CHAPTER

4

Internet Information Gathering

Never underestimate the importance of the information gathering phase in penetration testing. I admit that I used to underestimate it myself, but over the years, I have realized how vital this phase can be. Once I was working on a project that was not yet deployed into the production environment, so practically speaking, there was no information yet on the internet, right? Out of curiosity, I entered the test environment URL on Google, and it turned out that one of the developers accidentally copied the internal network URLs to GitHub. That's just one example of the horror stories that I have witnessed during my career. Speaking of horror stories, one of them happened with a company out there. The developer pushed to GitHub the credentials of the AWS cloud host, and a hacker took advantage of this and connected remotely to the server. Of course, you can guess the rest.

The focus of this chapter is on the primary methodology of the penetration testing phase. You shouldn't run scanners blindly without learning what you're looking for. One of the steps that we already discussed in the previous chapter is the search for subdomains. This task is part of passive information gathering, too (if you use the web as a data source to get your results). You can go back to the previous chapter if you need a refresher.

Here's what you will learn in this chapter:

- Use internet search engines to get your results
- Use Shodan

Kali Linux Penetration Testing Bible. Gus Khawaja

- Use Google queries
- See how to display information about domains using the Whois database
- See how the essential tools for passive footprinting work on Kali, including TheHarvester, Dmitry, and Maltego

Passive Footprinting and Reconnaissance

Let's define some terminology about this subject. Gathering information passively using the internet's public information has so many technical names. Some people call it *reconnaissance*, and others call it *passive footprinting* or *passive information gathering*. Sometimes you will hear the name OSINT, which stands for *open source intelligence*. Use any technical word you prefer, but make sure not to be confused with all these terms since they all mean the same thing: collecting data about your target using publicly available sources. Do not get confused about the different between *footprinting* and *fingerprinting* because the latter is used for identifying the operating system.

One of the most critical tasks in passive footprinting is to know what you're looking for. There are tons of tools that get the job done, but you need to understand what they do in the background. So, here are the items that you need to look for while executing your task of information gathering:

- Company subdomains
- Websites
- Public IP addresses (including the one on the cloud AWS/Azure)
- Leaked internal IP addresses
- Public DNS records (MX mail records, etc.)
- Leaked credentials (mainly on GitHub or Pastebin)
- Previous breaches
- Significant business change (e.g., acquisitions)
- Business financial information (this can reveal a secret partner)
- Business phone numbers (for social engineering)
- Employee public information (for social engineering)
- A company presence on social media (e.g., LinkedIn)

Internet Search Engines

The public search engines are your entry into the search for weaknesses linked to your target. In the next section, you will see in detail how to take advantage

of the Google search engine to get all the necessary information (leaks). That being said, there are plenty of search engines you can use besides Google. Those search engines will point you in the right direction after you tell them what you are looking for. They are a gold mine that will reveal what is leaking out there in the wild about your target. Here's a list of search engines that you should add to your arsenal kit:

- **Google search engine**: `google.com`
- **Shodan online scanner**: `shodan.io`
- **DuckDuckGo search engine**: `duckduckgo.com`

Shodan

Shodan is a great online tool that will scan the internet for you. There are limitless options that you can use on this monster. Here's a quick example: let's say you want to find Docker engines that are visible on the internet and listening on port 2375 (the default port number for a Docker daemon).

