403 Forbidden: A Global View of CDN Geoblocking

Allison McDonald
University of Michigan
amcdon@umich.edu

Matthew Bernhard
University of Michigan
matber@umich.edu

Luke Valenta
University of Pennsylvania*
lukev@seas.upenn.edu

Benjamin VanderSloot
University of Michigan
benvds@umich.edu

Will Scott
University of Michigan
willrs@umich.edu

Nick Sullivan
Cloudflare
nick@cloudflare.com

J. Alex Halderman
University of Michigan
jhalderm@umich.edu

Roya Ensafi
University of Michigan
ensafi@umich.edu

ABSTRACT

We report the first wide-scale measurement study of server-side geographic restriction, or geoblocking, a phenomenon in which server operators intentionally deny access to users from particular countries or regions. Many sites practice geoblocking due to legal requirements or other business reasons, but excessive blocking can needlessly deny valuable content and services to entire national populations.

To help researchers and policymakers understand this phenomenon, we develop a semi-automated system to detect instances where whole websites were rendered inaccessible due to geoblocking. By focusing on detecting geoblocking capabilities offered by large CDNs and cloud providers, we can reliably distinguish the practice from dynamic anti-abuse mechanisms and network-based censorship. We apply our techniques to test for geoblocking across the Alexa Top 10K sites from thousands of vantage points in 177 countries. We then expand our measurement to a sample of CDN customers in the Alexa Top 1M.

We find that geoblocking occurs across a broad set of countries and sites. We observe geoblocking in nearly all countries we study, with Iran, Syria, Sudan, Cuba, and Russia experiencing the highest rates. These countries experience particularly high rates of geoblocking for finance and banking sites, likely as a result of U.S. economic sanctions. We also verify our measurements with data provided by Cloudflare, and find our observations to be accurate.

ACM Reference Format:
Allison McDonald, Matthew Bernhard, Luke Valenta, Benjamin VanderSloot, Will Scott, Nick Sullivan, J. Alex Halderman, and Roya Ensafi. 2018. 403 Forbidden: A Global View of CDN Geoblocking. In *2018 Internet Measurement Conference (IMC '18), October 31-November 2, 2018, Boston, MA, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3278532.3278552

*Luke Valenta performed this work while at Cloudflare.

1 INTRODUCTION

Researchers have devoted significant effort to measuring and circumventing nation-state Internet censorship (e.g., [23, 39, 49]). However, censorship is not the only reason why online content may be unavailable in particular countries. Service operators and publishers sometimes deny access themselves, server-side, to clients from certain locations. This style of geographic restriction, termed *geoblocking* [45], may be applied to comply with international regulations, local legal requirements, or licensing restrictions, to enforce market segmentation, or to prevent abuse.

Geoblocking has drawn increasing scrutiny from policymakers. A 2013 study by the Australian parliament concluded that geoblocking forces Australians to pay higher prices and should be regulated [6], and in 2017, the European Union banned some forms of geoblocking in order to foster a single European market [16]. Moreover, some Internet freedom advocates argue that the harm posed by geoblocking extends beyond the financial: it contributes to the wider phenomenon of Internet "balkanization" [18], in which users from different regions have access to vastly different online experiences.

Although some instances of geoblocking may be justified, there is abundant anecdotal evidence that overblocking frequently occurs. For example, after the GDPR came into effect in May 2018, several major U.S.-based news sites blocked access from Europe entirely [7]. All sites built on Google App Engine are unavailable in Cuba and Iran, due to Google's interpretation of U.S. regulations [48]. Other companies block all users from regions that produce large volumes of abuse, such as comment spam, when alternative security measures might result in far less collateral damage [44]. We hope that quantifying geoblocking will help reduce such overblocking by highlighting the extent of its impact on users .

In this paper, we report the first global measurement study of website geoblocking. Comprehensively measuring geoblocking is challenging. The phenomenon takes many forms:

sometimes a whole website is blocked in entire countries, while in other instances only particular content items are unavailable. In many cases a site is reachable but refuses to accept payment from or ship goods to users in blocked regions. Services often do not disclose that they practice geoblocking, so the reason content is unavailable must be inferred and distinguished from other anti-abuse practices and from network-based censorship.

To make large-scale measurement tractable, we focus our investigation on one important means by which sites implement geographic restrictions: using features built into large CDNs and cloud providers. Many popular providers, such as Akamai and Cloudflare, allow sites to restrict their availability by country. Using these CDNs, we positively identify specific characteristics associated with services enabling geographic controls, and use those characteristics to identify a larger set of CDNs that enable customers to geoblock. Within a large set of popular sites, this allows us to characterize the types of content that are most likely to be unavailable and the places where unavailability can be attributed to legal requirements.

To measure the extent of CDN-based geoblocking worldwide, we used Luminati [36], a commercial platform that sells access to proxy servers operated by users of the Hola VPN service. To safeguard those users, we refrained from probing sites from high-risk categories as well as sites known to be censored by governments. (We discuss these and other safeguards and ethical considerations in Section 3.3.) We implemented a new probing tool, Lumscan, that greatly improves the reliability of the data.

We collected two principal data sets. First, we developed a semi-automated system for identifying geoblocking enacted through CDNs. We accessed a safe subset of the Alexa Top 10K sites from 177 countries globally, extracted and clustered possible block pages, and manually examined each cluster. From this, we identified 7 CDNs and cloud providers that facilitate widespread geoblocking and extracted signatures to recognize each blocking behavior. Next, we found the customers of these CDNs in the Alexa Top 1M, took a 5% sample of these domains, and tested them globally to find the relative rate of geoblocking of each of these CDNs customer sets.

Our results show that geoblocking is a widespread phenomenon, present in most countries globally. Of the 8,000 Alexa Top 10K domains we tested globally, we observed a median of 3 domains inaccessible due to geoblocking per country, with a maximum of 71 domains blocked in Syria. Of domains in the Alexa Top Million, we observed an overall rate of 4.4% of domains utilizing their CDN's geoblocking feature in at least one country. We observed countries that are currently under sanctions (Iran, Cuba, Syria, and Sudan) to be geoblocked at a significantly higher rate, and Shopping websites to be the most common type of service to geoblock by raw number of domains.

Beyond characterizing CDN-based geoblocking, our results show that server-side blocking can be a significant source of error for censorship measurement—an area of very active research (e.g., [24, 39, 49]). We find that 9% of domains on the Citizen Lab Block List [12], a widely used list of censored domains, returned a CDN block page in at least one country. This indicates that censorship measurement studies should take geoblocking into consideration before ascribing unavailable sites to network-based censorship. We also discuss the relationship between censorship and geoblocking.

Roadmap: Section 2 provides background on geoblocking and our methodology. Section 3 describes exploratory measurements that informed our study design. We report our Alexa Top 10K measurements in Section 4. Section 5 expands our investigation of the CDNs identified in Section 4 into the Alexa Top 1M. Section 6 reveals data provided to us by Cloudflare and validates our previous observations. Our discussion of the relationship between censorship and geoblocking, the role of CDNs, and our limitations can b found in Section 7. Section 8 reviews related work, and we conclude in Section 9.

2 BACKGROUND

2.1 Geoblocking

Websites may choose to restrict access to their content for many reasons. On forums seeking instructions for blocking Internet traffic by country (e.g., [47]), we see motivations that range from complying with legal restrictions, removing access from locations with many malicious login attempts, or simply reducing unwanted traffic.

Legal restrictions are an often cited reason that websites geoblock. U.S. export controls, managed by the Office of Foreign Asset Controls within the Department of the Treasury, limit both physical and intellectual property that U.S.-based entities can transfer to some nationalities without explicit authorization [19]. Similar institutions exist in many countries to enforce international sanctions and tariffs, and many websites may feel compelled to deny access because of embargoes. There can be significant unintended consequences to such broad enforcement of export restrictions.

Geoblocking can present itself in many ways. A website may choose to entirely block the network connection, resulting in timeouts when trying to access the site. Other sites might serve a custom block page that explains why access has been denied. Geoblocking may also happen at the application layer. A user may be able to load the main page of a website, but find that the login button has disappeared, or that some content is not available. These more nuanced

changes in content are of significant interest, but we leave these questions to future work. In this paper, we focus on detecting when websites entirely deny access.

