Identifying Originating Traffic to Anonymity Networks

Owen Searls

December 14, 2017

Abstract

Anonymity software such as Tor allows users to hide their identity, location, and activity while using the internet. Tor does not, however, hide the fact that a person is using Tor; It is possible to identify Tor traffic both as it is entering and exiting the network. This limitation of Tor is well understood, and in many cases does not make Tor as a tool any less useful. Because Tor is a rarely used tool, the ability to identify Tor users can give network administrators an idea of who might have something to hide. If someone who is attempting to identify the source of Tor traffic can make a guess as to what network that traffic originated from, they can reduce their list of suspects to only the people who were using Tor on that network at the time in question. It is easy to imagine a situation where the only people with the motive to execute an attack or the knowledge to become a whistleblower are the employees of a particular company, or the students of a particular university. The focus of this article will be on methods of identifying Tor users on a network, and how those users can avoid discovery, as well as the security and policy implications of the ease with Tor users can be identified.

Contents

Introduction

By the nature of the Internet Protocol it is generally possible to identify the source of any internet traffic, and with enough scrutiny, to connect that source to a particular person. This ability can be used by anyone from network administrators at schools or offices to government spy agencies to monitor the online behavior of internet users, and by tech companies to compile large databases of personal information which are often handed over to law enforcement with little judicial oversight. Considering the ease with which internet traffic can be monitored, it is clear why some users would want to remain anonymous online, and why this ability is especially important to those who face persecution from their governments or communities, find their freedom of speech restricted, or seek to share evidence of institutional wrongdoing.

One of the most widely used and most secure internet anonymity tools is Tor, a protocol and software package developed at the United States Naval Research Laboratory in the 1990s, and release as free and open source software in 2004. Tor operates using an onion routing technique, where data sent through the Tor network passes through a series of nodes, each of which removes a layer of encryption to reveal the address of the next node the data should be sent to. The Tor protocol guarantees that data sent through the network will pass through at least 3 Tor servers, and the principles of onion routing mean that no node will in the chain will know more than the addresses of the nodes directly before and after it, or if any data it receives is from a Tor user or another intermediate node. Only the last server in this chain will be able to identify it's location in the chain. This server, known as the exit node, will remove the final layer of encryption to reveal a destination address outside of the Tor network, and will be able to read any data originally sent to Tor in the clear. (*About Tor* n.d.)

The limitations of Tor are well documented, but perhaps not well understood by most users. Tor might be best described as a black box which can be inserted into the middle of a stream of internet traffic, thereby hiding the identity of a Tor user from the person they are communicating with, and preventing any potentially malicious observer from knowing who the Tor user is talking to. Tor does not attempt to hide that traffic on either side of that black box is going to or coming from Tor; In fact the IP addresses of Tor nodes are made publicly available. This means that it is generally simple to identify internet traffic coming from Tor as well as originating traffic to the Tor network. If it is possible to identify this traffic, it is also of course possible to block it, and in the case of originating traffic, it will also often be possible to identify the Tor user in the real world.

To the Community

In recent years, and especially since the 2016 elections, many Americans have become more suspicious of their government, and of government surveillance in particular. It is a common refrain that this type of surveillance doesn't affect people who have nothing to hide or aren't breaking the law, but when Americans concede that the government might sometimes pass laws that are worth breaking or target activists because of their political beliefs, the importance of maintaining privacy and enabling secure communication becomes especially

clear. Around the world, there are millions of people living under repressive governments that actively monitor internet users to clamp down on dissent and prevent the free flow of ideas. Tor is a valuable tool for activists, dissenters, and members of persecuted social groups to safely organize and protect their identities online, in other words for people who do have something to hide. When these individuals make up the majority, or even a large fraction of Tor traffic, the very fact that a person is using Tor can make them the target of an oppressive regime. It is possible to prevent identification by accessing the Tor network through an unlisted bridge relay and using some kind of protocol obfuscation, but when facing a significant enough opponent this becomes a cat and mouse game. Ultimately, the solution to this problem is to raise awareness of security issues and encourage wider adoption of tools like Tor among all types of internet users.

