Devashish Gosain*, Mayank Mohindra, and Sambuddho Chakravarty

Too Close for Comfort: Morasses of (Anti-) Censorship in the Era of CDNs

Abstract: Recent research claims that "powerful" nation-states may be hegemonic over significant web traffic of "underserved" nations (*e.g.,* Brazil and India). Such traffic may be surveilled when transiting (or ending in) these powerful nations. On the other hand, content distribution networks (CDNs) are designed to bring web content closer to end-users. Thus it is natural to ask whether CDNs have led to the localization of Internet traffic within the country's boundary, challenging the notion of nation-state hegemony.

Further, such traffic localization may inadvertently enhance a country's ability to coerce content providers to censor (or monitor) access within its boundary. On top of that, the obvious solution, *i.e.,* anti-censorship approaches, may sadly face a new dilemma. Traditional ones, relying on proxies, are easily discoverable. Whereas newer ones (*e.g.,* Decoy Routing, CacheBrowser, Domain Fronting and CovertCast *etc.*) might not work as they require accessing web content hosted outside the censors' boundary. We thus quantitatively analyzed the impact of web content localization on various anti-censorship systems.

Such analysis requires geolocating the websites. Thus we adapted a multilateration method, Constraint Based Geolocation (CBG), with additional heuristics. We call it as *Region Specific CBG (R-CBG)*. In more than 89% cases, R-CBG correctly classifies hosts as inside (or outside) w.r.t. a nation. Our empirical study, involving five countries, shows that the majority (61% − 92%) of popular country-specific websites are hosted within a client's own country. Further, additional heuristics classify the majority of them to be on CDNs.

Keywords: Anti-Censorship, CDN, IP Geolocation

DOI 10.2478/popets-2021-0023
Received 2020-08-31; revised 2020-12-15; accepted 2020-12-16.

1 Introduction

The Internet consists of more than 50,000 ASes that interact with one another through commercial business contracts (as *customers*, *peers*, or *providers* [39, 56]). In such arrangements, customer AS pays to a provider AS for Internet connectivity. Therefore, network researchers view the Internet as having a hierarchy *i.e.,* a few ASes appear in a very large fraction of network paths [22, 41].

These few ASes are headquartered in only a handful of "powerful" countries (*e.g.,* the US) [18]. It is believed that since these countries intercept a large fraction of network paths, they may be hegemonic over traffic originating from "underserved" countries. They likely surveil [33] (or censor [53]) *transitory (or terminal)* flows originating from these "underserved" countries. For instance, previous researchers [33] claimed that 95% of the Internet paths to Alexa top-1k websites originating from "undeserved" countries like India, either *transit or end* in "powerful" countries like US. Such observations hold good only *if* the client's Internet traffic crosses its national border.

However, popular web services rely on CDN infrastructure that ensures necessary redundancy (*e.g.,* Google Global Cache [8]) for providing high availability and performance [84]. Due to proliferation of CDNs, a country's traffic to popular destinations might not exit the country's boundary. This may challenge the claim that powerful nation states have a hegemony over the transitory (or terminal) Internet traffic. But, this may also inadvertently strengthen the ability of dictatorial regimes to coerce content providers for regulating web access within their boundaries [31, 85]. Hence, in this research, we answer questions like — *are popular websites hosted within the same country as that of the client*?

If the websites are hosted within the countries, unhindered web access using anti-censorship systems requires even more attention than earlier. Traditionally, such systems have been relying on publicly accessible proxies. But eventually adversaries discover such proxies and add them to their access blacklist. Recent attempts to disrupt this dynamic include Decoy Routing [44, 48, 80, 81], Cache Browser [43], Domain Fronting [37], CovertCast [60] *etc.* Unlike regular proxies, these

***Corresponding Author: Devashish Gosain:** IIIT Delhi, India E-mail: devashishg@iiitd.ac.in
Mayank Mohindra: IIIT Delhi, India E-mail: mayank15056@iiitd.ac.in
Sambuddho Chakravarty: IIIT Delhi, India E-mail: sambuddho@iiitd.ac.in

approaches rely on web content hosted outside the censor's boundary. Thus, we also answer questions like — *can newer anti-censorship systems seamlessly work even in the presence of CDNs?*

Addressing such concerns requires identifying if popular websites (*e.g.,* country specific Alexa top-1k) are located within a country's geographic boundary. Thus, we began by employing *Constraint Based Geolocation* (CBG) [42], a multilateration technique to geolocate Internet hosts. The process involves estimating the position of a target host using distance measurements from sufficient number of fixed reference points. These distances are estimated from RTTs between reference nodes and target hosts.