From the user's perspective, whole-site geoblocking will sometimes present itself as an HTTP status code `403`. This status code is defined in RFC 7231 as a tool for a server to inform the requester that it "understood the request but refuses to authorize it" [22]. Because some service denial occurs due to international sanctions, the HTTP status code 451 is also relevant. Defined in RFC 7725, this status code is intended to inform users that their request was denied for legal or policy reasons [8]. However, this status code has not yet seen wide adoption, and we only observed an HTTP 451 twice in the course of our experiments.

Content Delivery Networks (CDNs) will commonly provide customers with security control tools. With these, customers can block IPs or locations based on their own policies or using a reputation score of the request or IP, which, for example, might be calculated based on rate of HTTP requests from a particular client (e.g. DoS protections) [2]. In these cases, a user will receive a block page that has been generated by the CDN rather than the end server.

While private companies are free to restrict access to content as they see fit—and are sometimes legally required to do so—our interest in this phenomenon stems from wondering to what extent geoblocking is contributing to the fragmentation of access to content online, which can be detrimental to the participation of Internet users worldwide in the global online community. We hope that by shedding light on the scope and impact of geoblocking, we can induce companies and policymakers to more fully consider alternative methods for meeting regulatory and security needs that cause less exclusion of users online.

2.2 Methods of Data Collection

Collecting representative measurements of site availability is a long-standing challenge in the field of censorship measurement. We want to both know that our measurement method is diverse enough to capture the phenomenon across a region and, in the case of censorship measurement, that if a user is associated with the device conducting the measurement, they are not put in harm's way because of sensitive domains being requested from their device. Fortunately, this second consideration is one way in which our study differentiates from censorship measurement—we wish to see when servers refuse access to a user, not when a nation-state blocks access. This allows some additional flexibility in measurement technique, as discussed below.

We collected measurements from a range of vantage points in different locations and across types of networks. Our primary vantage points were residential user machines provided through the Luminati proxy service. Luminati leverages the user-base of Hola Unblocker [28] for client-based proxy nodes. Luminati users connect to a superproxy with configuration information, including the desired geographic location of an exit node, whether to use the same exit node for multiple requests, and whether to send traffic via HTTP or HTTPS. Our use of Luminati follows the characterization of the service presented by Chung et al. [11].

For validation process, we also used a set of VPSes in 16 selected countries. We selected 9 servers to span the GDP range of country wealth by selecting every 10th country down a list of relative GDP [31] until it was difficult to find legitimate VPS services. We selected an additional 7 countries based on researcher interest due to known sanctions or reputations for content unavailability. The 16 VPSes were located in Iran, Israel, Turkey, Russia, Cambodia, Switzerland, Austria, Belarus, Latvia, the United States, Canada, Brazil, Nigeria, Egypt, Kenya, and New Zealand.

The VPS providers we used were based on recommendations from local activists. We did not use certain popular VPN/VPS providers because of their malicious marketing strategies: Ikram et al. [30] observed that some VPN providers such as HideMyAss and SecureLine often manipulate their WHOIS records to influence how their vantage points are geolocated by third parties. We verified the location of each VPS by requesting a website we set up on Cloudflare, and examining the geolocational headers Cloudflare inserts. This gives us some confidence that the claimed location of each VPS is likely to match the data CDNs use in making geoblocking determinations.

3 EXPLORATION AND VALIDATION

For our initial exploration of geoblocking, we identified two services that offer geoblocking as a feature, Akamai and Cloudflare. Akamai's Content Delivery Network is one of the world's largest distributed computing platforms, with more than 233,000 servers in over 130 countries, peering with over 1,600 networks around the world. Cloudflare is another global CDN, with numerous connections to Internet exchange points worldwide [14]. These two CDNs combined provide services to more than 25,000 of the Alexa Top Million domains, and in total multiple millions of domains [9]. Since these CDNs make it easy to implement geoblocking, we reasoned that many of their customers likely enable it, making them ideal candidates for our exploratory measurements.

3.1 Identifying and Validating Signals of Geoblocking

We identified a subset of Alexa Top Million that are customers of Akamai and Cloudflare by examining the DNS server used by each domain. While this method only exposes a fraction of Akamai and Cloudflare customers, it gives us a subset of domains we can be confident use Akamai and Cloudflare. In all, we found 2,171 domains using Cloudflare and 4,111 using Akamai.

Fetching each of these domains from a VPS in Iran using curl, we observed 707 HTTP `403 Forbidden` responses, compared to 69 from a U.S.-based control server. Upon inspecting these pages in a browser, we found that both Akamai and Cloudflare present easily recognizable error pages that make them differentiable from other forms of blocking, such as block pages generated by Internet censorship in Iran [5]. It is important to note that the Cloudflare page specifically indicates that it is being served due to geoblocking, but the Akamai page is less specific and also appears when Akamai's abuse detection system is triggered.

To scale up these measurements, we used ZGrab [20] from our set of VPSes. We first validated the behavior of ZGrab by randomly selecting 50 domains and manually observing that the response received when requesting the home page through ZGrab was the same as when the home page was loaded interactively in a web browser set to use our VPS as a proxy. ZGrab was configured with the `User-Agent` string set to mimic Firefox on Mac OS X. We manually checked all responses returning status code 403 observed in Iran, Turkey, Israel, and the U.S. to confirm that the same status was returned when collected in a real web browser. We found that on the order of 30% of the Akamai 403s appeared to be false positives: the crawler request was flagged as a bot or otherwise was denied access while a real web browser request was able to load the page. The set of domains that resulted in false positives from ZGrab were nearly identical across countries, indicating that these false positives are not typically location dependent.

Next, we fetched the 6,282 Cloudflare and Akamai domains in the Alexa Top 1M from each VPS (100,512 domain-country pairs) and detected 1,068 domain-country pairs that resulted in block pages (19 for Cloudflare, 1,049 for Akamai). Once again, we manually verified each block page instance by visiting it in a web browser tunneled through the VPS. Of the 1,068 instances, 782 appeared to be genuine instances of geoblocking (19 for Cloudflare, 763 for Akamai), with 269 unique domains in total (12 for Cloudflare, 257 for Akamai).

Of the 1,068 instances of likely geoblocking across all domain and country pairs initially reported by our automated data classifier, 286 (27%) proved to be false positives upon manual inspection—all from Akamai.

3.2 Lumscan: Luminati Scanning Tool

The Luminati service is a collection of HTTP proxies that exit traffic at residential machines, enabling us to make requests from an end-users' vantage points within each country. However, Luminati is not perfect; since the residential IPs are VPN users, interference by the local network can be expected on those clients' requests. In order to overcome the challenges associated with getting raw data from an end-user connection, we developed a tool, Lumscan, to perform our measurements with a number of features to improve the reliability of our results.

The first improvement Lumscan performs is to verify connection to a known online page, in order to verify that the client has local web connectivity. We connect to a Luminati-controlled webpage that also returns the client's IP and geolocation information. Second, Lumscan repeats each failed request a configurable number of times. This reduces the impact of proxies on unreliable networks.

Lumscan also allows the user to specify HTTP headers that are sent in the final request. Merely setting `User-Agent` is insufficient to suppress bot detection, which likely contributed to the high false positive rate in our VPS study.

Finally, in order to support a high rate of requests, we implemented load balancing. We distribute requests across residential exit machines as well as across the Luminati superproxies that mediate our requests. We only perform 10 requests with a given exit machine before changing exit machine. This keeps us from consuming too many resources on any single end user's machine. It also allowed us to collect each of these datasets in a matter of hours rather than days, providing a single snapshot in time and minimizing the chance of observing policy changes.

3.3 Ethics

Our initial investigation of geoblocking used 16 vantage points in different geographic regions, all located in commercial hosting facilities. In all cases, the account for the server used an author's real name and university email address, and we complied with the terms of service and acceptable use policies of the hosting companies. This allowed us to probe availability of a broad range of content without imposing risks on end users.

Our broader data collection from residential IPs made use of the Luminati VPN service [36]. Luminati allows paid traffic to exit the computers of end-users who have installed their free VPN service. Luminati advertises itself as allowing competitive market research, but the company was supportive of our research during multiple Skype conversations.

