Internet of Things and Cloud Computing
2018; 6(3): 56-62
http://www.sciencepublishinggroup.com/j/iotcc
doi: 10.11648/j.iotcc.20180603.11
ISSN: 2376-7715 (Print); ISSN: 2376-7731 (Online)

Best Practices to Protect Your Privacy Against Search Engines Data Mining – A Review

Franck Seigneur Nininahazwe, Micheal Ernest Taylor

School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, China

Email address:
seigneurinuyasha777@yahoo.fr (F. S. Nininahazwe), delen007@live.com (M. E. Taylor)

To cite this article:
Franck Seigneur Nininahazwe, Micheal Ernest Taylor. Best Practices to Protect Your Privacy Against Search Engines Data Mining – A Review. *Internet of Things and Cloud Computing*. Vol. 6, No. 3, 2018, pp. 56-62. doi: 10.11648/j.iotcc.20180603.11

Received: August 30, 2018; **Accepted**: September 19, 2018; **Published**: October 13, 2018

Abstract: The internet has revolutionized the way people do things in the twenty first century. Most day to day activities are been done using the internet via search engines. Search engines are great tools which aids in surfing the web more easily. The most famous ones includes Google, Bing, Yahoo and many more others which provides users with directions to resources stored on the web and other websites. However these search engines pose several issues to users of which the most challenging of them, which has drawn the attention of many researchers is privacy. Search engines learn about the user while using them by gathering every bit of information such as identity, location and so on. In this paper, the authors discuss some protection methods and privacy techniques users can observe to prevent privacy invasion and data security whilst using these systems. The study further presents some comparisons between and gives detailed understanding to enable one have a clear idea. Further discussions and suggestions after the comparisons and a conclusion on the best ways to protect user's privacy on the internet whilst using these internet based search engines are also presented by the authors.

Keywords: Search Engines, Privacy, Tor Network, I2p

1. Introduction

In the twenty first century, a lot of people use the internet on a daily basis, and all these people use at least a search engine. These search engines belong to enterprises, in order to improve the quality and offer the best experience possible, the provider needs to know its users. The problem comes from the fact that nobody really knows what these search engines take, though there are regulations and user agreements, privacy policies that govern information of these systems. End users often do not check or read the terms of contract provided and inadvertently revealing personal information.

In order to minimize the impact of search engines, the author in [1] stipulated that companies should shoulder significant social responsibility and rules should be defined. Although, the idea is reasonable, the regulations provided presently do not tie down the obligations of these companies to participate in voluntary practices. Other studies state that, companies should develop some tools or algorithm to help protect the privacy of users [2]. This research ascertain those companies partly to be criticized for not securing the user's information, which means that suggestion may work only in order to prevent hackers from getting those information as well.

Furthermore, search engines privacy problem should be the responsibility of every user since the companies are nowhere near to stop spying on the regular users. There are many techniques such as VPN, the use of a proxy server and many others that can help. A final assessment and suggestions at the end of this review is given.

2. Related Works

On the internet, the matter of privacy, concerns many areas and there is still a long way to fixing all of them. Since the area is highly extensive, the authors focused on the privacy issue on search engines data mining. Many researchers all over the world are trying to resolve it. Dr. Steven Mintz [3] and L. Hinman [1] have proposed a code of ethics. Shaozhi Ye, Felix Wu, Raju Pandey and Hao Chen [2] suggest to add noise injection and many other propositions but as always each user will need to care about their privacy because many of the solutions proposed generally include the help of those

companies who own the search engines or any other service on the internet. And that is what this research seeks to achieve by providing pertinent solutions.

3. Privacy Issues

The specifics of search engines varies from one to another but they are mainly known as complex algorithm that scours the internet in search of the needed information according to pre-set settings usually key words or sentence. Other used criteria will depend on how much the search engine knows about the user, which can go from demographic area, age and or gender. As said earlier, search engines can use a simple word to help find what one wants, the user does not need to know exactly what he/she wants, just a simple idea will be enough. The internet would work without search engines but it would be a lot different. When surfing only 10% of users will look beyond the top of search engines results [3]. The search engines can do more than help find the information needed. They provide other services and one of them is advertisement. Some of these advertisement can be specific to a demographic set or even a person. It seems to be amazing but when one realize how they do it, then it becomes questionable, data mining algorithm are mostly used to track (individuals) on the internet and know almost everything about that user from the kind of music, where internet is used, the website visited mostly. In short, every single information about the individual is on the web [4], are gathered and used by these search engines.

