Measuring the Accessibility of Popular Websites While Using Mullvad VPN

Francine Biazin do Nascimento , Stefanie Roos
TU Delft

Abstract

There are many valid reasons for someone to choose to stay anonymous online, not least of which is the fact that online privacy is a human right. However, discrimination against users of anonymity networks from web-servers and content distribution networks on the grounds of defense against malicious users often means that genuine users are faced with excessive challenge-response tests and differentiated content, or even blocked altogether. This study has investigated the extent to which users of Mullvad VPN are blocked when accessing popular websites and it has also explored the nature of these blocks. No statistically significant difference was found when requesting only home pages from 3,000 domains, but this changed when classifying 1,000 domains and considering content beyond the home page. This impact on the user's experience is also reflected on the categories of website that most engage in blocking, with some essential services such as health and government presenting high blocking ratios. Given that more discerning ways of preventing access from malicious users without affecting genuine ones exist, this generalised blocking of Mullvad VPN users is unjustifiable.

1 Introduction

"Big Brother is Watching You."
— George Orwell, *1984*.

The right to privacy in the digital age is a topic that has been at the forefront of international debates on human rights, the surveillance state, and a myriad of related issues, particularly in a post-Snowden era [1], [2]. It has been the subject of a United Nations resolution that categorically establishes the right to online privacy as a human right [3]. This right should also — perhaps even especially — contemplate those living under censorship and various degrees of restrictions to freedom of speech and information [4], [5].

Indeed, ensuring the right to privacy of every individual is often the *raison d'être* of anonymity networks such as Tor and many virtual private network (VPN) services [6]–[10]. These networks allow users to access information in a secure and anonymous manner, which is not only a human right, but also a necessity for those who could be discriminated against or even persecuted by local authorities based on, *e.g.*, infringements of religious or moral legislation and customs [11]–[13]. Moreover, the privacy guaranteed by these networks is essential for the continuous advancement of human rights through the work of whistleblowers, activists, and journalists [14].

Anonymity networks have not only helped those living under strict censorship to safeguard their privacy, but also aided academic research into the various aspects of censorship, such as the types of blocking and filtering performed, what content it affects, and how it is executed [15]–[17]. Indeed, substantial research has been conducted on censorship both on a country level — where countries with notoriously restrictive political regimes such as China [16], [18]–[21] have received greater attention — and on a global scale, resulting in tools and methodologies that can be used in further research, such as ICLab [15].

However, the extent of restrictions on access to online content by users of anonymity networks outside of a censorship context remains, to the best of the author's knowledge, largely unexplored. These restrictions have been mentioned in Niaki, Cho, Weinberg, *et al.* [15] in the context of the VPNs used as vantage points in ICLab, and have been explained in Davidson, Goldberg, Sullivan, *et al.* [22] as the issue that Privacy Pass — a tool that content distribution networks (CDNs) can use to anonymously authenticate users and reduce the number of challenge-response tests that these users receive — aims to solve. Indeed, they explain that such restrictions are mainly due to how CDNs tend to block IP addresses with a 'poor reputation,' of which a large portion is made up of the shared IP addresses of anonymity networks because of malicious users [22].

This type of blocking constitutes server-side blocking and the only studies that explore it either focus on Tor exclusively, as is the case with Khattak, Fifield, Afroz, *et al.* [23], or treat VPNs insofar as they provide a foil for the more pervasive blocking of Tor, as in Singh, Nithyanand, Afroz, *et al.* [24]. It is precisely to widen the scope of and build on this scholarship that the present study was conducted. The VPN service chosen was Mullvad VPN due to both the author's familiarity with it, and its excellence in privacy [25]–[28]. The study aims to determine the extent to which users of Mullvad VPN

are blocked by popular websites outside of a censorship context, explore the nature of these blocks, and thus contribute to establish the scale of Internet censorship and restrictions to freedom of information. Moreover, it can shed light on the bias that current attack detection systems used by CDNs and other networks have against privacy-aware users, and its findings can prompt further development and adoption of alternative, more discerning systems that do not punish users for exerting their right to privacy.

In essence, the research questions are: to what degree do popular websites block users of Mullvad VPN, and what is the nature of these blocks. To this end, websites were requested regularly both whilst using Mullvad VPN and openly accessing the network, *i.e.*, using a control connection. The HTTP status code, a screenshot of the loaded page for successful requests, and any errors from both connections were logged and then compared to detect HTTP blocking (*e.g.*, `403` status code), timeouts (*i.e.*, sites that took longer than forty seconds to load), network errors (*e.g.*, connection refused, too many redirects), differentiated content (*e.g.*, a `200` status code is received, but a block page or CAPTCHA page is loaded), and partial blocking (*e.g.*, the home page loads, but some functionality such as login or search is blocked). The results show that there is no significant blocking of home pages, but that this changes when navigating to subpages. They also show that certain categories of websites, such as restaurants and shareware, tend to block more often than others.

Section 2 describes the methods and approaches used to address the research questions; the overall design and implementation of the crawler used to gather the data are detailed in Section 3; and the experiment design and setup are delineated in Section 4. The block classification is explained in Section 5, and the results are presented in Section 6 and discussed in Section 8, which also expounds on the study's limitations. The ethical ramifications of the study and its reproducibility are examined in Section 7. Finally, the conclusions and possibilities for future work are reported in Section 9.

2 Methodology

To establish to what extent users of Mullvad VPN are blocked by popular websites, careful considerations need to be made regarding the definitions of blocked access and popular websites used in the study, how and what data will be collected, and what will be done to minimise any bias in the data set.

2.1 Types of Server-Side Blocking

Restrictions to the access of a website experienced by users solely by virtue of the connection having been established through an anonymity network constitute server-side blocking. Research into the different types of blocking that these users might experience when accessing popular websites was somewhat restricted to the available literature on Tor [29]. Nevertheless, the results reveal that the most common types of blocks are human challenge-response tests (*e.g.*, CAPTCHAs), block pages, and restricted functionality (*e.g.*, the user can access the website's home page, but cannot login or use any search features) [22]–[24], [29].

The most widespread reasons for server-side blocking of Tor are explained by Singh, Nithyanand, Afroz, *et al.* [24] in terms of *reactive blacklisting* and *proactive blacklisting*. The former happens when (at least) one user who has been assigned a (shared) IP address conducts themselves in a manner that causes the (shared) IP address to be blacklisted; this form of blacklisting is also described by Davidson, Goldberg, Sullivan, *et al.* [22]. The latter constitutes preemptively blacklisting an IP address identified as that of a Tor exit node, even if there is no history of any misconduct associated with that IP address, and presumably to prevent undesired traffic [24].

This study assumes — based on the information presented by Davidson, Goldberg, Sullivan, *et al.* [22] — that at least *reactive blacklisting* also applies to VPNs, and measures the impact on users of Mullvad VPN as evidenced by the most common types of blocks listed above.

2.2 Identifying Server-Side Blocking

Further research was conducted into how these types of blocking can be identified [30]. In essence, the most straightforward ways of detecting blocks are at the Transport Layer and at the Application Layer level.

At the Transport Layer level, one can analyse the initial TCP three-way handshake and observe how the server responds to the client's SYN request: if the server returns a SYN-ACK, then it can be safely concluded that the request was successful and no blocking was performed at this layer; however, if a RST or no response at all is received, then the request was unsuccessful and more information is needed to distinguish between systemic failure (*e.g.*, accidental packet loss, network congestion, network outages) and intentional blocking [23]. This is why most studies argue for sending a number of successive requests and only concluding that there was any blocking or tampering in a TCP connection if a certain threshold of failures was reached [17], [23], [24], [31].

At the Application Layer level, blocking can be identified from HTTP status codes such as `403` *Forbidden* and `501` *Not Implemented*, or from differentiated content returned with a `200` *OK* status code, such as block pages and pages with features like login missing or disabled [23]. However, it is vital to note that it can only be concluded that a received non-`200` status code is evidence of blocking if the same request from a control connection does result in a `200` status code [23].

Furthermore, a `200` status code alone cannot be assumed to signify an unblocked request. This is because some servers will return a block page instead of the desired page [23], and detecting these blocks is a non-trivial task [32], [33]. Indeed, Khattak, Fifield, Afroz, *et al.* [23] mention this type of server-side blocking as one that their study fails to detect. However, Jones, Lee, Feamster, *et al.* [32] propose three automated identification methods: page length comparison, cosine similarity, and DOM similarity. Building on this work and that of Khattak, Fifield, Afroz, *et al.* [23], Singh, Nithyanand, Afroz, *et al.* [24] also present a method that compares screenshots of pages loaded with Tor and from a control using perceptual hashing, a technique that ensures that similar inputs result in similar hashes (*cf.* §5.1). Thus, given certain thresholds, some cases can be automatically assumed to constitute blocking or unblocking, while others require further investigation.

In this study, a combination of these methods is used to detect blocking (*cf.* §3, §4, §5).

2.3 Popular Websites

The results of research into the potential sources of lists of global popular domains show that the Alexa list of 'top sites' [34] is a suitable choice for the present study [35]. This is attributable to the ubiquitous use of the list in previous studies (*e.g.*, [15], [33], [36], [37]), and in spite of known instability issues [38], [39], which were mitigated by restricting the list of domains used to the Alexa Top 10K results [35], [40].

