CHAPTER

Foundation: Understanding the Basics

1

INFORMATION IN THIS CHAPTER

- Information overload
- What is internet
- How it works
- What is World Wide Web
- Basic underlying technologies
- Environment

INTRODUCTION

Information Age. The period of human evolution in which we all are growing up. Today internet is an integral part of our life. We all have started living dual life; one is our physical life and the other is the online one, where we exist as a virtual entity. In this virtual life we have different usernames, aliases, profile pictures, and what not in different places. We share our information intentionally and sometimes unintentionally in this virtual world of ours. If we ask ourselves how many websites we're registered on, most probably we won't be able to answer that question with an exact number. The definition of being social is changing from meeting people in person to doing Google hangout and being online on different social networking sites. In the current situation it seems that technology is evolving so fast that we need to cope up with its pace.

The evolution of computation power is very rapid. From an era of limited amount of data we have reached to the times where there is information overload. Today technologies like Big data, Cloud computing are the buzzwords of the IT industry, both of which deal with handling huge amount of data. This evolution certainly has its pros as well as cons, from data extraction point of view we need to understand both and evaluate how we can utilize them to our advantage ethically. The main obstacle in this path is not the deficiency of information but surprisingly the abundance of it present at the touch of our fingertips. At this stage what we require is relevant and efficient ways to extract actionable intelligence from this enormous data ocean.

Hacking Web Intelligence. http://dx.doi.org/10.1016/B978-0-12-801867-5.00001-X
Copyright © 2015 Elsevier Inc. All rights reserved.

Extracting the data which could lead toward a fruitful result is like looking for a needle in a haystack. Though sometimes the information which could play a game changing role is present openly and free to access, yet if we don't know how to find it in a timely fashion or worse that it even exists it would waste a huge amount of critical resources. During the course of this book we will be dealing with practical tools and techniques which would not only help us to find information in a timely manner but also help us to analyze such information for better decision making. This could make a huge difference for the people dealing with such information as a part of their daily job, such as pentesters, due diligence analysts, competitive intelligence professionals, etc.

Let's straightaway jump in and understand the internet we all have been using for so long.

INTERNET

Internet, as we know it has evolved from a project funded by DARPA within the US Department of Defense. The initial network was used to connect universities and research labs within the US. This phenomenon slowly developed worldwide and today it has taken the shape of the giant network which allows us to connect with the whole world within seconds.

DEFINITION

Simply said the internet is a global network of the interlinked computers using dedicated routers and servers, which allows its end users to access the data scattered all over the world. These interconnected computers follow a particular set of rules to communicate, in this case IP or internet protocol (IP) for transmitting data.

HOW IT WORKS

If you bought this book and are reading it then you must be already knowing how internet works, but still it's our duty to brush up some basics, not deeply though. As stated above, internet is a global network of interconnected computers and lots of devices collaboratively make the internet work, for example, routers, servers, switches with other hardware like cables, antennas, etc. All these devices together create the network of networks, over which all the data transmission takes place.

As in any communications you must have end points, medium, and rules. Internet also works around with these concepts. End points are like PC, laptop, tablet, smartphone, or any other device a user uses. Medium or nodes are the different dedicated servers and routers connected to each other and protocols are sets of rules that machines follow to complete tasks such as transmission control protocol (TCP)/IP. Some of the modes of transmission of data are telephone cables, optical fiber, radio waves, etc.

WORLD WIDE WEB

World Wide Web (WWW) or simply known as the web is a subset of internet or in simple words it's just a part of the internet. The WWW consists of all the public websites connected to the internet, including the client devices that access them.

It is basically a structure which consists of interlinked documents and is represented in the form of web pages. These web pages can contain different media types such as plain text, images, videos, etc. and are accessed using a client application, usually a web browser. It consists of a huge number of such interconnected pages.

FUNDAMENTAL DIFFERENCES BETWEEN INTERNET AND WWW

For most of us the web is synonymous to internet, though it contributes to the internet yet it is still a part of it. Internet is the parent class of the WWW. In the web, the information and documents are linked by the website uniform resource locators (URLs) and the hyperlinks. They are accessed by browser of any end device such as PC or smartphone using hypertext transfer protocol (http) and nowadays generally using https. HTTP is one of the different protocols that are being used in internet such as file transfer protocol (FTP), simple mail transfer protocol (SMTP), etc. which will be discussed later.

So now as we understand the basics of internet and web, we can move ahead and learn about some of the basic terminologies/technologies which we will be frequently using during the course of this book.