Search engine's data mining has limitations. Web users are most often receive advertisement and information which are not related to their search queries. A crucial example is when the user queries the search engine for adults' clothes and baby clothes are displayed. Such advertisement makes one wonder the sort of information these search engines may have gathered or if there isn't a hacker sitting trying to steal information.

Concerning collecting data they all do it, but the difference is in their privacy policies and how they treat ads. Some researchers searched how to extract data without compromising user privacy, some aim at individual privacy and others at corporate privacy [5].

Here are a small list of today's search engines and a small hint about their privacy policies.

3.1. Search Engines That Collect Data

(1) Google: when it comes to search engines, google is the GOD amongst them all. Google search engines has the intelligence in deciding which website is good and which is not. It also tops in data collection as well. It collects and tracks every bit of information, from your IP address to website visited, as well as purchases made online etc.
(2) Yahoo: it's historically one of the good search engines and also another major collector of data. It's also known to have been the victim of two major hacks which left users at risks.
(3) Bing: it's not a good one when it comes to results but can compete with google when it comes to data collection. For instance, when one reads the privacy policy, one will realize the extent of tracking even to the exact geographical location.
(4) Aol: The privacy policies states clearly the connectivity usage collected even with data powered by Bing.
(5) Ask.com: even if in the last years it received less and less information each year, they are still a big collector of data.
(6) Lycos: it's an out dated platform which collects a huge amount of data even if it's less than Google and Bing.
(7) Baidu: it is mostly used in China and can notice that it is the Chinese Google in both results and collecting data.

3.2. Search Engines That Do Not Usually Collect Data

(1) Ixquick: in the beginning it was 100% ads free but now it's enhanced by google so some ads are collected. The great feature it has to offer is the ability to open the result anonymously in a proxy window.
(2) DuckDuckGo: when it comes to privacy this one is very good, it does not collect data only queries and has no ads, and it has also a Tor service. The issue is that the search engines is oriented crowd-sourced and user generated content so the information might not be really up to date.

Major search engines are owned by corporations so their first concern is to make money. Then it is not hard to imagine the conflict that can arise between having profit and user interest. Some search engines are expanding, adding some other services to their catalogues like online music or shopping coupons. This creates some problem, for example when Google added online travel as one of their services, it started placing their flight services on the top of research results even before other major players in that business like Orbitz and Expedia [3].

Most of the search engines secrecy reside in the algorithm that major search engines use, the question is to know if this algorithm should be known by the public, some argue that they should stay secret but also that they should make their policies known by everybody and they also should follow them [3]. Unfortunately transparency does not make things clear, policies of giants like Google are known to be hard to understand.

In fact the results of an April 2012 survey by strategic branding firm Siegel and Gale showed that users do not understand how Google tracks, stores and shares their information. David Siegel said that to read all the terms of services received, it would take 76 days and that people do not really understand what they are agreeing to. This study can say that companies have the right but also the duty to keep their algorithm details private because not only what makes a difference with others. However, reveals everything put on the search engine is a great danger of manipulation by spammers.

In an august 2013 survey, 80% of the respondents agreed that there should be more control on how data are collected and used. Michael Walker ~~idea~~ perception ~~is~~ to create some kind of Hippocratic Oath for analytics professionals so they would avoid violating consumer privacy rights. Whether people are going to voluntarily follow it or being obliged to, search engines are going to keep doing what they are doing.

4. Privacy Protection

4.1. VPN and Proxy Server

In order to fight against those issues some solutions are suggested. This study considers how they are deployed, then existing privacy protection can be put into 3 categories: server side, network and user side [5]. It is also possible to use more than one category at the same time. The solutions suggested are related to either network or user side.

(1) VPN: A virtual private network (VPN) extends a private network across a public network and enables users to send and receive data across shared or public networks as if their computing devices were directly connected to the private network. Applications running across a VPN may therefore benefit from the functionality, security, and management of the private network [6]. A VPN is created by establishing a virtual point-to-point connection through the use of dedicated connections, virtual tunnelling protocols, or traffic encryption. A VPN available from the public. Internet can provide some of the benefits of a wide area network (WAN). From a user perspective, the resources available within the private network can be accessed remotely [7].