However, similar to others [76], we noted gross inaccuracies in CBG's geolocation — about 4000-5000 kms. Such errors often arise when the reference nodes and target hosts are far apart (at times even in different countries) [76]. We thus augment CBG, with additional heuristics, by selecting the reference nodes in the country under consideration and assuming that the target host also lies within the same country. It then predicts whether a host is located inside (or outside) the country, using the proposed heuristics. We call this *Region specific CBG (R-CBG).* For five countries under consideration *viz.,* India (IN), Iran (IR), Saudi Arabia (SA), Brazil (BR) and United States (US), R-CBG achieved more than 89% accuracy (for each) when tested against the ground truth (domiciles of RIPE nodes obtained from RIPE Atlas project [14]). Thereafter, we used R-CBG to geolocate Alexa top-1k websites. R-CBG reports 61% – 92% of country specific Alexa top-1k websites to be located in the same country as of the client. Moreover, we repeated our experiments for five consecutive months by selecting monthly snapshots of Alexa top-1k websites. We observed a similar trend *i.e.,* websites hosted inside (or outside) within the given country remained almost the same. This confirms that *web traffic generated by clients often terminates within their countries' boundaries* and web content is served to clients from caches located within their own country.

This traffic localization may inadvertently impact newer anti-censorship solutions. For instance, recent solutions (like Decoy Routing, CacheBrowser *etc.*) rely on web requests to popular (unfiltered) sites hosted outside the censors' boundary. Our measurements indicate that these solutions may also be severely impacted, as majority of country specific Alexa top-1k sites are hosted within the client's country. Furthermore, our heuristics reveal that, these sites either use anycast IPs or are hosted on non-CDN infrastructure.

Interestingly, we find that RTT by itself can be used to determine if a host is located inside a country or not. For targets outside a country, RTT is often greater than a threshold (and vice versa). We use this threshold to determine the relative position (w.r.t a country) of less popular sites (ranked above Alexa top-1k). The following is the summary for our research efforts and findings:

- We quantify what fraction of country specific Alexa top-1k websites are located within the same country as that of client, for Saudi Arabia (SA), India (IN), Iran (IR), Brazil (BR) and United States (US). This involved using R-CBG, a heuristic driven multilateration technique, that compensates for location estimation errors. In more than 89% of the cases R-CBG accurately judges if a target is hosted in a particular country or not.
- For countries under consideration, R-CBG reports 61% – 92% of country specific Alexa top-1k websites to be located in the same country as of the client. This has two major implications:
  - Challenging the earlier claims [33], that a few "powerful" countries may observe majority of transitory (or terminal) Internet traffic, particularly those that originate from "underserved" countries. Due to the presence of CDN front-ends, traffic originating from countries under test, does not exit their respective boundaries. (**Note:** We acknowledge that powerful countries (where the CDNs are headquartered) might coerce the CDN providers to fetch data of clients (residing in other countries), from globally distributed CDN front-ends. We consider this beyond the scope of this paper.)
  - Hindrance towards adoptions of anti-censorship solutions (*e.g.,* Decoy Routing *etc.*) which rely on popular web content, and require the traffic to leave the censor's boundary.
- For all five countries under consideration, we identify the type of CDN a website is using (anycast [30] or DNS based [78]). We find that a large fraction of country specific Alexa top-1k websites use anycast CDNs (rather than DNS based). *E.g.*, in US, 59% of the Alexa websites use anycast CDNs, 19% use DNS CDN and remaining 22% were hosted on non-CDN infrastructure.
- We observe that, by and large, RTT may itself be sufficient for identifying whether a website is located inside the country or not. Our results reveal a clear distinction in RTT for websites that are hosted inside, versus those that are not. *E.g.*, in Iran for

> 99% of the websites which were located inside, we observe RTT < 30 ms. Whereas, majority of those which were located outside had RTT > 100 ms. We further use RTT to classify 25k websites (Alexa top-5k for each of the five countries) as inside or outside.

