Measuring accessibility of popular websites while using Tor

Anant Pingle, **Stefanie Roos**
TU Delft

Abstract

Tor is an anonymity network used by a vast number of users in order to protect their privacy on the internet. It should not come as a surprise that this service is also used for abuse such as Denial of service attacks and other malicious activities because of the anonymity it provides. For protecting themselves from this abuse, websites block Tor in various ways. We investigate the extent and frequency of this kind of blocking by requesting the Alexa top 1000 websites with and without Tor with the objective of highlighting the differential treatment observed by privacy-minded users. We build upon existing studies by using diverse metrics to measure discrimination and by extending our search to three sub pages of websites for detecting sophisticated blocking. We find at least 25.8% of the Alexa top 1000 websites discriminating on the home page against Tor users as opposed to 20.03% observed in previous studies. This number rises to 31.7% after including the three sub pages. We also discover new types of blocks such as Tor users being served old or different versions of websites. We categorize the blocked websites and find that Online Shopping and Finance/ Banking categories discriminate most against Tor while Social Networking sites and Search Engines discriminate the least.

1 Introduction

Although communicating through encrypted messages enables confidentiality, it does not stop website providers or other third parties from knowing the identity of users. Tor [1] is an anonymity network that enables users to surf the web without revealing their identity. Users seek anonymity for a variety of reasons, including past experiences and life situations [2]. An example can be that anonymity enables whistleblowers to reveal information without the fear of them being in danger. Even commonly, there is a substantial number of users who wish to remain anonymous. Tor, however, has been abused for drug trafficking, copyright infringement [3] and other malicious activity such as sending spam and performing Denial of Service attacks [4]. As a result, it faces differential treatment [5]. A popular Content delivery network (CDN) provider Cloudflare published in an article "Due to the behavior of some individuals using the Tor network (spammers, distributors of malware, attackers, etc.), the IP addresses of Tor exit nodes may earn a bad reputation, elevating their Cloudflare threat score."[1]. We consider this differential treatment as blocks which range from users being asked to solve a simple puzzle (e.g. a CAPTCHA [6]) to them being denied access to the site. The extent and frequency of such blocking is still unknown. It is important that privacy-aware users know (and have the right to know) what kinds of obstructions, restrictions and tampering they face while accessing information using the anonymity network [7]. It is also essential that awareness is spread among the designers of attack detection systems, which usually associate anonymity to criminal activities.

While this blocking constitutes censorship, most studies about web censorship are focused on blocking enforced by the government, Internet service providers (ISP's) or other points of control [8; 9] and don't focus at the blocking done by a website. There have only been two studies so far that measure this type of (server-side) blocking [10; 5]. The study by Khattak *et al.*[10] being the first, it underestimates the frequency of blocking and only crawls the home page of websites as pointed out by Singh *et al.*[5]. The results from the study are also expected to have false positives owing to the use of a headless crawler which is often blocked for bot behavior. The study by Singh *et al.* builds up on the previous study by using a Selenium crawler with bot-detection avoidance strategies and comparing screenshots (as opposed to HTTP status codes which can not detect block pages that return a 200 OK status code), extending the crawl to login and search functionalities on home pages and measuring failed HTTP requests or TLS handshakes using privacy sensitive logging among other improvements. Both these studies lack an exact description of the extent of the blocks. Furthermore, the most recent research by Singh *et al.* was done in 2017, and we assume that a lot has changed concerning Tor's reputation as well as website policies in this period of time. Additionally, upon contacting one of the contributors to the Tor project, we found a similar project being carried in parallel

[1] https://support.cloudflare.com/hc/en-us/articles/203306930-Does-CloudFlare-block-Tor-

that also aims to detect discrimination against Tor users[2]. The project however, is in experimental stages and has inadequate documentation and code. This study builds on earlier work by expanding the search to three more pages in addition to the home page in order to discover more sophisticated blocks.

Aim: Through this study we aim to measure the extent to which popular websites block users accessing them using Tor. We also seek to discover the frequency of such blocking and compare the change in blocking over time. The study is performed by requesting the Alexa top 1000 websites with a residential connection in the Netherlands along with various Tor exit nodes and comparing the responses using diverse metrics.

Key findings: Through our analysis, we discovered that there has been an increase in the amount of home page blocking from 20.03% to 25.8% since the last study by Singh *et al.* This number changes to 31.7% after including 3 sub pages. We also discovered new types of blocks such as Tor users being served older versions of websites. We found a change in the categories of websites that block Tor users the most. These categories are Online Shopping and Finance/Business in which 40% of the websites were blocked for all our chosen Tor exit nodes.

Section 3 describes the method by which the research questions are answered. A detailed explanation of our experiment along with the techniques used for identifying discrimination are presented in Section 4. We then present our results in Section 5. Ethical and reproducibility aspects of the research are explained in Section 6. A comparison with prior work is done in Section 7. Finally, the conclusions and future work are presented in Section 8.

2 Tor

“The Tor network [1] is a low-latency overlay network, that anonymises transmission control protocol (TCP) streams ” [11]. The Tor project [3] explains that Tor protects users against traffic analysis wherein intermediaries can see what communication is going on over a public network by inspecting headers of internet data packets. Through this, they can infer information like source and destination of a packet which might cause distress in privacy-minded users. Tor helps mitigate this problem by routing data packets from source to destination while taking random pathways through three nodes/relays. Each relay along the path only knows from which relay a packet arrived and to which relay the packet must be sent. This is achieved using encryption. The entry relay is called the *guard*, which accepts connections from users. The relay that interacts with the destination on behalf of the user is known as the *exit* relay and the one in between the guard and the exit is called the *middle* relay. Figure 1 describes an interaction between a source and destination while using Tor.

