International Journal of Modern Agriculture, Volume 10, No.2, 2021
ISSN: 2305-7246

PASSIVE ENUMERATION IN WEB APPLICATION SECURITY

Abdul Muqsit Baig[1], Kishori Kasat[2]

[1,2]Symbiosis Institute of Computer Studies and Research,
Symbiosis International (Deemed University), Pune, India
[2] kishori.kasat@sicsr.ac.in

Abstract

In an age where websites need to be properly secured in order for them to avoid being a part of a major breach, Reconnaissance plays an important role in both attacking as well as defending a website. Everyone want to defend and secure assets as much as possible to avoid a possible cyber-attack on our organization.

The "reconnaissance" phase is the first phase of an attack. The main aim of research is to study the organization or an asset to defend which is also known as target and find out as much information about it as possible. From the most common details to the smallest ones, the study is required to note down every possible thing about target. This information will help to map out target and possibly find flaws in the Business structure or infrastructure.

This paper studies different passive techniques available for a pen tester or internal security team of an organization to map out the public facing assets or endpoints of the target application. The first step an Attacker would perform, when attacking an organization, would be to map out the organization's entire infrastructure and finding out as much information about their target as possible without engaging actively with the target. This paper tries to cover different techniques an attacker might use to find sensitive information about the target which is usually available publicly. These techniques are well known in the Information Security field and would try to help the defenders to be a step ahead of the attackers

Key words: Reconnaissance; penetration testing; passive asset mapping; security; web application security

Introduction

The main aim of what security tries to achieve is: *confidentiality, integrity, and availability* CIA – Confidentiality, Integrity, Availability (Uma & Ganapathi, 2013). This can be achieved by using techniques like deploying a firewall. A firewall mainly provides security by blocking or filtering packets, which are ingress/incoming to the network. This is achieved by applying policies or rules, which define what needs to be done for a specific packet when it is detected by the firewall.

The firewall may ACCEPT or REJECT such packets depending upon the policy. However, if an organization focuses more on the performance of it's product then this is where a firewall falls behind.

What is Reconnaissance (or Recon for short)? It is the art of knowing everything about your target before you carry out an attack. The first stage before a hacker attacks its target is finding out everything about the target (Bacudio, Yuan, Chu, & Jones, 2011). A hacker would usually identify a target he wants to attack and then find all information about the target. Usually in this phase, the attacker finds out different vulnerabilities and threats to the target (Samant, 2011).

Provide an understanding to the owner of an asset of what they might be exposing to the internet and if it is intentional. From production or staging environments to long forgotten endpoints that have been indexed by crawlers like Google, etc.

Materials and methods/Methodology

Existing articles and current knowledge focus on either width or depth, but not both. As discussed in Discovering Subdomains (Bugcrowd, n.d.) and Let's Recon (OWASP, n.d.), this paper rather focuses on both – depth as well as width of the subdomain enumeration phase. Not limited to this, we will further discuss how we can avoid manually scraping the different techniques described in this paper by automating them to an extent. This paper focuses on an in-depth understanding upon the enumeration phase (as discussed in this paper ahead) as well as all the minute details and protocols playing part in the enumeration phase.

This paper leans on passive side of the enumeration phase, where we will not be interacting or sending any HTTP Requests to the target. All we will be doing is gathering subdomains from as many different passive sources as possible. The problem with this is the lack of automation in doing this work and manually performing these techniques from every single source and then proceeding to merge the results.

In relation to this paper, researchers have decided to introduce a tool in python, namely **"subfetch"**. This tool solves the problem of manually scraping the web for results by automating this task.

To use **"subfetch",** you simply input the name of your target's **main domain** (*facebook.com for example*) and **Shodan's API Key** and it will automatically query the respective API's and fetch all subdomains for the same domain.

"**Subfetch"** is in its initial stages and at the time of writing this paper. The only sources which are scraped are:

- Crt.sh
- Shodan

It is also worth mentioning that this tool is designed for a Linux environment specifically. This is because the results obtained from **subfetch** can be directly piped into other tools on the *Linux Command Line.*

Reconnaissance

Recon is the art of knowing your target before you carry out an attack. This may include studying the entire target system, locate and map the entire network and all system devices which are connected to that network or services available on those systems via different ports, and vulnerabilities in those systems which we are testing (Uma & Ganapathi, 2013).

