IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 4, April 2018

Deep Web: An Exploration

Archana Sahai

Assistant Professor, Amity Institute of Information Technology, Amity University, Lucknow

Abstract: Deep web is the side of the internet that deals with all kind of black work that can be done on the Internet anonymously. This review paper discusses about the sides of the internet like deep web and surface net. It describes how the deep web can be accessed using TOR. It also explores about the various applications of deep web along with its deep dark side.

Keyword: Deep Web, Dark Internet, The Onion Router (TOR), vulnerability, socket layer(SSL) , Transport Layer Security(TLS), Deep Web Analyzer (DeWA)

I. INTRODUCTION

The world of internet was created by Time Berners-Lee in the year 1989. He first made a rough or a first time model application in the year 1990. The wide internet is usually known as net. Deep web the invisible web or the hidden web is a part of the World Wide Web[1]. The Deep Web actually includes any website that cannot be indexed by any search engine. This includes any page that cannot be detected by the 'crawlers' used by Google and its competitors to search the web for pages to fill its results pages. It consists primarily of database-driven websites, and any part of a website that's past a login page. It also includes sites blocked by local webmasters, sites with special formats, and transient sites. Google and other engines cannot reach these because it isn't programmed to fill out search forms and click on the submit button, rather, they must interact with the web server that's presenting the form, and send it the information that specifies the query and other data that the web server needs.

For this reason Deep web is also some time referred to as black internet and the hidden part of the internet.Deep web is a place which is out of reach of government. Deep web can be said as a first home to hackers and a place where one can commit any sort of internet crime.

In general search engines do not index the following types of Web sites:

- Proprietary sites
- Sites requiring a registration
- Sites with scripts
- Dynamic sites
- Ephemeral sites
- Sites blocked by local webmasters
- Sites blocked by search engine policy
- Sites with special formats
- Searchable databases

Proprietary sites require a fee for access. Registration sites require a login or password. A bot can index script code (e.g., Flash, JavaScript), but it can't always ascertain what the script actually does.

II. WEB SIDES

There is nothing in the world that has one side similarly Internet alsohas two sides:

(a) The Surface Web or clear web
(b) The Deep Web

2.1 Surface web

The Shallow Web, also known as the Surface Web or Static Web, is a collection of Web sites indexed by automated search engines. A search engine bot or Web crawler follows URL links, indexes the content and then relays the results back to search engine central for consolidation and user query. Ideally, the process eventually scours the entire Web, subject to vendor time and storage constraints.The part of the internet that is available for all, and is searchable to the various search Engines such as Google, Yahoo, and Bing[3]. These pages do not have to depend on a particular database for the content they have. This "searchable Internet" is called the Clear Web.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 4, April 2018

They reside on a server which can be retrieved. And they are the html files whose context never changes. If we wish to make any change then we need to directly change the html codes & a new version of that page is created. Thus whenever we say surface web, we mean the common website. Surface Web are the site whose name have a .com, .org, .net are similar extensions. [6]

Fig1. Two sides of Internet

2.2 Deep Web

Deep web is the invisible, the hidden part of the deep internet. This part of the deep Web is not visible to the general public [2]. If the user needs to access this part of the internet he needs to fill the form and submit its valid inputs. The Creators of deep Web named it so because the content of deep web is beyond the search and reach of the various search engines. It is a challenge to access such data in a data management community.

Deep Web Search Strategies [14]

- Be aware that the Deep Web exists.
- Use a general search engine for broad topic searching.
- Use a searchable database for focused searches.
- Register on special sites and use their archives.
- Call the reference desk at a local college if you need a proprietary Web site. Many college libraries subscribe to these services and provide free on-site searching.

III. ACCESSING DEEP WEB

To mine the Deep Web manually or through search engines would be an impossible task. There are currently a number of bots available that attempt to solve the problem. Such crawlers must be designed to automatically parse, process, and interact with form-based search interfaces that are designed primarily for human consumption. They must also provide input in the form of search queries, raising the issue of how best to equip crawlers with the necessary input values for use in constructing search queries. Stanford has built a prototype engine named the Hidden Web Exposer (HiWE). HiWE tries to scrape the Deep Web for information using a task-specific, human-assisted approach. Others that are publicly accessible include Infoplease, PubMed and the University of California's Infomine. There is another Big Data Mining tool known as BrightPlanet's. It is Deep Web Monitor, which allows to set a specific query such as a location or keyword, and harvest the entire web for relevant information. But the only disadvantage of TOR is that the web pages are extremely unreliable, they can be down for hours, days or can be permanently down. The webpages of TOR network can be really very slow as TOR routes our connectivity through other user's computer, so that our anonymity could be protected.

