Narrowing Down the Secrets of the Internet - A Review of Privacy Leakages and Prevention Methods

Bojin Chen[1(✉)], Congjun Xu[2], Wenxiang Cai[3], and Yuheng Xia[4]

[1] Wuhan No.2 High School, Wuhan, China
code004accepted@gmail.com
[2] School of Software Technology, Dalian University of Technology, Dalian, China
[3] Department of Computer Science, University of Sheffield, Sheffield, UK
[4] Department of Mathematics and Physics, North China Electric Power University, Beijing, China

Abstract. With the massive growth of Internet users, the demand for users to browse the web has also increased enormously. As a consequence of this, users leave a significant number of browser footprints online in addition to a significant amount of personal data. In addition, many third-party websites provide users with free personalised services by default, while collecting and recording users' private information without their knowledge. This data will likely be leaked to third-party companies for secondary purposes such as identity theft. This makes it more challenging to maintain internet privacy. In this paper, we have introduced the current mainstream HTTPS protocol and the process of browsing the Internet in detail. Additionally, we discuss several projects built to enhance privacy—Tor, I2P, and VPNs—explain how each improves online privacy and security. At the same time, we also introduce some common means of enhancing privacy and possible risks.

Keywords: Privacy · Internet · Network · Leakage · Security

1 Introduction

There is a rising risk for Internet users about the privacy of their personal information due to the growing number of daily activities conducted online. Due to the alarming rise in web information leakage, online browsing, banking, transacting, shopping, social network interaction, and any online collaboration or communication can jeopardise an individual's right to privacy. In particular, the users' information can be accessed, acquired, kept, mined, connected, shared, contracted, and perhaps sold, mostly without their knowledge or consent and primarily for financial gain. By endangering the privacy of both their online and private lives and posing various intriguing privacy concerns, the data trails and footprints left behind during Web browsing enable third-party corporations and aggregators to create digital dossiers of Internet users.

© The Author(s) 2023
R. B. B. M. Hussain et al. (Eds.): ICHSSR 2023, ASSEHR 765, pp. 87–95, 2023.
https://doi.org/10.2991/978-2-38476-092-3_13

Nearly everyone in today's world regularly uses the Internet and leaves digital footprints on many websites. Our privacy is more exposed as a result. Websites like Google, Amazon and Facebook keep track of our purchase patterns, query history, and other sensitive data. Large commercial enterprises (aggregators, analytics services, ad servers, etc.) continually monitor user activity, behaviour, and habits to grab customers' attention and deliver them with tailored advertising. The process of monitoring someone's online activities is not necessarily harmful, but it does have a significant influence on users and the network, in addition to increasing the money from marketing efforts [1].

Furthermore, there are insufficient safeguards in place to protect user data once these websites have collected it. Therefore, there is a high probability that these pseudo-anonymous data and personally identifiable information (i.e. email addresses, full names, addresses, phone numbers, fax numbers, credit card numbers, social security numbers, etc.) will be leaked by insiders to third-party companies, or be stolen by network hackers. This data might be exploited for secondary purposes such as identity theft, social engineering assaults, online and off-line tracking, etc. [2].

In a series of significant studies, information leakage has been recognised as a pervasive problem [3–5]. A study [5] that highlighted the seriousness of the issue found that 56% of the sites analysed (75% when considering users) directly leak sensitive and identifiable information to third-party aggregators. More recent statistics show that 63% of users agreed with a statement of concern for third parties monitoring activities [6]. Internet users are increasingly concerned about privacy due to extensive records of personal behaviour, the combination of anonymous clickstream data and personally identifiable information, and its disclosure to third-party websites without permission or consent. These problems require more effective solutions to protect privacy from violations.

The paper is organised as follows: we present the historical context of online privacy in Sect. 2 below, the iteration of the Hypertext Transfer Protocol (HTTP) to Hypertext Transfer Protocol Secure. Section 3 describes how information is transmitted and leaked over the Internet. In Sect. 4, we discuss three professional projects (Tor, I2P and virtual private network) built to enhance privacy. In addition, we describe several common measures to enhance privacy.