3 Methodology

Our measurements rely on requesting web pages with a control connection and various Tor exit nodes in diverse phases while combining techniques from the work of Khattak *et al.* and Singh *et al.* We extended these studies by expanding the search to three sub pages as blocks can also appear on sub pages [10]. The number three was chosen arbitrarily. We use a diverse set of metrics in order to classify a website as discriminating. We consider a website blocked for Tor when it is discriminating on the home page or any sub page for all Tor exit nodes in our experiment. It should be noted that for this study, we only consider discrimination by websites on the Application layer of the OSI model[4]. This is by choice and also because of additional factors mentioned in section 4 that we believe might add false positives to our results. We conducted our measurements at roughly the same time each day, usually starting at 10 PM CEST. We understand from the study of Khattak *et al.*, that the time of the day can add bias to our results as certain servers operate only at particular times of the day or particular days of the week and we investigate this claim with a subset of websites and highlight the dissimilarities. We now discuss some practical decisions taken for the experiment. The experiment is described in detail in section 4.

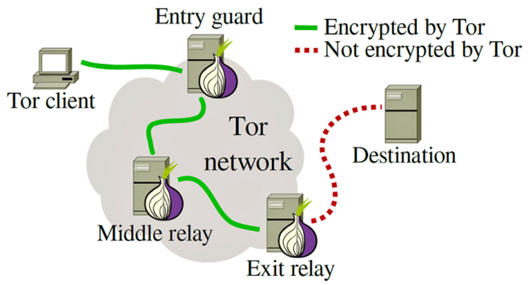


Figure 1: A typical three hop Tor circuit representing the communication between the Tor client and the destination.
[11, Figure. 1.1]

3.1 List of popular websites

We chose the Amazon Alexa list of websites as it is the most commonly used website list in security research[12]. It was also used in prior work related to this study[10; 5]. We only chose the top 1000 websites because of practical restrictions such as time and computing resources.

3.2 Control connection

The control connection acts as the baseline to compare the results of Tor nodes against. Western Europe is classified as an uncensored region[13]. The Netherlands being a country in this region and for reasons such as availability, a Dutch control connection with a residential connection provided by Ziggo[5] was chosen. The residential connection helps in simulating a regular user. Using an IP address of a data center or a cloud provider might have added false positives to our results.

[2] https://gitlab.torproject.org/woswos/CAPTCHA-Monitor/-/wikis/GSoC-2021

[3] https://www.torproject.org/

[4] https://www.forcepoint.com/cyber-edu/osi-model

[5] https://www.ziggo.nl/

3.3 Exit relays

We chose a set of four exit relays at random in the Netherlands whose exit policies allow outgoing connections on ports 80 (HTTP) and 443 (HTTPS). We only chose relays with Exit, Running and Valid flags according to Tor project [6] as we think they best suited our needs. Only Dutch exit nodes were chosen in order to eliminate false positives such as encountering blocks because of government censorship enforced in the country where the exit node is hosted. Examples of such countries are China, Iran, Russia which are classified as one of the worst internet censors[14]. Four exit nodes are chosen as per prior work [10] and also because of the limited time and computing resources available.

3.4 Libraries

We decided upon a combination of libraries used in previous work [10; 5] along with some adjustments to collect precise measurements for the experiment.

Requests

We chose the *Requests*[7] library for Python for requesting the home pages of websites. *Requests* allows users to send HTTP/1.1 requests, access responses, add custom headers, and connect to proxies (required to connect to the Tor proxy) among other functionalities in an easy manner. Other similar options at hand were for example, Scrapy[8] and BeautifulSoup[9] but they offered a lot more functionality than what was required for the corresponding phases of the experiment. Our only necessity for this library was to obtain the status codes of the responses received from the home page requests and we think that this Python library worked well.

Selenium

We also needed a library in our experiment which launched a web browser and requested the web pages for us (non-headless crawler) as it helped in simulating a real user. Libraries such as *Requests* can't be used as they don't launch a browser and neither do they load JavaScript. Differences in the JavaScript loaded by web pages can therefore not be compared using these libraries. We chose Selenium as it is considered one of the best browser automation tools for web scraping [10]. It has highly comprehensible documentation and has been used in several studies including the one from Singh *et al.*

4 Experiment design

We performed the experiment in phases with the initial phases being relatively quick compared to the later ones. For each phase, we performed two runs of measurements. We do this to verify if the first measurement was inconsistent. Inconsistencies can arise for example, due to problems in the network or problems with CDN's on which websites are hosted. Statistically significant differences in blocking between the two runs can reveal such inconsistencies. As the primary question of the research is to measure the extent and frequency of blocks, the rationale behind choosing the phase wise approach was to eliminate websites enforcing obvious blocks in the primary (fast) phases and having a small number of inputs for the later (slow) phases to check for advanced blocks. As an example, for the cases where a domain enforces a block on the home page in both runs of the experiment phase, it is ineffectual to check the sub pages because our definition of blocking requires only a single page to be blocked. Such cases are eliminated in the initial phases by our experiment. The flow diagram depicting the phases is seen in Figure 2.

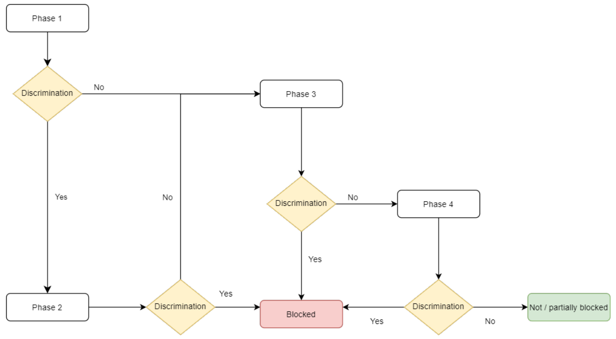


Figure 2: Phases of experiment

We used criteria used by prior studies by Khattak *et al.* and Singh *et al.* In particular, we used HTTP status codes to check for discrimination and used perceptual hashing to compare similarities between screenshots in an automated way. In addition, we used a new metric of comparing the structural similarity of two HTMLs. This method uses sequence comparison of the HTML tags to compute the similarity[11]. We now describe the phases of our experiment, as well as the metrics utilized in each phase, and explain why we chose them. We also explain for each phase, how we shortlisted websites for subsequent phases.