This list was downloaded on 10 May 2021 and used in its entirety, with the exception of four domains (`oeeee.com`, `taleo.net`, `tamin.ir`, `support.wix.com`) that were found to crash the browser used by the crawler in the early testing stages, and were therefore excluded.

2.4 Bias in the Data

Research reveals a myriad of potential sources of bias in a data set obtained from crawling a list of popular websites, which could result in false positives [41]. The main sources, however, can be divided into three categories: network connectivity issues, geoblocking, and crawler-related issues.

Network connectivity issues include, *e.g.*, accidental packet loss, unexpected interruptions or disconnections in the network, routing delays, connection issues on the client's side, and servers being down. Although most of these issues are completely out of the author's control, nevertheless there are still some measures that could be — and were — taken to mitigate their effect and rule out some false positives. These measures include: ensuring the list of popular websites used remains the same throughout the study to guarantee consistency in measurements (*cf.* §2.3), requesting the same website using the same settings more than once within a reasonable time period (*cf.* §2.2), and repeating these requests without the use of an anonymity network (*cf.* §2.2).

Issues related to geoblocking can manifest in servers denying or restricting access to users of certain geographical locations due to, *e.g.*, economic sanctions, censorship, or security. McDonald, Bernhard, Valenta, *et al.* [33] investigate these motives in the context of CDNs, and Tschantz, Afroz, Sajid, *et al.* [36] explore geoblocking in relation to non-compliance with the European Union's (EU) General Data Protection Regulation (GDPR). Since geoblocking *per se* is unrelated to blocking users of anonymity networks, the locations of the exit nodes chosen for the study were restricted to Sweden and the Netherlands in order to minimise bias (*cf.* §4.1).

Lastly, a rather pervasive source of bias can be the crawler itself. Indeed, the bias introduced by using automated crawls instead of human browsing when collecting data is the subject of a recent study by Zeber, Bird, Oliveira, *et al.* [42]. They argue that this bias is exacerbated by the fact that most crawls are only performed once, using specialised frameworks and implementations that vastly differ from human browsing [42]. Moreover, the widespread use and development of bot-detection techniques render the implementation of an undetectable crawler virtually infeasible [23], [33], [42]. Therefore, in order to mitigate at least in part the introduction of a plethora of false positives in the data due to bot-detection, each crawl was repeated five times and the crawler was carefully designed and implemented (*cf.* §3).

3 Crawler

In order to contemplate the necessary points delineated in Section 2 regarding the identification of server-side blocking and the minimisation of bias in the data set, the crawler needs to be designed and implemented accordingly (*cf.* Fig. 2).

3.1 Design

The crawler can perform home page requests to all the top 3,000 domains from the Alexa list, and request up to two subpages from the top 1,000 domains that were not found to be blocked. This is to ensure that partial blocks are detected. In order to guarantee consistency across different crawls and enable classification, the list of subpages was dynamically compiled once by using the crawler to access the links present in each home page, and was then used statically, without any alterations, for the crawls that gathered data.

Every request made using Mullvad VPN is also made from an open connection (*i.e.*, a control connection) in relatively close succession (*i.e.*, at most within a few minutes) to increase the certainty that any differences in the results were due to the use of the VPN.

Similarly, each request that does not succeed, *i.e.*, does not return any HTTP status code, is repeated up to two times to rule out network connectivity issues. The total number of attempts was chosen to be three because preliminary testing showed that to be an efficient number for minimising false positives due to network issues.

Moreover, any request times out after n seconds, where the value of n increases with the attempt number — at first, 30, then 35, and finally 40 seconds — to more accurately identify blocking by timeouts and accommodate pages with heavy resources. To minimise bot-detection, the crawler waits for five seconds before repeating the request.

Each successful request that returns a status code in the `200` range triggers a screenshot command. The screenshots made from VPN requests will later be compared to those from a control connection to identify block pages, human challenge-response tests, and differentiated content (*cf.* §5).

For each request, the crawler logs a unique ID number, a subpage index (if any), a timestamp, the duration of the request, the number of attempts, the original request domain, the resolved response domain, the IP address, the status code received, and any network errors. This will enable the necessary analysis to identify server-side blocking (*cf.* §5).

The crawler is also able to automatically accept most cookie consent requests so that a more accurate depiction of the page is recorded. It also blocks most advertisements and pop-up windows so that they do not interfere with the subsequent automated analysis of the screenshots taken and to reduce bandwidth usage.

3.2 Implementation

In order to best avoid false positives due to bot-detection, a framework capable of accurately and efficiently controlling a fully-fledged browser is needed. There are many options available that excel in different aspects of web crawling and automated testing, but Google's Puppeteer [43] — a Node.js application programming interface (API) developed to control Chrome or Chromium using the DevTools Protocol [44]

— is the most complete one, surpassing others in usability, reliability, and performance [45]. Moreover, it does not suffer from the same race conditions as Selenium does [46] nor does it need additional frameworks in order to take screenshots like Scrapy does. Indeed, Puppeteer was also selected in studies that require evading bot-detection [47] and was therefore chosen as the framework used to implement the crawler.

In order to leverage the benefits of type inference and minimise runtime errors, TypeScript was the language chosen for the crawler [48]. However, due to defective builtin types present in Puppeteer versions 7 and above, the more stable and fully compatible version 5 was used [49]. Additionally, in order to address an error related to handling requests that is still an open issue in the Puppeteer repository [50], the workaround proposed by a user in the same thread was followed: commenting out the asserts on lines 217, 268, and 314 in the `HTTPRequest.js` file in the Puppeteer source code.

Despite these issues and in addition to the aforementioned reasons for choosing Puppeteer, there are also the benefits of dedicated plugins and excellent compatibility with Chrome extensions. Indeed, the `puppeteer-extra` plugin [51] enables the use of two others: `puppeteer-extra-plugin-stealth`, which applies various techniques to avoid bot-detection and currently passes all public bot tests [52], and `puppeteer-extra-plugin-adblocker`, which is an efficient adblocker that also supports tracker blocking [53].

To address the automated cookie consent requirement, the `I don't care about cookies` Chrome extension was used [54]. However, the use of extensions is not supported by Puppeteer when running Chromium headless (*i.e.*, without opening a browser instance), which means that it had to be run in headful mode (*i.e.*, opening a browser instance). Since this was necessary for the cookie consent, another extension was used to mitigate the cases when `puppeteer-extra-plugin-adblocker` failed to block advertisements and pop-up windows: `AdBlock` [55], which blocks all advertisements that do not comply with the 'Acceptable Ads' programme [56].

4 Experiment

The crawler was both developed in the context of and deployed in the stages into which the overarching experiment of requesting websites and collecting data was subdivided.

4.1 Design

The experiment was designed as a series of stages, each informing the structure and setup of the next:

- **Stage 0**: 10 domains, only home pages, Finnish exit node, Dutch control connection;
- **Stage 1**: 100 domains, only home pages, Swedish exit node, Dutch control connection;
- **Stage 2**: 1,000 domains, only home pages, French exit node, Dutch control connection;
- **Stage 3**: 3,000 domains, only home pages, Swedish exit node, Dutch control connection;
- **Stage 4**: 1,000 domains, 2 subpages from each, Dutch exit node, Dutch control connection.

Since Stages 0-2 were mainly used to implement and test the crawler, ensuring reliability and scalability, their results will be omitted from this paper. Moreover, the network errors reported are likely biased due to the fact that the requests from the Mullvad VPN connection and from the control connection were made sequentially, rather than in parallel.

The results from Stage 3, however, were used to select 1,000 domains found not to block Mullvad VPN. This was done in order to better gauge the number of websites engaging in partial blocking in Stage 4 (*cf.* §5).

Locations of VPN exit nodes were restricted to EU countries in order to minimise any geoblocking bias (*cf.* §2.4). The specific countries were chosen randomly for Stages 0-2, and based on availability of Mullvad VPN servers and correspondence with the control connection for Stages 3 and 4.

4.2 Setup

The requests were run in parallel on separate machines with separate Internet connections. Control connection requests were made from a MacBook Air (1.4 GHz Dual-Core Intel Core i5, 4 GB RAM), using a Dutch residential broadband connection provided by Ziggo. Mullvad VPN connection requests were made from a MacBook Pro (2 GHz Quad-Core Intel Core i5, 16 GB RAM), using a Dutch 4G, mobile connection provided by T-Mobile.

This setup suffers from synchronisation issues due to different hardware processing and connection speeds. For Stage 3, the time differences between control and VPN requests has stayed within a ten-minute margin, which is acceptable because meaningful network issues are unlikely to have either developed or been resolved in that space of time, and should therefore not significantly influence the final results. However, for Stage 4, the requests became desynchronised by a margin ranging from two to three hours. Although these results are less reliable due to this gap between requests, nevertheless the repetition of the crawls on five different days should help somewhat mitigate any bias introduced.

