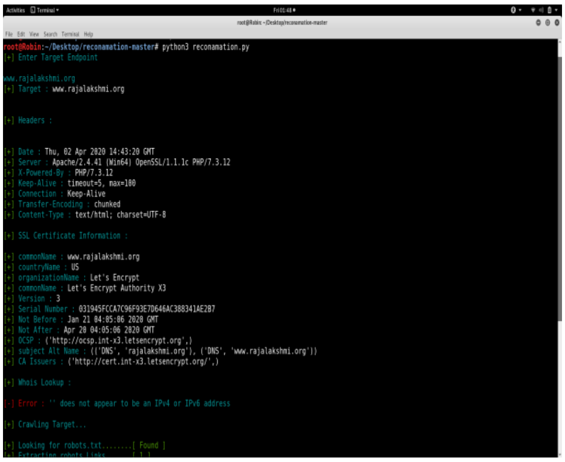
9. *Malicious Aggregate Index:*

Malicious Aggregate Index performs the task of analysing if the target is malicious across multiple malicious artifact databases and further calculates and presents a number in a scale of 1 to 10 where 1 denotes that the target is least likely to be malicious and 10 denotes that the target is highly likely malicious according to multiple artifacts database.

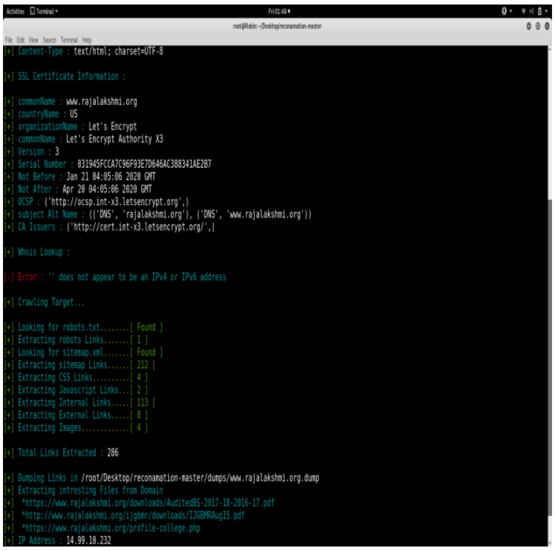
Process:

On execution of the handler which interacts with the UI components, the input (target) from the user is obtained, based on the nature of the target the handler invokes the model in a certain fashion based on the interoperability and relationship between these modules. Further the data obtained from the execution of either single or multiple interoperated modules is pushed into an elastic search which is a database cum search engine that allows further correlating data between the data points and also allows custom search and querying of the data.

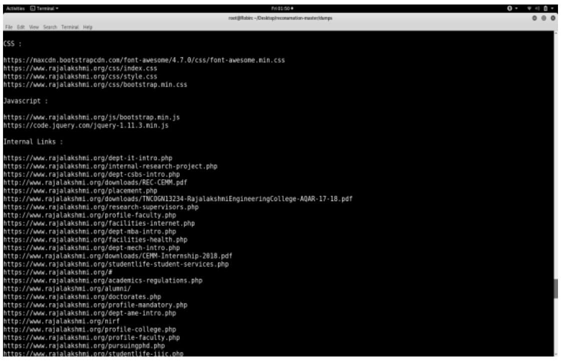
5. Sample Screenshot


Activities
Terminal
Fri 01:48
root@Robin: ~/Desktop/reconamation-master
File Edit View Search Terminal Help