4.1 Phase 1

In this fast phase, for each website in the Alexa top 1000 list, we first requested the home page with our control connection and then with one Tor exit node using the *Requests* library while recording the HTTP status code and the response content. We configured Tor as a SOCKS proxy and updated the Tor configuration (torrc) file with the IP address of our chosen exit node. We modified the *User-Agent* header to be that used by Chrome[12] (a popular browser in 2021[13]) in order to simulate a modern web browser. We did this modification because we found a few websites that refused access without these headers but allowed access with them. Singh *et al.* critique that doing such a modification can still lead to websites

[6] https://metrics.torproject.org/rs.html#search

[7] https://docs.python-requests.org/en/master/

[8] https://scrapy.org/

[9] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[10] https://limeproxies.netlify.app/blog/differences-between-browser-automation-tools/

[11] https://pypi.org/project/html-similarity/

[12] https://www.google.com/chrome/

[13] https://www.techadvisor.com/test-centre/software/best-web-browsers-3635255/

blocking the requests[5], however, we do not eliminate any website in this phase and neither do we use it as the sole measurement technique. This phase took around 30 minutes after a few optimizations.

Discrimination criteria:
We consider a website to be non discriminating if the control connection and Tor node both return a HTTP 2xx status code or if both return a non 2xx status code. The latter is considered non-discriminating as we believe it to have occurred due to the usage of a headless crawler which does not load JavaScript [15].

Selecting websites for Phase 1 and 2:
We performed two runs of the experiment on consecutive days due to the reasons mentioned before. Websites that were found to be discriminating in both runs were handled in Phase 2 and the remaining websites were promoted to Phase 3.

4.2 Phase 2

This quick elimination phase also involved requesting the home pages of the discriminating websites from Phase 1 using *Requests* while recording similar data. However, this time we used the three remaining exit nodes. We first probed the website using the control connection and then with the three exit nodes one after the other. We initially planned on using ExitMap[14] which was used in the study by Khattak *et al.* ExitMap enables users to run modules on a subset of all Tor exit nodes where modules can be tasks like fetching a web-page or uploading a file[14]. However, we were not able to configure it correctly due to unexpected errors. As a consequence, we ran three instances of Tor and configured them as SOCKS proxies on different ports. The *User-Agent* field was modified in a similar way as before. After optimizations, this phase ran for approximately 30 minutes on average.

Discrimination criteria:
We report discrimination if the control connection receives a HTTP 2xx status code and all three Tor nodes receive a non 2xx status code. The cases in which all three Tor exit nodes time-out are excluded as we suspect them occurring due to blocks enforced by governments or other points of control rather than by the websites themselves. For example, it is known that China blocks requests from Tor[8] and hence considering the website as discriminating might result in false positives. Such timeouts might also indicate discrimination at layers 3 and 4 of the OSI model as pointed by Khattak *et al.*, however, most websites we saw in this category were from China. We believe that government censorship is to be blamed here. Cases of partial time-outs and all other cases are considered in Phase 3.

Selecting websites for Phase 3:
As before, after two runs of the experiment, we first classified websites which were discriminating in both run 1 and run 2 as websites blocked for Tor. For the websites that were excluded from both the runs (because of time-outs), we took the common ones and requested the websites three times with time-outs of 300 seconds with the same three Tor exit nodes. All cases where the request timed out for all the three exit nodes in all three runs were excluded from further analysis in this study and all remaining websites were promoted to Phase 3.

4.3 Phase 3

Phase 3 involved crawling the home pages of the websites obtained from the previous phases using our control connection and all four chosen Tor exit nodes. This time, using Selenium in non-headless mode as opposed to *requests*. Here, we launched a full Chrome web browser and requested the sites. Analogous to Phase 2, we ran 4 instances of Tor on different ports as SOCKS proxies and requested the home pages on each of these nodes one after the other. We saved a screenshot and the page source after a 3 seconds delay in order to permit the JavaScript to load entirely. We could not find optimizations like concurrent requests in this phase and hence we do each request one node after the other on our 4 exit nodes, leaving at least 12 seconds between subsequent requests from the same node (at least 3 seconds per node). We consider this an advantage as websites would not be suspicious of bot activity from our nodes. We used uBlock Origin [15], an adblocker for Chrome as it reduced the page load times on websites. We chose it because of prior experience with it and based on reviews on the Chrome store. We also used the "I don't care about cookies" [16] extension for Chrome in order to automatically accept Cookie warnings during our crawls. A crawl of around 800 websites took approximately 13 hours to complete.

Discrimination criteria:
We use perceptual hashing (p-Hashing) as used by Singh *et al.* and HTML structural similarity in order to automatically classify websites as discriminating.
p-Hashing: p-Hashing[16] can be used to compare the similarities between images. p-Hashing produces similar hashes for similar images unlike cryptographic hashes which produce drastically different hashes for small differences in the input.[17]. We used the Imagehash[18] library for python which provides an implementation of the technique. The library can be used to generate a 64-bit p-Hash of an image and two images can be compared using the hamming distance between them. In order to decide on thresholds for classifying images as blocked or unblocked, we manually inspected 500 pairs of control and Tor screenshots and calculated their hamming distances. We then split the pairs in two groups of 300 and 200, analogous to training and test sets in Machine learning[17]. We agreed on a maximum manual classification rate of 10%. We then classified the images in the training group on a variety of values as upper and lower bounds. For each pair of upper and lower bound, we calculate the accuracy of the correctly classified samples and also the fraction of images that are to be classified manually. The plot of the manually classified fraction and the accuracy can be seen in Appendix A.