History of the Issue

The history of the Tor project is in some ways inextricably tied to the issue being considered in this paper. Tor was originally developed by the military for defense purposes, for example to allow law enforcement officials to hide their government affiliations while conducting operations online. The problem that this presents, of course, is that if only government agents use Tor any Tor traffic that someone who suspects they are under investigation receives will be easily identified as the work of law enforcement. Michael Reed, one of the original creators of Tor, wrote in the tor-talk mailing list that this problem is fundamentally why Tor became open source, saying that the fact that Tor would be used for other potentially malicious uses was "immaterial to the problem at hand" and in fact necessary to the parts of the government that wanted the abilities that Tor provides in the first place. (Reed 2011)

After Tor was made publicly available in 2004 the number of non-government users increased dramatically, but perhaps not enough that someone observing a network would expect to see anyone on that network using Tor. This fact has been exploited by law enforcement agencies to target Tor users for investigation, and in some cases prosecution. In one of the more famous examples of a Tor user being identified and linked to an online crime, in 2013 a Harvard University undergraduate sent a bomb threat to school officials using Tor and an anonymous email service named Guerilla Mail. During the course of the investigation FBI agents working with local police and Harvard campus IT were able to determine that only one student on the Harvard wireless network used Tor during the time the email was sent, and upon questioning that student quickly produced a confession. (Memmott 2013)In this instance it is particularly interesting that Tor did it's job properly. There was no way to connect the email received by Harvard to the students use of Tor, but once a suspect was identified conventional police techniques were sufficient to connect that chain. If the student had wiped their computer of any incriminating evidence and simply denied any wrongdoing it is unlikely that this case would have resulted in a conviction. If the student had used a network unaffiliated with Harvard or connected to Tor in a less identifiable way they may never have faced police scrutiny at all.

More generally, governments have used information on who is using Tor to target users for increased computer surveillance. Documents leaked by Edward Snowden in 2013 revealed that the NSA and GCHQ target Tor users with malware that can log keystrokes and internet

traffic, and give the agencies access to local files. With this information law enforcement should with some work be able to reliably connect the actions of the person under surveillance with the actions of the anonymous Tor user being identified, if they are in fact the same person. Again, it is significant that the basic Tor protocol remained secure against attacks by such well funded agencies as the NSA and GCHQ, with NSA documents revealing that the agency has never been able to reliably de-anonymize users at will. (Ball, Schneier, and Greenwald 2013) An attack that simply targets all Tor users, however, is potentially as dangerous, and carries the added risk of collateral damage to other Tor users.

While the legality of the NSA's spying is unclear, and their actions were widely condemned after the Snowden revelations, this type of profiling of Tor users was in some part validated by the Supreme Court of the United States in 2016. In April of that year, the court decided that magistrate judges could approve warrants for the seizure of computers and electronic devices regardless of their physical location if Tor had been used on that computer. Previously, judges had only been able to authorize this type of warrant in their own district. These warrants can also often be used to justify exactly the type of hacking that the NSA was revealed to be involved in three years earlier. (McLaughlin 2016) This enshrinement of the profiling of Tor users into federal law sets a precedent that a desire for anonymity suggests guilt, and gives significant new powers to the FBI. This type of development only raises the stakes for the issue of protecting the knowledge that a person is using Tor in addition to their identity.

Methods of Identifying Originating Traffic

Whenever a users network connection is being monitored they may find themselves at risk of being identified as using Tor, and perhaps profiled because of their desire for anonymity. This surveillance can occur at the level of a network administrator, ISP, or government agency.

Known Tor Relays

The simplest method of identifying originating originating traffic to Tor is to look for packets being sent to the IP addresses of listed Tor routers. To function as part of the Tor network these routers identify themselves to Tor directory authorities which then distribute the complete list of relays to Tor clients. This list is freely distributed to Tor clients without any attempt to block malicious users, and is also available on the Tor Project website. Additionally, the Tor Project also publishes archives of all Tor routers in use on any given date going back to October of 2007. While there are ways to avoid communicating directly with a publicly listed node, and therefore detection via this strategy, the vast majority of Tor users are vulnerable. This technique is also well within the capabilities of even a relatively unsophisticated network admin, and may be possible to execute after the fact from network log files. It is worth noting that many servers which operate Tor routers also host website or other types of internet services. This means that observing a user connecting to one of these servers is not sufficient to show that they are using Tor. Ideally this would provide some measure of protection against this technique, but in the worst case scenario this added measure of doubt would likely be of little consequence to an authoritarian government.