5 Block Classification

The data obtained from Stages 3 and 4 of the experiment were subsequently analysed in order to identify any server-side blocking and thus classify each request in one of the following categories: not blocked, blocked, maybe blocked, or presenting no discernible difference from the control connection. The requests classified as blocked were also subdivided into six different block categories: HTTP blocks, timeout blocks, network error blocks, differentiated content, block page, and challenge-response test. The last three were only identifiable from screenshots taken from successful requests (*i.e.*, a response with an HTTP status code in the `200` range).

5.1 Perceptual Hashing

The automated comparison of screenshots was performed with perceptual hashing, a technique that can be used to fingerprint various forms of multimedia due to its robustness against minor distortions caused by, *e.g.*, noise, small modifications, and compression technique [57]. The use of perceptual hashing to assess the level of similarity between two images has widespread adoption in academic studies (*e.g.*, [58],

[59]), and was also used by Singh, Nithyanand, Afroz, *et al.* [24] to compare screenshots.

In this study, the Python library ImageHash [60] was used to obtain the perceptual hashes of the screenshots of successful requests from the VPN and the control connections. The absolute value of the difference between the two hashes was calculated and stored for the classification process, which deemed difference values below 20 as not blocked, and all other values as requiring manual verification.

The threshold value of 20 was chosen by a process reminiscent of machine learning practices in the division of the data set into a training set and a validation set. The screenshots taken during the crawl performed on 21 May 2021 were all manually checked and the lowest difference value that presented a block was 20. The screenshots from 22 May 2021 were then automatically compared, with 20 as a threshold, and manually checked for validation: the lowest difference value found to present a block was 23. It was assumed that the data taken from one day would be representative of the data obtained on other days. This process provided some reassurance that 20 was indeed a reasonable threshold value, and that any potential false negatives in subsequent automated analysis (which were performed for the remaining data sets of Stage 3 and all data sets of Stage 4) should be minimal and therefore unlikely to be statistically significant.

5.2 Block Classification Pipeline

To classify each request from the Mullvad VPN connection, a thorough comparison was made with the response from the control counterpart. A flow diagram of this process, including the type of block assigned, can be found in Figure 3.

A request was only deemed not blocked when the Mullvad VPN connection returned a response with an HTTP status code in the `200` range and one of the following was true: (1) the control connection also obtained a status code in the 200 range and either the difference in the perceptual hashes of the respective screenshots was below 20, or the manual check confirmed there was no block; (2) the control obtained a status code outside of the `200` range, timed out, or resulted in a network error. This is because such cases were most likely attributable to Internet connectivity issues: they accounted for 30 to 34 out of the 3,000 requests made each day, of which roughly half were always the same websites. Subsequent manual checks confirmed no evidence of blocking and that a few took quite some time to load (significantly over the forty-second timeout threshold).

Comparatively, a request was only classified as blocked when the response from the control connection had a status code in the `200` range, but the VPN connection somehow failed (*i.e.*, HTTP status code outside of the `200` range, timeout, network error) or the manual check of the screenshots showed evidence of blocking through differentiated content (*i.e.*, broken HTML or more than $\approx 40\%$ of elements missing; *cf.* Fig. 5), challenge-response tests (*i.e.*, CAPTCHAs and similar challenges; *cf.* Fig. 6), or block pages (*i.e.*, a page refusing access for any reason and without the option of authentication through a challenge-response test; *cf.* Fig. 7).

Cases when the requests from both the Mullvad VPN and the control connections failed with the same kind of response (*i.e.*, both had an HTTP error, a timeout, or a network error) were categorised as presenting no difference between the two connections. This is because they might have been blocked for reasons other than the use of an anonymity network, such as geoblocking or bot-detection.

Lastly, cases when both requests failed, but with different kinds of responses (*e.g.*, the Mullvad VPN connection had an HTTP error whilst the control connection timed out), were classified as maybe blocked due to insufficient information. Indeed, a similar reasoning is presented by Niaki, Cho, Weinberg, *et al.* [15] for cases where censorship could neither be confirmed not denied. The majority of these instances in the present study were due to timeouts in the control connection.

Based on this classification of each request, the data from Stage 4 was also analysed at the domain level, where each of the 1,000 domains were classified according to the presence of home page blocks, subpage blocks, any potential blocks, presenting no difference, or no block. This further classification followed a similar logic as the one presented above and its flow diagram can be found in Figure 4.

6 Results

The data obtained from the crawls and block classifications were examined in relation to the research questions: to what degree do popular websites block users of Mullvad VPN, and what is the nature of these blocks.

6.1 Stage 3: Home Pages Only

The data gathered during Stage 3 are summarised in Table 1. The slight discrepancy in the number of successful responses between the VPN and the control connections is most likely due to the poor Internet connection used for the control. Nevertheless, the numbers stay reasonably consistent throughout the five days, suggesting that there is no statistical difference between them. Indeed, the results of a Chi-Squared test on the data under the null hypothesis of independence corroborate this suggestion: the p-value was 0.996 for the VPN connection and 0.820 for the control ($\alpha = 0.05$). This test was chosen because it is the most appropriate to check if the discrete variables are independent between samples of categorical data. Since the data from each day are independently distributed, they can be used as a single data set of 15,000 data points for each of the connections, which should increase the statistical power of the analysis.

This data was then run through the block classification pipeline (*cf.* §5.2) and the resulting classifications are summarised on Table 2. Similarly to the data itself, the block classifications are also reasonably consistent throughout the five days, and the Chi-Squared test executed under the null hypothesis of independence confirmed that there is no statistical difference between them (p-value $= 0.122$, $\alpha = 0.05$). Therefore, these data can also be treated as a single data set.

Before continuing with the analysis, it is important to discuss how to treat the requests classified as 'Maybe Blocked'. A Two-Sample Proportion test was run under the null hypothesis of independence to establish whether counting these requests as 'Blocked' or 'Not Blocked' would significantly impact the results and, with a p-value of 0.052 ($\alpha = 0.025$), the

Date	Total Requests	Status Code in 200 Range		Status Code Outside 200 Range		Timeouts		Network Errors	
		Mullvad VPN	Control	Mullvad VPN	Control	Mullvad VPN	Control	Mullvad VPN	Control
2021-5-21	3000	2760	2743	87	96	13	13	140	148
2021-5-22	3000	2752	2742	89	97	13	17	146	144
2021-5-23	3000	2751	2740	90	95	18	21	141	144
2021-5-24	3000	2748	2740	88	99	20	12	144	149
2021-5-25	3000	2748	2743	92	95	18	8	142	154
Averages (%)		91.73%	91.39%	2.97%	3.21%	0.55%	0.47%	4.75%	4.93%

Table 1: *Summary of data gathered during Stage 3 (requesting home pages only).*

Date	Total Requests	Not Blocked	Blocked	Maybe Blocked	No Difference	Types of Blocks					
						HTTP	Timeout	Network Error	Differentiated Content	Block Page	Challenge-Response Test
2021-5-21	3000	2755	20	10	215	9	5	1	2	1	2
2021-5-22	3000	2739	36	12	213	11	5	7	9	2	2
2021-5-23	3000	2746	24	1	229	10	5	4	3	0	2
2021-5-24	3000	2741	32	7	220	9	10	6	5	0	2
2021-5-25	3000	2743	34	5	218	15	10	4	3	0	2
Averages (%)		91.49%	0.97%	0.23%	7.30%	0.36%	0.23%	0.15%	0.15%	0.02%	0.07%

Table 2: *Summary of blocks identified during Stage 3 (requesting home pages only). The figures categorised into different types of blocks all come from the Blocked column.*

test confirmed no statistical difference. This test was chosen because it is the most appropriate to compare the proportions between two samples of categorical data. Therefore, whenever necessary to make a distinction, these requests will be counted as 'Blocked' because the prevalence of successful requests suggests that such issues solely on the VPN connection are more likely to be due to server-side blocking.

The results from the block classification show that HTTP status codes outside the 200 range are the most common types of blocks, followed by timeouts, and that, for successful requests, blocking is primarily experienced in the form of differentiated content. Nevertheless, the overall number of blocked requests is rather small, comprising only 146 out of the 15,000 requests performed (an average of 0.97%).

Indeed, in order to assess the degree of blocking, one could compare the number of requests that were blocked, maybe blocked or presented no difference when connecting from Mullvad VPN to the number of requests that resulted in HTTP status codes outside the 200 range, timed out or resulted in network errors when connecting from the control connection. A Two-Sample Proportion test performed under the null hypothesis of independence on these aggregated figures returned a p-value of 0.741 ($\alpha = 0.025$), which suggests that any blocking faced by Mullvad VPN users does not significantly degrade their experience of popular websites when compared to random network connectivity issues.

6.2 Stage 4: Subpages

The data gathered during Stage 4 of the experiment are summarised in Table 4. Similarly to Stage 3, a Chi-Squared test under the null hypothesis of independence was performed to confirm that there is no statistically significant difference between the data collected on the five days (p-value $= 0.923$ for Mullvad VPN and p-value $= 0.4$ for control; $\alpha = 0.05$). Therefore, the data sets were combined for each connection.