[14] https://github.com/NullHypothesis/exitmap

[15] https://ublockorigin.com/

[16] https://www.i-dont-care-about-cookies.eu/

[17] https://www.phash.org/

[18] https://pypi.org/project/ImageHash/

We then sorted the results based on the difference of accuracy and the fraction of manual classification and picked the top 10 values with manual classification fraction $< 10\%$. The results can be found in Appendix A. These values were then tried on the test group. We chose an upper bound of 21 and a lower bound of 17 based on the results. We did not choose the bounds with the largest difference on the test group as they might be biased towards our training group. We chose more flexible values. This resulted in a higher fraction of samples that we had to classify manually than planned, however, we think that it helped to mitigate the problem of the thresholds being biased and also helped reduce the false positives. Before classifying a website as discriminating, we check the structural similarity.

Structural similarity: This metric compares HTML tags in order to compute the similarity. We use html-similarity[19], a python library that provides implementation for the metric. It returns a value between 0 and 1 with 1 indicating 100% similarity. Through this technique websites with large differences in content can automatically be considered non-discriminating as the structure of their HTMLs would be almost perfectly identical. An example can be seen in Figure 3 where the p-hash distance is fairly high but the HTML structural similarity is almost 1.0. We chose the thresholds analogous to the ones for p-Hashing. Appendix A shows the plot of the manually classified fraction and accuracy. The Table shows the difference of accuracy and fraction of manual classification of the top 10 threshold pairs. We found an upper bound of 0.5 and a lower bound of 0.3 with which we were able to label majority of the samples in an automated fashion.

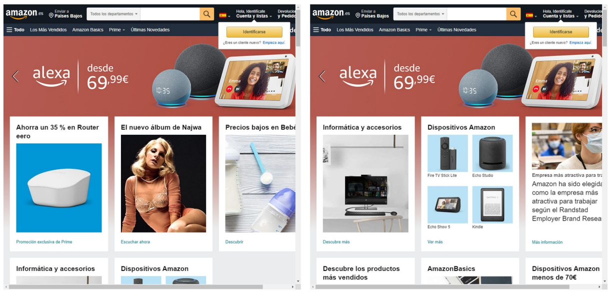


Figure 3: Two images differing largely in p-hash distances but having high structural similarity.

Cases of time-outs: As in the previous phase, the requests for which all Tor exit nodes time-out are excluded. Furthermore, cases where the control connection times out are also excluded as they act as our baseline and without them we can not check for discrimination. For cases where there are time-outs on a subset of the exit nodes, we consider them non discriminating as we suspect them to have occurred due to an exit node that had been recently used for abuse.

Selecting websites for Phase 4:

We again performed two runs of the experiment in this phase in order to account for inconsistencies mentioned previously in this section. We excluded all websites from further analysis which were excluded in both runs (intersection). All websites considered discriminating in both the runs were considered blocked for Tor. All remaining websites were handled in Phase 4.

4.4 Phase 4

In this final phase, we check the websites passed on by Phase 3 for non-obvious blocking. We do this by requesting the home pages as well as 3 other pages within the same domain using our control connection and all four Tor exit nodes. The configuration of Tor is done analogous to Phase 3. We once again stored the screenshots and page sources. In order to choose the three pages, we first located all “a” tags with the “href” attribute, then we filtered the ones within the same domain and finally chose three links at random. The crawl for this phase took over 30 hours to execute on our computing resources. Due to the limited time at hand and because of the assumption that most blocks were detected in the prior phases, we only performed one run of this phase.

Discrimination criteria:

We also use the same metrics as Phase 3 in order to compare the samples. We consider a website blocked for Tor if there is discrimination between the control connection and all the Tor exit nodes on at least one of the four pages we crawl. All other cases are considered non discriminating.

5 Results

We now present the results we obtained from each phase of the experiments. We also present the effect of the time of the day on the blocking rate, as well as perform a categorization of blocked websites.

5.1 Phase 1

We performed two runs on consecutive days on the Alexa top 1000 URLs in this phase and recorded our results. We classify the websites as discriminating or non discriminating based on the criteria mentioned in Section 4. The results from the two runs were not perfectly identical. We saw 779 and 785 non-discriminating websites in the first and second runs respectively while on the other hand 221 and and 215 websites discriminating in the two runs. To check if this inconsistency was statistically significant, we performed a Chi-squared test [18] under the null hypothesis of there being no relationship between the two runs performed on consecutive days and the blocking observed (independence). We chose this non-parametric test as it is ideal for comparing nominal data and also because our data satisfied all the requirements of the test.[20]. We could not use parametric tests like the student t-test because our data did not meet the condition of normality [21]. We calculated $X^2(1, N = 2000) = 0.1055, p = 0.74$ which indicates no relationship among the two runs on consecutive days and the blocking which in turn shows that the inconsistencies are not statistically significant. Through this

[19] https://pypi.org/project/html-similarity/

[20] https://libguides.library.kent.edu/spss/chisquare

[21] https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/paired-sample-t-test/

Table 1: HTTP responses received by the Control (C) and Tor (T) nodes in two runs of Phase 1.

Run	2xx	3xx	4xx	5xx	Miscellaneous
Run-1 C	855	0	28	9	108
Run-1 T	640	1	138	13	208
Run-2 C	851	0	29	8	112
Run-2 T	642	1	138	13	206

we can learn that network conditions have not hindered our results in any way. We can also learn that day of crawling is independent of the blocking rate.