In order to avoid identification by this method a Tor user must use what is known as a bridge node. These unlisted Tor entry nodes work in the same way as a normal Tor router, but have a much narrower distribution. Users who report that they face internet censorship when configuring the Tor browser will be prompted to connect through one of the bridges distributed with the browser, and bridges can also be distributed through other even more secure methods including email at the address bridges@bridges.torproject.org. There are a limited number of Tor bridges in operation, and those that do exist are always at risk of becoming too widely known. Together this means that users who connect to Tor through a bridge may face significantly slower connections, and that the number of bridges is certainly insufficient to support all or even most Tor users. Likely for this reason, the Tor Project only recommends bridges for users facing censorship from their ISP or government. In the page of security advice distributed with the Tor browser and on the download page users who are concerned that someone could discover that they are using Tor are advised that they could use a bridge, but that "ultimately the best protection is a social approach: the more Tor users there are near you and the more diverse their interests, the less dangerous it will be that you are one of them". (*Download Tor* n.d.) This true, and combined with more general advice such as only connecting to Tor on networks outside of your home, work, or school, solves many aspects of the problem. Unfortunately, the individuals who have the greatest need for online anonymity might often be part of communities where tools like Tor are rarely used, and cover traffic will be of little when Tor is simply blocked or otherwise censored.

Of course, if an adversary is able to discover the address of a Tor bridge, they can add it to there database of Tor relays and identify any traffic sent to that server using the same simple methods that can be used to identify Tor traffic that does not utilize a bridge. The most obvious way to identify Tor bridges is to make many requests that appear to originate from legitimate users. The Chinese government has employed this technique to block Tor bridges since shortly after its initial crackdown on Tor in late 2009. Initially, this involved making enough requests to the HTTPS bridge distribution server located at https://bridges.torproject.org/bridges. These requests came from enough unique IP addresses and subnets that the Tor Project was unable to discover that they were coming from a single attacker. In 2010, China again was able to block a significant number of bridges by overwhelming the email based bridge distribution system. (Dingledine 2011) Attackers have also identified Tor bridges by scanning for servers that respond to the TLS handshake in a way that identifies them as a Tor server. This requires some of the same protocol analysis tools that will be discussed in the next section, but is also fundamentally easier because there are less nodes to be identified, and each is responding to requests from many clients. In 2011, security researchers running a Tor server in Singapore found that their server was blocked inside China very soon connection from the country was made. Additionally, they detected a high volume of continuous scanning, much of it from a single IP address. (Winter and Lindskog 2012)

Protocol Analysis

The Tor protocol is based on a Transport Layer Security (TLS) connection, and by default is designed to look like normal HTTPS web traffic. The Tor Protocol Specification states that Tor clients should use similar TLS header fields as other common clients, but these

headers are still often unique to Tor, and can therefore be fingerprinted and used to identify Tor traffic. (Dingledine and Mathewson 2017) This technique can be used to identify traffic to Tor bridges as well as standard guard nodes, because both communicate using the same protocols. This technique is used in China to block Tor traffic, mostly focusing on the particular cipher list used in the TLS handshake by Tor clients. (Winter and Lindskog 2012) Tor traffic is also differentiated from normal HTTPS traffic in other ways, such as unusually long TLS sessions, self signed and randomized certificates, and unusually regular timing characteristics. In general, these techniques can be defeated by using a pluggable transport, several of which ship with the Tor browser. These protocols randomize and obfuscate Tor traffic to prevent fingerprinting attacks which rely on regular patterns of traffic, as well as attacks which rely on unique features of the Tor protocol. Some of these transports rely mostly on simple randomization, while others, such as SkypeMorph, aim to disguise traffic as another commonly used internet service, in this case the video chat app Skype. (Mohajeri Moghaddam et al. 2012) The pluggable transports distributed with Tor are generally effective against less sophisticated attacks, but researchers have found that with enough work adversaries using deep-packet inspection can defeat almost any attempt at obfuscation. (Wang et al. 2015) Ultimately, the feasibility of these methods may depend on the rates of false positives. In the context of censorship in which this issue is most often discussed too many false positives could lead to more damage to the portions of the internet that contribute to economic growth than is palatable to government leaders. In the context of identifying Tor users to target them for further scrutiny, false positives only create extra extra work for police and government spy agencies, which may not be an issue to an authoritarian regime.