(2) Proxy: In computer networks, a proxy server is a server (a computer system or an application) that acts as an intermediary for requests from clients seeking resources from other servers [8]. A client connects to the proxy server, requesting some service, such as a file, connection, web page, or other resource available from a different server and the proxy server evaluates the request as a way to simplify and control its complexity. Proxies were invented to add structure and encapsulation to distributed systems. Today, most proxies are web proxies, facilitating access to content on the World Wide Web, providing anonymity and may be used to bypass IP address blocking.

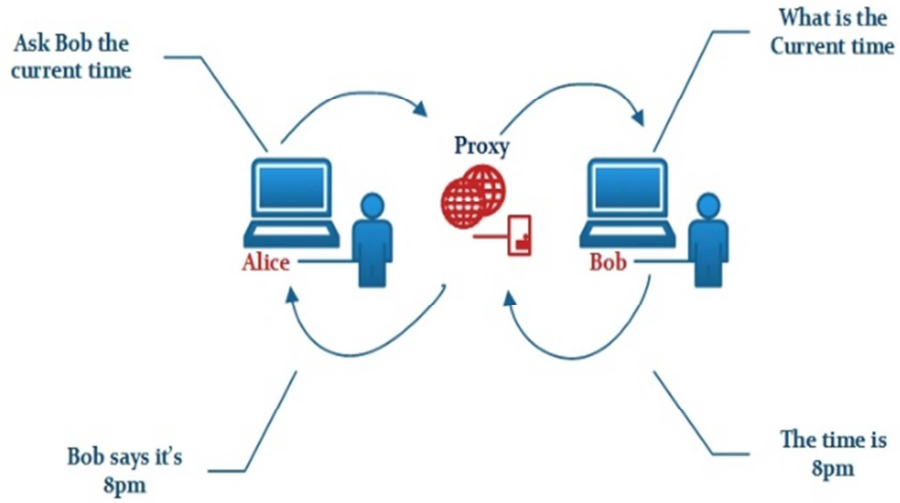

Figure 1. *Proxy server.*

This is an example of the use of a proxy. The two computers in blue are communicating with each other using the red node as a proxy (the one in the middle). Bob doesn't know whom to send the information, so the privacy of Alice is protected.

The comparison between a proxy service and a VPN service are shown in the Table 1.

Table 1. *Comparison between a Proxy Service and a VPN Service.*

	Compatibility	Speed	Setup	Stability	Security	Encryption
VPN service	Linux, windows, mac, ios, android	Fast speed on any network	Simple setup, easy connection and disconnection	as stable as your internet connection	Encryption on all your application at one time	Great encryption on the whole device
Proxy service	Only with application which support proxies	Fast speed on any app that support the proxy service	Depends on the app being configured	Good stability on any network	Encryption on one application at a time	Only when the proxy configuration support encryption

This study concludes that a VPN service is better than a proxy service.

4.2. Tor Network and I2p

Tor:

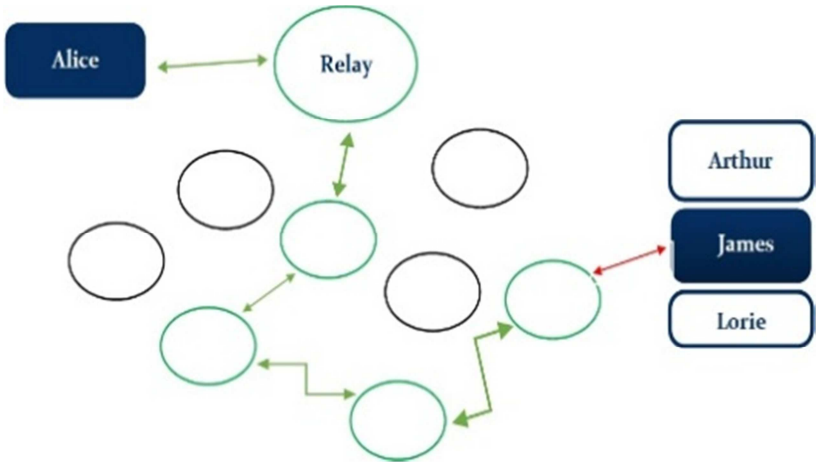

Figure 2. *Proxy server.*

Those relay in green represent one encrypted tor path.