Through this phase, we found 207 websites discriminating which we dealt with in Phase 2. The remaining websites were processed in Phase 3.

The responses received by the Control connection and the Tor node in the two runs can be seen in Table 1. We performed a Chi-squared test to compare the results between the control and the Tor node for both the runs. The $X^2(4, N = 2000) = 130.377, p < 0.00001$ for both run 1 and run 2 clearly show discrimination against Tor users.

5.2 Phase 2

For this phase, we also performed two runs of the experiment with the 207 websites filtered from Phase 1 using our control connection and the remaining 3 Tor exit nodes. We exclude all websites that timed out on all 3 exit nodes in either of the runs. We found the blocking rates in both runs to be exactly the same. We then included all the timed out websites, calculated for each website the fraction of Tor nodes being discriminated against and compared the time-outs for both runs. This process is mentioned in detail while presenting the results of Phase 3. We saw a difference of 1 request in both runs. As before, the Chi-squared test $X^2(4, N = 414) = 0.0274, p = 0.999$ confirmed that the difference was not statistically significant. The contingency table can be seen in Table 2.

Table 2: Contingency table representing the runs of Phase 2 and the fraction of blocked exit nodes.

Run	0	0.33	0.67	1	Excluded
Run-1	26	3	5	114	59
Run-2	27	3	5	114	58
Total	53	6	10	228	117

At the end of this phase, we marked 101 websites to be blocked for Tor as they returned non-2xx status codes on all Tor exit nodes, whereas a 2xx status code for the control connection. We excluded 58 websites from the study as they timed out 3 times on all exit nodes even after setting an increased time-out limit of 300 seconds.

5.3 Phase 3

Following the pattern of the previous phases, we performed two runs of the experiments in this phase on the 841 websites promoted from Phase 1 and 2 with our control connection and 4 Tor exit nodes. We first classified as excluded from both runs the requests in which either the control connection or all four Tor exit nodes timed out for reasons mentioned in Section 4. For each website, we first count the number of Tor nodes for which the request was successful (i.e. no time-outs or other miscellaneous errors) and then compute the fraction of discriminating Tor exit nodes. For example, if for a website, 2 nodes face no discrimination, 1 node faces discrimination and 1 node times out, we record the result as 1/3. In order to compare the pairwise results of the two runs, we only consider the requests where both runs have a result (i.e. no exclusion). We compared the means of the two runs to see if there was a statistically significant difference in the blocking rates. We used the paired t-test to compare the means as it is a strong test fit for the purpose and we met all the conditions for it. The distribution of the differences between the two runs was symmetrical but not perfectly normal, however, due to our large sample size and because of the central limit theorem [19], we were able to assume normality. The p-value of 0.942 indicates that the difference in blocking rate between the two runs was statistically insignificant. The number of excluded samples also differed in the two runs and the paired t-test could not account for these because we removed them to make our sample size equal. So as before, we chose the Chi-squared test for independence with which we could compare the categorical counts between the two runs to check for significant differences. We used the same data as the paired t-test but now also included the timed out samples. The $X^2(5, N = 1682) = 0.5720, p = 0.989$ clearly indicates no statistically significant difference which shows that the inconsistencies between the consecutive runs and the blocking categories are statistically insignificant. The contingency table for the phase can be seen in Table 3.

Table 3: Contingency table representing the runs of Phase 3 and the fraction of blocked exit nodes

Run	0	0.25	0.5	0.75	1	Excluded
Run-1	538	33	27	18	171	54
Run-2	538	29	28	16	178	52
Total	1076	62	55	34	349	106

As a result of this phase of the research, we found 157 websites blocked for all 4 exit nodes in both runs. We marked these websites blocked for Tor. Combined with the blocked websites of Phase 2, we see 258 confirmed front page blocks in total. We expect there to be at most 15 false positives because of reasons mentioned in Section 8.1. There were 44 websites that were in the list of excluded websites in both the runs and hence were excluded from the study. This, in combination with the excluded websites from Phase 2, results in 102 websites being excluded from this study. As discussed earlier, while this might be due to blocking on the Network layer, we believe that other factors also play a role and hence excluding them is the safest option to avoid false results.

5.4 Phase 4

Due to the limited time and computing resources we had at hand, we performed only one run of this phase which was aimed at detecting blocks on sub pages of the domains from the 640 websites from Phase 3. We use the same metrics of perceptual hashing and structural similarity to classify a page as discriminating. As before, we only consider websites where the control connection does not time out. We find at least 86 pages across 59 unique domains discriminating against all of our chosen exit nodes out of the 2313 links we crawled for this phase. This makes up a total of 317 front page + 3 sub pages blocks after considering results from prior phases.

5.5 Effect of time of the day on blocking

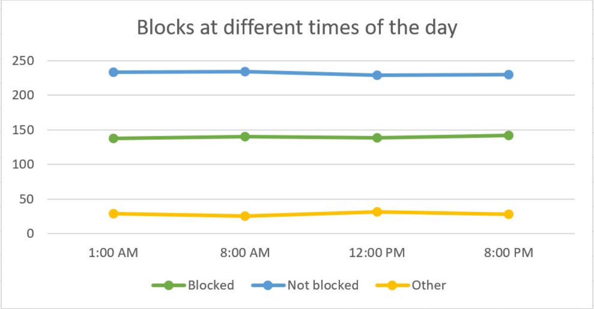


Figure 4: Blocking observed at different times of the day.