Fig 2- The Hidden part of Internet

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 4, April 2018

IV. TOR: THE ONION ROUTER

In the past technological barriers made it difficult to access the Deep Web, but it is now possible to overcome these barriers. Many individuals and institutions have compiled a list of invisible Web directories. By now, people start thinking that accessing deep web would be a real tough job but surprisingly it is not so, rather it is really simple to access deep web. All one need to do to access deep web is to download TOR browser. TOR allows us to connect to deep web and access the web pages. The TOR network is a group of volunteer-operated servers that allows people to improve their privacy and security on the Internet. TOR users employ this network by connecting through a series of virtual tunnels rather than making a direct connection, thus allowing both organizations and individuals to share information over public networks without compromising their privacy. TOR is an effective censorship circumvention tool, allowing its users to reach otherwise blocked destinations or content. It can also be used as a building block for software developers to create new communication tools with built-in privacy features [15]. The name of the original software from where TOR is derived is "The Onion Router ".TOR can be defined as free software that enables user to hide their search history, web mail, their web activities and their social posts. TOR also makes one untraceable. A person can also hide that in which country he/she is doing any kind of online activity. TOR also hides IP address of the user. This feature of TOR is advantageous for hackers, business man, journalists and activists.

Another advantage of using TOR is that it is a free service. TOR makes it really difficult to track internet activities. But activities cannot be tracked if and only if one follows all the right precautions. Various deep Web communities could be accessed through the TOR network because TOR is a strictly private, secret and anonymous network. Dark nets like TOR and I2P require dedicated software that acts as a proxy, while alternative DNS systems and rogue TLDs need the use of dedicated DNS servers to resolve an address [11].Thebiggest advantage of using TOR is that the surveillance organization will also not be able to observe your internet activities[4].

V. TOR NETWORK PROCESSING

The steps to run the TOR network.

Step 1: Packet Wrapping, The data that we enter is first converted in the form of encrypted packets. Then unlike a normal internet browser, TOR removes the header of the packet because it has the addressing information which is known as packet wrapping.

Step 2: Relays, Now, these encrypted data packets are passed through many of the servers which are known as relays. We must keep in mind that a packet does not directly reach a final destination [13]. A roundabout path is created through which a packet has to travel which helps in shaking the pursuer.

Step 3: In each relay two informations are decrypted. Firstly, from where a particular packet is being sent and second where a packet now needs to be send .Now the relay rewraps the entire packet into a new wrapper and send it.

Step 4: Mostly a data packet are encrypted through a protocol called a secure socket layer(SSL) or the data packets are also encrypted through a more stronger version called transport Layer Security(TLS)[8].

One thing to keep in mind before using a TOR browser is that one must keep in mind that the webpage has some SSL or TLS encryption, which can be seen if the webpage has an 'https' instead of 'http'. A page must have a SSL or a TLS encryption because if a page will not have it, then our data will no longer be encrypted and hence, it could be not secure any more to use such a browser

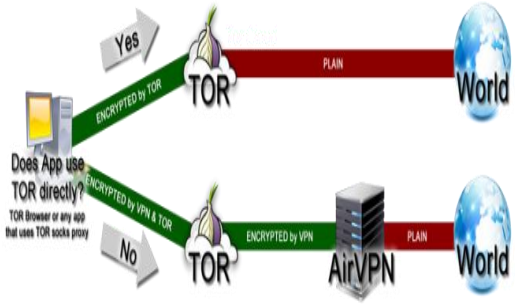

Fig 3- Working of TOR

VI. TOR OWNERS

The 'onion routing' was sponsored by the US Research Laboratory in 1990's and then TOR was itself developed in 2002 by navy and independent researchers.TOR is till date being updated and supported by the original creators of TOR. Tor project is a non-profitable, independent organization which is even now being funded by the US Government. One more advantage of tor is that it is open source software which means that anyone and everyone can view the coding and also incorporate it into their own software.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 4, April 2018

The TOR network has been extensively been revised. The TOR protocol has been examined by many researchers so the TOR browser offers the strongest possible security.

VII. ADVANTAGES OF DEEP WEB

The various advantages of using deep web,and its various applications are:

- Import and export of various kinds of drugs.
- Import and export of ammunitions and arms.
- Hiring assassins become accessible.
- Fake identifications and fake passports can be purchased because of deep web.
- Deep web is widely used by journalist to learn about "inside" information.
- Deep web is also responsible for all kinds of child pornography.
- Various kinds of experiments can be performed to humans
- Risk to confidential information.
- Easy access to all government information can also be done using deep web.
- Etc.

VIII. DARK SIDE AND LAW ENFORCEMENT ON DEEP WEB

Several international crimes occur on the surface webs which are a great challenge for Law enforcement agencies. But crimes on surface web are not as bigger as when deep Web comes into play. Crimes in deep web are comparatively much bigger and dangerous [14].