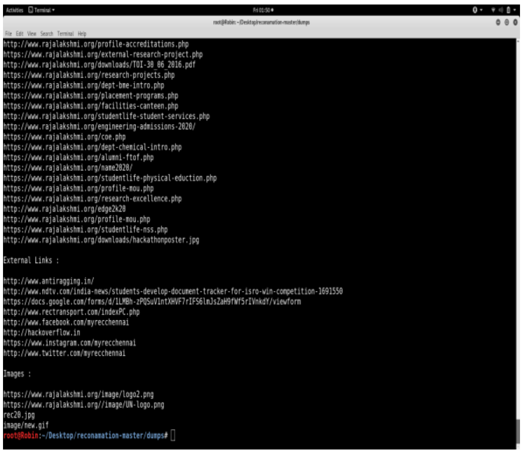
```
root@Robin:~/Desktop/reconamation-master# python3 reconamation.py
[+] Enter Target Endpoint

www.rajalakshmi.org
[+] Target : www.rajalakshmi.org


[+] Headers :


[+] Date : Thu, 02 Apr 2020 14:43:20 GMT
[+] Server : Apache/2.4.41 (Win64) OpenSSL/1.1.1c PHP/7.3.12
[+] X-Powered-By : PHP/7.3.12
[+] Keep-Alive : timeout=5, max=100
[+] Connection : Keep-Alive
[+] Transfer-Encoding : chunked
[+] Content-Type : text/html; charset=UTF-8

[+] SSL Certificate Information :

[+] commonName : www.rajalakshmi.org
[+] countryName : US
[+] organizationName : Let's Encrypt
[+] commonName : Let's Encrypt Authority X3
[+] Version : 3
[+] Serial Number : 031945FCCA7C96F93E7D646AC388341AE2B7
[+] Not Before : Jan 21 04:05:06 2020 GMT
[+] Not After : Apr 20 04:05:06 2020 GMT
[+] OCSP : ('http://ocsp.int-x3.letsencrypt.org',)
[+] subject Alt Name : (('DNS', 'rajalakshmi.org'), ('DNS', 'www.rajalakshmi.org'))
[+] CA Issuers : ('http://cert.int-x3.letsencrypt.org/',)

[+] Whois Lookup :

[-] Error : '' does not appear to be an IPv4 or IPv6 address

[+] Crawling Target...

[+] Looking for robots.txt........[ Found ]
[+] Extracting robots Links.......[ 1 ]
```


Activities
Terminal
Fri 01:48
root@Robin: ~/Desktop/reconamation-master
File Edit View Search Terminal Help

```
[+] Content-Type : text/html; charset=UTF-8

[+] SSL Certificate Information :

[+] commonName : www.rajalakshmi.org
[+] countryName : US
[+] organizationName : Let's Encrypt
[+] commonName : Let's Encrypt Authority X3
[+] Version : 3
[+] Serial Number : 031945FCCA7C96F93E7D646AC388341AE2B7
[+] Not Before : Jan 21 04:05:06 2020 GMT
[+] Not After : Apr 20 04:05:06 2020 GMT
[+] OCSP : ('http://ocsp.int-x3.letsencrypt.org',)
[+] subject Alt Name : (('DNS', 'rajalakshmi.org'), ('DNS', 'www.rajalakshmi.org'))
[+] CA Issuers : ('http://cert.int-x3.letsencrypt.org/',)

[+] Whois Lookup :

[-] Error : '' does not appear to be an IPv4 or IPv6 address

[+] Crawling Target...

[+] Looking for robots.txt........[ Found ]
[+] Extracting robots Links.......[ 1 ]
[+] Looking for sitemap.xml.......[ Found ]
[+] Extracting sitemap Links......[ 212 ]
[+] Extracting CSS Links..........[ 4 ]
[+] Extracting Javascript Links...[ 2 ]
[+] Extracting Internal Links.....[ 113 ]
[+] Extracting External Links.....[ 8 ]
[+] Extracting Images.............[ 4 ]

[+] Total Links Extracted : 286

[+] Dumping Links in /root/Desktop/reconamation-master/dumps/www.rajalakshmi.org.dump
[+] Extracting intresting Files from Domain
[+]  *https://www.rajalakshmi.org/downloads/AuditedBS-2017-18-2016-17.pdf
[+]  *http://www.rajalakshmi.org/ijgbmr/downloads/IJGBMRAug15.pdf
[+]  *https://www.rajalakshmi.org/profile-college.php
[+] IP Address : 14.99.10.232
```