The first phase of an attack is the "reconnaissance" phase (*Rossi, n.d.*)

The collection of all such data and information about our target, may lead to a Denial of Service (DoS) attack. Scanning the system to map the network is usually one of the steps carried out by pinging all available devices and then finding out which IP addresses are active and which are responsive. Reconnaissance plays a vital role in both attacking as well as defending a website (*Li & Xue, n.d.)*

From all this information, a hacker finds out which ports are open on which IP addresses and then which services are further running on which ports.

Types of Reconnaissance

Active Recon: In this type of recon, an attacker actually interacts with the target. This type of recon is much faster because it is manual mostly and more accurate. Again, it makes a lot more noise. Because we actively interact with the target to obtain all recon information, we are more likely to be caught in one of the firewalls.

Passive Recon: Passive recon usually requires no zero or no interaction with the target, so it is less likely to be detected by a firewall. Although that said, the information by using this type of recon is not totally accurate and is much slower.

Focus

The focus of this research paper would revolve around passive reconnaissance techniques. An attacker usually views the target's traffic and their publicly available assets to get more information and then studies and monitors the target based on this information.

Passive recon is a way to get information about our target organization or company without actively engaging and interacting with their systems. Both active and passive recon are a very important phase before carrying out a successful attack. The quality of the recon done then further makes it easy for carrying out a successful attack on the target occurs.

DNS

There are thousands and thousands of websites on the internet today. Again, each website has plenty of its own subdomains. If we did not have DNS we would have to remember the IP address to each website, we wanted to visit. Said that, it is practically impossible to remember the IP address of these websites. Hence, DNS makes it possible to enter a domain name in our own language using a particular URL format, and access the website.

The way this works is, you enter a domain name. Your browser then forwards the request to a DNS server to "resolve" this domain and returns an IP address back to us. Then using that IP, our request is forwarded to the server we want to reach. Such a system is called a DNS or Domain Name System.

A DNS database is a very popular, widespread database which contains mappings of domain names to their respective IP address

a. DNS Lookup

As we know that when we enter a website name, our request is first forwarded to a DNS server. So each DNS server contains mappings of different domains as well as sub-domains to their IP addresses. Hence, if we look closely, there are Root Servers of these DNS Servers, which contain IP addresses of the servers, which host the TLD or Top Level Domains.

The job of Root Servers is to help to access the servers that contain the database of the IP address of the domain name queried by the user. This is where a DNS lookup comes into picture. A DNS Lookup is the process of finding the IP address of a particular URL on the internet.

Enumeration

Enumeration is one of the initial phases of Reconnaissance. An attacker tries to enumerate (find out) all **assets** owned by the target organization. In the next sections, we will think from the point of view of an attacker and how an attacker might approach a target. This will help to identify assets, which should not be publicly available, and hopefully remediate issues related to the organization.

For the first steps, our goal is to identify and enumerate all domains/subdomains belonging to the target organization. We will achieve this in two ways:

1. Horizontal Enumeration

2. Vertical Enumeration

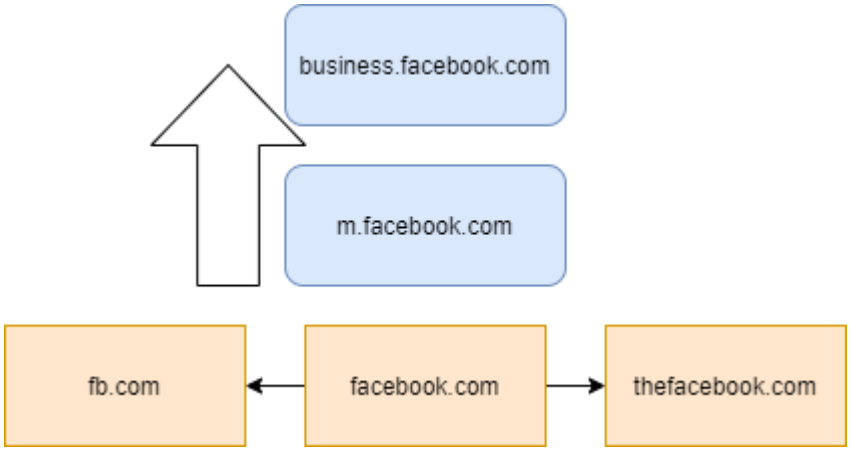


Figure 1: Enumeration workflow

1. Horizontal Enumeration

As shown in the above flowchart, horizontal enumeration is the phase of moving horizontally and finding out all **root domains** belonging to the target. In the above example, researchers have picked "facebook.com" as potential target for example purposes.

Facebook.com has plenty of root domains under its belt. In the above flowchart we can see that we have identified "fb.com" and "thefacebook.com" in the Horizontal Enumeration phase.

The importance of horizontal enumeration is to potentially expand an attack surface as an attacker. When one has a large attack surface to look at, more functionalities and more vulnerabilities are there. Also, this phase of enumeration will help to hopefully identify assets which have been forgotten by the organization's security team or have less security measures than the main target domain.

2. Vertical Enumeration

Vertical Enumeration, as shown in the above flowchart, is the phase of finding **subdomains** for a given root domain/host. This phase focuses on finding second-level domain names for the given domain.

As we are interested in increasing our attack surface as much as possible, it is important to find subdomains of domains and assets identified in the Horizontal Enumeration phase.

There are plenty of passive techniques to find subdomains for a target online. A few of them will be discussed in this research paper in the following sections.

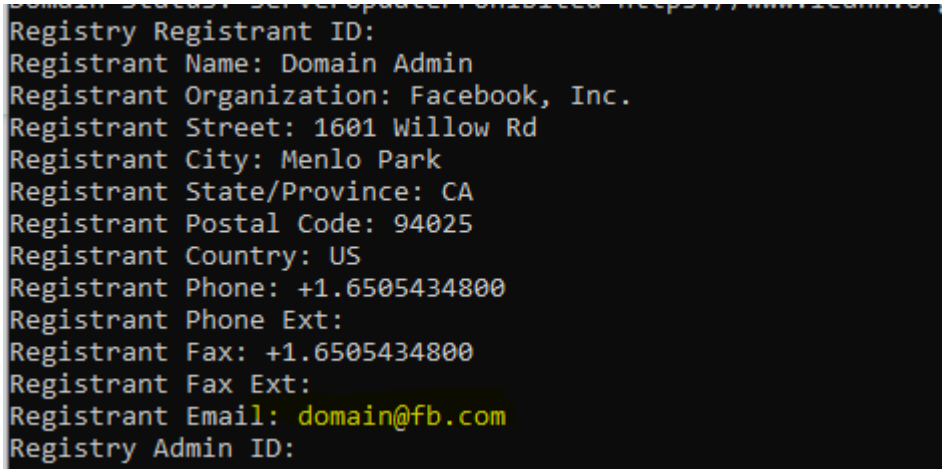
Horizontal Enumeration Techniques

As discussed in the above subsection, here researchers will discuss different passive techniques to horizontally enumerate our target and gain more assets belonging to target. Researchers have picked target as "facebook.com" for example purposes of this research paper.

1. Reverse WhoIS Lookup

This is a widely used way of finding domains owned by an organization. For this, one has to simply do a **whois lookup** of the main domain i.e. *facebook.com* in our case. One can use the "whois" tool in Ubuntu *(sudo apt install whois).*

Locate the "Registrant Email" field and note down the email address.



The email address of the Registrant is "**domain@fb.com**". Next visit any reverse DNS Lookup website. One such website is:

https://viewdns.info

Now paste the email address in the "Reverse Whois Lookup". It will return you with a list of domains belonging to facebook.com and registered by the Registrant Email: "domain@fb.com".

2. ASN Lookup

Autonomous system number or **ASN** for short is a unique number assigned to an autonomous system (AS) by the Internet Assigned Numbers Authority (**IANA**).