In order to reduce the risk that our research would not negatively impact the Luminati end users, we carefully limited

the set of sites that we probed. We removed several categories of sites: pornography, weapons, spam, and malicious content, as well as any sites that were uncategorized. We also removed domains that had been identified as censored by Citizen Lab [12].

We did not collect any personally identifiable information about Luminati users. Each probe result contained the geolocation information provided by Luminati and the HTTP and HTML data returned by the web request. Although Luminati returned IP addresses of the exit nodes within the geolocation information, we discarded these before doing any analysis on the data. As such, our IRB determined that the study was outside its regulatory purview.

4 ALEXA TOP 10K

Our early exploration gave us an insight into how two specific CDNs, Akamai and Cloudflare, allow their customers to geoblock. We want to now expand our understanding of the phenomenon across additional CDNs, as well as investigate whether we can observe other instances of geoblocking, potentially implemented in other ways. To do this, we explore geoblocking across the Alexa Top 10K most popular domains across 177 countries.

4.1 Methods

Our data was collected using our Lumscan tool with the Luminati Network. We extract possible block pages from our dataset, cluster them, and find new block pages. We then search the dataset for instances of these block pages and sample the domains again in countries where we saw them, in order to increase confidence in our observation. This methodology is described in more detail in this section, and a summary can be found in Table 1.

4.1.1 Initial Dataset. We choose to study the Alexa Top 10K domains as a set of popular websites with a global reach. Because we are requesting these domains from end-user devices, we first classify the 10,000 domains using FortiGuard and remove any dangerous or sensitive categories, such as Pornography, Weapons, and Spam. We also remove any domains that appear in any of the Citizen Lab censorship list for any country. This leaves us with 8,003 domains.

We began by sampling from 195 countries and kept countries that were able to respond to all of our requests. Each domain is sampled 3 times as a baseline measurement. 177 countries were able to respond to all our requests.

Overall, we observe 286 domains that never successfully respond to our request. Luminati itself blocks requests to some domains, which can be identified with the header `X-Luminati-Error`. These account for 13 of the inaccessible domains. The rest of the requests consistently timed out or tried to make more than our limit of 10 redirects. 90% of the domains we sampled saw less than a 11.7% error rate, where error indicates that we were unable to get a response from the site, either due to proxy errors or errors such as timeouts and lengthy redirect chains.

The errors are also not inordinately affecting only certain countries. From our initial 3 samples per country-domain, we have at least one valid response from between 89.2% and 93.9% of tested domains in each country. The one exception to this, Comoros, sees a response rate of 76.4%. This shows that an initial snapshot of 3 samples per country-domain pair gives us excellent coverage of domains in nearly every country.

4.1.2 Metrics for Identifying Outliers. From these samples, we want to extract pages that are likely block pages. Guided by the work of Jones, et al., we first explored whether page length is a good metric for finding outlier pages [32]. For each domain, we extract the longest observed instance of the page across countries and note that length as the likely size of the true page. We then compare the size of each individual sample for a domain to the representative size. If the length difference is greater than 30%, we extract this page as a possible block page for clustering.

However, we found that we were collecting enough data that potential block pages were too many to be clustered efficiently. In an early exploratory experiment, we had taken our list of Akamai and Cloudflare domains used in our VPS study to sample each domain 10 times in every country and ranked the countries by number of Akamai and Cloudflare block pages seen. With information from this data, we take the top 20 countries with the most block pages and find the representative sizes in our Alexa 10K data among those countries. We then extract the pages whose length is 30% or more shorter than our representative length for that domain. We find 24,381 samples are outliers, or 5.1% of our initial set.

4.1.3 Clustering & Identifying Page Signatures. After extracting the set of potential block pages, we cluster the HTML documents using single-link hierarchical clustering, which does not require that we know the number of clusters beforehand. We use term frequency-inverse document frequency with 1- and 2-grams to generate feature vectors using scikit-learn, a machine learning library in Python [42]. This resulted in 119 clusters, which we examined by hand and used to extract all pages that were potential signals of geoblocking. The CDNs that we identified were Akamai, Cloudflare, Amazon CloudFront, SOASTA, Incapsula, and Baidu. We also identified Google AppEngine, a hosting service, serving block pages. Our clustering method also identified three CAPTCHA services, namely Cloudflare, Baidu, and Distil Networks. We also identified the Cloudflare JavaScript challenge page. We chose also to include a fingerprint for the nginx `403 Forbidden` page and the Varnish `403 Forbidden`

Initial Domains	Safe Domains	Initial Samples	Clustered Pages	Clusters	Discovered CDNs and Hosting Providers
10,000	8,003	1,416,531	24,381	119	7

Table 1: Overview of data at each step in Methods. This table shows the data at each step of our geoblock page discovery process. "Initial Samples" consists of the 3 samples per domain in each of the 177 countries we examine.

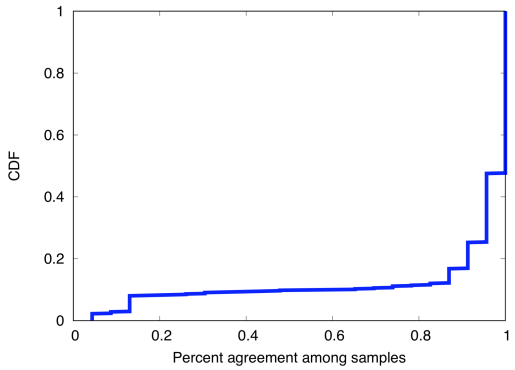

Figure 1: Consistency for various sample rates. This CDF shows the consistency of geoblocking for different sample rates for domain-country pairs were we expect to see a geoblock page. (see Section 4.2). A sample size of 20, which we use to confirm geoblocking, yielded only 3.9% of domain-country pairs with less than an 80% geoblocking rate.

page. Finally, a large cluster represented Airbnb, which states on its block page that it does not serve its website to users in Crimea, Iran, Syria, and North Korea. As an obvious example of geoblocking, we included this block page to see whether their stated blocking practices aligned with what we observed.

We identify 5 pages that are explicitly geoblocking: Cloudflare, Amazon Cloudfront, Baidu, Google AppEngine, and Airbnb.

4.1.4 Resampling Block Pages. Finally, we identify all instances of the above block pages in our entire dataset. We took the country-domain pairs where we saw at least one instance of an explicit block page and sampled them 100 additional times, in order to explore how consistently we observe the geoblock page with different size samples. From the population of 100 measurements per country-domain, we take 500 samples of each size and find how many returned the block page. The results of this experiment can be seen in Figure 1.

Therefore, in every country in which we observe any of our block pages, we sample that domain 20 times in order to gain a higher confidence that the signal was correct. We then set a threshold of 80% agreement for the domains we consider geoblocked.

Table 2: Recall for block pages and other content for Top 10K sites. Here are our recall rates for the 30% difference in length metric.

	Recalled	Actual	Recall
Akamai	1446	3313	43.7%
Cloudflare	406	433	93.8%
AppEngine	381	499	76.4%
Cloudflare Captcha	1181	1264	93.4%
Cloudflare JavaScript	664	1001	66.3%
Amazon CloudFront	36	95	37.9%
Baidu Captcha	128	139	92.1%
Baidu	3	3	100.0%
Incapsula	362	710	51.0%
Soasta	36	36	100.0%
Airbnb	49	49	100.0%
Distil Captcha	315	1028	30.6%
nginx	1524	2656	57.4%
Varnish	22	22	100.0%
Total	6553	11248	58.3%

4.1.5 Evaluating Metrics. After conducting measurements, we evaluate some of the heuristics we chose.

Page Length Heuristic. After extracting a set of 14 block pages from our clusters (see Section 4.1.3), we returned to this metric to evaluate its effectiveness. We found that the overall recall was only 58.3%. This metric was far more accurate for some block pages than others; the relative recalls are listed in Table 2. We also examine the difference in size between each sample and the length we had compared it against to find the difference, as shown in Figure 2. This shows that selection of length cutoff is relatively arbitrary between 5% and 50%—both will yield around 20% false negative across the whole dataset.

We chose percentages of page length following the methodology of Jones, but we note that other experimentation showed that using raw length differences is not as effective. Using percentages normalizes the lengths of pages, while raw length differences excessively penalize long pages. The purpose of

this metric is as a rough heuristic, so while it is fortunate it is effective, it is not critical that it be extremely discerning.