In the example shown in Figure 4.1, I used the following query:

```
port:2375 product: "Docker"
```

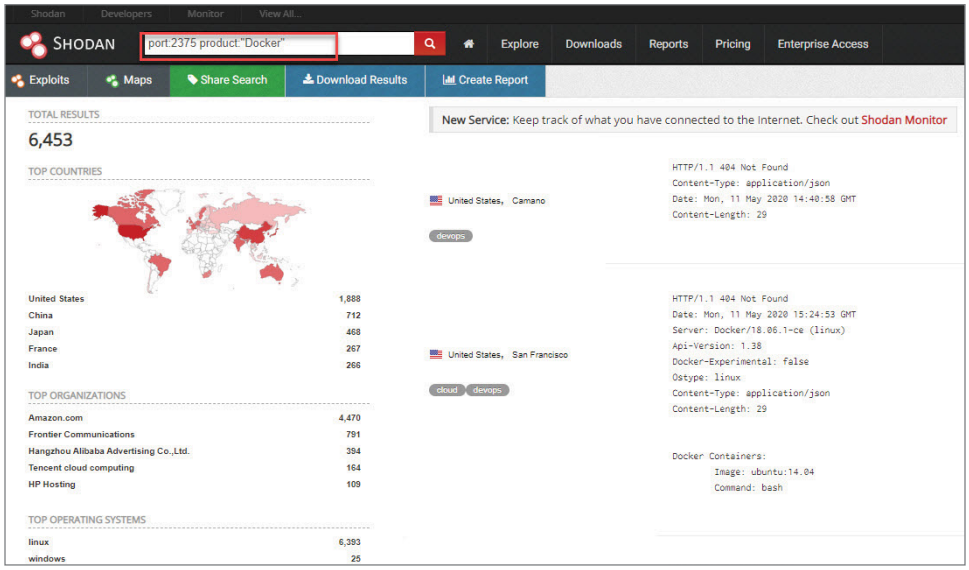

Figure 4.1: Shodan

This is not the only query criteria that you can use on the Shodan site. There are others too; Table 4.1 covers the most popular ones.

Table 4.1: The Most Common Query Criteria Used on the Shodan Site

SEARCH FILTER	DESCRIPTION	EXAMPLE
Port	Port number	`Port:80`
Product	Product name	`Product: "Apache"`
Org	Organization name	`Org: "Target company name."`
Country	Two-letter country name	`Country:CA`
City	City name	`City:Montreal`
hostname	Domain name	`Hostname: "domain-name.com"`
Server	Server name	`Server: "Linux"`
http.title	Web page title	`http.title:"Dashboard"`

You can check out some practical examples at `www.shodan.io/explore`.

Google Queries

Google is a powerful search engine that will allow you to find the information that you're looking for (e.g., leaked credentials about your client/employer) if you know how to use it properly. Google gives you the ability to query its database with advanced filter criteria. Some people call it the *Google hacking database* (GHDB), while others call it *Google dorks*.

Let's start with the first `site:` query that I always use at the beginning of my engagement, as shown in Figure 4.2. This specific query will allow you to look for all the web pages and sites associated with your target domain name (this query will reveal subdomains as well).