The results of the block classification are summarised in Table 5, and a Chi-Squared test under the null hypothesis of independence confirms no statistical difference between the five days (p-value $= 0.825$, $\alpha = 0.05$), resulting in the combination of the data. HTTP blocks remain the most common type of blocks, comprising 0.25% of the 15,000 requests. However, when requesting subpages, differentiated content comes as a close second, adding up to 0.22%. This illustrates the subtleties involved in server-side blocking.

The data was also classified at the domain level, where each of the 1,000 domains was classified as 'Home Page Blocked', 'Subpage Blocked', 'Maybe Blocked', 'No Difference', or 'Not Blocked' (*cf.* §5.2, Fig. 4). The results of this classification are summarised in Table 3. The outcomes of a Chi-Squared test performed under the null hypothesis of independence confirm that there is no statistical difference between the data collected on the five days (p-value $= 0.979$, $\alpha = 0.05$) and the data sets were therefore combined.

Noticeably, there were no requests classified as 'Maybe Blocked' in this stage. Both this and the low incidence of blocks could be due to the fact that the domains requested were obtained from those classified as 'Not Blocked' in the previous stage. Indeed, a Two-Sample Proportion test performed under the null hypothesis of independence confirmed that the number of failed or blocked requests out of 15,000 from the VPN connection is not significantly different from the total number of failed requests from the control connection (p-value $= 0.722$, $\alpha = 0.025$).

However, when a similar test was performed on the data further classified at the domain level, where the number of domains classified as 'Home Page Blocked', 'Subpage Blocked', or 'No Difference' (*cf.* Table 3) was compared to the number of domain requests that did not result in an HTTP status code in the 200 range from the control connection, the difference is significant (p-value $=$ 1.334e-18, $\alpha = 0.025$). This suggests that there is a meaningful proportion of websites that block access to certain content beyond the home page when a user connects through Mullvad VPN.

6.3 Blocks by Categories

To further explore the nature of the blocks encountered in both stages, the domains requested were categorised using the McAfee URL categorisation service [61].

Date	Domains	Not Blocked	Home Page Blocked	Subpage Blocked	Maybe Blocked	No Difference
2021-6-8	1000	964	6	7	0	23
2021-6-9	1000	965	4	6	0	25
2021-6-11	1000	962	7	6	0	25
2021-6-12	1000	958	8	10	0	24
2021-6-13	1000	963	9	6	0	22
Averages (%)		96.24%	0.68%	0.70%	0.00%	2.38

Table 3: *Summary of domain blocks identified during Stage 4 (requesting two subpages from each of 1,000 domains).*

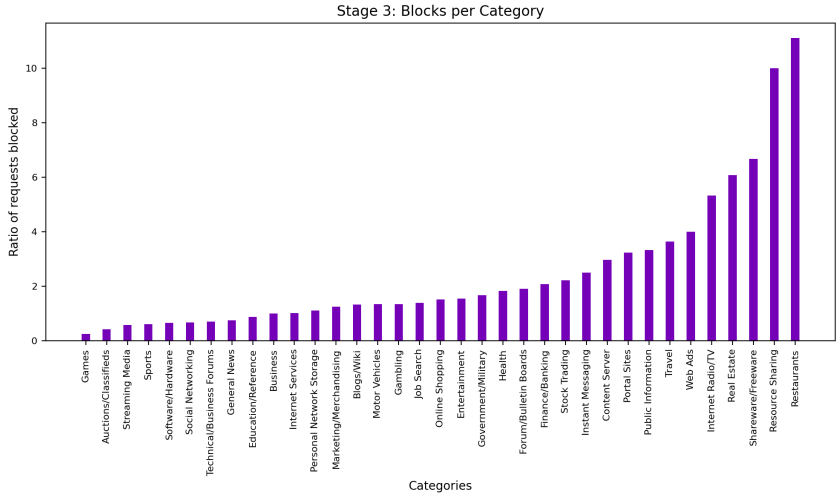

Figure 1: *Graph illustrating the ratio of blocked requests identified per category of website during Stage 3 (requesting home pages only).*

The results of the categorisation of requests from Stage 3 are summarised in Table 6, and the ratio of blocked requests for each category that presented some form of blocking is illustrated in Figure 1. From this data, it can be concluded that 'Restaurants' block access from Mullvad VPN users the most, whilst 'Games' block the least. Indeed, this almost tenfold discrepancy in blocked ratios was confirmed by a Chi-Squared test performed under the null hypothesis of independence (p-value $= 6.805$e-13, $\alpha = 0.05$).

The results of the categorisation of requests from Stage 4 are summarised in Table 7, and the ratio of blocked domains for each category that presented some form of blocking is illustrated in Figure 8. This shows a more dramatic disparity between the ratios of 'Remote Access' and 'Finance/Banking'. Upon manual inspection, this high ratio was found to be due to a single domain (`anydesk.com`) that blocked two out of five requests. Indeed, this difference was confirmed to be statistically significant by a corresponding Chi-Squared test (p-value 1.26e-11, $\alpha = 0.05$). However, the fact that the p-value from Stage 4 is greater than that from Stage 3 indicates that the high ratio of around 40% for 'Remote Access' in Stage 4 did not dramatically influence the results of the test for that data set.

7 Responsible Research

The key ethical concerns of this study are related to the crawler and `robots.txt` files, whereas its reproducibility can be explored in relation to the VPN service used, real-world changes, and code documentation and publication.

7.1 Ethics

The primary purpose of `robots.txt` files is to instruct search engine crawlers as to which pages it can and cannot request for a particular domain with a view to control traffic and prevent the server from being overburdened [62]. Secondary purposes can be fundamentally subdivided into data copyright issues and perceived endorsement from content creators of data divulged by content users [63]. Indeed, the legal and ethical ramifications surrounding the use and misuse of `robots.txt` files are manifold [63]–[65].

Nevertheless, the author's decision to implement a crawler that does not consult such files before making requests is based on the premise that the purpose of these files does not apply directly to the use case in question.

Firstly, `robots.txt` files are part of the Robots Exclusion Protocol, a work in progress whose latest draft mention the use case of crawlers that access a website's entire uniform resource identifier (URI) space [66]. Since the maximum number of unique pages from a single domain requested by the crawler for each crawl in this study is three (one of them being the home page), it can be argued that the use cases are fundamentally different.

Secondly, the maximum number of requests made to each domain is nine (considering cases when a request fails and is repeated at most twice), with at the very least a six-second delay between each unique URL and a five-second delay between each retry in case of failure. In practice, the total delay is usually larger due to the download time of each page and the crawler typically takes between nine and ten hours

to crawl 3,000 domains, or 1,000 domains and two subpages from each. With these figures, it is virtually impossible to overburden a server which is capable of handling enough traffic to feature in the Alexa Top 10K sites worldwide.

Thirdly, the data downloaded in each request is mostly discarded (*cf.* §3.1). Only links readily available on each domain's home page are requested and the data saved are solely made available as statistics (*e.g.*, Table 1) or visual examples of blocked pages (*e.g.*, Fig. 7), which, by definition, do not expose significant content.

Lastly, it has been argued that the legal standing of `robots.txt` files is somewhat tenuous and that a more pragmatic approach is to follow the structure of the fair use exception in US copyright law [63]. In essence, each use case needs to be considered individually and all parties involved should reflect on the legal and ethical ramifications of their actions [64], as has been done in this study.

7.2 Reproducibility

The main hindrance to the reproducibility of the study is the fact that Mullvad VPN is a paid service. However, since the scope of the project is limited to crawls that run for a total of eleven days and Mullvad VPN costs €5 per month [67], this should still be within reach of most people interested in reproducing the results.

It is important to note that the reproducibility of the results are also somewhat limited by the nature of the study itself: since it measures network behaviour in the real world, the results obtained are inherently dependent on any changes in, *e.g.*, server policies, network configurations, or political and economical developments, that might happen over time.

Nevertheless, the most important design and implementation decisions for the crawler at each stage of the experiment and for the analysis of the data have been thoroughly documented in this paper (*cf.* §3, §4, §5, §6), and the code used for all stages of the experiment is readily available at the author's GitHub repository under an MIT license [68].

8 Discussion

The results from home page requests show that approximately 0.97% of requests made from a Mullvad VPN connection were somehow blocked (1.2% if 'Maybe Blocked' instances are counted as blocked), a number which does not significantly degrade the experience of users when accessing popular websites (*cf.* §6.1). Indeed, when compared to the $\approx 3.67\%$ of Tor requests blocked reported by Khattak, Fifield, Afroz, *et al.* [23] or to the even larger $\approx 20\%$ reported by Singh, Nithyanand, Afroz, *et al.* [24], this number seems even less significant.

However, when looking at domains that also block certain subpages, this number rises to 1.38%, which is enough to make a significant difference in how Mullvad VPN users experience popular websites (*cf.* §6.2). Moreover, some of the categories of websites that presented a greater ratio of server-side blocking offer essential services, such as 'Public Information' and 'Health' in Stage 3 data, and 'Education/Reference' and 'Government/Military' in Stage 4 data (*cf.* §6.3).