Initial Sample Size. With the dataset in hand, we look at the probability of seeing a block page with only three initial samples per country. To do this, we take the set of explicit geoblocking domain and country pairs in order to measure how likely it is that we would *not* see the block page with different sample sizes. Because we expect the block page to be served every time, we are measuring the rate of other failures, for example proxy errors, transient network failures, and local filtering like a corporate firewall.

We sampled each of these domain-country pairs 100 times. From each set of samples, we then selected 500 random combinations of different sample sizes to detect how many combinations would not yield a block page. For a sample size of 3, only 1.7% of our country-domain pairs did not yield at least one block page. The relationship between sample rate and false negatives can be seen in Figure 3.

4.2 Results

Overall we observe 596 instances of geoblocking by 100 unique domains in 165 countries. We were served explicit geoblock pages from Cloudflare, Baidu, Amazon Cloud Front, Google App Engine, and Airbnb. We also detected several other kinds of content, including Captchas and nginx error pages, but we restrict our analysis only to pages that explicitly signal that they are blocking due to geolocation.

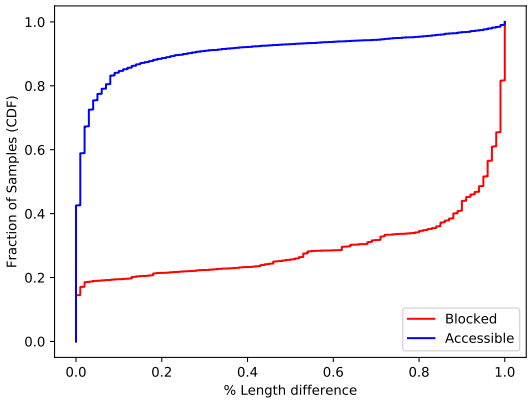

Figure 2: Relative sizes of block pages and representative pages. After selecting the longest observed instance of each domain across the top 20 geoblocking countries, we compare this page length with each sample and plot the length difference. The blocked pages are the samples that match one of our block page fingerprints.

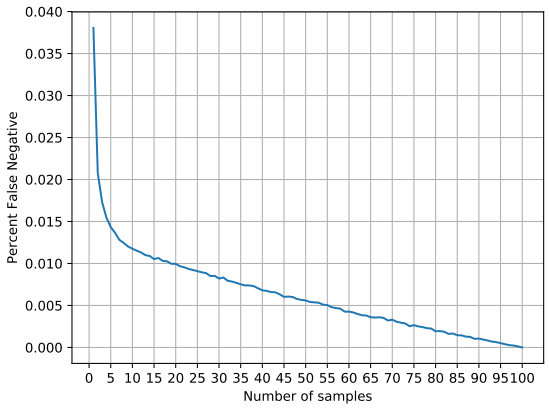

Figure 3: False negative rate for known geoblockers. This graph shows the rate of false negatives, where we see no instance of the block page, for different sampling rates of known geoblocking domain and country pairs.

Table 3: Most geoblocked categories by CDN. We show the top 10 cateogories of the geoblocked domains by CDN.

	Cloudflare	AppEngine	CloudFront	Total
Shopping	18	10	0	28
Business	9	3	1	13
Techonology	0	9	0	9
News/Media	3	6	0	9
Advertising	1	7	0	8
Job Search	4	0	0	4
Newsgroups	0	4	0	4
Sports	1	2	0	3
Education	1	1	0	2
Entertainment	1	1	0	2
Other	5	1	4	10
Total	43	44	5	92

Even for the domains that explicitly signal geoblocking, we do not observe the block page in 100% of samples in a country for just under half of all domain-country pairs. Some of these discrepancies can possibly be attributed to local connection interference, which might be observed if the Luminati device is inside of a corporate firewall. One domain, `http://makro.co.za`, returned a block page for each of our 3 initial measurements in 33 countries, but did not display any geoblocking when we sampled that domain again 20 times in each country several days later, suggesting that we may have observed a change in policy away from geoblocking. Other, smaller discrepancies may yet be attributed to geolocational errors. We limit our analysis to those

country-domain pairs that yield a block page at in at least 80% of the total 23 samples, which eliminates 77 instances, or 11.4%, in order to account for some of these transient errors. The distribution of the sample agreement is shown in Figure 4.

Table 4 displays the categories in which we saw geoblocking. Shopping, Travel, and Business all appear at the top of the list, indicating that consumer market segmentation may be a common motivation for geoblocking. We also see many other categories which are more prevalent in blocking, including Advertising and Job Search. Child Education tops the list in terms of fraction of category blocked, but this is a small category of sites that we tested, and only one geoblocks (`pbskids.com`, which as a U.S. site possibly blocks due to federal sanctions).

Table 5 shows the TLDs which geoblock the most. Sites using `.com` geoblock the most by a wide margin, which is likely just a simple reflection of the prevalence of `.com` sites in the Top 10K. Notably, outside of `.net` and `.org`, all other TLDs were country based. Although there were multiple sites with country TLDs that practiced geoblocking, this does not appear to be a major indication of policy within the Top 10K sites.

The most commonly geoblocked countries are also shown in Table 5. The top four countries are Syria, Iran, Sudan, and Cuba, by a wide margin. These are notably all countries sanctioned by the United States. Nigeria, China, and Russia are also more commonly geoblocked as compared to other countries.

4.2.1 Geoblocking by CDNs. The largest set of geoblocking websites are served by content distribution networks, three

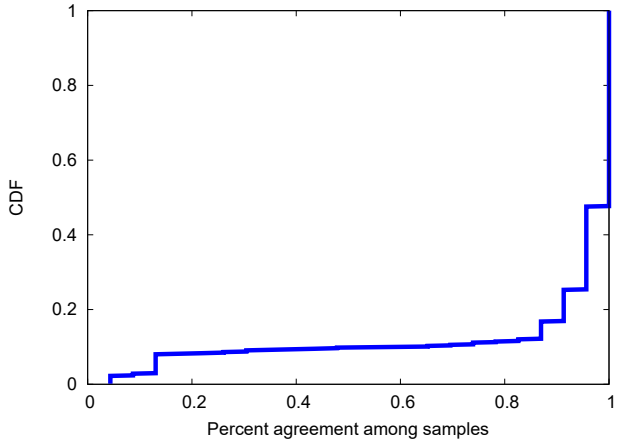

Figure 4: Consistency of geoblocking observations. The CDF of the number of probes of a given site before seeing a non-geoblock page. For the vast majority of sites seen geoblocking, the block page was seen in >80% of probes.

Table 4: Geoblocked sites by category. We show the 20 categories of tested sites in the Alexa 10k Luminati data. "Geoblocked" is the number of unique sites we observed being blocked in at least one country.

Category	Tested	Geoblocked
Child Education	8	1 (12.5%)
Advertising	120	8 (6.7%)
Job Search	97	4 (4.1%)
Shopping	787	29 (3.7%)
Travel	168	6 (3.6%)
Newsgroups and Message Boards	143	4 (2.8%)
Web Hosting	41	1 (2.4%)
Business	758	13 (1.7%)
Sports	179	3 (1.7%)
Personal Vehicles	78	1 (1.3%)
Reference	176	2 (1.1%)
Health and Wellness	92	1 (1.1%)
News and Media	938	9 (1.0%)
Freeware and Software Downloads	115	1 (0.9%)
Information Technology	1,239	9 (0.7%)
Games	348	2 (0.6%)
Entertainment	442	2 (0.5%)
Finance and Banking	454	2 (0.4%)
Education	583	2 (0.3%)
Total	6,766	100 (1.6%)

Table 5: Top TLDs and geoblocked countries for Top 10K sites where we detected geoblocking.

TLD	Count
`.com`	70
`.net`	3
`.org`	3
`.fr`	2
`.it`	2
`.jp`	2
`.in`	2
`.au`	1
`.br`	1
`.sg`	1
Other	13
Total	100

Country	Count
Syria	71
Iran	67
Sudan	66
Cuba	66
China	11
Nigeria	11
Russia	10
Brazil	8
Iraq	6
Pakistan	5
Others	275
Total	596

of which meet our criteria as explicit geoblockers: Cloudflare, Google AppEngine, and Amazon CloudFront. Table 3 shows the categories of geoblocking sites on each CDN. Shopping is the most prevalent on Cloudflare and AppEngine, but AppEngine hosts more geoblocking Information Technology, News, Advertising, and Message Board sites than Cloudflare. We see only one site on CloudFront from the top categories,

Table 6: Geoblocking among Top 10K sites, by country. These countries experienced the most geoblocking.