Activities
Terminal
Fri 01:50
root@Robin: ~/Desktop/reconamation-master/dumps
File Edit View Search Terminal Help

```
CSS :

https://maxcdn.bootstrapcdn.com/font-awesome/4.7.0/css/font-awesome.min.css
https://www.rajalakshmi.org/css/index.css
https://www.rajalakshmi.org/css/style.css
https://www.rajalakshmi.org/css/bootstrap.min.css

Javascript :

https://www.rajalakshmi.org/js/bootstrap.min.js
https://code.jquery.com/jquery-1.11.3.min.js

Internal Links :

https://www.rajalakshmi.org/dept-it-intro.php
https://www.rajalakshmi.org/internal-research-project.php
https://www.rajalakshmi.org/dept-csbs-intro.php
http://www.rajalakshmi.org/downloads/REC-CEMM.pdf
https://www.rajalakshmi.org/placement.php
http://www.rajalakshmi.org/downloads/TNCOGN13234-RajalakshmiEngineeringCollege-AQAR-17-18.pdf
https://www.rajalakshmi.org/research-supervisors.php
http://www.rajalakshmi.org/profile-faculty.php
https://www.rajalakshmi.org/facilities-internet.php
https://www.rajalakshmi.org/dept-mba-intro.php
https://www.rajalakshmi.org/facilities-health.php
https://www.rajalakshmi.org/dept-mech-intro.php
http://www.rajalakshmi.org/downloads/CEMM-Internship-2018.pdf
https://www.rajalakshmi.org/studentlife-student-services.php
https://www.rajalakshmi.org/#
https://www.rajalakshmi.org/academics-regulations.php
http://www.rajalakshmi.org/alumni/
https://www.rajalakshmi.org/doctorates.php
https://www.rajalakshmi.org/profile-mandatory.php
https://www.rajalakshmi.org/dept-ame-intro.php
http://www.rajalakshmi.org/nirf
http://www.rajalakshmi.org/profile-college.php
https://www.rajalakshmi.org/profile-faculty.php
https://www.rajalakshmi.org/pursuingphd.php
https://www.rajalakshmi.org/studentlife-iiic.php
```


Activities
Terminal
Fri 01:50
root@Robin: ~/Desktop/reconamation-master/dumps
File Edit View Search Terminal Help

```
http://www.rajalakshmi.org/profile-accreditations.php
https://www.rajalakshmi.org/external-research-project.php
http://www.rajalakshmi.org/downloads/TOI-30_06_2016.pdf
https://www.rajalakshmi.org/research-projects.php
https://www.rajalakshmi.org/dept-bme-intro.php
https://www.rajalakshmi.org/placement-programs.php
https://www.rajalakshmi.org/facilities-canteen.php
http://www.rajalakshmi.org/studentlife-student-services.php
http://www.rajalakshmi.org/engineering-admissions-2020/
https://www.rajalakshmi.org/coe.php
https://www.rajalakshmi.org/dept-chemical-intro.php
https://www.rajalakshmi.org/alumni-ftof.php
https://www.rajalakshmi.org/name2020/
https://www.rajalakshmi.org/studentlife-physical-eduction.php
https://www.rajalakshmi.org/profile-mou.php
https://www.rajalakshmi.org/research-excellence.php
http://www.rajalakshmi.org/edge2k20
http://www.rajalakshmi.org/profile-mou.php
https://www.rajalakshmi.org/studentlife-nss.php
http://www.rajalakshmi.org/downloads/hackathonposter.jpg

External Links :

http://www.antiragging.in/
http://www.ndtv.com/india-news/students-develop-document-tracker-for-isro-win-competition-1691550
https://docs.google.com/forms/d/1LMBh-zPQSuV1ntXHVF7rIFS6lmJsZaH9fWf5rIVnkdY/viewform
http://www.rectransport.com/indexPC.php
http://www.facebook.com/myrecchennai
http://hackoverflow.in
https://www.instagram.com/myrecchennai
http://www.twitter.com/myrecchennai

Images :

https://www.rajalakshmi.org/image/logo2.png
https://www.rajalakshmi.org//image/UN-logo.png
rec20.jpg
image/new.gif
root@Robin:~/Desktop/reconamation-master/dumps#
```

6. Conclusion and Future Scope

The application can be extended by implementing continuous monitoring and periodic assessment of the target domain to keep track of the variation of the data from various data points, this assists the soc and threat hunting/ intelligence teams in large organizations to continuously monitor for any alarming events. Further providing a dashboard with customizable visualization patterns and graphs provides much more insights of the data gathered than browsing through huge dumps of data.

To conclude, reconnaissance and osint which are the crucial parts not only in penetration testing or security assessments but also they act in helping the defensive security teams (Blue) to safeguard your company's information, henceforth it indispensable to have an automation tool to execute all the tedious manual efforts involved in gathering data and further providing higher accuracy rate.

References

[1] https://www.hackerone.com/blog/how-to-recon-and-content-discovery

[2] Open Source Intelligence Methods and Tools: A Practical Guide to Online Intelligence by NihadA.Hassan, Rami Hijazi

[3] Open Source Intelligence Techniques: Resources for Searching and Analyzing Online Information by Michael Bazzell

[4] Web Mining for Open Source Intelligence

[5] Open source intelligence base cyber threat inspection framework for critical infrastructures

[6] https://medium.com/bugbountywriteup/whats-tools-i-use-for-my-recon-during-bugbounty-ec25f7f12e6d

[7] https://www.hackerone.com/blog/how-to-recon-and-content-discovery