2 The Historical Background of Internet Privacy

2.1 Related Concepts of HTTP

HTTP. HTTP is Hyper Text Transfer Protocol. It stipulates the communication rules between the browser and the server and is the basic protocol for network transmission widely used at present. At first, the HTTP protocol was only used to transmit HTML documents. However, with the development of the Internet, the HTTP protocol has encountered more and more problems, mainly concerning performance and security. Performance: the original HTTP protocol can only send one request, so SPDY and HTTP 2.0 have been developed. Security: the communication of HTTP protocol uses plaintext, and the content is extremely vulnerable to eavesdropping. The identity of the communicating party does not need to be verified so that it may be masqueraded.

SSL and TSL. SSL is Secure Socket Layer. TSL is Transport Security Layer. SSL protocol was developed by Netscape Communications Corporation and later taken charge by IETF and renamed TSL. SSL protocol uses cryptography technology to ensure the security of the communication process.

HTTPS. HTTPS is Hyper Text Transfer Protocol Secure. Constructed by HTTP and SSL, it is not a new protocol. Instead, it only replaces the HTTP communication interface with SSL. Using HTTPS in the communication process can effectively ensure the security of communication.

2.2 Interaction Process of HTTPS Protocol

Apply for SSL Certificate. The server applies for a public key to the CA (certificate authority). CA will review the information provided by the applicant. After the review is passed, CA will use its private key to digitally sign the public key and bind the public key to the public key certificate. Then assign this certificate to the applicant. When a client posts an HTTPS request, the server will send the public key certificate to the client for encrypted communication. The client receiving the certificate will use the CA's public key to verify the certificate's digital signature. When the authentication is passed, the server identity is valid.

Information Interaction. The client enters the HTTPS URL and establishes a connection with the server through the TCP three-way handshake. Then, the client sends a Client Hello message, including the SSL protocol version number supported by the client, the encryption method and Client random. The server sends a series of messages, including confirming the encryption protocol version, the public key certificate and the Server random. The client confirms that the digital certificate is valid. Then it generates a Pre-master secret, encrypts it with the public key in the digital certificate, and sends it to the server. The client sends a Change Clipper Spec message to remind the server that the subsequent communication adopts Pre-master secret key encryption, and the server decrypts the message with its own private key to obtain the Pre-master secret sent by the client. The Finished message contains the overall check value of all messages connected so far. The server will also send Change Clipper Spec and Finished messages. After that, the client and server generate Session Key by the three random numbers and use it to encrypt the entire conversation. The client and server then begin to transmit encrypted information.

3 The Process of Browsing the Internet

When a user browses the Internet, they send a network request, including the destined website, the credential information of their account, and extra data, including settings of their desktop environment and their IP address, which can be used to interpret their physical location.

This information is first sent by the browser or Internet application on their device towards their router, which provides the Wi-Fi service. Then, the request goes through

the user's Internet Service Provider (ISP) and is sent across the globe to the location of the destination website. The network request then passes the ISP of the requested site and arrives at the website itself, providing it with the users' information to deal with.

If the connection uses a secure HTTPS protocol, the username and password, along with the extra data the user sends, will be encrypted throughout the process and only be unencrypted once it reaches the destination. This way, hackers connecting to the user's Wi-Fi will not be able to learn the information sent, leaving them with only the destination website and the user's physical location. Similarly, the ISPs on both sides, along with eavesdroppers in between, can only interpret the website browsed and the physical location of the user.

An unencrypted HTTP connection, however, could lead to severe information leakage. As everything is left unencrypted, hackers, eavesdroppers, and ISPs can learn everything about the user's connection, including their passwords and their extra data sent, not to mention their physical locations and the website being browsed.

4 Professional Projects Built to Enhance Privacy

One may notice that the location of the user is revealed even when the user is using HTTPS. This can be prevented by using projects built to hide users' IP addresses and assist them in protecting their privacy. The following are some examples.