An AS consists of blocks of IP addresses that have a particular policy for accessing external networks and are administered by an organization.

This is another way of finding different domains belonging to an organization. You simply gather all ASNs (autonomous system number) belonging to an organization. Simply go to:

https://bgp.he.net

Now type the organizations name, "facebook" and get a list of all ASN belonging to facebook. Now using these ASNs go to the below website:

https://www.ultratools.com/tools/asnInfo

Enter the ASNs here and retrieve the list of hosts assigned to that AS.

Vertical Enumeration Techniques

The main aim of vertical enumeration is to find the subdomains or 2nd level domains of the main domain. So for facebook.com an attacker would go ahead and make a list of all its subdomains i.e. m.facebook.com, business.facebook.com, etc.

There are plenty of techniques to enumerate subdomains of a domain. As mentioned earlier, researchers would be focusing on the passive ones in this research paper.

1. Certificate Transparency Logs

This is arguably the most efficient and reliable way of finding out the subdomains of a domain/host. When an organization deploys a new host/website on the internet, they also install an SSL certificate to it.

Certificate Transparency is an open framework that monitors and audits TLS/SSL certificates. Certificate transparency logs are a way to maintain the records of all publicly trusted digital certificates. These certificates contain all kinds of information – public key, information about the issuer, organization, etc.

Attackers take advantage of these Certificate transparency logs to retrieve assets which have been assigned an SSL certificate from a particular organization. One such website is,

https://crt.sh

Simply go to the above website and type the domain name for which you want to retrieve subdomains for. It will return you a list of hosts and additional information about their SSL/TLS certificates.

Certificates

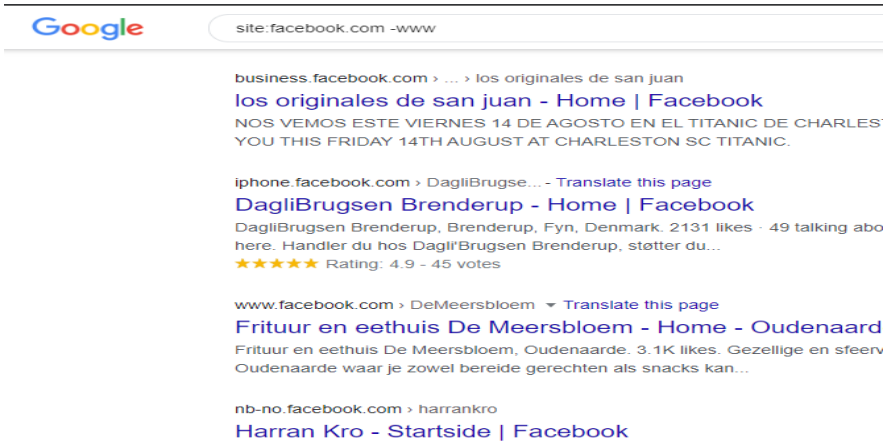
crt.sh ID	Logged At	Not Before	Not After	Matching Identities	Issuer Name
3205981337	2020-08-08	2020-08-08	2020-11-06	*.secure.facebook.com secure.facebook.com	C=US, O=DigiCert Inc, OU=www.digicert.com, CN=DigiCert SHA2 High Assurance Server CA
3192156486	2020-08-04	2020-06-04	2020-09-01	*.up.facebook.com up.facebook.com	C=US, O=DigiCert Inc, OU=www.digicert.com, CN=DigiCert SHA2 High Assurance Server CA
3190726759	2020-08-04	2020-08-04	2020-11-02	presto.vip.facebook.com www.presto.vip.facebook.com	C=US, O=DigiCert Inc, OU=www.digicert.com, CN=DigiCert SHA2 High Assurance Server CA
3173471350	2020-08-01	2020-08-01	2020-10-29	*.v6.facebook.com v6.facebook.com	C=US, O=DigiCert Inc, OU=www.digicert.com, CN=DigiCert SHA2 High Assurance Server CA
3167073033	2020-07-30	2020-07-30	2020-10-28	*.up.facebook.com up.facebook.com	C=US, O=DigiCert Inc, OU=www.digicert.com, CN=DigiCert SHA2 High Assurance Server CA
3165810519	2020-07-30	2020-07-30	2021-07-31	*.f-facebook.com f-facebook.com	C=US, O=DigiCert Inc, OU=www.digicert.com, CN=Encryption Everywhere DV TLS CA - G1
3143733640	2020-07-26	2020-07-21	2020-10-12	*.facebook.com facebook.com *.m.facebook.com	C=US, O=DigiCert Inc, OU=www.digicert.com, CN=DigiCert SHA2 High Assurance Server CA
3135574051	2020-07-24	2020-07-24	2020-10-22	llama-ztp.corp.facebook.com	C=US, O=DigiCert Inc, OU=www.digicert.com, CN=DigiCert SHA2 High Assurance Server CA
3119819734	2020-07-21	2020-07-21	2020-10-19	*.facebook.com facebook.com *.m.facebook.com	C=US, O=DigiCert Inc, OU=www.digicert.com, CN=DigiCert SHA2 High Assurance Server CA
3119786038	2020-07-21	2020-07-21	2020-10-12	*.facebook.com facebook.com *.m.facebook.com	C=US, O=DigiCert Inc, OU=www.digicert.com, CN=DigiCert SHA2 High Assurance Server CA
3104949769	2020-07-18	2020-07-09	2020-10-07	*.z.facebook.com z.facebook.com	C=US, O=DigiCert Inc, OU=www.digicert.com, CN=DigiCert SHA2 High Assurance Server CA
3100769402	2020-07-17	2020-07-07	2020-10-05	*.alpha.facebook.com alpha.facebook.com *.fb.alpha.facebook.com fb.alpha.facebook.com *.fb.m.alpha.facebook.com	C=US, O=DigiCert Inc, OU=www.digicert.com, CN=DigiCert SHA2 High Assurance Server CA