	Cloudflare	CloudFront	AppEngine	Total
Syria	20	3	44	71
Iran	20	3	37	67
Sudan	20	2	44	66
Cuba	20	2	44	66
China	8	2	0	11
Nigeria	10	1	0	11
Russia	7	3	0	10
Brazil	8	0	0	8
Iraq	5	1	0	6
Pakistan	3	2	0	5
Other	127	148	0	275
Total	248	167	169	596

with the others being dispersed through less common categories.

Table 6 shows the breakdown of countries blocked by each CDN. Google AppEngine explicitly blocks Cuba, Iran, Syria, Sudan, Crimea, and North Korea due to sanctions [25], and we can observe the effects of this. We do not see AppEngine blocking in any other country. We see a similar increase in geoblocking for these countries in Cloudflare and CloudFront's sites, but AppEngine sites block these countries at a much higher rate than the other two. Geoblocking from Cloudflare sites is overall much more visible than the from the other two CDNs.

We use the methods described in Section 5.1 to obtain the number of sites in the Alexa Top 10K using each of these CDNs or hosting providers. We find 1,394 Cloudflare fronted domains, 364 Cloudfront domains, and 108 Google AppEngine domains. Google AppEngine has by far the highest rate of geoblocking, with 40.7% of its customers inaccessible in at least one country. Comparatively, only 3.1% of Cloudflare customers geoblock, and only 1.4% of Amazon Cloudfront customers geoblock in at least one country.

4.2.2 Other observations. While in general we observe geoblocking to be a country-wide phenomenon, we have observed a counterexample to this. The website geniusdisplay.com served an nginx block page for most of our measurements in Russia, but we received some AppEngine block pages for it (few enough that it did not meet our threshold value for considering the site geoblocked). Upon manual inspection, we noticed that we only received the AppEngine page when attempting to access the site from IPs in Crimea, suggesting that at least Google AppEngine is displaying geoblocking at a finer granularity than country-wide.

We observed two other explicit geoblocking pages that we do not include above. We observed one website (fasttech.com) in China that served a Baidu blockpage, which is nearly identical Cloudflare blockpage in content. We also observed 347 instances of geoblocking from Airbnb on various geographic TLDs, exclusively blocking Iran and Syria.

In addition to explicit geoblock pages, we also observed several other block pages that were either not geoblocking but may contribute to the overall discrimination against certain countries, e.g. captchas, or ambiguous pages for which we cannot confidently say whether blocking is based on geolocation. This set of 200,417 observations includes captchas from Cloudflare, Baidu, Distil, a JavaScript challenge page from Cloudflare, and ambiguous block pages from SOASTA, Akamai, and Incapsula.

5 ALEXA TOP 1M

In this section we expand our results from Section 4 to look at the prevalence of geoblocking of five services in the Alexa Top 1M: the CDNs Cloudflare, Amazon Cloudfront, Akamai, Incapsula, and the hosting provider Google AppEngine.

5.1 Methods

5.1.1 Identifying CDN Population. In order to find the rate of geoblocking in the Top 1M by CDN or hosting provider that we identified in the previous section, we first needed to find the population of domains using each service from which we could sample.

Several CDNs were simple to identify; when requesting a site fronted by Cloudflare, Amazon Cloudfront, and Incapsula, a special header is appended to the response: `CF-RAY`, `X-Amz-Cf-Id`, and `X-Iinfo`, respectively. We used ZGrab to request all domains in the Top 1M and identified all domains that returned these headers anywhere in the redirect chain. With this method, we found 109,801 Cloudflare, 10,856 Amazon Cloudfront, and 5,570 Incapsula domains in the Alexa Top 1M.

To identify Akamai domains, we send a Pragma header [1] to all domains in the Alexa Top 1M, which triggers the Akamai edge server to insert cache-related headers into the response. If we saw these Akamai cache headers anywhere in the redirect response chain, we considered the domain to be fronted by Akamai, because at some point during the request there would be an opportunity for Akamai to block the request. We discovered 10,727 Akamai domains in the Top 1M.

Finally, we consider Google AppEngine. According to Google forums [26], Google AppEngine traffic will stem from IPs that are discoverable by doing a recursive lookup on `_cloud-netblocks.googleusercontent.com`. Using this method, we found 65 IP blocks and 16,455 domains in the Top 1M hosted on AppEngine.

In total, we found 152,001 unique domains in the Alexa Top 1M that use one of these services. 1,408 domains showed signs of using two services. For example, `zales.com` contained both the Incapsula and Akamai headers. Because these domains have the potential to be blocked by either service, we consider them to be customers of both.

5.1.2 Sampling. We categorized these domains using FortiGuard. We then excluded the same risky categories as before: categories relating to pornography, violence, drugs, malware, dating, censorship circumvention, and meaningless or unknown categories. We also eliminated any domains found in the Citizen Lab Block List. This left us with 123,614 domains. Finally, we took a 5% random sample of these domains to create a test list of 6,180 domains.

Using the same method as in Section 4.1, we first sample each domain 3 times in each country. For our explicit geoblockers (Cloudflare, Amazon Cloudfront, and Google AppEngine), for each country-domain pair where we see at least one block page, we sample again 20 times. For our non-explicit geoblockers (Akamai and Incapsula), for every domain where we see a block page in any country, we sample the domain again 20 times in every country.

5.1.3 Dataset. Of the 6,180 domains we sampled, 26 never successfully responded to our requests. We saw 3 domains indicating that Luminati would not complete the request via the `X-Luminati-Error` header. This is only 0.05% of our sample, compared to 0.2% of Alexa Top 10K domains, indicating that more popular websites are more protected by Luminati. Furthermore, 90% of the domains we investigate saw an error rate of 3.0% or less, where error here indicates that we were unable to get a response from the site, either due to proxy errors or errors such as timeouts and lengthy redirect chains. This is markedly lower than our Alexa Top 10K study.

5.2 Results

We find that geoblocking in the Alexa Top 1M follows similar patterns to those in the Alexa Top 10K, as we will report in this section.

5.2.1 Explicit Geoblockers. / Looking first at the providers that explicitly geoblock (specifically Cloudflare, Amazon Cloudfront, and Google AppEngine), we see 1,565 instance of geoblocking across 176 countries—all countries except Seychelles. This accounts for 238 unique domains in 176 countries. The most geoblocked countries are shown in Table 7. The median number of sites blocked in each country is 4, indicating that most countries have at least a few domains preventing access by their residents.

Table 7: Geoblocking among Top 1M sites, by country. These countries experienced the most geoblocking.

	Cloudflare	CloudFront	AppEngine	Total
Iran	64	7	107	178
Sudan	55	2	112	169
Syria	55	3	110	168
Cuba	50	3	112	165
China	24	10	0	34
Russia	18	10	0	28
Ukraine	18	4	0	22
Nigeria	12	5	0	17
Brazil	12	5	0	17
Romania	13	3	0	16
Other	527	224	0	751
Total	848	276	441	1,565

In the Alexa Top 1M, AppEngine remains the provider with the most geoblocked domains; of the 667 Google AppEngine domains in our 5% sample of Top 1M CDN sites, 112 domains displayed the geoblock page, or 16.8%. All but 5 domains were blocked in each of Syria, Sudan, Iran, and Cuba; 5 domains were censored in Iran, preventing us from measuring geoblocking, and 2 were censored in Syria. This is much lower than the rate of 40.7% of AppEngine domains in the Top 10K that showed signs of geoblocking. Amazon Cloudfront had 16 of its 512 domains practicing geoblocking, a rate of 3.1%, which is slightly higher than we observed in the Top 10K. Finally, Cloudflare saw geoblocking on 110 of 4,283 Cloudflare domains at a rate of 2.6%, which is comparable to what we observed in the previous study.

As can be seen in Table 7, Iran, Syria, Sudan, and Cuba are the countries experiencing the most geoblocking in this dataset by raw number of inaccessible domains. We also see other large countries such as Russia and China appearing in the top ten.