Indeed, since some servers actively block users accessing them from locations inside the EU in order to avoid issues related to GDPR compliance,[1] the use of VPN services with exit nodes outside the EU could prove vital for users who depend in any way on the services offered by these websites. Although it can be conjectured that this number is likely to have decreased as servers adapt to GDPR requirements in order not to lose business in the European market, nevertheless similar situations might arise in the future. Moreover, the issue of geoblocking is more pervasive than mere GDPR compliance [33], [36], not to mention the situations where government-enforced censorship comes into play (*cf.* §1).

Therefore, it is paramount that servers continue to explore more discerning ways of protecting themselves against malicious users and undesired traffic that do not involve the categorical blocking of users of anonymity networks. Indeed, Cloudflare has shown that this is possible with the development and implementation of Privacy Pass [22], [69], [70]. Similar alternatives exist, such as the secure group anonymous authentication protocol (GAAP) developed by Agrawal, Bu, Del Rosario, *et al.* [71] which also uses zero-knowledge proofs; the Practical Anonymity at the Network Level (PANEL) solution based on hardware switching proposed by Moghaddam and Mosenia [72]; and CACTI, the Captcha Avoidance via Client-side Trusted Execution Environment (TEE) Integration based on rate-proofs created by Nakatsuka, Ozturk, Paverd, *et al.* [73].

8.1 Limitations

The study is inherently limited by time and resource constraints. Regarding time, it had to be designed and performed in its entirety within ten weeks. Therefore, each stage in the experiment could only be run for a limited time, which could have had an impact on the reliability and statistical significance of the data collected. Regarding resources, the author only had access to a sub-par home Internet connection and attempted to mitigate this by also utilising a 4G mobile connection and repeating failed requests (*cf.* §3, §4). Neither one of the connections is ideal for network measurement studies and could also have impacted the reliability of the data, as it is likely to have done during Stage 4 when the desynchronisation between the crawlers ranged from two to three hours.

9 Conclusion and Future Work

Although the right to online privacy constitutes a universal human right, server-side discrimination against users of anonymity networks means that those who choose to exert that right will receive an inferior service. This discrimination can take many forms, such as excessive challenge-response tests, differentiated content, block pages, HTTP errors, network errors, and timeouts.

The present study has investigated the extent to which users of Mullvad VPN are blocked in these ways when trying to access popular websites and what is the nature of these blocks. The experiment first looked at the top 3,000 domains from the

[1] Tschantz, Afroz, Sajid, *et al.* [36] report that 74 domains from the Alexa Top 500 list engaged in differentiated treatment of EU users once GDPR went into effect on 25 May 2018.

Alexa Top 10K sites and concluded that the requests limited to home pages only experience 0.97% blocks (1.2% if unsure data is counted as blocked), which does not constitute a statistically significant degradation in service when compared to general failures in a control connection.

However, once this was extended to requesting two subpages from each of the top 1,000 domains that had not engaged in blocking in the previous stage, the new figure of 1.38% of domains partaking in home page or subpage blocking did present a statistically significant difference. Indeed, this deterioration in how Mullvad VPN users experience popular websites is also reflected in the categories of websites that present a high ratio of blocks, such as health and government, and which constitute essential services.

Regardless of how this discrimination might be justified on the grounds of self-protection against malicious users who tend to operate through anonymity networks, the fact that there are more discerning alternatives available for authenticating genuine users without compromising their anonymity means that there is no legitimate reason for jeopardising their online experience.

Future work should consider running each stage of the experiment for longer periods of time and requesting more domains for greater statistical significance. It could also investigate if faster and more reliable Internet connections eliminate cases when there could be no certainty of a block, and if overall failures decrease significantly. It might prove fruitful to explore whether the time of day bears any correlation with the number of failures and blocks experienced. Lastly, the study could benefit from being conducted with more VPN exit node locations, perhaps including sites outside of the EU with the appropriate control connections.

Acknowledgments

The author would like to thank Dr Stefanie Roos for essential guidance and feedback throughout the research process. Paula Iacoban, Willemijn Tutuarima, Jurgen Mulder, and Anant Pingle for support, helpful suggestions, and shared research. She would also like to thank Dr Fokko van de Bult and Dr Christophe Smet for answering her questions regarding statistical analysis and providing valuable advice. Lastly, she would like to thank Bruna Louzada, Dixit Sabharwal, Emilija Zlatkutė, Kevin Tjiam, Evaldas Latoškinas, Shruti Arora, Tim Anema, Jonathan Dönszelmann, and Timea Nagy for their help categorising URLs.

References

[1] M. Milanovic, "Human rights treaties and foreign surveillance: Privacy in the digital age," *Harvard International Law Journal*, vol. 56, p. 81, 2015.

[2] S. Kulhari, "Data protection, privacy and identity: A complex triad," in *Building-Blocks of a Data Protection Revolution*, Nomos Verlagsgesellschaft mbH & Co. KG, 2018, ch. III, pp. 23–37.

[3] A. Chander and M. Land, "United nations general assembly resolution on the right to privacy in the digital age," *Int'l Legal Materials*, vol. 53, p. 727, 2014.

[4] R. Shandler, "Measuring the political and social implications of government-initiated cyber shutdowns," in *8th USENIX Workshop on Free and Open Communications on the Internet (FOCI 18)*, 2018.

[5] C. M. Wong, "Internet at a crossroads how government surveillance threatens how we communicate," *Human Rights Watch, World report 2015: events of 2014*, pp. 14–26, 2015.

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

[7] Tor, *History*, (Accessed on 04/19/2021). [Online]. Available: https://www.torproject.org/about/history/.

[8] Mullvad, *Privacy is a universal right - blog*, (Accessed on 04/19/2021), Dec. 2016. [Online]. Available: https://mullvad.net/en/blog/2016/12/5/privacy-universal-right/.

[9] ProtonVPN, *About*, (Accessed on 04/19/2021). [Online]. Available: https://protonvpn.com/about.

[10] HotspotShield, *What is a vpn?* (Accessed on 04/19/2021). [Online]. Available: https://www.hotspotshield.com/what-is-a-vpn/.

[11] 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*, 2013, pp. 2657–2666.

[12] E. Jardine, "Tor, what is it good for? political repression and the use of online anonymity-granting technologies," *New media & society*, vol. 20, no. 2, pp. 435–452, 2018.

[13] M. T. Khan, J. DeBlasio, G. M. Voelker, A. C. Snoeren, C. Kanich, and N. Vallina-Rodriguez, "An empirical analysis of the commercial vpn ecosystem," in *Proceedings of the Internet Measurement Conference 2018*, 2018, pp. 443–456.

[14] Tor, *Who uses tor?* (Accessed on 04/19/2021). [Online]. Available: https://2019.www.torproject.org/about/torusers.html.en.

[15] A. A. Niaki, S. Cho, Z. Weinberg, N. P. Hoang, A. Razaghpanah, N. Christin, and P. Gill, "Iclab: A global, longitudinal internet censorship measurement platform," in *2020 IEEE Symposium on Security and Privacy (SP)*, IEEE, 2020, pp. 135–151.

[16] R. Ensafi, P. Winter, A. Mueen, and J. R. Crandall, "Analyzing the great firewall of china over space and time," *Proceedings on privacy enhancing technologies*, vol. 2015, no. 1, pp. 61–76, 2015.

[17] T. K. Yadav, A. Sinha, D. Gosain, P. K. Sharma, and S. Chakravarty, "Where the light gets in: Analyzing web censorship mechanisms in india," in *Proceedings of the Internet Measurement Conference 2018*, 2018, pp. 252–264.

[18] Anonymous, "Towards a comprehensive picture of the great firewall's dns censorship," in *4th USENIX Workshop on Free and Open Communications on the Internet (FOCI 14)*, 2014.

[19] J. R. Crandall, D. Zinn, M. Byrd, E. T. Barr, and R. East, "Conceptdoppler: A weather tracker for internet censorship.," in *ACM Conference on Computer and Communications Security*, 2007, pp. 352–365.

[20] O. Farnan, A. Darer, and J. Wright, "Poisoning the well: Exploring the great firewall's poisoned dns responses," in *Proceedings of the 2016 ACM on Workshop on Privacy in the Electronic Society*, 2016, pp. 95–98.

[21] X. Xu, Z. M. Mao, and J. A. Halderman, "Internet censorship in china: Where does the filtering occur?" In *International Conference on Passive and Active Network Measurement*, Springer, 2011, pp. 133–142.

[22] A. Davidson, I. Goldberg, N. Sullivan, G. Tankersley, and F. Valsorda, "Privacy pass: Bypassing internet challenges anonymously.," *PoPETs*, vol. 2018, no. 3, pp. 164–180, 2018.

[23] S. Khattak, D. Fifield, S. Afroz, M. Javed, S. Sundaresan, V. Paxson, S. J. Murdoch, and D. McCoy, "Do you see what i see? differential treatment of anonymous users," Internet Society, 2016.

[24] 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)*, 2017, pp. 325–341.

[25] Mullvad, *Mullvad vpn - privacy is a universal right*, (Accessed on 06/25/2021). [Online]. Available: https://mullvad.net/en/.