2 Relevant Research

2.1 Proliferation of CDNs

Content Distribution Network (CDN) is a distributed architecture, that relies on *replica* servers to minimize end-users' access latency. It acts as an intermediary between the content publishers (site owners) and the end-users. Thus, it caches content at its edge servers (called *front-ends* [36]).

In general, there are two types of CDNs — *anycast* and *DNS* based CDNs [26, 30]. In anycast based CDNs [30] (*e.g.,* Cloudflare), a single IP address is announced through multiple BGP advertisements, often from different geographic locations. A client's web request is directed to the closest possible front-end, based on BGP policies of the client's ISP. However, in DNS based CDNs [78] (like Akamai), same website is resolved to different IP addresses, depending on the client's location. Whenever a client initiates a DNS query for a website, the resolver responds with an IP address which is (likely) closest to it. In general, DNS based CDNs maintain a separate mapping system to direct clients to their nearest front-ends.

In 2011 Ager *et al.* [19] conducted a measurement study to identify the extent of web content replication across different parts of the globe. They resolved Alexa popular websites, but (i) considered only DNS based CDNs in their study and (ii) relied on Maxmind database for geolocating the IP addresses which are known to be erroneous [40]. They reported that, at least 46% of the popular domains were served from North America, 20% from Europe and 18% from Asia; the other three continents *viz.,* Africa, Oceania and South America did not serve much of the content. Additionally, when content was requested from North America 58.2% of the content was served from the same continent, while this number was only 26% for Asia.

To further study the proliferation of CDNs, many researchers focused on mapping the complex ecosystem of individual CDNs. For example, in 2013, Calder *et al.* [25] reported that Google had front-ends in over 100 countries and 768 ASes. Böttger *et al.* [24] studied Open Connect, the CDN owned by Netflix. Authors reported that IXPs play a vital role in large-scale content delivery for Open Connect's world-wide customer base. Further, global CDNs have partnered with local CDNs of China to cater to their growing user base [82].

In 2018, Yeganeh *et al.* [84] studied the NetFlow data of a stub AS to understand the "locality of Internet traffic". They assumed RTT as distance metric for locality. Websites with low RTT were considered closer than those with higher. They reported that 90% of the traffic for the top 13 content providers was delivered within a 60 ms RTT. Their results indicate that attempts made by different CDNs to bring content closer to the edge of the network are probably successful.

However, Scott *et al.* [69] took a different direction, rather than mapping a specific CDN, authors present a joint analysis of CDNs and Internet censorship. They reported that 20% of the Alexa top-10K websites were using CDNs and found 4,819 instances of ISP level DNS hijacking in 117 countries for the same.

2.2 Anti-Censorship Approaches

Free and open communication over the Internet, and its censorship, is a widely debated topic. There are numerous evidences of large scale Internet censorship [20, 27, 35, 66, 83] by various regimes. Thus, censorship circumvention systems have been devised [50, 73, 74]. Traditional systems rely on proxies; however their IP addresses are eventually discovered and blacklisted.

Systems like Decoy Routing, Cache Browser *etc.* are designed to avoid being discovered. All these approaches do not have any trivially identifiable protocol signatures (*e.g.,* already known IP addresses). Rather, they primarily rely on web content hosted outside the adversaries' control. We now briefly explain them.

Decoy Routing [34, 44, 48, 70, 80, 81] employs routers (rather than end hosts) as proxies. Web requests carrying steganographic tags, sent to an apparent "overt" destination, are en route intercepted by the Decoy Router (DR) hosted beyond the censor's control. Based on the signatures, the DR identifies the packets and diverts them towards the intended "covert" destination. DR requires the "overt" destination to be an unfiltered site, positioned outside the censor's control. This assumption may not always hold true if such sites are hosted on CDNs, located inside censored countries. The decoy routed packets may never reach the DR.

CacheBrowser [43] (and its successor CDNReaper [85]) leverages the fact, that it is hard for an adversary to opt-in for IP filtering to censor CDN hosted content. This is because: (1) both censored and uncensored content is shared by the CDN infrastructure; IP filtering could inadvertently cause collateral damage by blocking the uncensored content as well. (2) It is hard for the adversary to enlist all possible IP addresses associated with the CDN hosted website, as the content is replicated on multiple front-ends spread across the globe. (3) Mapping CDN front-ends to clients is highly dynamic *i.e.,* IP addresses of the front-ends that are returned to the clients are updated every few minutes. It is hard for censors to enumerate all IP addresses for a specific censored content. As a consequence, censors can opt for DNS manipulation for blocking censored content (*i.e.,* it could respond with incorrect or bogon IP address for the blocked websites). Thus authors propose, if a client somehow learns a legitimate IP address located outside the adversaries' control (via an out-of-band channel), he can access the website.