DEFINING THE BASIC TERMS

IP ADDRESS

Anyone who has ever used a computer must have heard about the term IP address. Though some of us might not understand the technical details behind, yet we all know it is something associated with the computer address. In simple words, IP address is the virtual address of a computer or a network device that uniquely identifies that device in a network. If our device is connected to a network we can easily find out the device IP address. In case of Windows user it can simply be done by opening the command prompt and typing a command "ipconfig". It's near about the same for a Linux and Mac user. We have to open the terminal and type "ifconfig" to find out the IP address associated with the system.

IP address is also known as the logical address and is not permanent. The IP address scheme popularly used is IPv4, though the newer version IPv6 is soon catching up. It is represented in dotted decimal number. For example, "192.168.0.1". It starts from 0.0.0.0 to 255.255.255.255. When we try to find out the IP address

associated with our system using any of the methods mentioned above, then we will find that the address will be within the range mentioned above.

Broadly IP address is of two types

1. Private IP address
2. Public IP address

Private IP address is something that is used to uniquely identify a device in a local area network, let's say how our system is unique in our office from other systems. There are sets of IP address that is only used for private IP addressing:

10.0.0.0–10.255.255.255
172.16.0.0–172.31.255.255
192.168.0.0–192.168.255.255

The above mentioned procedure can be used to check our private IP address.

Public IP address is an address which uniquely identifies a system in internet. It's generally provided by the Internet Service Provider (ISP). We can only check this when our system is connected to the internet. The address can be anything other than private IP address range. We can check it in our system (despite of any OS) by browsing "whatismyipaddress.com".

PORT

We all are aware of ports like USB port, audio port, etc. but here we are not talking about hardware ports, what we are talking about is a logical port. In simple words, ports can be defined as a communication point. Earlier we discussed how IP address uniquely identifies a system in a network and when a port number is added to the IP address then it completes the destination address to communicate with the destination IP address system using the protocol associated with the provided port number. We will soon discuss about protocol, but for the time being let's assume protocol is a set of rules followed by all communicating parties for the data exchange. Let's assume a website is running on a system with IP address "192.168.0.2" and we want to communicate with that server from another system connected in the same network with IP address "192.168.0.3". So we just have to open the browser and type "192.168.0.2:80" where "80" is the port number used for communication which is generally associated with http protocol. Ports are generally application specific or process specific. Port numbers are within the range 0–65535.

PROTOCOL

Protocol is a standard set of regulations and requirements used in a communication between source and destination system. It specifies how to connect and exchange data with one another. Simply stated, it is a set of rules being followed for communication between two entities over a medium.

Some popular protocols and their associated port numbers:
- 20, 21 FTP (File Transfer Protocol): Used for file transfer
- 22 SSH (Secure Shell): Used for secure data communication with another machine.
- 23 Telnet (Telecommunication network): Used for data communication with another machine.
- 25 SMTP (Simple Mail Transfer Protocol): Used for the management of e-mails.
- 80 HTTP (Hyper Text Transfer Protocol): Used to transfer hypertext data (web).

MAC ADDRESS

MAC address is also known as physical address. MAC address or media access control address is a unique value assigned to the network interface by the manufacturer. Network interface is the interface used to connect the network cable. It's represented by hexadecimal number. For example, "00:A2:BA:C1:2B:1C". Where the first three sets of hexadecimal character is the manufacturer number and rest is the serial number. Now let's find MAC address of our system.

In case of Windows user it can simply be done by opening the command prompt and typing a command either "ipconfig/all" or "getmac". It's near about the same for a Linux and Mac user. We have to open the terminal and type "ifconfig-a" to find out the MAC address associated with the system. Now let's note down the MAC address/physical address of our network interface of our system and find out the manufacturer name. Search for the first three sets of hexadecimal character in Google to get the manufacturer name.

E-MAIL

E-mail is the abbreviation of electronic mail, one of the widely used technology for digital communication. It's just one click solution for exchanging digital message from sender to receiver. A general structure of email address is "username@domainname.com". The first part which comes prior to @ symbol is the username of any user who registered himself/herself for using that e-mail service. The second part post @ symbol is the domain name of the mail service provider. Apart from all these, nowadays every organization which have website registered with a domain name also creates mail service to use. So if we work in a company with domain name "xyz.com" our company e-mail id must be "ourusername@xyz.com". Some popular e-mail providers are Google, Yahoo, Rediff, AOL, and Outlook, etc.