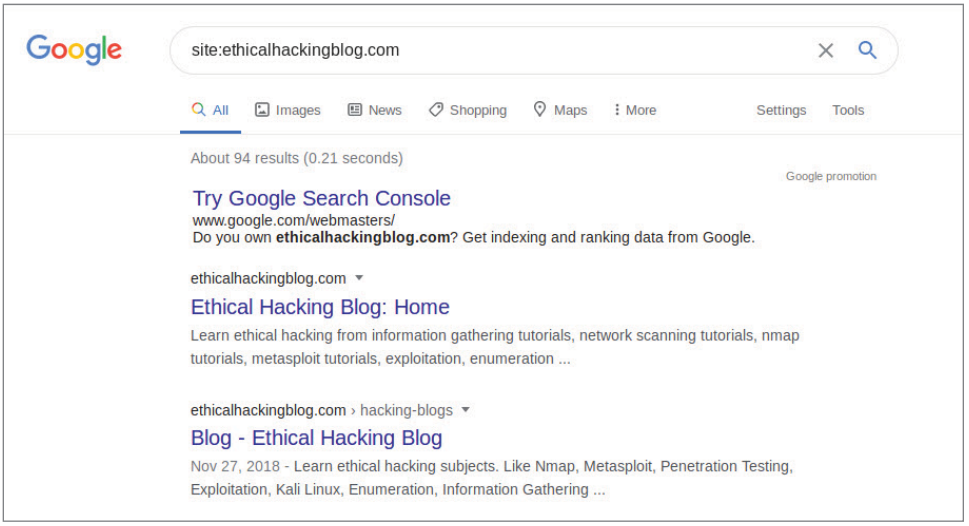

Figure 4.2: Google Dork Site Filter

These days everything is published publicly on GitHub. Here's an interesting query that you can use on Google to look for juicy information posted on GitHub. Figure 4.3 shows a few of the results the query pulled when searching `GitHub.com` using the keywords `Gus Khawaja`:

```
site:github.com [keywords]
```

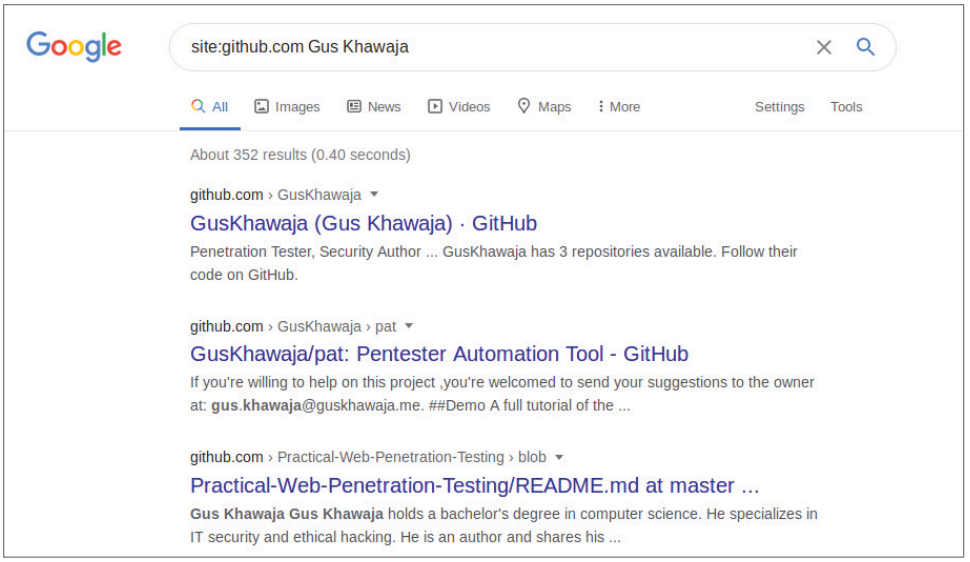

Figure 4.3: Google Dork Site Filter with Description

That's just a simple example, but you will be surprised by the number of times you'll discover leaked information on GitHub during an engagement. Developers tend to use GitHub without understanding the consequences, and you, as a professional in this field, can take advantage of this flaw. Table 4.2 lists some other exciting queries that you can use in Google.

Table 4.2: Google Dorks Common Queries

QUERY	DESCRIPTION	EXAMPLE
`inurl:[criteria]`	Search for text inside a URL	Search for SQLi candidates in a website: `site:[domain] inurl:?id=`
`intitle:[criteria]`	Search for text inside the title of a web page	Search for CCTV: `intitle:"index of" "cctv"`

Continues

Table 4.2 (*continued*)

QUERY	DESCRIPTION	EXAMPLE
`filetype:[file extension]`	Search file types by using their extensions	Search for files connected to a domain that belongs to your target: `Site:[domain] filetype:"xls | xlsx | doc | docx | ppt | pptx | pdf"`

You can use the GHDB queries on Exploit-db to visualize the latest ideas. Check out `www.exploit-db.com/google-hacking-database`, as shown in Figure 4.4. (Exploit-db belongs to the Offensive Security team, the founders of Kali Linux.)