4.1 Tor

The Onion Router, or Tor for short, is a project built by the U.S. Naval Research Lab (NRL). It aims to provide Internet connections that do not reveal their origins or destinations. Relying on a decentralised network operated by thousands of volunteers around the globe, the Tor project sends a user's request through a series of proxies, known as relays, before it reaches its destination, the website the user tries to visit.

While using a typical proxy, where the service provider can surveillant and monitor the complete traffic, learning both the user's physical location and the website they are trying to visit, the multiple relays in a Tor connection process are randomly selected among nodes in various countries belonging to various organisations or individuals. Besides, the user's client establishes a temporary key with each relay in the connection circuit, allowing only the final and exit relay to read the user's traffic details. Therefore, information leakage with Tor is nearly impossible. Although the first node may still learn the user's IP address, and the last node can interpret the destination of the traffic and even the data inside if the user does not use encryptions like HTTPS, this vague information is unlikely to convey any privacy of the user.

Built upon the Extended Support Release of the famous Firefox browser, the Tor Browser consists of features that help increase the browser's security. These features include disabling JavaScript for potentially dangerous websites, making the privacy-focused DuckDuckGo search engine its default option and more. Also, it assists users in censored areas in establishing a connection with a Tor relay by providing bridges and other censorship circumvention services.

Despite its sophisticated structure and an enormous number of relays, Tor, like all other proxy services, is vulnerable to correlate timing attacks. An adversary can correlate an outgoing connection from a user's device to an ingoing connection towards websites that the user browses by their relevant timings, causing the connection to lose anonymity potentially. However, since the probability of this event is near zero, we can still regard Tor as one of the most privacy-protecting services.

4.2 I2P

The I2P (Invisible Internet Project) is another project which focuses on enhancing privacy and protecting anonymity. With every client also working as a node, the I2P tunnels every connection through a series of nodes, making attackers unable to tell which node a user is.

In I2P, every client shares bandwidth and becomes a node in the network. It fetches a list of initial nodes from a server called "Reseed Server", before using these as peers to reach more nodes and fully join the client into the network. Then, the client connects to these peers each time it starts. Clients also establish encrypted and unidirectional tunnels with each other to ensure the data transmitted between them cannot be monitored in any way.

Different from Tor where routes between nodes are predetermined by a centralised service that records information about all nodes, each node in an I2P network records the performances of its peers and tunnels incoming connections through these nodes based on their performance rank. The out proxy, similar to the Exit Node in a Tor circuit, is another centralised service provided by volunteers.

Tunnels in I2P are short-lived compared to Tor. This tactic increases the network's security by providing fewer samples for an attacker to launch an attack [7]. Besides, this provides extra protection for a user by switching their connections more often and decreasing the possibility of the user being monitored by a malicious node. Also, while circuits in a Tor connection usually go through only three nodes, connections through I2P transport many more nodes before reaching the target of a request. The selection procedure introduced earlier ensures the connection's speed, eliminating nodes with low connection speeds or unstable performance.

As mentioned previously with Tor, I2P is also vulnerable to the timing correlation attack. However, its security is still maximised by the enormous number of clients, all of which also perform as nodes, and the complicated process for a request sent by a client to reach our everyday Internet.

4.3 VPN

The Virtual Private Network (VPN) is a tool mainly used to hide the IP address, and therefore the physical location, of a user, by tunnelling a network request through the VPN service provider before it reaches the destination. This confuses the receiver of the network request by leading them to think that the request was sent from the location of the VPN service provider instead of the user.

To access a VPN service provided by organisations, users need to install their dedicated client software and correctly configure them. Then, the data of the user is first

encrypted locally and sent to the server of the VPN service provider, which decrypts the data and forwards it to its destined website. This increases the security of the user by adding an extra layer of encryption initially and the privacy of the user by preventing leakage of the user's IP address.

