**John Schriner**
**Monitoring the Dark Web**

Contrary to what one may expect to read with a title like *Monitoring the Dark Web*, this paper will focus less on how law enforcement works to monitor hidden web sites and services and focus more on how academics and researchers monitor this realm. The paper is divided into three parts: Part One discusses Tor research and how onion services work; Part Two discusses tools that researchers use to monitor the dark web; Part Three tackles the technological, ethical, and social interests at play in securing the dark web.

## Part One: Tor is Research-Driven

Tor (an acronym for 'the onion router' now stylized simply 'Tor') is an anonymity network in which a user of the Tor Browser connects to a website via three hops: a guard node, a middle relay, and an exit node. The connection is encrypted with three layers, stripping a layer at each hop towards its destination server. No single node has the full picture of the connection along the circuit: the guard knows only your IP but not where the destination is; the middle node knows the guard and the exit node; the exit node knows only the middle node and the final destination. The website knows only the exit node's IP, so it can't determine where the Tor Browser user is located. Tor Project also works on onion services, which are location-hidden services like web servers of varied content, IRC servers, and alarmingly, a growing number of command and control interfaces for botnets.

Tor is research driven[1]. Tor Project relies on academics around the world to ethically attack and defend the network. From its start as a project of the Navy, it has been accepting U.S. government funding. The Tor Project has been making an effort to

1 https://research.torproject.org/

move away from government funding, seeking diversification across many interests and through crowd-funding [1]. In fact, according to the current director of the Tor Project, Shari Steele, Tor Project is both *funded* and *attacked* by parts of the State Department. Grants and support from the National Science Foundation as well as the U.S. Naval Research Laboratory continue to fund and contribute to Tor research [1][2].

Much of the research dealing with Tor is discovering how nation states block Tor with traffic analysis and deep packet inspection. This guides more research into obfuscating the Tor protocol to resemble SSL traffic or even Skype [3]. Naturally, blocking these protocols entirely on the national level would be unpopular to justify if not technically impossible. These "pluggable transports"—as they're called—continue to be developed, discovered by censoring nations, and improved; it becomes a cat and mouse game that has continued for over a decade.

Philip Winter of the Department of Mathematics and Computer Science at Karlstad University in Sweden has done excellent work in how Tor is now seen as a censorship-resistance tool; his research involves discovering how the Great Firewall of China attempts to block Tor, and how users employ Tor bridges to evade the censors [3]. Censoring ISPs or nations are able to block access to the Tor network by blocking access to known Tor nodes. Bridges are non-publicly listed entry nodes given to users just a few at a time by simply sending an email or visiting the BridgeDB.[2]

Web monitoring and censorship in general is beyond the scope of this paper, but in one country's instance, given a unique data dump of 600 GB of Blue Coat logs in Syria, we are able to analyze Internet filtering "in the wild." One study of the logs found that blocking Tor was inconsistent but aggressive blocking could be correlated to government aggression and opposition protests (e.g. July-August 2011) [4]. It's important to note in moving forward in the discussion, that if Tor is successfully blocked,

2 https://bridges.torproject.org/

the sole method for accessing Tor's onion services is stymied.

Tor Project provides insight into conducting ethical research[3]. Given that Tor is used by many to evade censorship and/or browse the web anonymously, engaging in research that potentially deanonymizes users is unethical. The Tor Research Safety Board encourages researchers to use a small test network if possible; to attack your own traffic; only collect data that's safe to make public; the benefits should outweigh the risks. They encourage researchers to ask for assistance and Tor Project will supply in-person help. When Carnegie Melon University (CMU) was reportedly paid to deanonymize Tor users for the FBI and are yet to help Tor Project fix the vulnerability, the question of ethics comes into discussion: as Shari Steele notes, it's problematic too because CERT[4] comes out of CMU [1].

Defining the dark web is a point of inconsistency across the clearnet. The dark web is often confused with the Deep Web. The Deep Web is all non-public web material that isn't indexed by search engines [5]: this includes corporate intranet sites, paywalled material, and government and corporate databases. The dark web is included in this Deep Web but it's something different: these are web sites that can only be accessed using special software like Tor Browser or I2P. Although these websites often use popular web server software like Apache or Nginx, their IP—and hence their location—is kept hidden.