However, in an anycast CDN, all its front-ends (local as well as foreign) would use the same IP address. Thus it could become trivial for the censor to blacklist the IP address (but at the cost of collateral damage). Further, anycasting often directs web requests to a front-end likely in same country as that of client. This could facilitate coercion by the censor on CDN provider to restrict access to blocked content on front-ends within its boundary, as already reported by the authors.

Domain Fronting [37] is a circumvention scheme that hides the actual destination of a communication from the censor. It leverages CDNs and cloud services that host multiple domains behind common front-end servers (*e.g.,* Google App engine, Amazon Cloudfront *etc.*). The apparent HTTPS communication of the client with a popular site (the front-end), bears a request to the filtered domain in the HTTP Host field, hidden from the censor by HTTPS encryption. The front-end on receiving the request, decrypts the HTTPS request, and forwards this request to a proxy server hosted on such cloud services. The adversary can only inspect the innocuous HTTPS request header, which contains no information about the proxy server and thus keeps it hidden from the censor. Domain Fronting enables accessing proxies hosted on cloud platforms and thus it can also be used as an entry point to facilitate access to other anti-censorship schemes such as Tor, Psiphon *etc.* For instance, Meek [37] is a Tor pluggable transport [17] which utilizes Domain Fronting to enable access to the otherwise blocked Tor network. However, if the front-end of the CDN is located inside the censor's boundary, censor can coerce the CDN provider to abandon the support for Domain fronting [1, 6, 15].

CovertCast [60] relies on sending the content of blocked websites via popular real-time encrypted video streaming services (*e.g.,* YouTube). It is believed that censors are generally unwilling to block such services en-masse. However, the authors acknowledge that the censors may coerce the service operators to block specific accounts associated with CovertCast. Interestingly, the presence of CDN hosted streaming sites located inside censorious countries, may facilitate such coercion.

Further, it must be noted that there exist other recent peer-to-peer (p2p) circumvention schemes like DeltaShaper [21], FreeWave [45], and SnowFlake [57] *etc.* that sends web traffic over p2p systems like Skype. The impact of traffic localization on such schemes is an important part of our future work.

2.3 Geolocation Techniques

Existing methods for geolocating Internet hosts are broadly classified as *active* and *passive*.

Passive methods mostly involve database lookups. These databases (*e.g.,* Maxmind [12]) are populated from various sources like reverse DNS lookups, RIRs, and ISPs [28]. Though widely used, these databases are notoriously erroneous [40].

Active methods involve estimating geolocations from RTT measurements. The initial efforts [62] involved mapping nodes with unknown locations (also called *targets*), proximus to known locations. Later approaches mostly relied on Internet *multilateration* [42, 49, 75, 76]. They involve estimating the position of a target host, using distance measurements from sufficient number of fixed reference points. The process involves plotting circles on the world-map with centers as the reference nodes and the distances as the radii. We call the region enclosed by such a circle as the *Probe Coverage Region (PCR)*. In theory multiple PCRs should intersect at exactly the target's location. However, due to measurement errors, rather than intersecting at one point, they often form an intersection area (ref. Fig. 2). The target possibly lies in this intersection region.

There are various ways for estimating the distance a packet may travel. An example is the *Speed of Light (SOL)* constraint. It is largely believed that packets on fiber network cannot travel faster than two-thirds the speed of light (c) [42]. This principle provides an upper bound ($(2/3)c * OWD$, where OWD is one-way delay) on the estimated distance the packets may travel. These distances may be used for multilaterating a target.