2. Google dorks

Another easy and quick way to find out the subdomains for a host is to simply search on google. It's as easy as that. But, instead of normally searching on google, we will use "Google Dorks" or "Google Hacking" which are nothing but special queries which Google understands.

Attackers are well versed with different types of google dorking techniques, which gives them an edge over the defending team. To find subdomains using this method we simply enter the below query into google.com,

site: facebook.com –www



3. Shodan

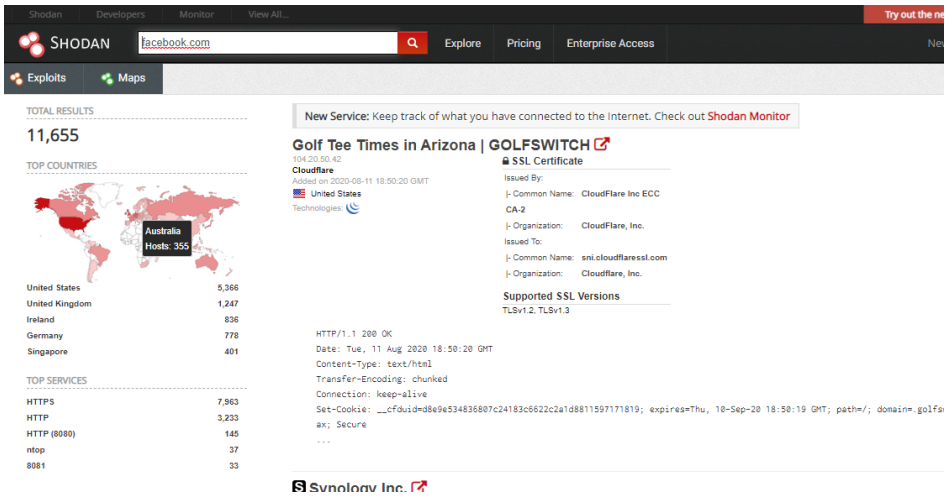
Shodan is well known among hackers. It is the "hackers search engine". Shodan lets you find specific types of

computers, routers, servers, etc. Connected to the internet. Shodan continuously scans the internet for these types of devices and maintains a database.