[26] V. C. Perta, M. V. Barbera, G. Tyson, H. Haddadi, and A. Mei, "A glance through the vpn looking glass: Ipv6 leakage and dns hijacking in commercial vpn clients," *Proceedings on Privacy Enhancing Technologies*, vol. 2015, no. 1, pp. 77–91, 2015.

[27] S. Li Cain and A. Hardy, *Mullvad vpn review 2021*, (Accessed on 06/25/2021). [Online]. Available: https://www.forbes.com/advisor/business/software/mullvad-vpn-review/.

[28] R. Rimkienė, *Mullvad vpn review*, (Accessed on 06/25/2021). [Online]. Available: https://cybernews.com/best-vpn/mullvad-vpn-review/.

[29] A. Pingle and S. Roos, "Measuring accessibility of popular websites while using tor," 2021.

[30] W. Tutuarima and S. Roos, "Measuring accessibility of popular websites when using protonvpn," 2021.

[31] R. Ramesh, R. S. Raman, M. Bernhard, V. Ongkowijaya, L. Evdokimov, A. Edmundson, S. Sprecher, M. Ikram, and R. Ensafi, "Decentralized control: A case study of russia," in *Network and Distributed Systems Security (NDSS) Symposium 2020*, 2020.

[32] B. Jones, T.-W. Lee, N. Feamster, and P. Gill, "Automated detection and fingerprinting of censorship block pages," in *Proceedings of the 2014 Conference on Internet Measurement Conference*, 2014, pp. 299–304.

[33] A. McDonald, M. Bernhard, L. Valenta, B. VanderSloot, W. Scott, N. Sullivan, J. A. Halderman, and R. Ensafi, "403 forbidden: A global view of cdn geoblocking," in *Proceedings of the Internet Measurement Conference 2018*, 2018, pp. 218–230.

[34] Amazon, *Alexa - top sites*, (Accessed on 04/22/2021). [Online]. Available: https://www.alexa.com/topsites.

[35] J. Mulder and S. Roos, "Measuring the blocking of an.on users by popular websites through web scraping," 2021.

[36] M. C. Tschantz, S. Afroz, S. Sajid, S. A. Qazi, M. Javed, and V. Paxson, "A bestiary of blocking: The motivations and modes behind website unavailability," in *8th USENIX Workshop on Free and Open Communications on the Internet (FOCI 18)*, 2018.

[37] R. Sundara Raman, P. Shenoy, K. Kohls, and R. Ensafi, "Censored planet: An internet-wide, longitudinal censorship observatory," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020, pp. 49–66.

[38] Q. Scheitle, O. Hohlfeld, J. Gamba, J. Jelten, T. Zimmermann, S. D. Strowes, and N. Vallina-Rodriguez, "A long way to the top: Significance, structure, and stability of internet top lists," in *Proceedings of the Internet Measurement Conference 2018*, 2018, pp. 478–493.

[39] E. Kirda, "Getting under alexa's umbrella: Infiltration attacks against internet top domain lists," in *Information Security: 22nd International Conference, ISC 2019, New York City, NY, USA, September 16–18, 2019, Proceedings*, Springer Nature, vol. 11723, 2019, p. 255.

[40] W. Rweyemamu, T. Lauinger, C. Wilson, W. Robertson, and E. Kirda, "Clustering and the weekend effect: Recommendations for the use of top domain lists in security research," in *International Conference on Passive and Active Network Measurement*, Springer, 2019, pp. 161–177.

[41] I. P. Iacoban and S. Roos, "Measuring accessibility of popular websites while using the i2p anonymity network," 2021.

[42] D. Zeber, S. Bird, C. Oliveira, W. Rudametkin, I. Segall, F. Wollsén, and M. Lopatka, "The representativeness of automated web crawls as a surrogate for human browsing," in *Proceedings of The Web Conference 2020*, 2020, pp. 167–178.

[43] Google, *Puppeteer*, (Accessed on 05/27/2021). [Online]. Available: https://github.com/puppeteer/puppeteer.

[44] ——, *Chrome devtools protocol*, (Accessed on 05/27/2021). [Online]. Available: https://chromedevtools.github.io/devtools-protocol/.

[45] E. Persson, *Evaluating tools and techniques for web scraping*, (Accessed on 05/27/2021), 2019. [Online]. Available: https://www.diva-portal.org/smash/record.jsf?pid=diva2:1415998.

[46] M. Pennisi, *A day at the races: Avoiding random failures in selenium ui tests*, (Accessed on 06/25/2021). [Online]. Available: https://bocoup.com/blog/a-day-at-the-races.

[47] S. Wiefling, N. Gruschka, and L. L. Iacono, "Even turing should sometimes not be able to tell: Mimicking humanoid usage behavior for exploratory studies of online services," in *Nordic Conference on Secure IT Systems*, Springer, 2019, pp. 188–203.

[48] Microsoft, *Typescript*, (Accessed on 05/27/2021). [Online]. Available: https://www.typescriptlang.org.

[49] *[bug] typescript issues with puppeteer v7 ("no exported member" error, etc) #428*, (Accessed on 05/27/2021). [Online]. Available: https://github.com/berstend/puppeteer-extra/issues/428.

[50] *Unhandledpromiserejectionwarning: Error: Request is already handled! #5334*, (Accessed on 05/27/2021). [Online]. Available: https://github.com/puppeteer/puppeteer/issues/5334.

[51] *Puppeteer-extra*, (Accessed on 05/27/2021). [Online]. Available: https://github.com/berstend/puppeteer-extra/tree/master/packages/puppeteer-extra.

[52] *Puppeteer-extra-plugin-stealth*, (Accessed on 05/27/2021). [Online]. Available: https://github.com/berstend/puppeteer-extra/tree/master/packages/puppeteer-extra-plugin-stealth.

[53] *Puppeteer-extra-plugin-adblocker*, (Accessed on 05/27/2021). [Online]. Available: https://github.com/berstend/puppeteer-extra/tree/master/packages/puppeteer-extra-plugin-adblocker.

[54] D. Kladnik, *I don't care about cookies*, (Accessed on 05/27/2021). [Online]. Available: https://www.i-dont-care-about-cookies.eu.

[55] *Adblock*, (Accessed on 05/27/2021). [Online]. Available: https://getadblock.com.

[56] *Acceptable ads*, (Accessed on 05/27/2021). [Online]. Available: https://acceptableads.com.

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

[58] B. Yang, F. Gu, and X. Niu, "Block mean value based image perceptual hashing," in *2006 International Conference on Intelligent Information Hiding and Multimedia*, IEEE, 2006, pp. 167–172.

[59] V. Zakharov, A. Kirikova, V. Munerman, and T. Samoilova, "Architecture of software-hardware complex for searching images in database," in *2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, IEEE, 2019, pp. 1735–1739.

[60] *Imagehash 4.2.0*, (Accessed on 06/03/2021). [Online]. Available: https://pypi.org/project/ImageHash/.

[61] McAfee, *Customer url ticketing system*, (Accessed on 06/14/2021). [Online]. Available: https://www.trustedsource.org.

[62] Google, *Introduction to robots.txt*, (Accessed on 06/05/2021). [Online]. Available: https://developers.google.com/search/docs/advanced/robots/intro.

[63] M. Schellekens, "Robot.txt: Balancing interests of content producers and content users," *Bridging Distances in Technology and Regulation*, p. 173, 2013.

[64] M. Thelwall and D. Stuart, "Web crawling ethics revisited: Cost, privacy, and denial of service," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 13, pp. 1771–1779, 2006.

[65] Z. Gold and M. Latonero, "Robots welcome: Ethical and legal considerations for web crawling and scraping," *Wash. JL Tech. & Arts*, vol. 13, p. 275, 2017.

[66] M. Koster, G. Illyes, H. Zeller, and L. Harvey, "Robots Exclusion Protocol," Internet Engineering Task Force, Internet-Draft draft-koster-rep-05, Jun. 2021, Work in Progress, 10 pp. [Online]. Available: https://datatracker.ietf.org/doc/html/draft-koster-rep-05.

[67] Mullvad, *Pricing*, (Accessed on 06/05/2021). [Online]. Available: https://mullvad.net/en/pricing/.

[68] F. Biazin do Nascimento, *Cse3000-research-project*, (Accessed on 06/05/2021). [Online]. Available: https://github.com/francinebiazin/CSE3000-research-project.

[69] N. Sullivan, *Cloudflare supports privacy pass*, (Accessed on 06/06/2021), Nov. 2017. [Online]. Available: https://blog.cloudflare.com/cloudflare-supports-privacy-pass/.

[70] A. Davidson, *Supporting the latest version of the privacy pass protocol*, (Accessed on 06/06/2021), Oct. 2019. [Online]. Available: https://blog.cloudflare.com/supporting-the-latest-version-of-the-privacy-pass-protocol/.

[71] R. Agrawal, L. Bu, E. Del Rosario, and M. A. Kinsy, "Design-flow methodology for secure group anonymous authentication," in *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, IEEE, 2020, pp. 1544–1549.

[72] H. M. Moghaddam and A. Mosenia, "Anonymizing masses: Practical light-weight anonymity at the network level," *arXiv preprint arXiv:1911.09642*, 2019.