DOMAIN NAME SYSTEM

Domain name system (DNS) as the name suggests is a naming system for the resources connected to the internet. It maintains a hierarchical structure of this naming scheme through a channel of various DNS servers scattered over the internet.

For example, let's take google.com it's a domain name of Google Inc. Google has its servers present in different locations and different servers are uniquely assigned with different IP addresses. It is different for a person to remember all

the IP address of different servers he/she wants to connect, so there comes DNS allowing a user to remember just the name instead of all those IP address. In this example we can easily divide the domain name into two parts. First part is the name generally associated with the organization name or purpose for which domain is bought as here Google is the organization name in google.com. The second part or the suffix part explains about the type of the domain such as here "com" is used for commercial or business purpose domain. These suffixes are also known as top level domains (TLDs).

SOME EXAMPLES OF TLDS:

- net: network organization
- org: non-profit organization
- edu: educational institutions
- gov: government agencies
- mil: military purpose

One of the other popular suffix class is country code top level domain (ccTLD). Some examples are:

- in: India
- us: United States
- uk: United Kingdom

DNS is an integral part of the internet as it acts as yellow pages for it. We simply need to remember the resource name and the DNS will resolve it into a virtual address which can be easily accessed on the internet. For example, google.com resolves to the IP address 74.125.236.137 for a specific region on the internet.

URL

A URL or uniform resource locator can simply be understood as an address used to access the web resources. It is basically a web address.

For example, http://www.example.com/test.jpg. This can be divided into five parts, which are:

1. http
2. www
3. example
4. com
5. /test.jpg

The first part specifies the protocol used for communication, and in this case it is HTTP. But for some other case other protocols can also be used such as https or ftp. The second part is used to specify whether the URL used is for the main domain or a subdomain. www is generally used for main domain, some popular subdomains are blog, mail, career, etc. The third part and forth part are associated with the domain

name and type of domain name which we just came across in DNS part. The last part specifies a file "test.jpg" which need to be accessed.

SERVER

A server is a computer program which provides a specific type of service to other programs. These other programs, known as clients can be running on the same system or in the same network. There are various kinds of servers and have different hardware requirements depending upon the factors like number of clients, bandwidth, etc. Some of the kinds of server are:

Web server: Used for serving websites.
E-mail server: Used for hosting and managing e-mails
File server: Used to host and manage file distribution

WEB SEARCH ENGINE

A web search engine is a software application which crawls the web to index it and provides the information based on the user search query. Some search engines go beyond that and also extract information from various open databases. Usually the search engines provide real-time results based upon the backend crawling and data analysis algorithm they use. The results of a search engine are usually represented in the form of URLs with an abstract.

Apart from usual web search engines, some search engines also index data from various forums, and other closed portals (require login). Some search engines also collect search results from various different search engines and provide it in a single interface.

WEB BROWSER

A web browser is a client-side application which provided the end user the capability to interact with the web. A browser contains an address bar, where the user needs to enter the web address (URL), this request is further sent to the destination server and the contents are displayed within the browser interface. The response for the request sent by client contains of raw data with associated format for the data.

Earlier browsers had limited functionality, but nowadays with various features such as downloading content, bookmarking resources, saving credentials, etc. and new add-ons coming up every day, browsers are becoming very powerful. The advent of cloud-based applications has also hugely contributed in making browsers the most widely used software.

VIRTUALIZATION

Virtualization can be described as the technique of abstracting physical resources, with the aim of simplification and utilization of the resources with ease. It can consist

of anything from a hardware platform to a storage device or OS, etc. Some of the classifications of virtualization are:

Hardware/platform: Creation of a virtual machine that performs like an original computer with an OS. The machine on which the virtualization takes place is the host machine and the virtual machine is the guest machine.
Desktop: Concept of separating the logical desktop from the physical machine. The user interacts with the host machine over a network using another device.
Software: OS level virtualization can be described as hosting of multiple virtualization environments within a single OS instance. Application virtualization is hosting of individual applications in an environment separated from the underlying OS. In service virtualization the behavior of dependent system component is emulated.
Network: Creation of a virtualized network addressing space within or across network subnets.

WEB BROWSING—BEHIND THE SCENE

So now as we have put some light on some of the technological keywords that we will be dealing with in later chapters, let's dive a little deeper and try to understand what exactly happens when we try to browse a website. When we enter a URL in a browser it divides the same into two parts. Let's say we entered "http://www.example.com". The two parts of this URL will be (1) http and (2) example.com. The reason for doing so is that to identify the protocol used and domain name to resolve it to an IP address. Let's again assume that the IP address associated with the domain name example.com is "192.168.1.4" then browser will process it as "192.168.1.4:80" as 80 is the port number associated with protocol HTTP.