We mostly performed our crawls at night. However, they ran for more than 10 hours for Phase 3 and 4. We performed an experiment in order to investigate if the time at which we took the measurements affected the results in any way. We randomly chose 400 URL's from our list of Alexa top 1000 websites and crawled the front pages while recording the screenshots and page sources using our control connection and the Tor exit node used in Phase 1. The crawls were conducted at 1AM, 8AM, 12PM and 8PM CEST and lasted around 2 hours each. We used discrimination criteria identical to the ones used in Phase 3. We classified the URLs in one of 3 categories. Blocked, Not blocked and Other. The "Other" category comprises of the results in which either the control connection or the Tor node either time-out or present other non-application layer related errors. We performed a Chi-squared test [18] under the null hypothesis that there is no relationship between the time of the day and blocking imposed by websites. We found $X^2(6, N = 1600) = 0.7882, p = 0.99238$ which indicates no relationship between the time of the day and blocking. Our results for the experiment can be seen in Figure 4.

5.6 Extent of blocking

We saw most blocks similar to the ones discovered in prior work [5; 10; 20]. We however also saw a new type of blocking not mentioned before. We now present the most common types of blocks we saw.

- Blocking of the entire domain: These are the websites that deny access to Tor users completely. These can be spotted as block pages usually with the phrase "Access Denied".
- CAPTCHA's or 2FA: The websites in this categories present a puzzle to the users to solve, most commonly a CAPTCHA[6].
- Blocking specific functionalities: The websites in this category blocked users from accessing certain functionalities. For example, Google presented CAPTCHA's to Tor users only while searching. T-mobile [22] did not allow Tor users to log in by presenting an error page.
- Serving old or different versions of websites: We found two cases where we noticed that Tor users were being served older or different versions of websites as compared to regular users. We found these during manual verification in Phase 3. We believe there are a lot more of such cases, however, we can not report the exact number as the majority of the discrimination comparison was done in an automated fashion.

5.7 Categorization of blocked websites

Among the websites we found blocked on the front page, we used the McAfee URL categorization service [23] in order to categorize the websites. We used this as it was also used in previous studies [5] and would help us in comparing our results. We consider only the categories in which we have at least 20 websites. We find that the Online shopping and Finance/Banking categories are the most discriminating with around 40% of the websites in these categories being blocked. Social Networking and Search Engines are least discriminating with around 9% and 3.6% of the websites being blocked in these categories. The results can be seen in Figure 5.

6 Research Integrity

We carried out our study whilst respecting research integrity. We now highlight the ethical implications and reproducibility aspects of the research.

6.1 Ethical implications

Most ethical concerns arise from our use of automated software for accessing websites.

Bandwidth usage

We performed our crawls from a residential IP address shared with the other people. We took consent from all people sharing the IP address and ensured to run the crawls at night. As for the Tor exit nodes in question, the advertised bandwidth of the relays we used ranged from 15-31 MiB/s. We believe that requesting at most four web pages one after the other in intervals of at least 12 seconds should not consume a substantial portion of the bandwidth. Furthermore, because Tor is a free and open source software, there were no Terms of Service agreement disallowing our activities.

[22] account.t-mobile.com

[23] https://www.trustedsource.org/

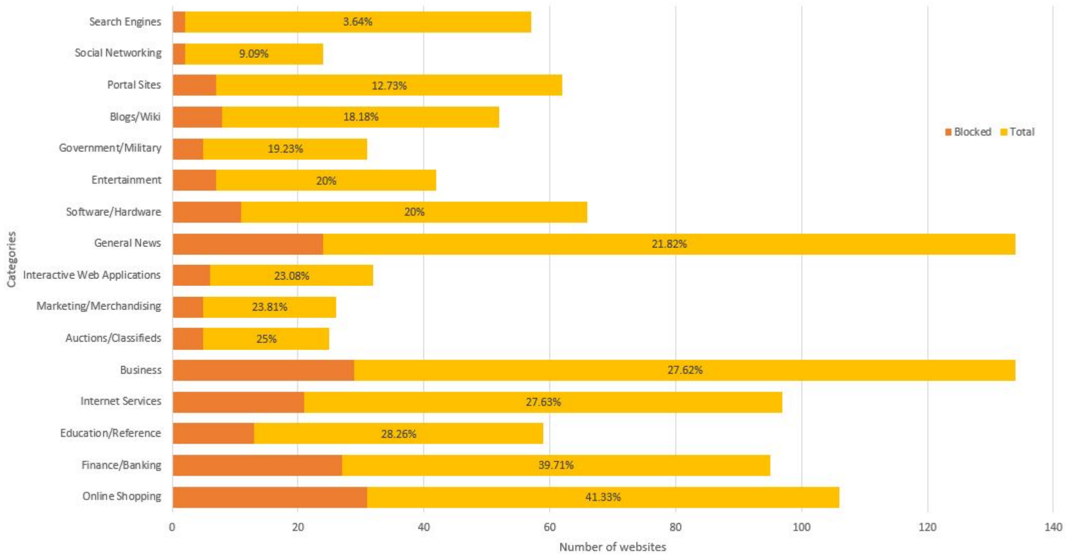


Figure 5: Categories with at least 20 websites from the Alexa top 1000 in them and their discrimination percentage.

Robots.txt

According to the documentation of Google, the robots.txt file tells search engines the URL's that a crawler can access on the website[24]. This is done in order avoid overloading websites with requests. The robots.txt or Robot Exclusion Protocol [21] however, is not an official Internet standard[25] yet. We did not consider this file because we believe that visiting four pages of websites cannot create substantial load on websites in any way, and also since the REP is mostly applicable to major search engines.

6.2 Reproducibility

In order to facilitate reproducibility, we have made all our code available on the TU Delft EWI Gitlab repository[26]. The procedure for performing our experiments is explained README.md file. In addition, we have also included the screenshots, HTML files and other data that we collected in the same repository. It should be emphasized, however, that the findings may change depending on when the experiment is repeated. This might occur as a result of network problems or a change in the blocking enforced on Tor.