If a user is accessing Tor without a bridge, they are almost certain to be identified as using Tor by almost any adversary with the ability to monitor their internet connection. At the same time, most countries which have made serious attempts at blocking Tor use some form of protocol analysis in addition to blocking Tor entry nodes, so this issue is very significant to those users, and will only increase in importance in the United States and other countries where unrestricted internet access is the norm if more people chose to access Tor through an unlisted node in order to protect themselves from profiling. Additionally, if even a minority of users in a country which censors Tor access a bridge server using a Tor client without a pluggable transport or using a compromised transport it will make it significantly easier for that country to identify the server as a bridge and block it for all users.

Traffic Analysis

A potentially more sever threat to Tor users anonymity exists when an attacker is able to monitor broad swaths of the internet, potentially allowing them to match up traffic entering and exiting the Tor network, thus de-anonymizing and subverting the entire function of Tor. This attack could be carried out by an ISP, but is more dangerous in the hands of a national government with the ability to monitor all internet traffic within a country. As described in the last two sections of this paper, it is trivial to identify traffic coming from Tor, and very often possible to identify traffic going to Tor. When an ISP or government agency has the ability to monitor both ends of these connections, they may be able to connect Tor clients to their anonymous actions. Such attacks generally make use of some form of traffic analysis

and statistical methods, which can match up internet traffic on either side of the Tor network by its timing characteristics, numbers and sizes of packets, and bandwidth variations. In laboratory conditions, traffic analysis attacks using data from the Cisco's NetFlow software have been shown to be highly effective. (Chakravarty et al. 2014) The NSA was also revealed to have experimented with this type of attack, but with little meaningful success. (Ball, Schneier, and Greenwald 2013) A national spy agency even more willing to collect data on its citizens internet usage, however, would have a significant advantage in performing a traffic analysis attack. Tor traffic analysis also becomes significantly easier as the attacker controls more Tor routers, giving them intermediate data from inside the Tor network, and reducing the distance that data must travel before it can be de-anaonymized.

Conclusion

In the United States, it seems like very few people are concerned that they could be identified as using Tor, despite the very real risk of being tied tied back to the actions they thought were anonymous through circumstantial evidence and conventional police work, or simply targeted by law enforcement because of their desire for anonymity. Much of this attitude is certainly attributable to the general lack of awareness of security issues that afflicts most internet users, but even the issue of originating traffic to Tor being identified is rarely mentioned even among dedicated Tor users who should be expected to be highly interested in security and privacy issues. Perhaps the ambivalence about this issue is due to the trust that most Americans still have in their government. After all, countries like China and Iran have been actively blocking Tor for over ten years while the Tor Project still receives funding from the Federal Government via the National Science Foundation and the State Department. In many ways, however, the danger to Tor users is much greater when their supposedly anonymous behavior is monitored than when access to Tor is blocked all together, and the FBI has periodically targeted innocent political and civil rights activists for surveillance for it's entire existence. Many Tor users would also probably be more likely to take steps to hide the fact that they are using Tor if better technologies and systems existed to enable them. A push to deploy more Tor bridges and the development of technologies to keep those bridges hidden would make it much easier to connect to the Tor network silently. It may also be worth investigating other protocols that may be fundamentally more secure than Tor, such as a peer to peer system that removes the possibility of identifying any small group of servers to which all users must connect. There is also a much more optimistic view of Tor users behavior surrounding this issue. Maybe most users do not want to hide their usage of Tor, and are in fact happy to show their support for a protocol that they believe to be necessary to protect internet privacy into the future while at the same time creating cover traffic that can protect other Tor users who are at risk of persecution today. Today the fact that a person chooses to use Tor means that they might have something to hide, but if enough innocent people use Tor that knowledge would tell you absolutely nothing.