Tor Project has developed onion services (formerly "hidden services") since 2002 [6]. To visit an onion service, a user needs to be proxied through the Tor network. Web services routed through Tor belong to the Dark Web. The Tor Browser connects a user to an onion service (website) and keeps both entities from learning the IP of the other. This is done through introduction points and the negotiation of a Rendezvous Point[5]. Hidden services, as we have made clear, are location-hidden and thus very difficult to take down.

---

3 https://research.torproject.org/safetyboard.html
4 https://www.cert.org/
5 https://www.torproject.org/docs/hidden-services.html.en

Alongside its portal for Tor Research, Tor Project has a portal for Tor Metrics[6]. This portal keeps track of the number of volunteer relays, the location and number of exit nodes, bandwidth-use, and extrapolated onion service metrics like the number of .onion sites and onion service traffic. Traffic to Tor onion services is still a very small percentage of total Tor traffic; statistics released by Tor Project extrapolated data from a small set of volunteer relays to estimate between 2% - 5% of all Tor traffic involves onion services [7][8].

The anonymity network I2P (Invisible Internet Project)[7] is lesser-known than Tor but shares some of the fundamental properties like wrapped layers of encryption (sometimes called garlic routing because it bundles messages together to evade traffic analysis) as well as hiding the location of services like web servers, integrated clients for peer-to-peer networks, Internet Relay Chat, and gateways to the clearnet. I2P also relies on academic research to further its mission of providing anonymity. I2P documentation refers researchers to the ethics outlined by the Tor Project [9]. This decentralized approach to providing anonymity and censorship-resistant services will be the future of the internet; in fact, Brewster Kahle of the Internet Archive foresees a shift to a decentralized web and reminds us that the way we use the internet is not set in stone—we can invent new protocols and new ways of hosting and sharing content [10].

We will discuss content expansively in Part Three, but Onion Services using webservers with hidden locations may host a wide variety of content from illegal drug markets, gambling, child abuse image forums, contract killers, to political opposition web sites, chans (image forums), a gateway to Facebook (if it's blocked in your native country), or repositories of academic papers, like Sci-Hub. Some of these websites are only available through Tor whereas some have clearnet versions that can succumb to DNS-takedowns, TLD seizures, hosting-takedowns, or ISP/government monitoring.

---

6 https://metrics.torproject.org/
7 https://geti2p.net/en/

**Part Two: Tools to Monitor the Dark Web**

OnionScan[8] is a tool written in Go that checks for website vulnerabilities, misconfigurations, and operational security holes in onion services. The purpose of the project is to help secure services, but it's also a tool to monitor and track Dark Web sites. By making research and investigation into these sites easier and public, the project hopes to "create a powerful incentive for new anonymity technology" [11]. Indeed, OnionScan notes that government agencies scan the Dark Web but they don't release their findings [12]; OnionScan provides data and regular reports of its findings.[9]

OnionScan tests various problems with the an onion site's configurations[13]: one of the largest misconfigurations is *mod_status* if the server is running Apache. *mod_status* is useful for debugging but isn't enabled by default. If an administrator enabled it and forgot about it, anyone who queries an Apache server's *mod_status* sees myriad leaked information about the server, including real IP address, other hosted Dark Web and clearnet sites, and secret areas of the website [14].

OnionScan checks for open directories: instead of an index.html, one may be able to get an index of images in /images/ or even backup copies of the site or databases in /backups/. OnionScan checks to see whether or not the server strips EXIF data from images. EXIF data may include the make of a camera or phone and even geolocation data. This metadata can be used fingerprint a user if images properties are identical or even find the location of cocaine for sale on the Silk Road, one of the largest earlier drug markets [15].

OnionScan identifies and captures Google Analytics scripts as webmasters will haphazardly use the same unique code on several sites they administrate, including clearnet sites. Onionscan identifies cryptocurrency addresses that can correlated a shared

8 https://github.com/s-rah/onionscan
9 https://onionscan.org/

Bitcoin or Litecoin wallet on another site.

The newest version of OnionScan has a Correlation Lab[10] which links previously-scanned sites to new ones using a number of recorded properties including the aforementioned *mod_status* server identity leak, correlated clearnet versions of a site, or mistakenly-shared personally-identifying information like an email address.