[73] Y. Nakatsuka, E. Ozturk, A. Paverd, and G. Tsudik, "Cacti: Captcha avoidance via client-side tee integration," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021.

A Appendices

A.1 Crawler Design State Diagram

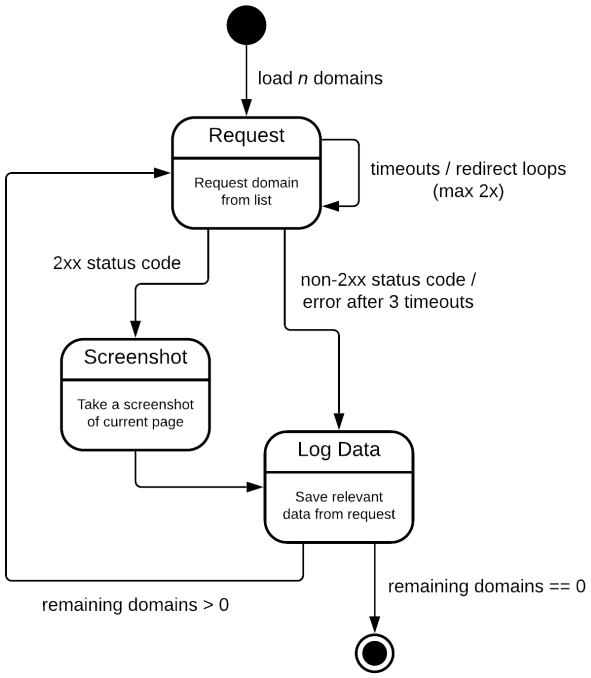

Figure 2: *State diagram illustrating the logic of the crawler.*

A.2 Block Classification Flow Diagram

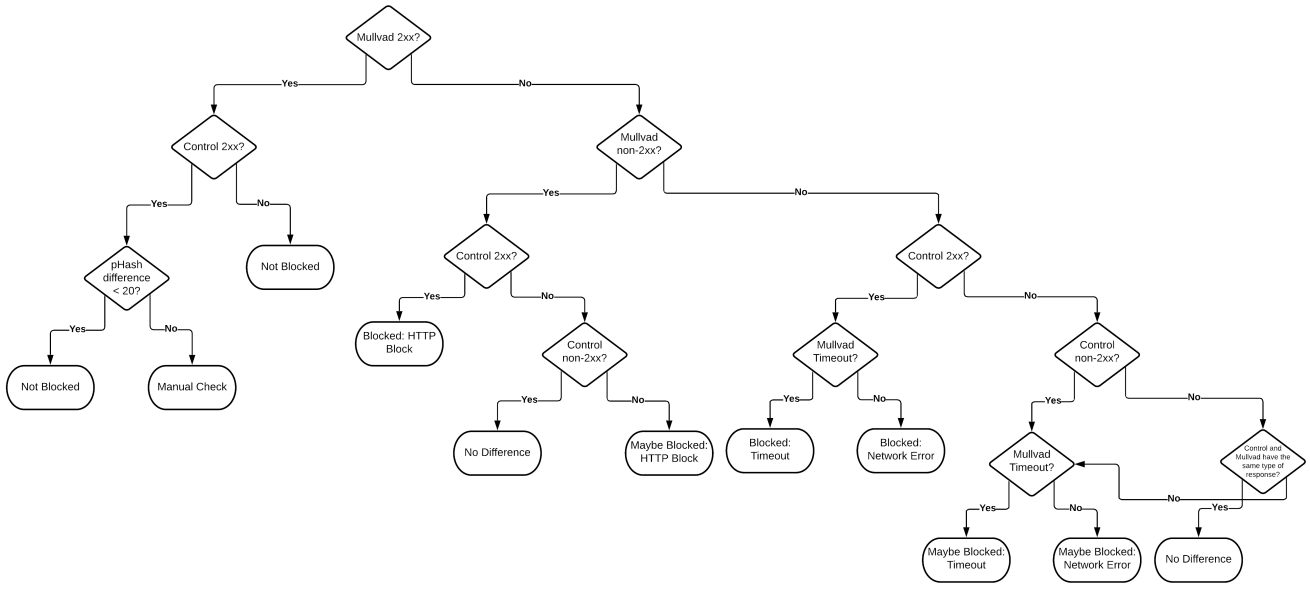

Figure 3: *Flow diagram illustrating the block classification process based on the responses obtained from both the Mullvad VPN and the control connections.*

A.3 Subpage Block Classification Flow Diagram

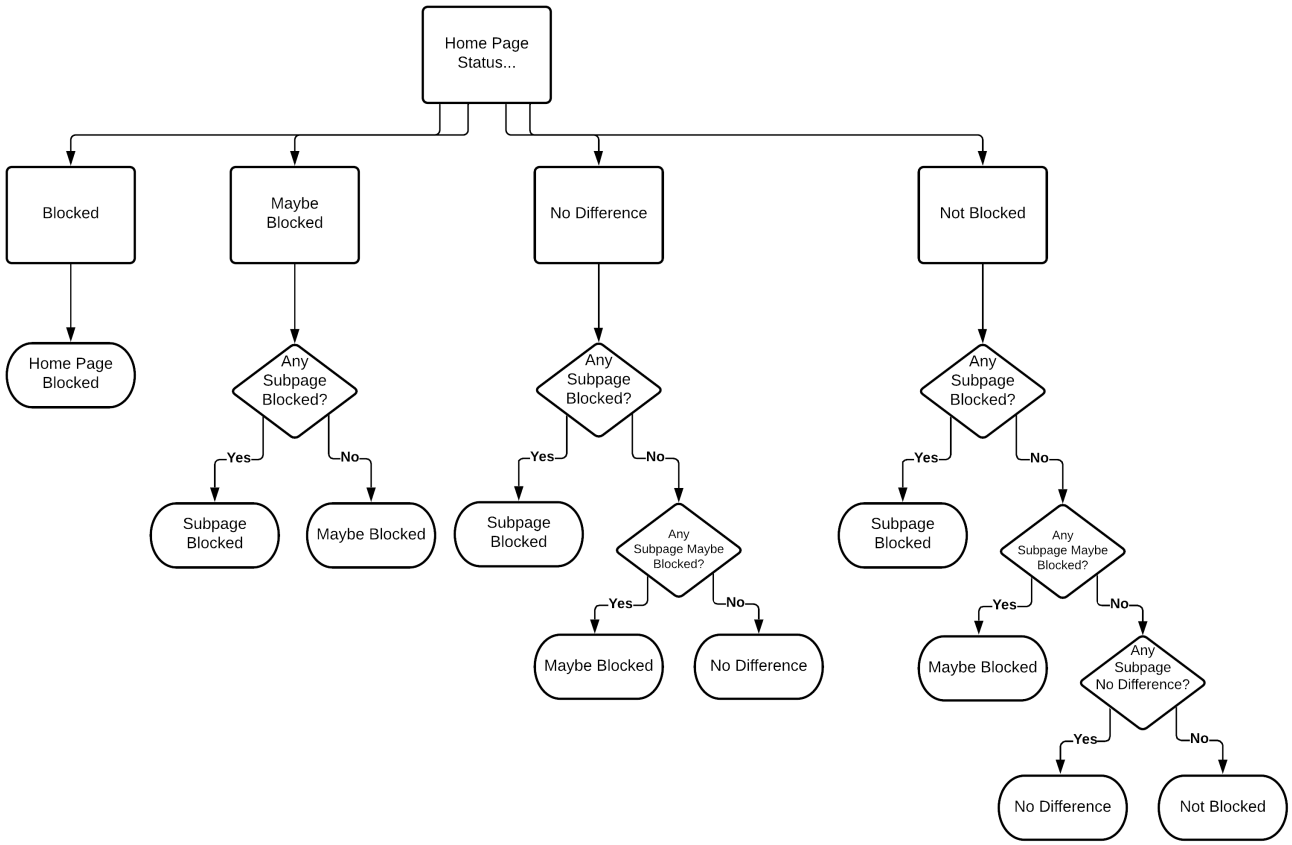

Figure 4: *Flow diagram illustrating the subpage block classification process based on the responses obtained from the general request block classification.*

A.4 Examples of Content Blocking

Figure 5: *Differentiated content (broken HTML) served by* `gimy.app` *when connecting from Mullvad VPN, 24 May 2021.*

Figure 6: *Challenge-response test (CAPTCHA) served by* `anjuke.com` *when connecting from Mullvad VPN, 25 May 2021.*

Figure 7: *Block page served by* `cdiscount.com` *when connecting from Mullvad VPN, 22 May 2021.*

A.5 Stage 4: Summary of Results

Date	Total Requests	Status Code in `200` Range		Status Code Outside `200` Range		Timeouts		Network Errors	
		Mullvad VPN	Control	Mullvad VPN	Control	Mullvad VPN	Control	Mullvad VPN	Control
2021-6-8	3000	2960	2944	24	28	5	14	11	14
2021-6-9	3000	2960	2951	28	23	6	13	6	13
2021-6-11	3000	2955	2953	28	24	8	15	9	8
2021-6-12	3000	2954	2955	31	23	7	14	8	8
2021-6-13	3000	2956	2944	34	25	4	9	6	22
Averages (%)		98.57%	98.31%	0.97%	0.82%	0.20%	0.43%	0.27%	0.43%