Supporting Material

The supporting material for this project takes the form of a python library which can identify both originating and exiting Tor traffic using router lookup and protocol analysis techniques. A program that uses this library and can easily detect Tor traffic from a packet capture file or network interface is also included.

The source code and documentation for this project is publicly available at https://github.com/owensearls/Toriginator, and will be submitted alongside this paper in the Toriginator directory.

While the techniques used by government to restrict access to the Tor network are certainly much more sophisticated than anything within the scope of this project, I believe that Toriginator is the most complete standalone open source tool designed to detect originating Tor traffic, and that it is certainly useful as a reference and to raise awareness of the issue.

References

About Tor (n.d.). URL: https://www.torproject.org/about/overview.html.en.

Ball, James, Bruce Schneier, and Glenn Greenwald (2013). "NSA and GCHQ target Tor network that protects anonymity of web users". In: *The Guardian*. URL: https://www.theguardian.com/world/2013/oct/04/nsa-gchq-attack-tor-network-encryption.

Chakravarty, Sambuddho et al. (2014). "On the Effectiveness of Traffic Analysis Against Anonymity Networks Using Flow Records". In: *Proceedings of the 15th International Conference on Passive and Active Measurement - Volume 8362*. PAM 2014. Los Angeles, CA, USA: Springer-Verlag New York, Inc., pp. 247–257. ISBN: 978-3-319-04917-5. DOI: 10.1007/978-3-319-04918-2_24. URL: http://dx.doi.org/10.1007/978-3-319-04918-2_24.

Dingledine, Roger (2011). "Ten ways to discover Tor bridges". In: *Tor Tech Report*.

Dingledine, Roger and Nick Mathewson (2006). *Design of a blocking-resistant anonymity system*. Tech. rep. 2006-1. The Tor Project.

— (2017). *Tor Protocol Specification*. Tech. rep. The Tor Project, Inc. URL: https://gitweb.torproject.org/torspec.git/tree/tor-spec.txt.

Download Tor (n.d.). URL: https://www.torproject.org/download/download.html.en.

McLaughlin, Jenna (2016). "Supreme Court Gives FBI More Hacking Power". In: *The Intercept*. URL: https://theintercept.com/2016/04/28/supreme-court-gives-fbi-more-hacking-power/.

Memmott, Mark (2013). "Student Is Charged In Harvard Bomb Scare". In: *NPR*. URL: https://www.npr.org/sections/thetwo-way/2013/12/18/255198081/student-is-charged-in-harvard-bomb-scare.

Mohajeri Moghaddam, Hooman et al. (2012). "SkypeMorph: Protocol Obfuscation for Tor Bridges". In: *Proceedings of the 2012 ACM Conference on Computer and Communications Security*. CCS '12. Raleigh, North Carolina, USA: ACM, pp. 97–108. ISBN: 978-1-4503-1651-4. DOI: 10.1145/2382196.2382210. URL: http://doi.acm.org/10.1145/2382196.2382210.

Reed, Michael (2011). *Iran cracks down on web dissident technology*. URL: https://lists.torproject.org/pipermail/tor-talk/2011-March/019913.html.

Wang, Liang et al. (2015). "Seeing Through Network-Protocol Obfuscation". In: *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*. CCS '15. Denver, Colorado, USA: ACM, pp. 57–69. ISBN: 978-1-4503-3832-5. DOI: 10.1145/2810103.2813715. URL: http://doi.acm.org/10.1145/2810103.2813715.

Winter, Philipp and Stefan Lindskog (2012). "How the Great Firewall of China is Blocking Tor". In: *Presented as part of the 2nd USENIX Workshop on Free and Open Communications on the Internet*. Bellevue, WA: USENIX. URL: https://www.usenix.org/conference/foci12/workshop-program/presentation/Winter.