Three more aspects that make the law enforcement even more problematic when it comes to deep web are-

1. **Encryption**- Anything and everything in deep web is encrypted which means that in deep web trapping or monitoring the criminal is a bit more tough because they are more aware . Criminals use encryption as the first countermeasure so that they could evade detection [9].
2. **Attribution**- as we know that everything in deep web happens on the 'onion domain' thus it becomes extremely difficult for attribution to be determined.
3. **Fluctuation-** As we have already discuss about how dynamic deep web is, an online forum can fluctuate its location, i.e. it can be at a specific URL today but it may be not there the next day. One more problem with deep web is that the information we gathered a month ago is no more valid today because there are often changes observed in the naming and the address schemes. Thus considering the time frame the laws must be rigorously documented so that any online activity must be time-stamped and the cases are prevented from being invalid [15].

THE ROLE OF SECURITY VENDORS

Although for a normal Internet user who till date is only aware of surface internet will never find the use of deep web but still security vendors must be good enough to protect their customers from the worldwide cybercriminal activities. TOR is increasingly used by the malwares today to access confidential data. So our security vendors should be able to create detections means and the countermeasures against all these attacks and threats [13].

On contrary, there are users for whom there are legitimate reasons for which they need to visit the deep web. These reasons are that the supply of certain prescription drugs is much easier, in deep web they can access recreational drugs which are also illegal in some areas, Deep web allows discussion on socially banned topics, and deep web also allows journalists to share information. Thus in these cases it is the responsibility of the security vendor to protect their customers[7].

Fig 4- Law and deep web

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 4, April 2018

But because today the trading of illegal goods is the one and the most dominant deep web activities it is very essential for the details of worldwide sellers and buyers to be hidden[6].

IX. DEEP WEB ANALYZER

The Deep Web Analyzer (DeWA) has been designed with the goal of supporting investigations in tracking down malicious actors, exploring new threats and extracting meaningful data from the Deep Web, e.g. new malware campaigns[11].

DeWA consists of the following 5 modules:

- A Data Collection module, responsible for finding and storing new URLs from multiple sources.
- A Universal Gateway, which allows to access the hidden resources in darknets like TOR and I2P, and to resolve custom DNS addresses.
- A Page Scouting module, responsible for crawling the new URLs collected.
- A Data Enrichment module that takes care of integrating the scouted information with other sources.
- A Storage and Indexing module, which make the data available for further analysis.
- Visualization and analytic tools.

X. CONCLUSION AND FUTURE SCOPE

In the future we will see that deep web will be in much demand in near future it will become a necessity for users to be familiar with the deep web because of the various limitations of the search engine.Deep web is still an ambiguous part for our digital world despite of the fact that how huge amount of information is being stored in it. There are still large amount of internet users who have still not heard about this another part of the internet and for them Google is all what the web has for us. On the other hand for others it is only a crime world.It may be true that the public awareness can increase the use of deep web, but facing the reality currently the users have not got enough reasons to migrate and discover deep net in place of the surface net and to anonymyzing some other software's in the near future.But very soon an unknown cold war or rather a race will be seen between the 'extreme libertarians' and the 'law enforcement agencies' where one will be finding more and more new ways to become more and more untraceable and anonymous to other. Running TOR will prevent data collectors and powerful aggregators like Google ads and Acxiom to perform traffic analysis. TOR will also prevent these data collectors to gather confidential data and frequent internet habits of the end users.

REFERENCES

[1] https://www.airsassociation.org/services-new/airs-knowledge-network-n/airs-articles/item/16323-8-best-deep-web-people-search-engines-updated
[2] AgriSurf http://www.agrisurf.com/agrisurfscripts/agrisurf.asp?index=_25
[3] AltaVista http://www.altavista.com/
[4] Bluestone [formerly http://www.bluestone.com]
[5] http://www.fastcolabs.com/3026989/an-up-to-date-laymans-guide-to-accessing-the-deep-webhttp://www.forbes.com/sites/marcochiappetta/2016/04/29/access-the-deep-web-and-protect-your-privacy-online-with-the-anonabox/#72cd0dc337c2
[6] http://www.techspot.com/guides/1292-web-security-anonymizer-primer/
[7] Northern Light http://www.northernlight.com/
[8] Open Directory Project http://dmoz.org
[9] http://gjcat.com/Issue/GJCAT_2012_0106.pdfSubmitted to American Public University System on 2016-09-26
[10] Securities and Exchange Commission http://www.sec.gov
[11] U.S. Census Bureau http://www.census.gov
[12] http://www.webpages.uidaho.edu/~mbolin/iffat-sami.htm
[13] http://www.trendmicro.co.uk/media/wp/exploring-the-deep-web-whitepaper-en.pdf
[14] https://www.computerworld.com/article/2548609/networking/mining-the-deep-web--search-strategies-that-work.html
[15] https://www.torproject.org/about/overview.html.en