Table 4: *Summary of data gathered during Stage 4 (requesting two subpages from each of 1,000 domains).*

Date	Total Requests	Not Blocked	Blocked	Maybe Blocked	No Difference	Types of Blocks					
						HTTP	Timeout	Network Error	Differentiated Content	Block Page	Challenge-Response Test
2021-6-8	3000	2953	20	0	27	5	2	6	5	2	0
2021-6-9	3000	2951	16	0	33	6	1	0	7	2	0
2021-6-11	3000	2946	23	0	31	5	4	5	7	2	0
2021-6-12	3000	2945	26	0	29	9	4	4	7	2	0
2021-6-13	3000	2944	27	0	29	12	1	2	7	4	1
Averages (%)		98.26%	0.75%	0.00%	0.99%	0.25%	0.08%	0.11%	0.22%	0.08%	0.01%

Table 5: *Summary of blocks identified during Stage 4 (requesting two subpages from each of 1,000 domains). The figures categorised into different types of blocks all come from the Blocked column.*

A.6 Stage 4: Block Ratio per Category

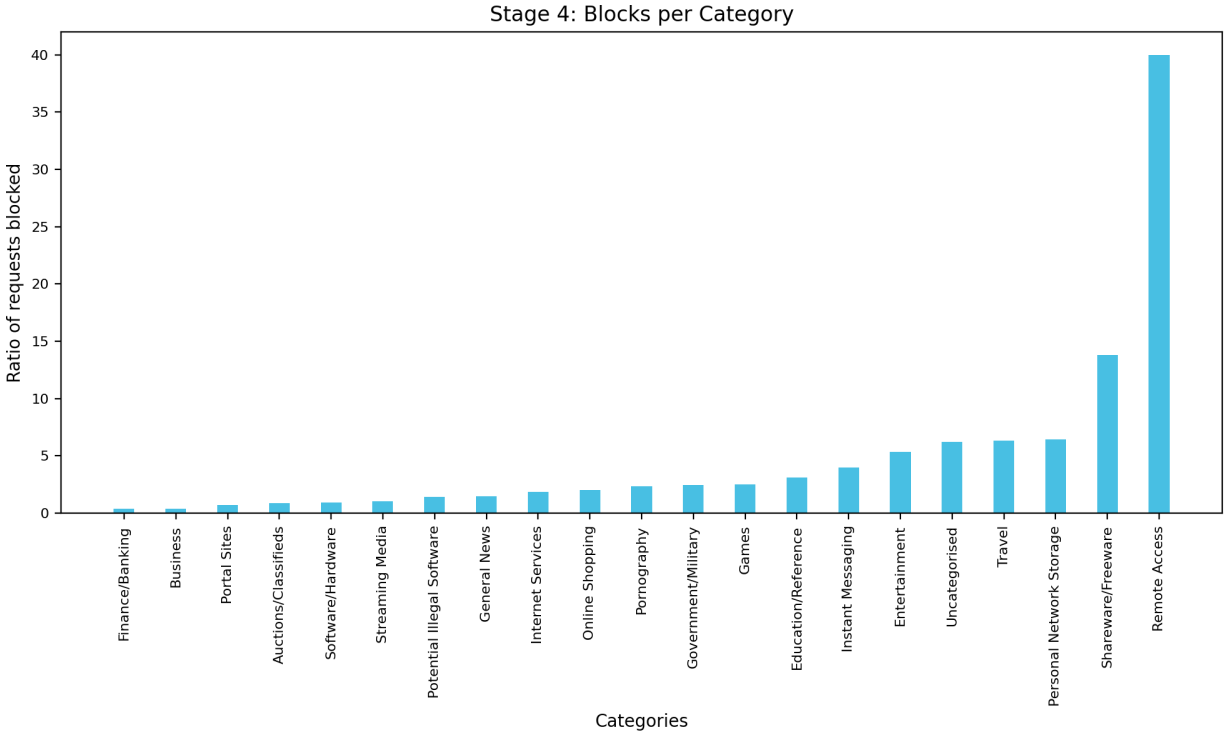

Figure 8: *Graph illustrating the ratio of blocked domains identified per category of website during Stage 4 (requesting two subpages from each of 1,000 domains).*

A.7 Stage 3: Categories

Category	Blocked	Other
Anonymizers	0	15
Anonymizing Utilities	0	15
Art/Culture/Heritage	0	10
Auctions/Classifieds	1	244
Blogs/Wiki	6	449
Business	10	1000
Chat	0	20
Consumer Protection	0	5
Content Server	4	131
Dating/Personals	0	25
Education/Reference	6	684
Entertainment	9	576
Fashion/Beauty	0	185
Finance/Banking	19	896
Forum/Bulletin Boards	2	103
Gambling	1	74
Gambling Related	0	30
Game/Cartoon Violence	0	5
Games	1	414
General News	12	1588
Government/Military	6	354
Health	2	108
Humor/Comics	0	5
Information Security	0	10
Instant Messaging	1	39
Interactive Web Applications	0	230
Internet Radio/TV	4	71
Internet Services	10	975
Job Search	2	143
Major Global Religions	0	20
Malicious Sites	0	30
Marketing/Merchandising	6	474
Media Downloads	0	65
Media Sharing	0	70
Messaging	0	5
Mobile Phone	0	35
Motor Vehicles	1	74
Non-Profit/Advocacy/NGO	0	35
Online Shopping	16	1039
P2P/File Sharing	0	15
Parked Domain	0	75
Personal Network Storage	1	89
Personal Pages	0	20
Pharmacy	0	15
Politics/Opinion	0	20
Pornography	0	325
Portal Sites	17	508
Potential Illegal Software	0	275
Professional Networking	0	25
Public Information	6	174
PUPs (potentially unwanted programs)	0	50
Real Estate	7	108
Recreation/Hobbies	0	40
Religion/Ideologies	0	10
Remote Access	0	5
Resource Sharing	1	9
Restaurants	5	40
School Cheating Information	0	5
Search Engines	0	375
Shareware/Freeware	6	84
Social Networking	1	149
Software/Hardware	4	606
Spam URLs	0	5
Sports	2	328
Stock Trading	2	88
Streaming Media	1	174
Technical Information	0	105
Technical/Business Forums	2	283
Text Translators	0	10
Travel	6	159
Uncategorised	0	295
Visual Search Engine	0	10
Web Ads	1	24
Web Mail	0	25
Web Meetings	0	35
Web Phone	0	5
Total	181	14,819

Table 6: *Summary of blocks identified per category of website during Stage 3 (requesting home pages only). Requests classified as 'Blocked' or 'Maybe Blocked' are counted in the Blocked column, and all others ('Not Blocked' and 'No Difference') are counted in the Other column.*

A.8 Stage 4: Categories

Category	Blocked	Other
Anonymizing Utilities	0	5
Auctions/Classifieds	1	115
Blogs/Wiki	0	135
Business	1	259
Chat	0	5
Consumer Protection	0	5
Content Server	0	5
Dating/Personals	0	5
Education/Reference	6	185
Entertainment	10	176
Fashion/Beauty	0	35
Finance/Banking	1	264
Forum/Bulletin Boards	0	40
Gambling	0	15
Gambling Related	0	10
Games	3	115
General News	9	593
Government/Military	2	79
Health	0	40
Humor/Comics	0	5
Information Security	0	5
Instant Messaging	1	24
Interactive Web Applications	0	111
Internet Radio/TV	0	20
Internet Services	5	267
Job Search	0	35
Major Global Religions	0	10
Marketing/Merchandising	0	114
Media Downloads	0	20
Media Sharing	0	45
Messaging	0	5
Mobile Phone	0	5
Motor Vehicles	0	10
Online Shopping	8	385
P2P/File Sharing	0	10
Parked Domain	0	20
Personal Network Storage	2	29
Personal Pages	0	10
Pharmacy	0	5
Politics/Opinion	0	5
Pornography	2	83
Portal Sites	2	279
Potential Illegal Software	1	70
Professional Networking	0	15
Public Information	0	61
PUPs (potentially unwanted programs)	0	5
Real Estate	0	55
Recreation/Hobbies	0	10
Remote Access	2	3
Restaurants	0	5
Search Engines	0	244
Shareware/Freeware	4	25
Social Networking	0	105
Software/Hardware	2	213
Sports	0	144
Stock Trading	0	35
Streaming Media	1	94
Technical Information	0	45
Technical/Business Forums	0	85
Text Translators	0	5
Travel	4	59
Uncategorised	2	30
Visual Search Engine	0	10
Web Ads	0	5
Web Mail	0	10
Web Meetings	0	10
Total	69	4,931

Table 7: *Summary of blocks identified per category of domains during Stage 4 (requesting two subpages from each of 1,000 domains). Requests classified as 'Home Page Blocked' and 'Subpage Blocked' are counted in the Blocked column, and all others ('Not Blocked' and 'No Difference') are counted in the Other column.*