We can also see that Google AppEngine only geoblocks in Iran, Syria, Sudan, and Cuba, which is consistent with the list of countries Google claims to block. North Korea had no Luminati hosts for us to probe from, and we may miss Crimea due to exploring geoblocking at a country granularity rather than regionally. This is one way in which our study may be expanded.

The distribution of domains that geoblock in at least one country across categories can be seen in Table 8. By raw numbers, Shopping is still the most geoblocked category, followed by Business, Information Technology, Personal Vehicles, and News and Media. By ratio of tested domains in each category, Personal Vehicles and Shopping each show that at least 10% of domains in that category practice geoblocking in at least one country, along with Auctions, which

Table 8: Geoblocked sites by top category. We show the top 15 of 25 geoblocked categories of *explicit* geoblocking sites by number of domains in the Alexa 1M Luminati data. "Geoblocked" is the number of unique sites we observed being blocked in at least one country.

Category	Tested	Geoblocked
Shopping	418	59 (14.1%)
Business	1,176	51 (4.3%)
Information Technology	1,016	34 (3.3%)
Personal Vehicles	79	16 (20.3%)
News and Media	345	12 (3.5%)
Society and Lifestyle	148	7 (4.7%)
Health and Wellness	146	5 (3.4%)
Travel	153	5 (3.3%)
Personal Websites and Blogs	176	4 (2.3%)
Education	239	4 (1.7%)
Games	206	4 (1.9%)
Sports	121	4 (3.3%)
Reference	81	4 (4.9%)
Job Search	42	4 (9.5%)
Finance and Banking	108	4 (3.7%)
Other	1,008	21 (2.1%)
Total	5,462	238 (4.4%)

is not in the top 15. This is a significant number of domains in each of these categories that are potentially inaccessible to users.

5.2.2 Non-Explicit Geoblockers. Akamai and Incapsula are also CDNs that offer their customers the opportunity to geoblock. However, both services display the same block page for other errors, making it more difficult to distinguish geoblocking from bot detection or other server errors. Because we do not have multiple hosts in each country with which we can manually check whether a domain is blocked, it is not possible for us to say with complete confidence which domains are geoblocking. However, here we can reason about what metrics possibly indicate geoblocking for these CDNs.

Intuitively, we are looking for domains that consistently send a block page in some countries and consistently do not in others. One metric we look at here is thus the consistency of block page within country. For all domains where we see an Akamai or an Incapsula block page, we consider any country receiving the block page at least 80% of the time to be consistent; each domain then has an overall consistency score that is the percent of countries that are consistent for each domain. For some domain where only two countries are blocked 100% of the time and the rest of the countries never see a block page, this would be a consistency score of 100%. Alternatively, if a domain had three countries each seeing 90% of samples returning a block page and one domain with 20% block pages, it would have a consistency score of 75%.

Applying this metric to our explicit geoblockers, we see that they each only have a consistency rate of 100% about 85% of the time. For Akamai and Incapsula, the rate is much lower; they have a 100% consistency rate only 13.9% and 15.9%, respectively. This is a good verification that Akamai and Incapsula are noisy block pages. Therefore, in order to be conservative, we will discuss only those domains that do not show a block page in all countries and that have 100% consistency.

We find 201 instances of geoblocking with Akamai and 200 instances with Incapsula. This encompasses only 14 of 101 domains that returned the block page at least once for Akamai and 17 of 107 domains for Incapsula. Both sets of domains see China, Russia, Cuba, Iran, Syria, and Sudan as the most blocked countries, indicating that we are indeed isolating geoblocking in these domains, if not exhaustively across all Akamai and Incapsula domains.

6 CLOUDFLARE VALIDATION

Cloudflare provided us with data that confirmed our measurements and gave further insight into the practice of geoblocking.

Cloudflare offers customers the ability to set specific access rules for their domains via the Firewall Access Rules feature [15]. These rules allow customers to whitelist, challenge, or block visitors based on IP address, country, or AS number. These rules give the site owner more fine-grained control over the visitors that Cloudflare allows to access their site. For instance, a website that is under attack might enable rules to present CAPTCHAs to visitors to cut down on bot traffic, or a retail site that only ships to certain countries may wish to block visitors based on geolocation.

The ability to block visitors by country is reserved for Cloudflare's Enterprise customers. Free-, Pro-, and Business-level customers still have the ability to present challenges by countries, but a human visitor from those countries could still access the site by completing a challenge. However, due to a regression, the country-blocking feature was enabled for customers of all tiers from April to August 2018. Cloudflare was able to provide us with a July 2018 snapshot of all active country-scoped rules set by their customers, which falls during the regression period. Each rule in the dataset includes the rule action (block, whitelist, challenge, js_challenge), the target country, the number of affected zones, the zone customer tier, and the rule activation date. Cloudflare zones are roughly defined as a domain and all its subdomains. We publish aggregates of the data to avoid revealing any individual customer information.

Table 9: Most geoblocked countries by Cloudflare customers, by account type. These countries experienced the highest rates of geoblocking by Cloudflare customers. "Baseline" gives the percentage of zones for each account type that have geoblocking enabled against any country.

Country	All	Enterprise	Business	Pro	Free
Baseline	1.93%	37.07%	2.69%	2.56%	1.72%
Russia	0.22%	4.90%	1.14%	0.44%	0.19%
China	0.22%	3.11%	1.16%	0.46%	0.20%
North Korea	0.20%	16.50%	0.38%	0.17%	0.10%
Iran	0.18%	15.57%	0.39%	0.13%	0.09%
Ukraine	0.18%	3.89%	0.71%	0.38%	0.15%
Romania	0.14%	3.63%	0.49%	0.24%	0.12%
India	0.14%	4.18%	0.48%	0.23%	0.11%
Brazil	0.13%	3.87%	0.43%	0.16%	0.11%
Vietnam	0.13%	3.08%	0.33%	0.16%	0.11%
Czech Rep.	0.11%	3.66%	0.40%	0.15%	0.09%
Indonesia	0.11%	2.24%	0.39%	0.12%	0.10%
Iraq	0.10%	3.99%	0.32%	0.09%	0.08%
Croatia	0.10%	3.44%	0.24%	0.13%	0.08%
Syria	0.10%	13.74%	0.17%	0.06%	0.02%
Estonia	0.10%	3.28%	0.32%	0.14%	0.08%
Sudan	0.10%	13.57%	0.12%	0.04%	0.02%

The data displayed in Table 9 shows that the scale of geoblocking we observed for Cloudflare was roughly accurate. Because North Korea had no Luminati vantage points, we were unable to measure how extensively it was blocked. This is one observation we gain from the Cloudflare data that was not visible in our measurements: North Korea is the third most geoblocked country, and the most blocked country of enterprise customers. This is particularly telling of the motivations of companies to geoblock. North Korea is under U.S. sanctions, but likely poses little to no other risk to companies because of the country's relatively low access to the Internet and virtual absence from international commerce, indicating that compliance with sanctions alone is a primary driving force of geoblocking for larger customers.

As the customers of Cloudflare with Business, Pro, and Free accounts were unable to use geoblocking features until April 2018, we can see in Table 9 that a significant number of accounts activated geoblocking in the last 3 months, especially considering that there are far more domains on Business, Pro, and Free accounts than for Enterprise accounts. This suggests that where the functionality is available, many websites will opt to use the feature. Additionally, we see that the free tier customers block China and Russia at a higher rate than other countries, including countries under sanctions, suggesting that the motivation for blocking these sets

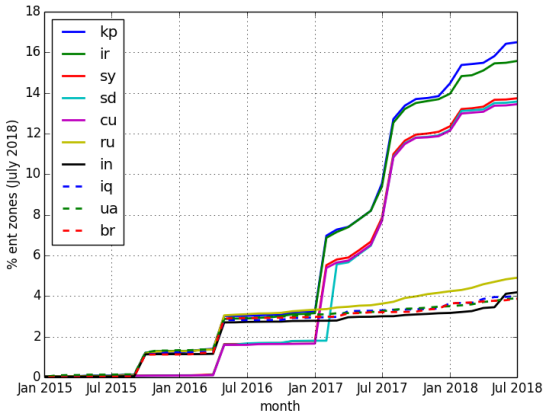

Figure 5: Cloudflare Enterprise customers' activation of geoblocking over time, by geoblocked country. This graph shows the activation dates of blocking rules over time for the Enterprise customers (ent) who had country-scoped geoblocking rules active in July 2018.

of countries might be different for sites without an enterprise-level contract with Cloudflare. Notably, Iran and North Korea experience significantly lower rates of geoblocking relative to other geoblocked countries for business, pro, and free accounts.