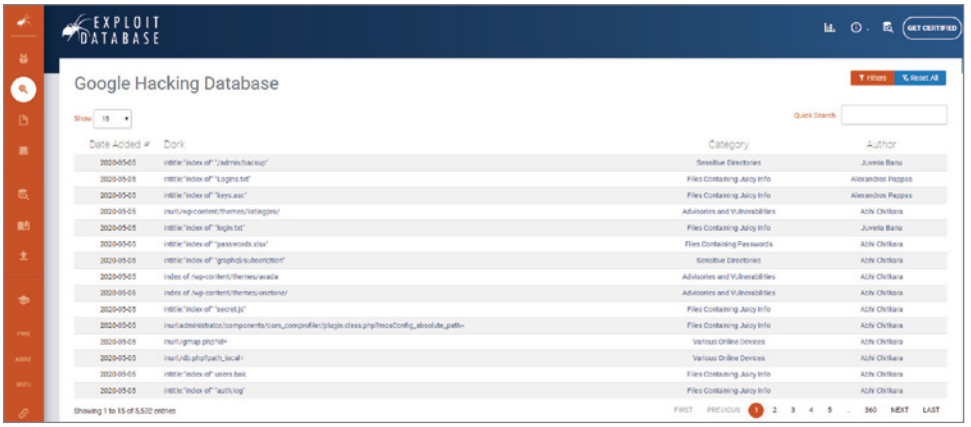

Figure 4.4: Google Hacking Database

Information Gathering Using Kali Linux

There are many tools preinstalled on Kali Linux that you can use for passive information gathering (Figure 4.5).

A lot of tools have common functionalities, some of which are completely free (e.g., Dmitry), while others have limited access to features unless you paid a yearly subscription (e.g., Maltego). In the upcoming sections, you will learn about the typical applications you can use for passive scanning. Still, the primary purpose of this chapter is to show you the fundamentals so that you don't use these scanners blindly without understanding their goals.

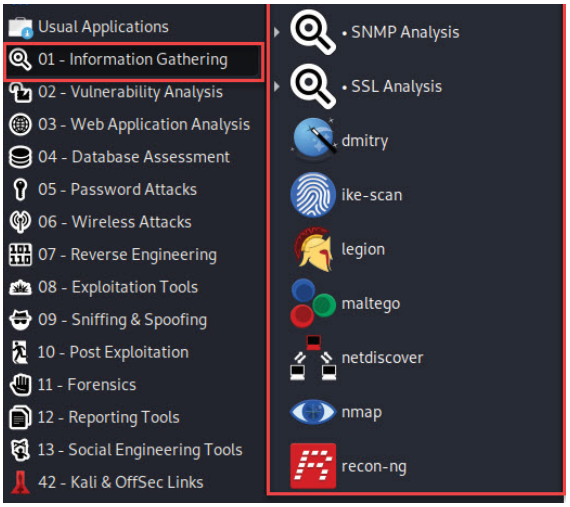

Figure 4.5: Kali Menu – Information Gathering

Whois Database

Every entity (a company or a single person) who buys a domain name is required to enter its personal information prior to the registration process. Most of this information will be published publicly in the Whois database. Now don't get too excited about this because domain providers (e.g., GoDaddy) will allow you to secure your personal information on the Whois database by charging fees (this is what I did for my blog website `ethicalhackingblog.com`). To use the Whois database weakness, you can use the Whois command on your Kali Linux.

```
$whois [domain]
```

Here's a sample of the `whois` query output (the following information is fictitious):

```
root@kali:~# whois [domain-name.com]
  Domain Name: acme.com
  Registry Domain ID: 2425408_DOMAIN_COM-VRSN
  Registrar WHOIS Server: WHOIS.test.COM
  Registrar URL: WWW.test.COM
  Updated Date: 2020-03-20T07:43:10.00Z
  Creation Date: 1991-04-17T04:00:00.00Z
  Registrar Registration Expiration Date: 2021-04-18T04:00:00.00Z
  Registrar: Ecorp, INC.
  Registrar IANA ID: 48
  Domain Status: clientTransferProhibited https://www.icann.org/
epp#clientTransferProhibited
```