The health of a project like this is crucial to maintaining continued scanning of the Dark Web and improvement of the tools. OnionScan, and projects like it, rely on contributors who can test, code, write documentation, and find bugs. OnionScan is still a new project and currently has two active developers. One good indication of a project's health is the number of issues: OnionScan currently has 68 issues, 41 that are closed. Lastly, another good indication of health from the GitHub page is that OnionScan has been forked 114 times, possibly leading to independent projects in monitoring the Dark Web. When discussing open source projects it's good to remember that projects rely on others to function: for example, the secure linux operating systems Tails and Qubes rely on the Tor protocol for access to the internet.

The Open Observatory of Network Interference (OONI)[11] is a project whose aim is to provide software tests to determine if clearnet websites, VPNs, Tor, or messaging software are being blocked by one's ISP or government[16]. It also checks to see if packets are being filtered or tampered with. Volunteers around the globe run the toolset and report the findings; the results are historic and current data on censorship at the ISP and national level. Only by monitoring the monitors do we find out what they are attempting to block or manipulate.

10 https://github.com/s-rah/onionscan/blob/master/doc/correlation-lab.md
11 https://ooni.torproject.org/

**Part Three: Monitoring the content of the Dark Web**

There are now several papers and studies that have categorized the content of the Dark Web available via Tor. One such study, and the first to systematically look at the content of onion services, found three indexes of onion service websites[12] and classified the 1171 sites by content (e.g. drugs, hacking, child abuse, pornography, etc.)[6]. An analysis of the 2165 posts is made on the material after one month of collection. The posts are categorized as subversive, ethical, or unethical. The researcher found that the vast majority of posts exhibited malign (unethical) disinhibition–the removal of self-censorship online–and that onion services are "used predominantly to evade repression from the state, but for matters that are highly unethical"[6]. The study found that very little content was political or socially subversive material. The author questions the role of onion services: "currently, the hidden services act as a protector of unethical content rather than the promoter of a censor-free place for ethical content" [6]. There are obvious shortcomings with the study: the author treats all pornography as "unethical, as it raises moral concerns" and excludes topics discussing surveillance entirely as "they belong rather to the field of politics and governance than morality" [6]. Alongside these two questionable assertions, the sample size of 2165 posts in the time frame of one month may be an interesting snapshot but may lack the content to really gauge its value. Providing context, 4chan, a clearnet image board where users may post as 'anonymous,' suffers from the same pithy statements and hate-speech alongside redeeming content–and boasts 900,000 posts each day[17]. One can clearly see the similarities of posting 'anonymous' on 4chan and actual anonymity using Tor; there are signs of both malign and benign disinhibition. The author concludes that because of the unethical content the best solution would be to stop the development of Tor onion services. This would not affect using Tor Browser to visit websites anonymously, but there would be no .onion sites.

12 The three indexes were The Hidden Wiki, Snapp BBS, and Ahmia.fi

As academics look at the conversation of academics as 'scholarly communication' we may note how Tor onion services have been a topic at every hacker convention. At the Chaos Communications Congress (31c3, 2014) Gareth Owen discussed the results of his research into content of onion services[18]. Owen and his team set up 40 high-speed relays for six months that had the HSDir flag set—after 25 hours they had a place in maintaining the Distributed Hash Table (DHT) and could thus monitor the onion service requests being made. This work is similar to Donncha O'Cearbhaill's research into mapping the DHT [19] and foundational work examining the bug that these researchers would use to discover onion service requests and collect destination URLs [20]. After collecting the requests at over 45,000/day, Owen's team scraped the website's root page for text only and categorized the content. They found that the top 40 spots in popularity belonged to botnet command and control servers[13]. Most of these requests would fail, however, because, as in the case of the Skynet botnet, the admin had been arrested and the C&C removed. This doesn't stop the zombie machines from still trying to connect, however. They found that the majority of successful connections/visits/lookups were to sites categorized as child abuse. It's unclear if these requests are by bots or humans—this troubling detail is something we cannot presently know. Tor Project suggests that protective services crawl the sites regularly in an effort to identify abused children. Nick Mathewson, co-founder of Tor Project, notes that directory requests do not give an accurate number of actual visits, and the actual traffic doesn't go through the HSDir at all but rather through rendezvous points, as mentioned above. Further, Mathewson describes how data derived from hidden service directory requests are skewed:

13 This was similar to O'Cearbhaill findings around the same time, noting that many were bitcoin-mining botnets

> "a methodology that looks primarily at hidden service directory requests will over-rate services that are frequently accessed from a Tor client that hasn't been there recently, and under-rate services that are used via tor2web, and so on. It also depends a lot on how hidden services are configured, how frequently Tor hidden service directories go up and down, and what times of day they change introduction points in comparison to what time of day their users tend to be awake. The greater the number of distinct hidden services a person visits, and the less reliable those sites are, the more hidden service directory requests they will trigger." [21]

We will continue to see sites like Sci-Hub mirror their clearnet domains with onion services. Sci-Hub is a controversial academic paper repository much discussed in the academic world. It has been called "the PirateBay for scientists." Sci-Hub calls to question whether the copyright system in the United States is really made to "promote science and the useful arts." It has surpassed over 50 million academic articles, reportedly containing a large percentage of Elsevier content and other for-profit vendors in academic publishing [22]. When the domain sci-hub.org was seized my a court injunction in New York, Sci-Hub moved to sci-hub.io, a TLD that could not easily be seized coupled with servers located in Russia, outside of U.S. courts' jurisdiction. Of course the Sci-Hub onion service server[14] could be located anywhere and it's now incredibly difficult to take down. In fact, Sci-Hub now offers a bot using Telegram messaging app that instantly provides academic articles when given the title or the DOI of the academic article [23].

Tools like OnionScan, by showing correlated sites as mirrors of the same operation, can help us to attain better data about content. If an online casino, for example, has 10 onion service sites that are found to be run by the same operator, this may lead us to see that it's not that online gambling is rampant but rather they are just mirrors. OnionScan's reports may lead us to finding out why some sites have longevity while others disappear very quickly. OnionScan will be forked into projects that help us shine a light on, and gather great amounts of data from, onion services.

The intrinsic problem with monitoring content of Tor onion services is that Tor

---

14 scihub22266oqcxt.onion

onion services and sites on I2P are purposely difficult to monitor; there is no way to get conclusive data without conflating or surmising. Roger Dingledine, a lead developer for Tor Project notes that they often think about scrapping onion services [24]. He is apprehensive about this though: he notes that they're very early in the development and there are some really compelling use cases like Facebook's onion service, or activist blogs that are DDoS-resistant. “We need to keep improving performance, consistency, and ease-of-use if we want to get beyond the very early adopters and see these use-cases take off” [24].

As more sites offer onion services for those living in blocked countries, maybe we'll see a change in the content. With the rise of ransomware command and control servers, drug markets popping up to replace ones taken down, FBI serving up malware to identify users–we can look forward to continuing the cat and mouse game we've been playing for a time now. Then again, with such a small percentage of Tor traffic going to onion services (2%-5%), perhaps the early adopters *are* the problem and we're just witnessing the pains of technology, law, and society workout their coexistence. By improving our tools to monitor the dark web, we'll have a better grounding to see in what direction we're headed.

References

[1] Farivar, C. (2016, January 10). Two months after FBI debacle, Tor Project still can't get an answer from CMU. Retrieved December 9, 2016, from http://arstechnica.com/security/2016/01/going-forward-the-tor-project-wants-to-be-less-reliant-on-us-govt-funding/

[2] Goulet, D., Johnson, A., Kadianakis, G., & Loesing, K. (2015). Hidden-service statistics reported by relays. Retrieved December 8, 2016, from https://www.nrl.navy.mil/itd/chacs/goulet-hidden-service-statistics-reported-relays

[3] Winter, P., & Lindskog, S. (2012). How China Is Blocking Tor. Retrieved December 2, 2016 from https://arxiv.org/pdf/1204.0447v1.pdf

[4] Chaabane, A., Chen, T., Cunche, M., De Cristofaro, E., Friedman, A., & Kaafar, M. A. (2014). Censorship in the Wild: Analyzing Internet Filtering in Syria. *arXiv:1402.3401 [Cs]*. Retrieved from http://arxiv.org/abs/1402.3401

[5] Greenberg, A. (2014, November 19). Hacker Lexicon: What Is the Dark Web? Retrieved December 2, 2016, from https://www.wired.com/2014/11/hacker-lexicon-whats-dark-web/