Nowadays, people also use VPN services to circumvent censorship and regain Internet freedom by pretending to be users in a distant country, therefore fooling both the destination server and the censorship methods established by governments or organisations. Although circumvention of censorship is not the initial purpose of VPN, it has become the most crucial factor attracting people to pay for such a service.

4.4 Other Projects

There are various other projects which focus on protecting users' privacy. Tails is a portable operating system built upon the Debian distribution of Linux. Tor, as its default networking application, provides the same level of protection as Tor system-wide by directing all Internet traffic throughout the system through the Tor network. Besides, Tails enhances its safety by protecting users against cold boot attacks, removing metadata in files with the mat2 tool, and using a live pattern to leave no traces on devices after a session is closed. The persistent storage in this USB-dedicated system is encrypted with LUKS, the default disk encryption technique used across popular Linux distributions, making it impossible to fetch any information without the passphrase.

Subgraph OS is another operating system based on the Tor network. Separating different applications into different containers and banning unnecessary containers from accessing the Internet provides extra protection against malicious software leaking users' physical locations and other information. It also uses full disk encryptions, cold boot attack prevention methods, and a Grsecurity-enabled hardened kernel, protecting its users from adversaries while providing security enhancement for everyday software.

DNSCrypt is an open-source project which can encrypt and anonymise the communications between users and DNS resolvers. It encrypts and authenticates the DNS traffic, a previously unencrypted traffic, and uses substantial amounts of servers to support the protocol, providing extra protection against DNS spoofing. Specifically designed for DNS, it defends against diverse types of DNS attacks, providing a safe browsing environment for the user. There are multiple open-source client implementations for DNSCrypt, making it an easy option to enhance security.

There are also multiple pieces of software with Tor integrated to provide extra anonymity, such as OnionShare, an open-source file transfer tool, and Whonix, a virtual machine friendly operating system based on the Debian distribution of Linux. These projects also protect the users' privacy by obfuscating network traffic, therefore anonymising the connection between the client and the Internet.

5 Common Measures to Enhance Privacy

Besides these professional projects, we have some common measures to enhance privacy. In this paper, we especially emphasise one method to do it. It refers to proxy software or server and can also be considered a network access method. It performs other operations

that things do not want or cannot perform. For example, when we operate on the database, the agent can record our operations after we operate on the database. It is a kind of firewall that works in the application layer, characterised by two connections. It can copy a copy of the files and web page data accessed remotely in the proxies at the near end. If the proxy is set online, it will check whether other people have visited the same website before each time when connecting to the web page. If so, it can directly send back the data without connecting to the outside. Its function is to obtain network information on behalf of network users. It is the relay station of network information. Generally, when we use a web browser to connect to other Internet sites to obtain network information, we directly contact the destination site server. Then the destination site server sends the information back [8].

Because all users of the intranet access the outside world through the proxy server, only one IP address is mapped, so the outside world cannot directly access the intranet. At the same time, IP address filtering can be set to restrict the access rights of the intranet to the outside; In addition, two internal networks without interconnection can also be interconnected through a third-party proxy server to exchange information. As mentioned above, all users only use one IP address externally, so it is not necessary to rent too many IP addresses to reduce the maintenance cost of the network. In this way, many machines in the LAN that are not connected to the external network can be connected to the external network through a proxy server of the internal network, greatly reducing the cost. Of course, there are also its disadvantages. For example, many network hackers hide their real IP addresses in this way to escape surveillance. Moreover, it has a small bandwidth and is connected to the target host through a proxy with large bandwidth. Moreover, the proxy server usually sets a large hard disk buffer. When there is external information passing through, it also saves it in the buffer. When other users access the same information again, the information is directly taken out from the buffer and transmitted to the user to improve the access speed.

The proxy server is another server between the browser and the web server. With it, the browser does not directly go to the webserver to retrieve the web page but sends a request to the proxy server. The signal will be sent to the proxy server first, and the proxy server will retrieve the information required by the browser and send it to our browser [9].