In Figure 5, we see the accumulation of blocking rules over time. Notably, North Korea, Iran, Syria, Sudan, and Cuba follow the same pattern, indicating that customers who activate blocking rules tend to treat these countries similarly—although notably not exactly the same, as Iran and North Korea experience more geoblocking than the other three.

7 DISCUSSION

In this section, we will discuss the insights we gained from this study and the remaining challenges and opportunities in studying geoblocking.

7.1 Geoblocking and Censorship

We want to consider how censorship and geoblocking may be related. First, we consider how censorship measurement may have been impacted by the presence of geoblocking. Beyond our active measurements, we looked for evidence of the block pages we identified appearing in censorship measurement datasets. We turn to the existing OONI measurement corpus [23]. OONI measurements are user submitted reports from around the world, created with a provided software client. The software draws URLs to test from the Citizen Lab list [12], a widely used list of censored domains, and

records the full body and headers of the response. All together, domains have been tested 87 million times by OONI clients.

We find 8,313 cases in 139 countries where OONI responses match the explicit signals of geoblocking we describe in Section 4. Instances occur in all of the 12 countries where OONI identifies state censorship. More importantly, these cases occur at least once for 97 domains, or 9%, of the global test list, indicating that geoblocking could be a significant confounding factor in censorship measurement.

This is also only a conservative estimate of how often server blocking is experienced in OONI measurements. The OONI methodology compares client measurements against a control, but that control is often made over Tor, which is also subject to blocking [43]. Saved reports only include the status and headers of the control measurement that is used for comparison, and not the actual contents of that request. As such, it is difficult for us to retroactively understand if a request is made at a time when a site was truly unavailable, or whether the control measurement was also blocked. This is a sizable effect. For example, there were 36,028 OONI measurements to sites using Akamai and Cloudflare infrastructure where the control measurement returned a `403` status code, compared to 14,380 requests where the local measurement was blocked but the control succeeded. The majority of these block pages in the OONI database, more than 30,000 in all, are correlated to blocking of the control request rather than the in-country probe data.

Conversely, it is certainly possible that censorship may disguise itself as geoblocking. A censor could inject or redirect to a mimicked geoblock page as a way to censor content without taking responsibility. To our knowledge, we did not observe any instances of this. Furthermore, we believe the incentive for companies to enable full-page geoblocking to be misaligned with traditional censorship, in which a government dictates what content should not be available to its own constituents. If a government were to demand a website enable geoblocking for its own IP space, the website would have little reason to comply. Large companies with multiple domains may need to cooperate with a censorious country, but we have seen in practice that this cooperation more frequently appears as selective access and content filtering rather than full denial of access [21, 38].

7.2 CDNs Enable Geographic Discrimination

Content Distribution Networks offer a valuable service for websites by decreasing latency to their users worldwide while providing a baseline level of security and protection that is otherwise nearly impossible to implement at the hosting server. With increase in value added by CDNs coupled with falling costs (Cloudflare, for example, offers basic services for free), more websites are opting into the service. While this makes security features accessible to far more website than previously, this trend increases centralization and enables more sites to use CDNs to control what content is served to which users on the Internet. CDNs are also incentivized to implement tools that add value to their big ticket customers, and they may choose to expose this functionality to most or all of their other users. This gives even the smallest websites the ability to enact fine-grained control over the ways in which their content is served, and to whom.

This is exactly what occurred with Cloudflare between April and August 2018. The ability for Business, Pro, and Free customers to use the geoblocking feature was enabled for 4 months, where it had previously only been available to Enterprise customers. By July 2018, when Cloudflare provided us a snapshot of data about blocking rules, customers in different tiers had already extensively utilized the geoblock feature. Although Cloudflare reverted to its former access model, in which only Enterprise customers could geoblock, we were provided with a valuable insight into what unrestricted geoblocking might look like. In the end, website operators seem to be fairly liberal in activating features that harshly restrict access to content. Customer access to these tools should be limited.

7.3 Limitations & Future Work

Positive identification of geographic blocking provides an exciting first step in studying this phenomenon, but also opens many more doors for future work. While we believe that this approach gives us insight into the most extreme form of geoblocking, namely the total denial of access, our method may miss custom geoblock pages that are not significantly different from the typical page. We also observed consistent timeouts for certain websites in only some countries; an exploration of whether timeouts are another method of geoblocking would be useful, although much more difficult to differentiate from censorship. Additionally, for instances in which a block page is also used for other access control like bot detection, our technique does not provide access to verify our observation through an interactive browser.

In this vein, additional work could be done to simulate a real browser in the automated requests from VPSes. As mentioned in Section 3, early experiments show that adding a full set of headers on ZGrab can reduce the rate of false positives significantly. While this does not eliminate the need for manual validation for non-explicit geoblockers, it does reduce the amount of manual work required to gain enough confidence in the data to automate the classification process.

Finally, discrimination that is not explicitly communicated to users is also important and much harder to measure. Prices

are often different when a site is viewed from different locations, or some features may be removed. We do not capture these effects in our current blockpage-based discrimination measurements, and further work into automatically detecting geographic differences in functionality or access is vital to understanding geographic discrimination.

8 RELATED WORK

The Internet is becoming increasingly regionalized due to sanctions, financial regulations, copyright and licensing rights, perceived abuse, or a perceived lack of customers. This issue is known to policy makers. The EU Parliament recently adopted a regulation to ban geoblocking for most types of online content to give users access to goods and services at the same terms, all over the EU [3]. The majority of the news has been either on geoblocking of multimedia products or geolocation-based price discrimination. We lack a global perspective on the extent of this phenomenon.

"Supply-side" censorship was noted as an important component of the censorship landscape in the initial announcement of the Open Net Initiative [39] but has remained relatively uncharacterized. Rather, the vast majority of studies focused on understanding and circumventing nation-state censorship, specifically in China [13, 27, 49], Iran [5], Pakistan [34, 37], and Syria [10]. These studies often illuminate a wide variety of censorship mechanisms such as country-wide Internet outages [17], the injection of fake DNS replies [4, 35], the blocking of TCP/IP connections [41], and HTTP-level blocking [18, 32, 40]. Relevant to our work, Jones et al. designed an automated way of detecting censorship block pages, which inform the user that an access to the web page is unsuccessful. Their fingerprinting technique uses page length and frequency vectors of words as features. In our study, we show that these features are not sufficient for detecting blocking by service providers.

This category of measurement studies, including ours, face a major hurdle: obtaining vantage points in target countries. There are research systems for this purpose, though they are often limited to network diagnostic tests. For example, RIPE Atlas has more than 10,000 vantage points but does not allow HTTP requests. ICLab provides 1,000 vantage points from popular VPN providers, but the VPN providers have their own customized policies and malicious marketing behaviors. A wide-ranging tool for censorship detection is provided by OONI [23]. OONI runs an ongoing set of censorship measurement tests from volunteer's devices and doesn't provide a platform for exploratory experiment design in order to mitigate risk to participants.

For gaining a representative residential platform, the Luminati platform has proven to be useful. The downside to Luminati is the cost of running measurements from their network. Chung et al. used Luminati to analyze End-to-End Violations in the Internet [11]. Huang used it to detect HTTP middleboxes [29].

Directly related to our work is a study by Khattak et al., which systematically enumerates and characterizes the blocking of Tor users by service providers [33]. The authors explored how much of the differential treatment received by users of the service was due to an explicit decision to block Tor versus the consequence of "fate sharing"—being blocked because of abuse. In work conducted concurrently with our study, Tschantz et al. explored the space of blocking and argued that different forms of blocking, including geoblocking, warrant more research [46]. Our study attempts to understand the role played by private companies in controlling access to different contents from different locations.

9 CONCLUSION

In this paper we have presented the first wide-scale measurement study of the extent of website geoblocking. We found that geoblocking is occurring in a broad number of countries and that many CDN customers utilize the geoblocking services they provide. Furthermore, across the Alexa Top 10K websites, we are able to observe a wide variety of block pages using a semi-automated technique, which helped us discover new CDNs and services that enable geoblocking. We have further explored the extent to which this form of content discrimination can affect censorship measurement, and find that a significant portion of a major list of censored domains contains domains that we have observed to practice geoblocking. While geoblocking is a diverse phenomenon with many different instantiations, we believe that this first study has shown that the phenomenon is both significant and empirically tractable.