Tor or the onion router is a network that allows the user to stay anonymous on internet, it protects against any location tracking, traffic analysis, network spying…

-How to access it: there are many ways of accessing and using the Tor network. The first one and most known is the Tor browser, it consists of a modified Mozilla Firefox ESR web browser, and uses HTTPS everywhere, also has the Tor proxy. It can operate on windows, Linux and macOS. Or one can use the Tor messenger which is an instant messaging program, or Orbot which is a Tor application for android. There are many application created by the Tor project society and can be found on the Tor network official website [10].

-How it works: user data are first encrypted and then transferred through different relays present in the tor network, which create a multi-layered encryption. An encryption layer is decrypted any time it gets to a tor relay and the data will then be followed to the next random relay until it reaches the destination. Each relay knows only its predecessor and successor [9]. The last node appears as the source of the data which makes identity tracking a nightmare. The Tor network can also provide anonymity to servers, website and can even be used to download torrent files with a configured P2P application.

[10] Tor network uses some good methods to protect the users. It uses a 128 bit AES in counter mode, RSA with 1024 bit keys and a fixed exponent of 65537, a Diffie-Hellman with a generator (g) of 2 for the modulus (p) whose hex representation is:

"FFFFFFFFFFFFFFFFC90FDAA22168C234C4C6628B80DC1CD129024E08"
"8A67CC74020BBEA63B139B22514A08798E3404DDEF9519B3CD3A431B"
"302B0A6DF25F14374FE1356D6D51C245E485B576625E7EC6F44C42E9"
"A637ED6B0BFF5CB6F406B7EDEE386BFB5A899FA5AE9F24117C4B1FE6"
"49286651ECE65381FFFFFFFFFFFFFFFF"

For building circuits Tor uses a modified Pareto distribution formula [10]. Where " s" is the total number of completed circuit they have seen, "x_i" as their "i-th" completed circuit time, "x_max" as the longest observed completed circuit, "u" as the number of unobserved timeouts without a record of their exact value and "n" as "u+s" as the total of circuit which has timeouts or completed and finally using the log laws they compute the following as the sum of logs in order to avoid overview and ln(1.0+epsilon) precision issues:

$$alpha_m = s/(u * \ln(x\_max) + Sum\_s\{\ln(x\_i)\} - n * \ln(Xm)) \quad (1)$$

For the connections between two tor relays or between a client and a relay, it uses TLS/SSLv3 for link authentication and encryption. And those are just few of the methods that the Tor network uses. The main vulnerability was found at tor exit point where the security is very low compared to the rest of the Tor network. Tor offers 7000 relays, around 2000 entry nodes and around 1000 exit nodes so the odds of being detected are one in two million (1/2000 x 1/1000) give or take. Because when it comes to computer nothing is 100% sure.

Tor faces criticism due to the fact that it's also used by criminals and terrorist to communicate between them and make their tracking hard to execute.