7 Discussion

Comparing our results to the last study done by Singh *et al.* in 2017, we see an increase in the home page blocks from 20.03% to 25.8%. Furthermore, we see new blocking methods such as Tor users being served older versions of websites and them being denied access to certain functionalities from our results. As for the blocking rate per category, according to the McAfee URL categorization service, we see a drop in the blocking rate in the social networking category from 30% to 9% and the News category from 53% to 22%. We see that blocking is most common in Online Shopping and Banking/Finance websites with 40% of the websites blocking all our chosen Tor relays. Following these are Education, Internet services and business websites which all share the same rate of blocking.

Singh *et al.* performed their experiment using more exit nodes compared to us. This might have lead to some differences in the results. In addition, their study considered only the Alexa top 500 websites while ours considered the top 1000. Despite these external variables that might explain the difference in blocking, we believe that there has been a degradation of the reputation of Tor since 2017. We assume the high blocking rates in Online shopping and Banking/Finance categories are due to the fear of Denial of Service or BOTNET attacks [22]. These websites are often used by regular users and we believe that denying access through Tor is rather unfair. Furthermore, Singh et al. mention, studies show that Tor users are as likely to make purchases from revenue generating websites as non-Tor users.[5].

8 Conclusions and Future Work

Through this study, we aimed to measure the extent and frequency to which popular websites blocked Tor users and compare the changes over time. From our results we see that a significant amount of websites block Tor users in different ways. The extent of this blocking we observed ranges from users being asked to solve CAPTCHA's, users being served old or different versions of websites, users being blocked access to certain functionalities all the way to them being denied access to the entire domain. As for the frequency of blocking, we see a blocking rate of around 25.8% among the Alexa top 1000 websites home pages. This, when combined with 3 sub pages, shows a rate of 31.7%. This rate is 5.77% higher com-

[24] https://developers.google.com/search/docs/advanced/robots/intro

[25] https://developers.google.com/search/blog/2019/07/rep-id

[26] https://gitlab.ewi.tudelft.nl/cse3000/2020-2021/rp-group-22/rp-group-22-apingle

pared to prior studies and indicates that the reputation of Tor has not improved but rather deteriorated in the last four years. We also observe changes in the blocking rates among various categories of the McAfee URL categorization service, finding that Online Shopping and Finance/Banking mistreat Tor users the most. We would like to again draw the attention of website operators, designers of attack detection systems and other parties to carefully develop ways to combat abuse while keeping privacy-aware individuals in mind.

8.1 Limitations

We now highlight some limitations observed during the study that could not be accounted for. Despite these, it is still clear that Tor users are being blocked significantly in a number of ways. These limitations can be taken into consideration while further building up on this study.

1. Time-out pages on Tor nodes in Phases 3 and 4: We detected time-outs on Selenium when the webdriver raised an exception. While this always worked with the control connection, Tor nodes sometimes rendered time-out pages instead of raising exceptions, causing us to flag the website for discrimination against Tor. We however think that this number is at most 15 for the home page blocks as these were the cases of partial (not all exit nodes) time-outs in Phases 1 and 2. Cases of complete time-outs were already excluded in Phases 1 and 2. This leads us to believe that the blocking rate might be around 24% which still shows an increase in blocking since 2017.
2. Number of Tor exit nodes used: While our definition of blocking required a page to be discriminating for all our chosen Tor exit nodes, there are about 1200 exit nodes at the time of writing this paper and we chose only four of them.
3. Number of sub pages crawled: We only checked three sub pages for this study but this can easily be extended to detect uncaught blocks.
4. Websites checked: We use as popular websites, the Alexa top 1000 list. In the future, this number can be increased to check more websites. Along with this, other lists of popular websites, for example Moz[27] can also be used.

9 Acknowledgements

We thank Francine Biazin do Nascimento, Ioana Paula Iacoban, Jurgen Mulder and Willemijn Tutuarima who carried out similar research with other anonymity networks for sharing their resources and ideas throughout the course of the study.

References

[1] R. Dingledine, N. Mathewson, and P. Syverson, "Tor: The second-generation onion router," tech. rep., Naval Research Lab Washington DC, 2004.

[2] R. Kang, S. Brown, and S. Kiesler, "Why do people seek anonymity on the internet? informing policy and design," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2657–2666, 2013.

[3] T. Minárik and A.-M. Osula, "Tor does not stink: Use and abuse of the tor anonymity network from the perspective of law," *Computer Law & Security Review*, vol. 32, no. 1, pp. 111–127, 2016.

[4] Z. Ling, J. Luo, K. Wu, W. Yu, and X. Fu, "Torward: Discovery of malicious traffic over tor," in *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, pp. 1402–1410, IEEE, 2014.

[5] R. Singh, R. Nithyanand, S. Afroz, P. Pearce, M. C. Tschantz, P. Gill, and V. Paxson, "Characterizing the nature and dynamics of tor exit blocking," in *26th {USENIX} Security Symposium ({USENIX} Security 17)*, pp. 325–341, 2017.

[6] L. Von Ahn, M. Blum, N. J. Hopper, and J. Langford, "Captcha: Using hard ai problems for security," in *International conference on the theory and applications of cryptographic techniques*, pp. 294–311, Springer, 2003.

[7] S. Burnett and N. Feamster, "Making sense of internet censorship: a new frontier for internet measurement," 2013.

[8] P. Winter and S. Lindskog, "How china is blocking tor," *arXiv preprint arXiv:1204.0447*, 2012.

[9] S. Burnett and N. Feamster, "Encore: Lightweight measurement of web censorship with cross-origin requests," in *Proceedings of the 2015 ACM conference on special interest group on data communication*, pp. 653–667, 2015.