Continues

(continued)

```
  Registrant Name: John Doe
  Registrant Organization: Ecorp Inc
  Registrant Street: 1234 Coney Island
  Registrant City: Brooklyn
  Registrant State/Province: NY
  Registrant Postal Code: 888999
  Registrant Country: US
  Registrant Phone:
  Registrant Phone Ext:
  Registrant Fax:
  Registrant Email: https://tieredaccess.com/contact/4355f620-51f6-44cc-
bab5-cda7d58313c4
  Admin Name: Mr. Robot
  Admin Organization: ECorp Inc
  Admin Street:
  Admin Street:
  Admin City:
  Admin State/Province:
  Admin Postal Code:
  Admin Country:
  Admin Phone:
  Admin Phone Ext:
  Admin Fax:
  Admin Email:
  Tech Name:
  Tech Organization:
  Tech Street:
  Tech Street:
  Tech City:
  Tech State/Province:
  Tech Postal Code:
  Tech Country:
  Tech Phone:
  Tech Phone Ext:
  Tech Fax:
  Tech Email:
  Name Server: DNS.ECORP.COM
  Name Server: NS1.ECORP.COM
  Name Server: NS2.ECORP.COM
  DNSSEC: unsigned
```

In summary, here's what Whois will reveal as public information:

- Registrant name
- Contact phone number
- E-mail address
- Entity physical address
- Domain expiry date
- NS (Name Servers) servers

TheHarvester

Finding e-mail addresses is crucial if you're going to conduct a social engineering attack or a phishing exercise. It's not hard to figure out the e-mail addresses if you're an insider since you already know what the e-mail format looks like (e.g., `firstname.lastname@domain.tld`).

There are many tools out there in the wild for this purpose. One is an old tool that's still doing a great job; it's called theHarvester. This tool will use popular search engines like Google to filter out the results:

```
$theHarvester -d [domain] -b [online sources] -s

root@kali:~# theHarvester -d ethicalhackingblog.com -b all -s
table results already exists

*******************************************************************
*  _   _                                             _            *
* | |_| |__   ___     /\  /\__ _ _ ____   _____  ___| |_ ___ _ __  *
* | __|  _ \ / _ \   / /_/ / _` | '__\ \ / / _ \/ __| __/ _ \ '__| *
* | |_| | | |  __/ / __  / (_| | |   \ V /  __/\__ \ ||  __/ |    *
*  \__|_| |_|\___| \/ /_/ \__,_|_|    \_/ \___||___/\__\___|_|    *
*                                                                 *
* theHarvester 3.1.0                                    *
* Coded by Christian Martorella                                   *
* Edge-Security Research                                          *
* cmartorella@edge-security.com                                   *
*                                                                 *
*******************************************************************


[*] Target: ethicalhackingblog.com

[*] Searching Yahoo.
[*] Searching CertSpotter.
        Searching results.
[*] Searching Intelx.
An exception has occurred in Intelx search: [Errno 2] No such file or
directory: 'api-keys.yaml'
[*] Searching Google.
        Searching 0 results.
        Searching 100 results.
        Searching 200 results.
        Searching 300 results.
        Searching 400 results.
        Searching 500 results.
[*] Searching Hunter.
[*] Searching VirusTotal.
        Searching results.
[*] Searching Baidu.
```

Continues

(continued)

```
[*] Searching SecurityTrails.
[*] Searching Threatcrowd.
        Searching results.
[*] Searching Bing.
[*] Searching Twitter usernames using Google.

[*] Users found: 3
--------------------
@GusKhawaja
@keyframes
@media
[...]
```