Constraint Based Geolocation (CBG):

For geolocating Internet hosts, direct application of SOL constraint is not recommended. RTT (or OWD) is directly proportional to congestion (queuing delay). Even when geodesic distance between two hosts is small, RTT between them can be large (due to congestion). However, SOL constraint being completely ignorant of such factors, over-estimates the distance. Thus, Gueye *et al.* [42] proposed Constraint Based Geolocation (CBG), to incorporate these factors. It aims to compute distances between reference nodes (in known locations) and the target host, using RTT measurements. Assuming the SOL constraint, CBG generates a "baseline" (ref. Fig. 1) depicting the linear relationship between delays and distances, ignoring factors like queuing delays *etc.* Thus, SOL almost always over-estimates the actual distance a packet travels. Hence CBG, introduces a *calibration phase* to compensate for queuing delays which is ignored by the SOL constraint. In this phase, all the reference nodes ping each other and generate a scatterplot between delay (RTT) and distance (distance to other reference hosts is already known). From this plot a "bestline" is computed in a manner that: (1) All the data points (RTTs) are above this line and (2) This line is closest to all the data points in the plot (ref. Fig. 1). Distance is estimated from RTT using the "bestline", instead of the "baseline." This distance is always expected to be less that what is estimated using the SOL constraint.

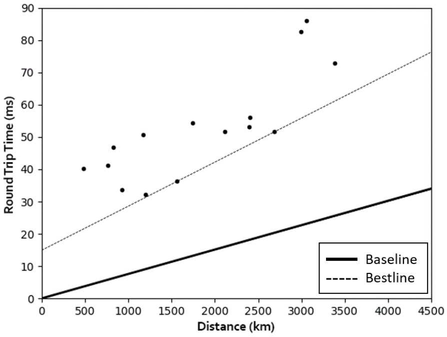

Fig. 1. Baseline and Bestline for a RIPE probe.

With reference nodes as centres and these distances as radii, CBG creates PCRs on world map and computes the intersection region of PCRs. The target's location is assumed to be the centroid of this intersection region. Authors validated the locations predicted using CBG against the ground truth — nodes (with known locations) spread across US and Western Europe.

3 Methodology

Our research involved identifying what fraction of Alexa top-'n' websites reside within (or outside) the said country using active geolocation techniques. Recently, Weinberg *et al.* [76] empirically demonstrated that CBG outperforms other active methods like Octant [79] and Spotter [52]. Thus, we began by using CBG to geolocate Internet hosts. It is known that reference nodes far from the target do not contribute much to the geolocation process [49, 76]. Thus, using recommendations from previous efforts [76], we selected reference nodes in the same continent as the target. However, our initial study for geolocating RIPE nodes as targets (with known locations) resulted in large errors — upto 4000 kms. Similar errors (≈ 5000 km) were also reported by Weinberg *et al.* [76]. Hence, to reduce such errors in geolocation, we augment CBG by selecting reference nodes closer to the target (likely in the same country). We call this approach as *Region Specific CBG (R-CBG)* and used it to identify whether an IP address is positioned inside (or outside) the desired geographic area.

3.1 R-CBG: Improving Accuracy of CBG

We now explain how we improved the geolocation accuracy for CBG using R-CBG. Further, we also explain how it multilaterates anycasted IP addresses.

Our initial observation in geolocating RIPE probes with CBG, resulted in large errors, even when reference nodes were selected in the same continent as the target. Thus, we went a step ahead and individually selected the reference nodes, such that they were evenly distributed and *located either inside, or close to the geographic boundary*, of the country under consideration. This vital step resulted in high accuracy, predicting the host as inside/outside the country.

As already mentioned in §2.3, to multilaterate an IP address, the reference nodes create probe coverage regions (PCRs) on the world map to produce an intersection region. The IP is expected to be located at the the centroid of this intersection of PCRs. To correctly identify, whether a node resides inside (or outside) the country, we present the following ***four-point heuristic:***

1. Intersection region of PCRs is completely within the boundary of the country (ref. Fig. 2)[1].
2. Intersection region of PCRs cuts through the country's boundary with centroid of region inside the country.
3. Intersection region of PCRs cuts through the country's boundary with centroid of the region outside the country.
4. Intersection region of PCRs is very large and subsumes the entire country's boundary (ref. Fig. 3).

[1] The maps shown in all the figures are just for representation. They do not represent the actual geopolitical boundaries of the countries.