10 ACKNOWLEDGMENTS

The authors thank Florian Schaub, Michael Hornstein, Nina Taft, and our anonymous reviewers for their help and constructive feedback. We are also grateful to Cloudflare for providing data. This material is based upon work supported by the National Science Foundation under grants CNS-1409505, CNS-1518888, and CNS-1755841.

REFERENCES

[1] Akamai. Akamai pragma header. https://community.akamai.com/customers/s/article/Akamai-Pragma-Header?language=en_US.
[2] Kona DDoS Defender. https://www.akamai.com/us/en/resources/cdn-ddos.jsp.
[3] amp/aw. EU parliament bans geoblocking, exempts Netflix and other streaming services, Feb 2018. http://www.dw.com/en/eu-parliament-bans-geoblocking-exempts-netflix-and-other-streaming-services/a-42475135.
[4] 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), 2014.

[5] S. Aryan, H. Aryan, and J. A. Halderman. Internet censorship in Iran: A first look. In *3rd USENIX Workshop on Free and Open Communications on the Internet (FOCI)*, 2013.

[6] Austrlian House Standing Committee on Infrastructure and Communications. At what cost? IT pricing and the Australia tax, July 2013. https://www.aph.gov.au/Parliamentary_Business/Committees/House_of_Representatives_Committees?url=ic/itpricing/report.htm.

[7] BBC News. GDPR: US news sites unavailable to EU users under new rules, May 2018. https://www.bbc.com/news/world-europe-44248448.

[8] T. Bray. An HTTP status code to report legal obstacles. RFC 7725, Feb. 2016.

[9] BuiltWith. Akamai usage statistics. https://trends.builtwith.com/cdn/Akamai.

[10] A. Chaabane, T. Chen, M. Cunche, E. D. Cristofaro, A. Friedman, and M. A. Kaafar. Censorship in the wild: Analyzing Internet filtering in Syria. In *14th ACM Internet Measurement Conference (IMC)*, 2014.

[11] T. Chung, D. Choffnes, and A. Mislove. Tunneling for transparency: A large-scale analysis of end-to-end violations in the Internet. In *16th ACM Internet Measurement Conference (IMC)*, 2016.

[12] Citizen Lab. Block test list. https://github.com/citizenlab/test-lists.

[13] R. Clayton, S. J. Murdoch, and R. N. M. Watson. Ignoring the Great Firewall of China. In *Privacy Enhancing Technologies Symposium (PETS)*. Springer, 2006.

[14] Cloudflare. https://www.cloudflare.com.

[15] Cloudflare. How do I control IP access to my site?, Aug. 2018. https://support.cloudflare.com/hc/en-us/articles/217074967-How-do-I-control-IP-access-to-my-site-.

[16] Council of European Union. Council regulation (EU) no A8-0172/2017, 2017.

[17] A. Dainotti, C. Squarcella, E. Aben, K. C. Claffy, M. Chiesa, M. Russo, and A. Pescapé. Analysis of country-wide Internet outages caused by censorship. In *11th ACM Internet Measurement Conference (IMC)*, 2011.

[18] J. Dalek, B. Haselton, H. Noman, A. Senft, M. Crete-Nishihata, P. Gill, and R. J. Deibert. A method for identifying and confirming the use of URL filtering products for censorship. In *14th ACM Internet Measurement Conference (IMC)*, 2014.

[19] U.S. Department of the Treasury. Office of Foreign Assets Control (OFAC). https://www.treasury.gov/about/organizational-structure/offices/Pages/Office-of-Foreign-Assets-Control.aspx.

[20] Z. Durumeric, D. Adrian, A. Mirian, M. Bailey, and J. A. Halderman. A search engine backed by Internet-wide scanning. In *22nd ACM Conference on Computer and Communications Security (CCS)*, pages 542–553, 2015.

[21] Facebook. Content restrictions based on local law, 2017. https://transparency.facebook.com/content-restrictions.

[22] R. Fielding and J. Reschke. Hypertext transfer protocol (HTTP/1.1): Semantics and content. RFC 7231, June 2014.

[23] A. Filastò and J. Appelbaum. OONI: Open Observatory of Network Interference. In *2nd USENIX Workshop on Free and Open Communications on the Internet (FOCI)*, 2012.

[24] Freedom House. Freedom on the net 2016, November 2016.

[25] Google. Countries or regions with restricted access. https://support.google.com/a/answer/2891389.

[26] Google. Google appengine faq. https://cloud.google.com/appengine/kb/#static-ip.

[27] GreatFire.org. https://en.greatfire.org.

[28] Hola unblocker. https://hola.org.

[29] S. Huang, F. Cuadrado, and S. Uhlig. Middleboxes in the Internet: a HTTP perspective. In *Network Traffic Measurement and Analysis Conference (TMA)*, pages 1–9. IEEE, 2017.

[30] M. Ikram, N. Vallina-Rodriguez, S. Seneviratne, M. A. Kaafar, and V. Paxson. An analysis of the privacy and security risks of android vpn permission-enabled apps. In *16th ACM Internet Measurement Conference (IMC)*, pages 349–364, 2016.

[31] International Monetary Fund. World economic outlook database. http://www.imf.org/external/pubs/ft/weo/2014/02/weodata/index.aspx.

[32] B. Jones, T.-W. Lee, N. Feamster, and P. Gill. Automated detection and fingerprinting of censorship block pages. In *14th ACM Internet Measurement Conference (IMC)*, 2014.

[33] S. Khattak, T. Elahi, L. Simon, C. M. Swanson, S. J. Murdoch, and I. Goldberg. SOK: Making sense of censorship resistance systems. *Proceedings on Privacy Enhancing Technologies*, (4):37–61, 2016.

[34] S. Khattak, M. Javed, S. A. Khayam, Z. A. Uzmi, and V. Paxson. A look at the consequences of Internet censorship through an ISP lens. In *14th ACM Internet Measurement Conference (IMC)*, pages 271–284, 2014.

[35] G. Lowe, P. Winters, and M. L. Marcus. The great DNS wall of China. *MS, New York University*, 21, 2007.

[36] Luminati. https://luminati.io.

[37] Z. Nabi. The anatomy of Web censorship in Pakistan. In *1st USENIX Workshop on Free and Open Communications on the Internet (FOCI)*, 2013.

[38] New York Times. Google, seeking a return to china, is said to be building a censored search engine, Aug. 2018. https://www.nytimes.com/2018/08/01/technology/china-google-censored-search-engine.html.

[39] OpenNet Initiative. Survey of government Internet filtering practices indicates increasing internet censorship, May 2007. https://cyber.harvard.edu/newsroom/first_global_filtering_survey_released.

[40] J. C. Park and J. R. Crandall. Empirical study of a national-scale distributed intrusion detection system: Backbone-level filtering of HTML responses in china. In *International Conference on Distributed Computing Systems (ICDCS)*, pages 315–326. IEEE, 2010.

[41] P. Pearce, R. Ensafi, F. Li, N. Feamster, and V. Paxson. Augur: Internet-wide detection of connectivity disruptions. In *38th IEEE Symposium on Security and Privacy*, May 2017.

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[43] 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*, pages 325–341, 2017.

[44] The Spamhaus Project. https://www.spamhaus.org/statistics/countries.

[45] M. Trimble. Geoblocking, technical standards and the law. In *Scholarly Works*, chapter 947. 2016. http://scholars.law.unlv.edu/facpub/947.

[46] 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)*, Baltimore, MD, 2018. USENIX Association.

[47] Z. Wallace. How to block entire countries from accessing your website. SitePoint, Apr. 2015. https://www.sitepoint.com/how-to-block-entire-countries-from-accessing-website/.

[48] M. Xynou, A. Filastò, and S. Basso. Measuring internet censorship in cuba's parknets. https://ooni.torproject.org/post/cuba-internet-censorship-2017/, 2017.

[49] J. Zittrain and B. Edelman. Internet filtering in China. *IEEE Internet Computing*, 7(2):70–77, 2003.