Moreover, its function can be summarised in the following points:

1. Improve internal access speed;
2. Act like a firewall;
3. Visit some websites that cannot be directly accessed;
4. Improve the security of Internet access
5. Some related access restrictions
6. Traffic restrictions that affect bandwidth too much
7. Block some access requests that affect the internal flow of the company
8. Set restrictions on ports to prevent illegal attacks

It also has some possible risks. When using the proxy server of a professional network service provider, our route and process may be recorded. If the network management is willing, it can even completely monitor our entire process and visit all information including Internet access time, route, various applications submitted, feedback, etc.,

which may lead to the leakage of secrets. For other users and destination servers on the Internet, we are safe, but for the proxy server itself, we are at a glance [10]. The administrator of the proxy server or the person who has the authority to manage the proxy server through other means can easily have our secret.

6 Conclusion

This paper analyses and explains the methods to enhance our privacy. In general, we start with the historical background of internet privacy, mentioning some necessary and well-known protocols and other common and basic knowledge in this background. In this part, we show that the client and the server often have privacy information interaction, and the third party can often intercept the information to obtain the client's private information. In order to strengthen privacy protection, people use the HTTPS protocol so that the user's credential information will be encrypted. We have also studied multiple projects built to satisfy such needs. A brief introduction of the basics of daily network protocols and data transferring policies is included to demonstrate what various projects have done to enhance our privacy and introduce their strengths and weaknesses. The result indicated that currently, there is no practical solution against time correlation attacks, as all proxies, VPNs, and projects, including Tor and I2P, fail to obfuscate timings, which is impossible in cryptographically secure connections. Finally, we discussed the common methods people could use to help them to protect their privacy.

Acknowledgment. Bojin Chen, Wenxiang Cai, Congjun Xu, and Yuheng Xia contributed equally to this work and should be considered co-first authors.

References

1. "Q3 2016 Internet Ad Revenues Hit $17.6 Billion, Climbing 20% Year-Over-Year, According to IAB", IAB, Dec. 28, 2016. https://www.iab.com/news/q3-2016-internet-ad-revenues-hit-17-6-billion-climbing-20-year-year-according-iab/ (accessed Sep. 2, 2022).
2. D. Perito, C. Castelluccia, M.A. Kâafar and P. Manils, "How unique and traceable are usernames?", 11th International Symposium on Privacy Enhancing Technologies, Waterloo, ON, Canada, July 27–29, 2011, pp. 1–17
3. D. Irani, S. Webb, K. Li and C. Pu, "Large online social footprints – an emerging threat", CSE '09, International Conference on Computational Science and Engineering, vol. 3, 2009, pp. 271–276.
4. B. Krishnamurthy and C.E. Wills, "Privacy leakage in mobile online social networks", Proceedings of the 3rd Conference on Online Social Networks, WOSN'10, USENIX Association, Berkeley, CA, USA, 2010, p. 4
5. B. Krishnamurthy, K. Naryshkin and C.E. Wills, "Privacy leakage vs. protection measures: the growing disconnect", Web 2.0 Security and Privacy Workshop, 2011.
6. C.E. Wills and M. Zeljkovic, "A personalized approach to web privacy – awareness, attitudes and actions", Information Management & Computer Security, 19 (1), 2011, pp. 53-73

7. “I2P Compared to Tor - I2P”, I2P Anonymous Net-work, Nov. 1, 2016. https://geti2p.net/en/comparison/tor (accessed Sep. 2, 2022).
8. “What Is Pay As You Go Internet?”, Netinbag, Jun. 2014. https://www.netinbag.com/en/internet/what-is-pay-as-you-go-internet.html (accessed Sep. 3, 2022).
9. M Mambo, K Usuda and E Okamoto, “Proxy Signatures for Delegating Signing Operation”, CCS ’96, Proceedings of the 3rd ACM Conference on Computer and Communications Security, New Delhi, India, March 14–16, 1996.
10. C Pei, S Irani, “Cost-aware WWW proxy caching algorithms”, USENIX Symposium on Internet Technologies and Systems, 1997.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.