[6] Guitton, C. (2013). A review of the available content on Tor hidden services: The case against further development. *Computers in Human Behavior*, *29*(6), 2805– 2815. https://doi.org/10.1016/j.chb.2013.07.031

[7] asn. (2015, February 26). Some statistics about onions | The Tor Blog. Retrieved December 15, 2016, from https://blog.torproject.org/blog/some-statistics-about-onions

[8] Kadianakis, G., & Loesing, Karsten. (2015, January 31). Extrapolating Network Totals from Hidden-Service Statistics. Retrieved December 12, 2016, from https://research.torproject.org/techreports/extrapolating-hidserv-stats-2015-01-31.pdf

[9] Academic Research - I2P. (n.d.). Retrieved December 19, 2016, from https://geti2p.net/en/research

[10] Internet Archive. (2016). *Brewster Kahle – "Locking the Web Open – a Call for a New, Decentralized Web."* Retrieved from http://archive.org/details/decentralizedwebsummit2016-brewsterkahle

[11] Lewis, S. J. (n.d.). s-rah/onionscan. Retrieved December 9, 2016, from https://github.com/s-rah/onionscan

[12] OnionScan on Twitter: "We know governments scan darknets; they don't share the results! You can help support independent darkweb research https://t.co/qvhUmLNX6J." (November 12, 2016). Retrieved December 5, 2016, from https://twitter.com/OnionScan/status/797469011571134464

[13] Lewis, S. J. (n.d.). What is scanned for? Retrieved December 9, 2016, from https://github.com/s-rah/onionscan/blob/master/doc/what-is-scanned-for.md

[14] Lewis, S. J. (2016, July 6). Thwarting Identity Correlation Attacks. Retrieved December 9, 2016, from https://mascherari.press/thwarting-identity-correlation-attacks/

[15] How drug listings on the dark net may have revealed sellers' locations. (n.d.). Retrieved December 19, 2016, from https://www.washingtonpost.com/news/the-

switch/wp/2016/09/15/how-drug-listings-on-the-dark-net-may-have-revealed-sellers-locations/
[16] Tor Project. (n.d.). Tor at the Heart: The OONI project | The Tor Blog. Retrieved December 9, 2016, from https://blog.torproject.org/blog/tor-heart-ooni-project
[17] Advertise - 4chan. (n.d.). Retrieved December 19, 2016, from http://www.4chan.org/advertise
[18] CCCen. (2015). *Tor: Hidden Services and Deanonymisation [31c3]*. Retrieved from https://www.youtube.com/watch?v=oZdeRmlj8Gw
[19] Donncha. (2013, May 15). Trawling Tor Hidden Service – Mapping the DHT. Retrieved from https://donncha.is/2013/05/trawling-tor-hidden-services/
[20] Biryukov, A., Pustogarov, I., & Weinmann, R. (2013). Trawling for Tor Hidden Services: Detection, Measurement, Deanonymization. *Security and Privacy (SP), 2013 IEEE Symposium on,* 80-94.
[21] Mathewson, N. (2014, December 30th). Some thoughts on Hidden Services | The Tor Blog. Retrieved December 19, 2016, from https://blog.torproject.org/blog/some-thoughts-hidden-services
[22] Bohannon, J. (2016, April 28). Who's downloading pirated papers? Everyone | Science | AAAS. Retrieved August 10, 2016, from http://www.sciencemag.org/news/2016/04/whos-downloading-pirated-papers-everyone
[23] Sci-Hub, The Repository Of "Infringing" Academic Papers Now Available Via Telegram. (n.d.). Retrieved December 19, 2016, from https://www.techdirt.com/articles/20160515/01471134445/sci-hub-repository-infringing-academic-papers-now-available-via-telegram.shtml
[24] Mathewson, N. (2014, December 30). Tor: 80 percent of ??? percent of 1-2 percent abusive. | The Tor Blog. Retrieved December 5, 2016, from https://blog.torproject.org/blog/tor-80-percent-percent-1-2-percent-abusive

An accompanying presentation of 'Monitoring the Dark Web' may be found at: https://github.com/johnschriner/presentations

*A note on citations:* I've used APA citation style but also added footnotes for material I would call 'quick links'; these are optional background materials and not necessarily materials to be cited.