- `-d` is for specifying your target's name. (If you use the `ethicalhackingblog.com` domain name like shown in the figure, don't be surprised if you find no results because the domain is secured for such attacks.)
- `-s` is to search on the Shodan web engine.
- `-b` is the online data source name; in the figure, I've chosen all of them. Here's the list from which you can choose (use the help command `-h` for more options):
  - Baidu
  - Bing
  - bing API
  - Certspotter
  - Crtsh
  - DnsDumpster
  - Dogpile
  - Duckduckgo
  - Github-code
  - Google
  - Hunter
  - Intelx
  - Linkedin and Linkedin_links
  - Netcraft
  - Otx
  - SecurityTrails
  - Threatcrowd
  - Trello

- Twitter
- Vhost
- VirusTotal
- Yahoo
- All (executes all the preceding data sources)

DMitry

DMitry, which stands for "deepmagic information gathering tool," is another application that does multiple things at the same time:

- `-w`: Perform a Whois lookup.
- `-n`: Retrieve records from `Netcraft.com` about the target.
- `-s`: Look for subdomains.
- `-e`: Search for e-mail addresses.
- `-p`: Scan for TCP open port (this is not passive).

```
root@kali:~# dmitry -wnse [domain-name.com]
```

Maltego

Maltego is great for passive information gathering. One could write a whole book about this tool because it contains everything you need to get the job done. If you want to use the complete functionalities, then you have to pay for an annual license. If you're a professional, this tool is a must, but you can still use the limited version and get some decent results.

Transform Hub

Transform Hub, shown in Figure 4.6, is a collection of sites where Maltego will go to fetch the data (e.g., Shodan, Kaspersky, Virus Total, etc.).

By default, most of these data sources are not installed; you have to click each of the ones that you want to install. There are multiple types of data sources:

- **Paid separately**: You will need to pay directly for these web services; they're not included in the Maltego license fees.
- **Free**: Lots of these data sources are completely free.
- **Authenticated API**: Some of these services will require you to open an account; then they will give you an API token to authenticate yourself and use their data.

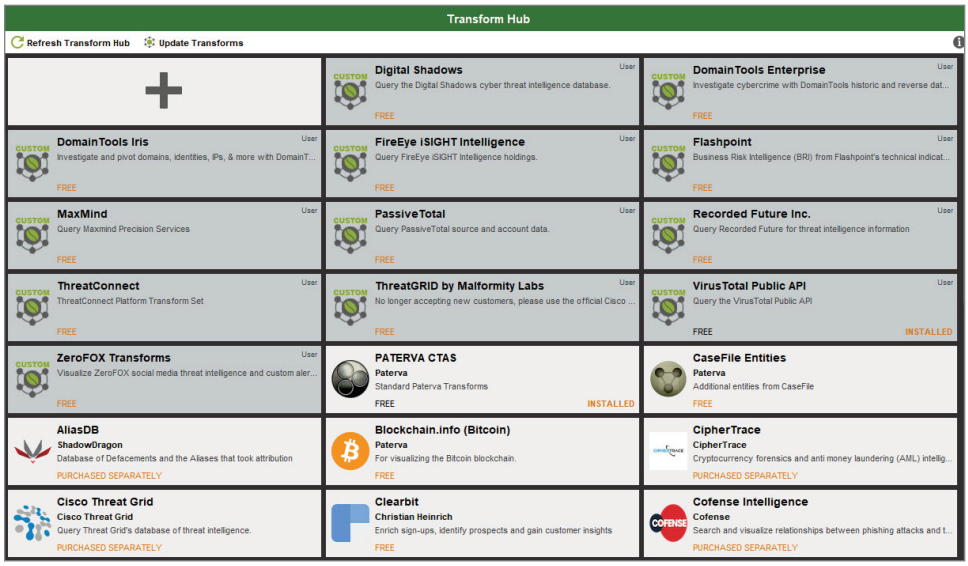

Figure 4.6: Maltego Transform Hub

Creating a Graph

The graph is the centerpiece of Maltego. In this section, you will execute your passive scans. Before you start a scan, you will need to select an entity first (e.g., a person, a company, a domain name, etc.). You can start with the domain name or the company name, and from there, you can choose what you want to scan. Let's take a look at a practical example; we will select the Ethical Hacking Blog DNS as an entity, as shown in Figure 4.7.