From paragraph which contains details about DNS we already came across that it is used to resolve the domain name into IP address but how? It depends whether we are visiting a site for first time or we often visit this site. But still for both the case the procedure remains quite same. First DNS lookup starts with browser cache to check if there is some records present or not or checks whether we visited this site earlier or this is the first time. If the browser cache does not contain any information the browser does a system call to check whether OS is having any DNS record in its cache or not. Similarly if not found then it searches the same DNS info in router cache if not found the ISP DNS cache then finally if not found any DNS record in these places starts a recursive search from root name server to top level name servers to resolve the domain name. The thing which we need to think about is that some domain names are associated with multiple IP addresses such as google.com in that case also it returns with only one IP address based on the geographic location of the user who intent to use that resource. The technique is also known as geographic DNS.

In above paragraph we understood how DNS lookup searches for information from browser cache but that is only for sites which are static, because dynamic sites contains dynamic contents that expires quickly. However, the process is quite same for both the cases.

After DNS resolution, browser opens a TCP connection to the server and sends a hypertext request based on the protocol mentioned in the URL as it is HTTP in our case browser will send an HTTP GET request to the server through TCP connection. Then browser will receive an HTTP response from the server with status code. In simple words, status codes define the server status for the request. There are different types of status codes, but that is a huge topic on its own; hence just for our understanding I will include some of the popular status codes that a user might encounter in browsing

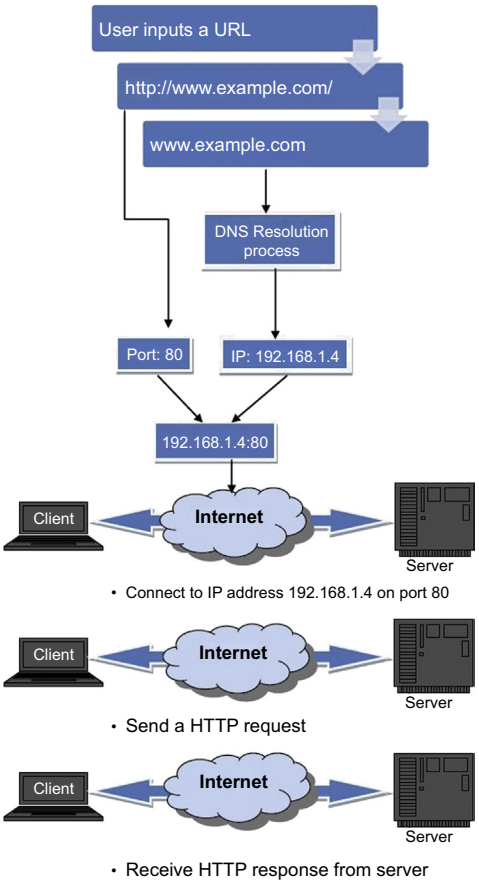

FIGURE 1.1

Web browsing—behind the scene.

HTTP STATUS CODE CLASSES

They lie between 100 and 505 and are categorized as different classes according to its first number.

- 1xx: Informational
- 2xx: Successful
- 3xx: Redirection
- 4xx: Client-error
- 5xx: Server-error

Some popular status codes:

- 100: continue
- 200: ok
- 301: moved permanently
- 302: found
- 400: bad request
- 401: unauthorized
- 403: forbidden
- 404: not found
- 500: internal server error
- 502: bad gateway

If browser gets any error status code then it fails to get the resources properly if not then it renders the response body. The response body generally contains html codes for the page contents and links to other resources, which further undergo the same process. If the response page is cacheable it will be stored in cache. This is the overall process takes place in the background when we try to browse something in internet using a browser.

LAB ENVIRONMENT

As we have discussed the basic concepts, now let's move ahead and understand the environment for our future engagements.

OPERATING SYSTEM

For a computer system to run we need basic hardware such as motherboard, RAM, hard disc, etc. but hardware is worthless until there is an OS to run over it. An operating system basically is a collection of software which can manage the underlying hardware and provide basic services to the users.

Windows

One of the most widely used OS introduced by Microsoft in 1985. After so many years it has evolved to a very mature stage. The current version is Windows 8.1. Though it has had its fair share of criticism yet it holds a major percentage of the market share. The ease of usability is one of the major features of this OS which makes it widely acceptable.