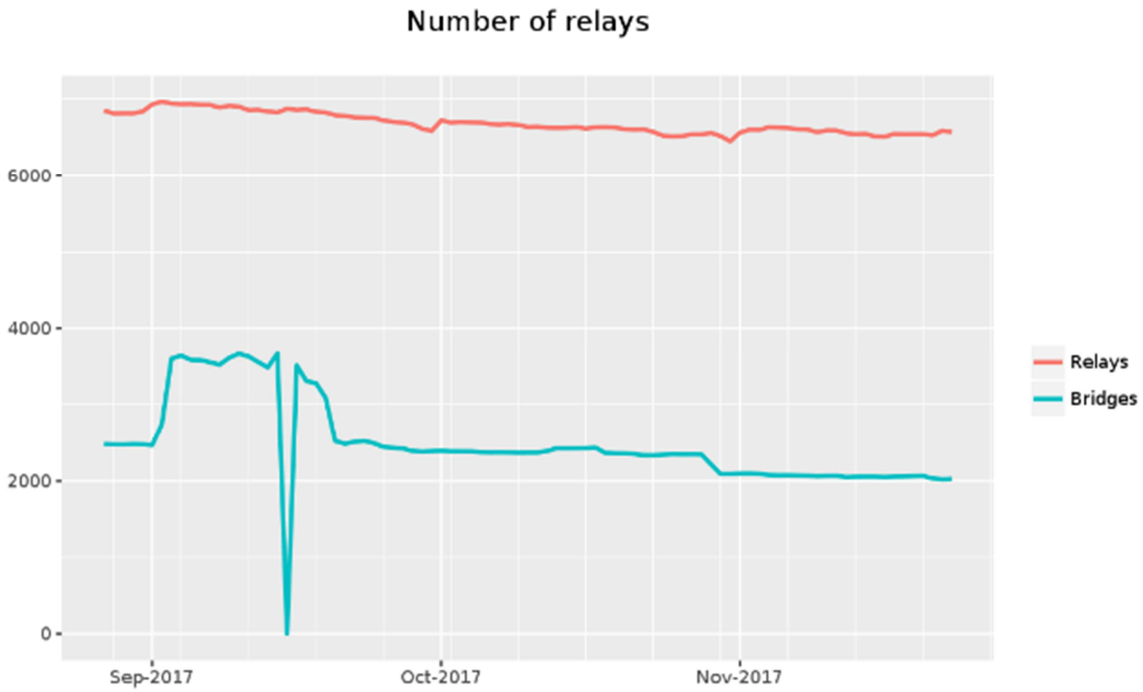

Figure 3. *Used Tor relays and bridges.*

I2p:

I2p or the invisible internet project is an anonymous network which is somehow similar to the Tor network but the network itself is strictly message based but there is a library available to allow reliable streaming communication on top of it [11].

-How to access it: the first thing you need to do if you are on windows is to install the i2p installer from the official website or if you are using a debian based operating system you can add the repositories. One can also use the tor browser to access the i2p network but will need to configure a little bit the browser.

It is used with i2p bote which is an email service which focuses on secure and anonymous email or for IRC (Internet Relay Chat) or other things such as eepsites (websites hosted on the i2p network) or for torrents.

-How it works: all communication is end to end encrypted using four layers of encryption even the end points are cryptographic identifiers, essentially a pair of public keys. To anonymize the message sent each client application use their i2p "router" to build few inbound and outbound "tunnels". When a client wants to send a message to another client, the client uses one of their outbound tunnels targeting one of the other client's inbound tunnels, eventually reaching the destination. According to their needs each client chooses the length of these tunnels.

The first time a client wants to contact another client, they make a query against the fully distributed network database, a custom structured distributed hash table (DHT) which is based on the kadmelia algorithm. This system is in place in order to allow the user to find the other client inbound tunnels easily. For the end to end encryption the i2p network uses AES with 256 bit keys and 128 bit blocks in CBC mode and also uses common primes for 2048 ELGamal encryption and decryption and uses garlic routing method to transport the message. According to JP Timpanaro [12] in order to perform a successful Sybil attack an attacker would need to generate in average more than 12K fake routing keys before finding the appropriate number to perform the attack. Considering K=3 and N~=4000:

$$fake\_routing\_keys = K * 2^{[log_2 N]+1} \quad (2)$$

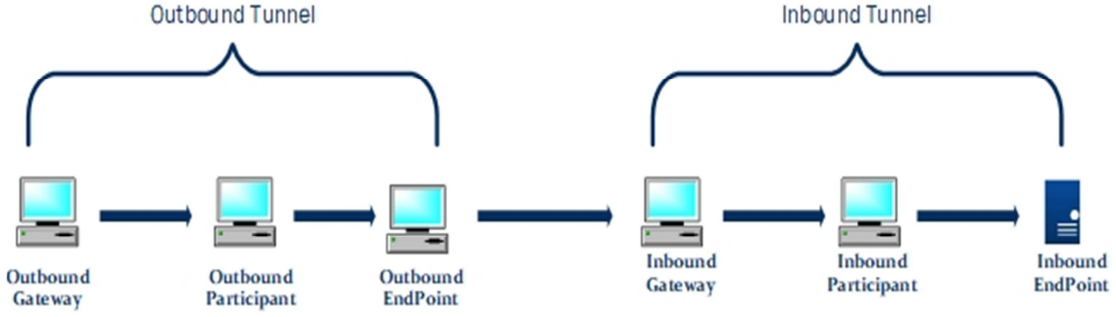

Figure 4. *I2p network.*

Though Tor network and I2p are similar they are not identical [13]. The following tables show the main differences.