If condition (1) or (2) holds good, R-CBG predicts the target as an ***inside****. Whereas if condition (3) or (4) holds good, R-CBG predicts the target as an* ***outside****.*

Rationale Behind Four-Point Heuristic: All multilateration approaches assume that the target IP is located inside the intersection region of PCRs. Thus, it is obvious that if the entire intersection region is completely inside a given country, the target IP is also located inside, validating heuristic 1. Next, if the intersection region lies partially within the country's boundary with its centroid positioned inside the country, it implies a large portion of the intersection region is located inside the country. This further indicates that likely the target IP is close to the border and hosted within the country, supporting our heuristic 2. Similarly, if the centroid is outside, likely the target IP is close to the border but located outside the country which is the rationale behind heuristic 3. Lastly, heuristic 4 represents the scenario where target is likely positioned very far from the country. This is intuitive — assuming that reference nodes are located inside the country (*e.g.,* India) and the target is located outside (*e.g.,* SA), the radii computed will be large for all the reference nodes, producing a large intersection area (shown in Fig. 3).

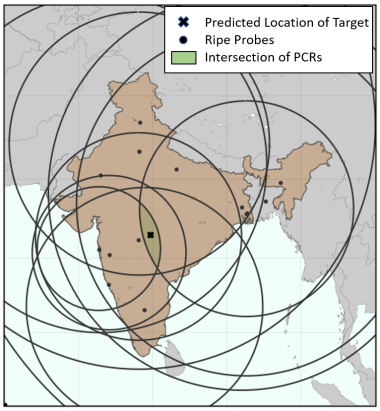

Fig. 2. IP address located inside India. The intersection area is well contained within the country.

Further, to test our heuristics, we use R-CBG for multilaterating RIPE nodes as targets, whose domiciles were known apriori. We tested it for five countries *viz.,* IN, IR, SA, BR and US. For each country, it resulted in high accuracy in correctly disambiguating inside targets from the outside targets. We present the details in §4.1.

Multilaterating Anycasted IP Addresses (within a country): CDN providers like Cloudflare use *anycast* architecture. In anycast, a particular IP prefix is announced from multiple geographic locations (called as *anycasted sites*). This helps serve web content through redundant caches at various anycasted sites. Based on the BGP policies of the client's ISP, the web request is redirected to any one of the sites (likely the closest).

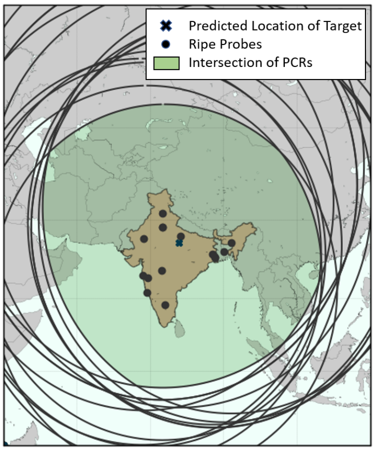

Fig. 3. An IP address located outside India. The entire country (India) is well contained in the intersection area.

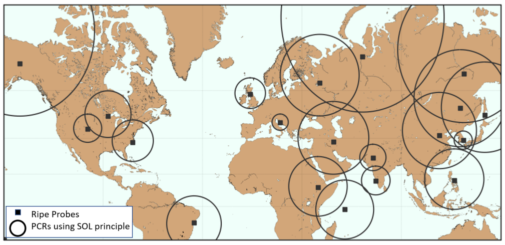

Fig. 4. Detecting IP anycasting. Dots represent the RIPE probes and circles represent the distance estimated to the IP address based on speed of light constraint.

It is non-trivial to use native CBG, with globally distributed reference nodes, for identifying if an anycasted website is located within a country or not. This is because each reference node ends up probing the same anycasted IP address, but at different sites. This results in multiple non-intersecting PCRs on the world map (ref. Fig. 4). Since, CBG predicts the location of target as the centroid of the intersection region of PCRs, in this case it fails due to the lack of such a region.

However, for R-CBG, we select the reference nodes inside the country under consideration. Thus, applying R-CBG to multilaterate an anycasted site might lead to majority of reference nodes pinging the same host. This would result in multiple PCRs forming an intersection region. Thereafter the four-point heuristic is applied to check if the target resides inside the country or not.

In cases where a single IP address is anycasted at multiple sites *within the same country* (*e.g.,* US), most

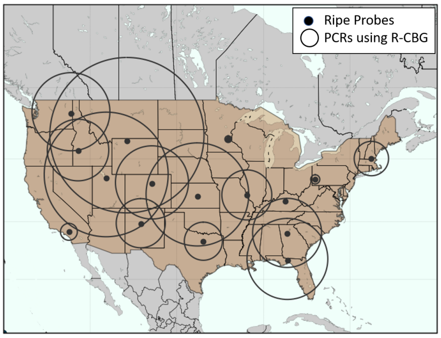

Fig. 5. A single IP address is anycasted at multiple locations within US itself.

of the PCRs might not intersect (as shown in Fig. 5). But, since majority of these circles are within the country's boundary, we ascribe this IP address as "inside".