Though during the writing of this book we were using Windows 7 64 bit, any version above 7 will also be fine and would function more or less in similar fashion.

Linux

Popular as the OS of geeks, this OS is available in many flavors. Mostly it is used for servers due to the stability and security it provides, but is also popular among developers, system admins, security professionals, etc. Though it surely seems a bit different as well as difficult to use for an average user, yet today it has evolved to a level where the graphical user interface (GUI) provided by some of its flavors are at par with Windows and Mac interfaces. The power of this OS lies in its terminal (command line interface), which allows to utilize all the functionality provided by the system.

We will be using Kali Linux (http://www.kali.org/), a penetration testing distribution during this book. It is based on Debian, which is a well-known, stable flavor of Linux. Though, other flavors such as Ubuntu, Arch Linux, etc. can also be used as most of the commands will be similar.

Mac

Developed by Apple this series of OS is well known for its distinctively sleek design. In the past it has faced criticism due to the limited options available at software front, but as of today there is a wide range of options available. It is said to be more secure as compared to its counterparts (in average use domain), yet it has faced some severe security issues.

Mac provides a powerful command line interface (GUI) as well as CLI which makes it a good choice for any computing operation. Though we were using Mac OS X 10.8.2 during the writing of this book, any later version will also be fine for practice.

Most of the tools which will be used during the course of this book will be free/open source and also platform independent, though there will be some exceptions which will be pointed out as and when they come into play. It is recommended to have a virtual machine of a different OS type (discussed above) apart from the base system.

To create a virtual machine we can use the virtualization software such as VirtualBox or VMware Player. Oracle VirutalBox can be downloaded from https://www.virtualbox.org/wiki/Downloads. VMware Player can be downloaded from http://www.vmware.com/go/downloadplayer/.

PROGRAMMING LANGUAGE

A programming language is basically a set of instructions which allows to communicate commands to a computing machine. Using a programming language we can control the behavior of a machine and automate processes.

Java

Java is a high-level, object-oriented programming language developed by Sun Microsystems, now Oracle. Due to the stability provided by it, it is heavily used to develop

applications following client–server architecture. It is one of the most popular programming language as of today.

Java is required to run many browser-based as well as other applications and runs on a variety of platforms such as Windows, Linux, and Mac.

The latest version of Java can be downloaded from: https://www.java.com/en/download/manual.jsp

Python

A high-level programming language, which is often used for creating small and efficient scripts. It is also used widely for web development. Python follows the philosophy of code readability, which means indentation is an integral part of it.

The huge amount of community support and availability of third party libraries makes it the preferable language of choice for most of the people who frequently need to automate small tasks. Though this does not mean that Python is not powerful enough to create full-fledged applications and Django, a Python-based web framework is a concrete example of that. We will discuss Python programming in detail in later chapter.

The current version of Python is 3.4.0, though we will be using the version 2.7 as 3.x series has had some major changes and it is not backward compatible. Most of the scripts we will be using/writing will be using the 2.7 version. It can be downloaded from https://www.python.org/download/releases/2.7.6/

BROWSER

As discussed above, a browser is a software application which is installed at the client's end and allows to interact with the web.

Chrome

Developed by Google and it is one of the most widely used browser. First released in 2008, today this browser has evolved to a very stable release and has left the competition way behind. Most of its base code is available online in form of Chromium (http://www.chromium.org/Home).

Today Chrome is available for almost all devices which are used for web surfing, be it a laptop, a tablet, or a smartphone. The ease of usability, stability, security, and add-on features provided by Chrome clearly makes it one of the best browsers available. It can be downloaded from https://www.google.com/intl/en/chrome/browser/.

Firefox

Firefox is another free web browser and is developed by Mozilla Foundation. The customization provided by Firefox allows to modify it to your desire. One of the greatest features of Firefox is the huge list of browser add-ons, which allows to tailor it for specific requirements. Similar to Chrome it is available for various platforms. It can be downloaded from https://www.mozilla.org/en-US/firefox/all/.

In this book we will mainly be using Chrome and Firefox as our browsers of choice. In a later chapter we will be customizing both to suit our needs and will also try out some already modified versions.

So in this chapter we have understood the basic technologies as well as the environment we will be using. The main motivation behind this chapter is to build the foundation so that once we are deep into our main agenda i.e., web intelligence, we have a clear understanding of what we are dealing with. The basic setup we have suggested is very generic and easy to create. It does not require too much installations at the initial stage, the tools which will be used later will be described as they will come into play. In the forthcoming chapter we will be diving deep into the details of Open source intelligence.