Table 2. *Terminology differences between Tor and I2p.*

I2p	Tor
Message	Cell
Router or client	Client
Tunnel	Circuit
NetDb	Directory
Floodfill Router	Directory Server
Fast Peers	Entry Guards
Inproxy	Entry Node
Outproxy	Exit Node
Hidden service, Eepsite or Destination	Hidden Service
LeaseSet	Hidden Service Descriptor
Inbound Gateway	Introduction point
Router Hidden	Node
I2p Tunnel Client	Onion Proxy
service, Eepsite or Destination	Onion Service
Router	Relay
Inbound Gateway + Outbound Endpoint	Rendezvous Point
RouterInfo	Router Descriptor
Router	Server

Table 3. *Features comparison between Tor and I2p.*

	I2p	Tor
User base	Small	Much bigger
Academic visibility	Less visibility	Much more visibility
Hacker community	Less visibility	Much more visibility
Funding	Less	Significant
Developers	Less	Much more
Scalability	Less better	Much better
Documentation	Poorly documented	Well documented
DOS attack	Less vulnerable	More vulnerable
Exit nodes	Less number of them	Large number of them
Memory usage	Inefficient	More efficient
Documentation in different languages	Unavailable	Available
Website	Good	Better
Software language	Java	C
Latency	High	Low
Bandwidth overheard	Very high	Very low
Throughput	Lower	Higher
Sybil attack	Not vulnerable	Vulnerable
Control	Distributed	Centralized
Switch method	Packet switched	Circuit switched
Directional	Uni-directional tunnel	Bi-directional circuit
Life of tunnels/circuit	Short	Long
Nodes selection criteria	Continuously profiling and ranking performance	Trusting claimed capacity
TCP/UDP transport	Both	TCP
Directory servers/floodfill peers	Varying and un-trusted	Trusted and hard coded

The tables above shows that, Tor network is clearly better than I2p but still have some flaws. The researchers of this study stipulate that the best way to reduce those flaws is to add a VPN, and by doing so use Tor over VPN. For the most amateur end users, it's just using the Tor browser and a VPN. The Tor browser will configure automatically the minimum needed configurations.

Figure 5. *TOR over VPN.*

5. Conclusion

In this paper, the researchers suggest some few solutions to deal with the privacy issues in search engines. Though there are other methods, the study focuses on what the users can do to protect themselves and the most utilized methods. In all of the methods suggested, Tor network is the best if a user want to use only one method, and this study recommends the use of Tor network over a VPN since that seem to be the most absolute way to protect user privacy on the web. In addition, there is a need to remember nothing is 100% surety and that one can only use the best way available.

References

[1] Hinman, L. M. (2005) "Esse est indicato in Google: Ethical and political issues in search engines." International Review of Information Ethics, 3(6), pp19-25.

[2] Ye, S., Wu, F., Pandey, R., & Chen, H. (2009) "Noise injection for search privacy protection." In Computational Science and Engineering, 2009. CSE'09. International Conference on Vol. 3. IEEE, pp1-8.

[3] Dr. Mintz, S. (2015), Is it time for a Code of Ethics for SEOs? Orfalea College of Business at Cal Poly San Luis Obispo.

[4] Clifton, C., Jiang, W., Muruguesan, M., & Nergiz, M. E. (2008) "Is privacy still an issue for data mining?" NGDM, Taylor and Francis.

[5] Singh, D. K., & Swaroop, V. (2013) "Data security and privacy in data mining: research issues & preparation." International Journal of Computer Trends and Technology, 4(2), pp194-200.

[6] Mason, Andrew G. (2002). Cisco Secure Virtual Private Network. Cisco Press. p. 7.

[7] "Virtual Private Networking: An Overview". Microsoft TechNet. September 4, 2001.

[8] World-Wide Web Proxies, Ari Luotonen, April 1994

[9] Tor metrics, https://metrics.torproject.org/

[10] Tor network project, https://www.torproject.org/

[11] I2p anonymous network, https://geti2p.net/

[12] Timpanaro, J. P., Cholez, T., Chrisment, I., & Festor, O. (2015) "Evaluation of the anonymous I2P network's design choices against performance and security." In Information Systems Security and Privacy (ICISSP), 2015 International Conference on. IEEE, pp1-10.

[13] Ali, A., Khan, M., Saddique, M., Pirzada, U., Zohaib, M., Ahmad, I., & Debnath, N. (2016) "TOR vs I2P: A comparative study." In Industrial Technology (ICIT), 2016 IEEE International Conference on. IEEE, pp1748-1751.