Shodan can be used to find assets which were either forgotten by the developers or they don't maintain it anymore. A simple Shodan search gives out thousands of results. Simply visit,

https://shodan.io/

One can either use "dorks" as discussed above or just type "facebook.com" as shown below:



As you can see, Shodan returned with almost **12k results** belonging to **facebook.com**

There are many ways how Shodan can be used. This research paper just gives an overview of this service.

4. Subfetch

As mentioned above, to solve the problem of automating the tasks discussed in this paper, "**subfetch**" was introduced. This tool is a simple python-based tool to take advantage of the APIs provided by the sources discussed above (crt.sh, Shodan, etc.).

This tool currently has only two sources:

- Crt.sh
- Shodan

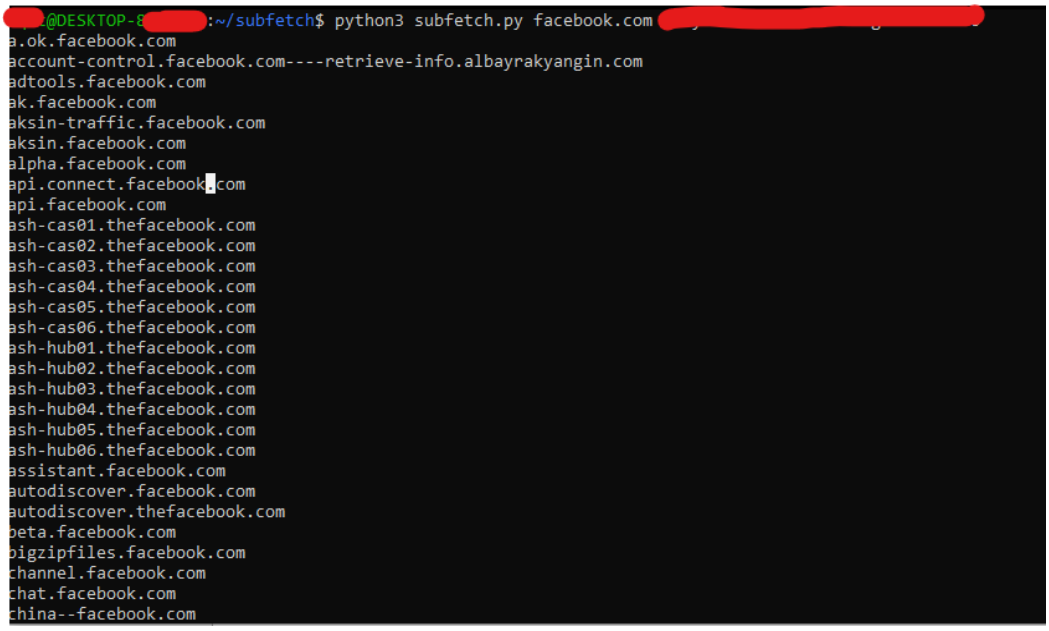
More sources will be added gradually to increase the results obtained by passively scraping the web.

To use this tool you would need to sign up at https://shodan.io and get an *API Key*. After you have done that, you can simply run this tool in a Linux based environment (preferable). Once everything is setup, you can use the below command to retrieve subdomains from the sources mentioned above:

python3 subfetch.py <target.com> <API_KEY>

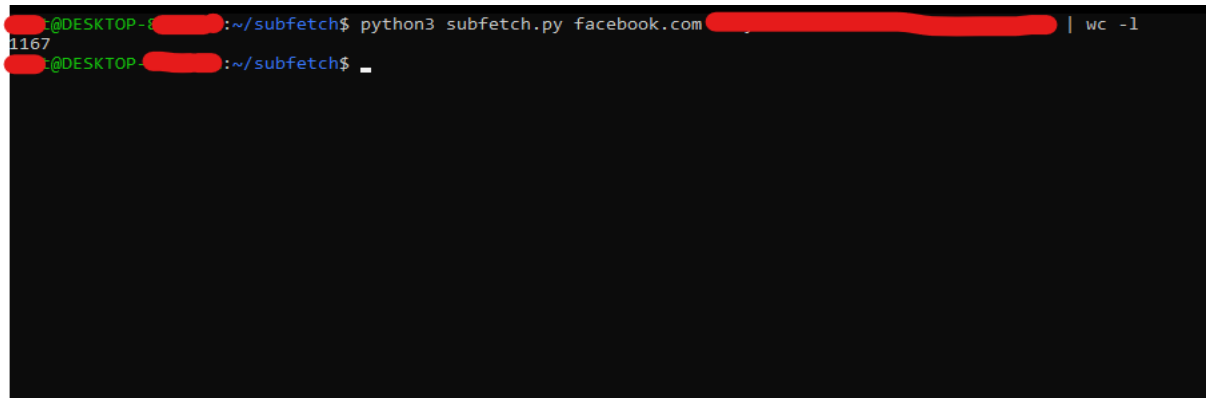
Below is an example to retrieve results for "*facebook.com*":