[10] S. Khattak, D. Fifield, S. Afroz, M. Javed, S. Sundaresan, V. Paxson, S. Murdoch, and D. Mccoy, "Do you see what i see? differential treatment of anonymous users," 02 2016.

[11] P. Winter, *Measuring and circumventing Internet censorship*. PhD thesis, Karlstads universitet, 2014.

[12] J. Mulder, "Measuring the blocking of an.on users by popular websites through web scraping." Unpublished manuscript, 2021.

[13] B. Warf, "Geographies of global internet censorship," *GeoJournal*, vol. 76, no. 1, pp. 1–23, 2011.

[14] C. Bitso, I. Fourie, and T. J. Bothma, "Trends in transition from classical censorship to internet censorship: selected country overviews," *Innovation: journal of appropriate librarianship and information work in Southern Africa*, vol. 2013, no. 46, pp. 166–191, 2013.

[15] D. Doran and S. Gokhale, "Web robot detection techniques: Overview and limitations," *Data Mining and Knowledge Discovery*, vol. 22, pp. 183–210, 06 2011.

[16] X.-m. Niu and Y.-h. Jiao, "An overview of perceptual hashing," *Acta Electronica Sinica*, vol. 36, no. 7, pp. 1405–1411, 2008.

[27] https://moz.com/top500

[17] T. M. Mitchell, “Does machine learning really work?,” *AI magazine*, vol. 18, no. 3, pp. 11–11, 1997.

[18] M. L. McHugh, “The chi-square test of independence,” *Biochemia medica*, vol. 23, no. 2, pp. 143–149, 2013.

[19] B. P. Roe, *The Central Limit Theorem*, pp. 107–118. New York, NY: Springer New York, 2001.

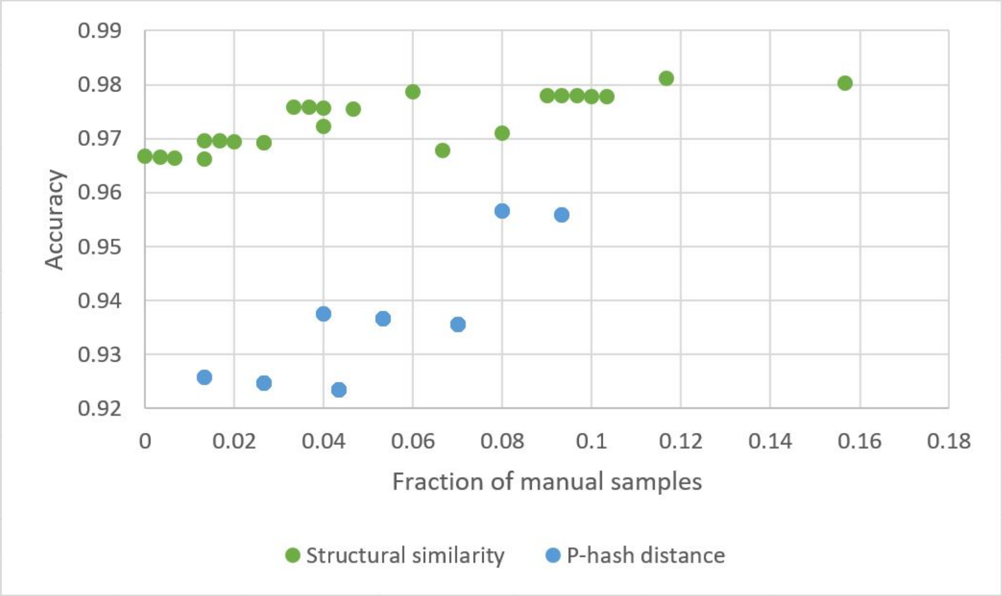
[20] C. Tran, K. Champion, A. Forte, B. M. Hill, and R. Greenstadt, “Tor users contributing to wikipedia: Just like everybody else,” *arXiv preprint arXiv:1904.04324*, 2019.

[21] P. Jha, S. Goyal, T. Kumari, and N. Gupta, “Robots exclusion protocol,” *Internation journal of emerging science and engineering*, vol. 2, no. 5, 2014.

[22] Y. D. Mane and U. P. Khot, “Botnet detection in low latency anonymous communication network: a branch of knowledge,” in *2018 International Conference on Smart City and Emerging Technology (ICSCET)*, pp. 1–6, IEEE, 2018.

A Thresholds for p-Hashing and structural similarity

Fraction of manual samples vs Accuracy of p-hash distance and structural similarity thresholds on the training group.



Top 8 upper and lower bounds for p-Hash distance and structural similarity by difference of accuracy and fraction of manual samples (Diff(A, F)) on training group.

	Upper bound	**Lower bound**	**Fraction manual**	**Accuracy**	**Diff(A, F)**
p-Hash distance	20	19	0.0133	0.9256	0.9123
	21	19	0.0133	0.9256	0.9123
	20	17	0.0266	0.924	0.8979
	20	18	0.0266	0.924	0.8979
	21	17	0.0266	0.924	0.8979
	21	18	0.0266	0.924	0.8979
	22	19	0.04	0.937	0.8975
	23	19	0.04	0.937	0.8975
	22	17	0.05	00.936	0.883
	22	18	0.05	0.936	0.883
Structural similarity	0.4	0.4	0.0	0.967	0.967
	0.5	0.4	0.003	0.966	0.963
	0.6	0.4	0.006	0.966	0.960
	0.4	0.3	0.013	0.969	0.956
	0.7	0.4	0.013	0.966	0.953
	0.5	0.3	0.0166	0.969	0.953
	0.6	0.3	0.02	0.969	0.949
	0.4	0.2	0.033	0.975	0.942
	0.7	0.3	0.026	0.969	0.942
	0.8	0.4	0.026	0.969	0.942