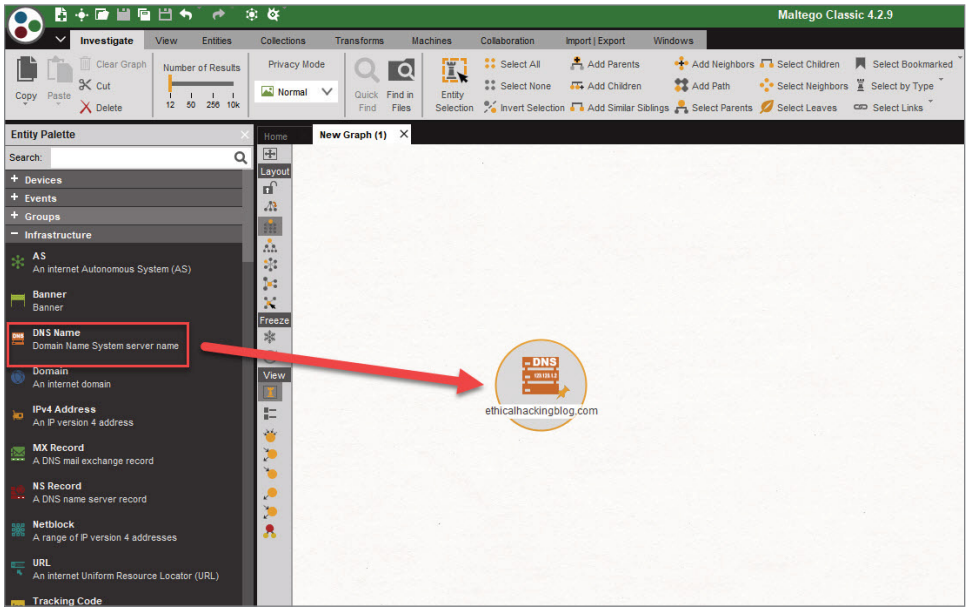

Figure 4.7: Maltego Entities

By right-clicking the entity, you can see all the types of scans that you can execute (Figure 4.8). Beginners are always tempted to click All Transforms (I used to think that way when I started using these tools many years ago). You should instead run the transform scans one by one to evaluate each scan separately.

Figure 4.8: Maltego Transforms

Next, click Convert to Domain and run the To Domains scan. (Click the play arrow on the right, as shown in Figure 4.9, to execute the scan.)

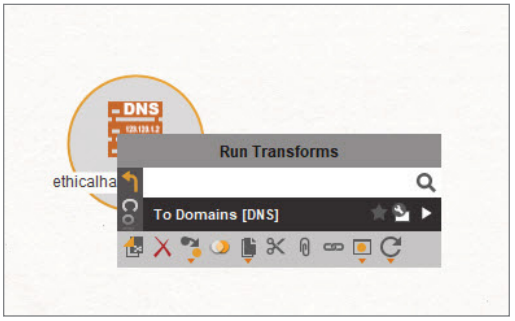

Figure 4.9: Maltego To Domains Tranform

At this stage Maltego will visually display the domain name associated to the DNS (Figure 4.10).

Perfect. Now when you right-click the domain name entity, you should see more transform options (Figure 4.11).