python3 subfetch.py facebook.com <API_KEY>



As mentioned above, this tool simply outputs a list of subdomains so that it is easier to pipe this output to another command. Hence, to check how many results are retrieved one can pipe the output from **subfetch** to the **wc** command:

python3 subfetch.py facebook.com <API_KEY> | wc -l



We have 1167 unique subdomains for *facebook.com* from this tool. As and when more sources are added to this tool, there will be more results.

Conclusion:

The techniques discussed above are just the tip of the iceberg when it comes to passive-ish reconnaissance techniques, but these techniques usually yield the maximum results compared to any other sources on the internet.

Researchers start off by understanding DNS and how DNS lookup works as well as gaining knowledge about how reconnaissance. This gives more understanding about the techniques which are followed after this section.

Certificate Transparency logs are the best sources of finding and mapping SSL/TLS certificates belonging to an organizations asset.

Shodan is also a flexible service which scans the entire internet and takes advantage of Certificate transparency along with other components (e.g open ports, services on these ports, etc) Hence, we've mentioned Shodan in this paper as it is very reliable while writing this paper. There are many more sources of finding different types of assets belonging to a particular organization which are out of the scope of this paper.

This paper introduces a small tool [**subfetch**] which helps to automate the manual task of enumerating subdomains in a way. This tool is in its early stages, hence, as more sources are added to, it will yield even more results.

Subfetch simply queries the APIs of these sources and scrapes subdomains after filtering out the duplicate ones. It outputs a list of subdomains without printing any extra output so that this list of subdomains can then be piped to other commands or stored. A person would require to have basic Linux CLI knowledge for this.

To wrap it up, the main aim of this paper is to focus more on the passive sources of information available on the internet. These sources give us the maximum output without us even interacting with our target. These techniques are heavily used by penetration testers (or pentesters) as well as hackers to gain intel about their target

References

1. Bacudio, A. G., Yuan, X., Chu, B. T., & Jones, M. (2011). An Overview of Penetration Testing. International Journal of Network Security & Its Applications.

2. BGP Tools. (n.d.). Retrieved from BGP: https://bgp.he.net/

3. Bugcrowd. (n.d.). Discovering subdomains. Retrieved from Bugcrowd: https://www.bugcrowd.com/blog/discovering-subdomains/

4. Certificate Transparency Logs. (n.d.). Retrieved from crt.sh: https://crt.sh

5. Hasan, A., & Meva, D. (n.d.). Web Application Safety by Penetration Testing.

6. Li, X., & Xue, Y. (n.d.). A Survey on Web Application Security.

7. OWASP. (n.d.). Let's Recon. Retrieved from OWASP: https://owasp.org/www-chapter-coimbatore/assets/files/Lets%20Recon.pdf

8. Rossi, B. (n.d.). Information Age. Retrieved from 7 steps hackers take to execute a succesful cyber attack: https://www.information-age.com/7-steps-hackers-take-execute-successful-cyber-attack-123460872/

9. Samant, N. (2011). AUTOMATED PENETRATION TESTING. SJSU ScholarWorks.

10. Shodan. (n.d.). Retrieved from https://shodan.io

11. Ultratools. (n.d.). Retrieved from https://www.ultratools.com/tools/asnInfo

12. Uma, M., & Ganapathi, P. (2013). A survey on various cyber-attacks and their classification. International Journal of Network Security.