Next, click the double arrows beside the DNS From Domain item to execute all the subdomains tests. After the scan has finished, it will display all the subdomains found under `ethicalhackingblog.com` (Figure 4.12):

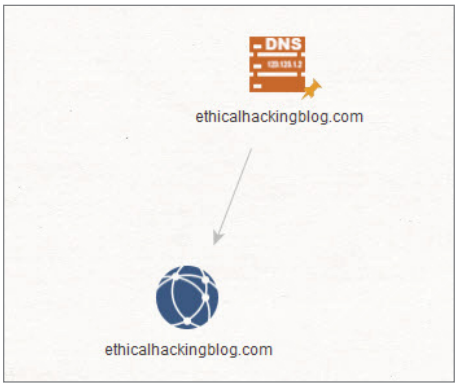

Figure 4.10: Maltego Domain Name / DNS

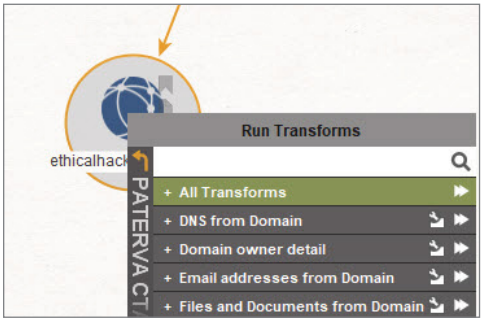

Figure 4.11: Domain Name Tansforms

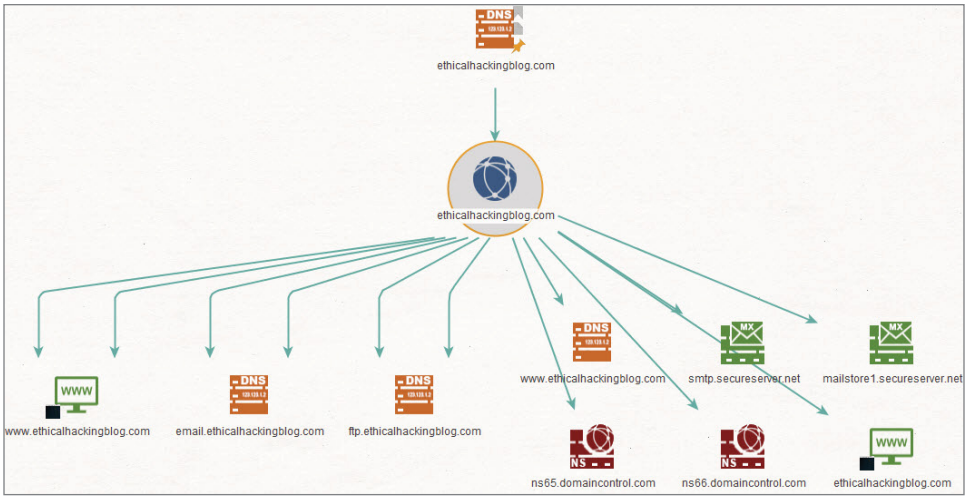

Figure 4.12: Maltego Subdomains Graph

The options are endless here: you can choose every entity type and right-click it to visualize the different kinds of information that you can query from the internet. Take note that this demo shows the paid version of Maltego 4.2.9, but you can perform most of the scenarios in the free edition as well.

Some security folks use free tools like recon-ng that do a job similar to the one executed in Maltego. (recon-ng is a Python scanner for information gathering that uses web API services to fetch its data.) For educational purposes, it is not harmful to try these free tools. However, if an organization is counting on your work, then money should not be an obstacle, and it is, in this case, recommended to take advantage of the yearly license. This is applicable not only for Maltego, but for most of the security tools out there (e.g., Nessus, Burp Suite Pro, etc.). If you want to show professional results, you must pay the price accordingly to get the job done correctly.

Summary

Information gathering is one of the main components during an engagement. Even if you're not going to conduct a social engineering attack, this phase will give you a different angle about the domain/company that you're targeting. The web is always hiding some secrets such as compromised passwords, confidential data, etc. With all the knowledge that you acquired in this chapter, you should be able to start conducting reconnaissance like the pros.