3.2 Applying R-CBG for Different Countries

We selected five different countries *viz.,* IN, IR, SA, BR and US for our analysis. Our goal was to identify what fraction of Alexa top-'n' websites reside within these five countries by using R-CBG. For each country under consideration, we enumerate the steps taken for the same:

1. We selected country specific Alexa top-1k websites, resolved them from a RIPE node located inside the country under consideration, and recorded their IP addresses.
2. Next, we individually selected atleast 15 reference nodes, such that they were distributed and located either inside, or close to, the geographic boundary of the country. (The rationale behind selecting atleast 15 reference nodes is further explained in §6.4).
3. The "bestline" (ref. §2.3) is computed using the RTT between the reference nodes (probes) and their (already known) geodesic distances. The probes ping each other and the corresponding RTTs are recorded.
4. Next, the reference nodes pinged the targets (*i.e.,* Alexa websites), and the corresponding RTTs were also recorded.
5. We use the "bestline" and the recorded RTT (from the previous step 4) to estimate the distance between the reference nodes and the targets.
6. Using the reference nodes as centers and distances as radii (from the previous step 5), PCRs are drawn on world map. The intersection region of these PCRs and its centroid is computed.
7. We use the intersection region of PCRs, the centroid and the country's boundary coordinates, along with the four-point heuristic, (ref. §3.1), to decide if the target is positioned "inside" the country or not.

For cases where all PCRs do not intersect, we select the intersection region formed by maximally intersecting PCRs. There are two possible explanations. Firstly, a reference node experiencing heavy congestion may underestimate the distance to the target, leading to a smaller PCR [76]. Secondly, as already explained, this may be a case where the target IP is anycasted within the same country. Maximally intersecting PCRs make R-CBG agnostic to such pitfalls.

3.3 Selection of Reference Nodes

We rely on RIPE Atlas for selecting our reference nodes. It offers two types of nodes — *anchors* and *probes*. Anchors are stable machines that regularly ping one another. On the contrary, probes are machines which may (or may not) be available all the time and might not respond to ping requests. Hence, we preferred selecting anchor nodes. However due to their scarcity in our tested countries, we also included a few stable probes.

While selecting probes we ensured the following: (i) Assuming target is within the country, the probes are evenly distributed, potentially surrounding the target. (ii) The probes are stable and respond to pings. To select the stable probes, we pinged them from our university machine for one week. We tested their connectivity 5 times a day, each time with 5 ping packets. Those probes which responded to more than 90% of the ping requests, qualified as stable reference nodes.

4 Data Collection and Results

4.1 Validating R-CBG

The aim of R-CBG is to determine if a target IP address resides in a country or not. To gauge its accuracy, we compared the outcomes of R-CBG against the domiciles of RIPE probes (known *a priori*). To that end, we first individually selected reference nodes in and around the country under consideration (ref. §3.3). Further, for the targets, we considered all the RIPE nodes upto 5000 km radius, from the country's approximate geographic center. To check their connectivity we pinged all of them. Those which responded were finally selected as targets, with their domiciles as the ground truths. Overall this ensured that we have sufficient targets both inside and outside the country.

Following the steps mentioned in §3.2, we multilaterated these RIPE nodes (selected as targets). More

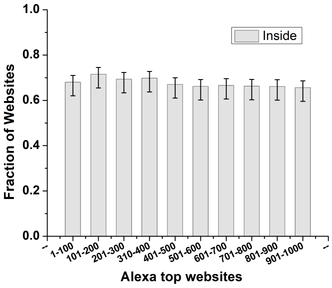

Fig. 6. Fraction of websites located inside/outside Iran.

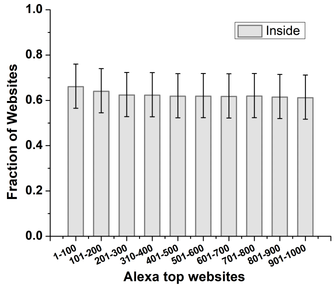

Fig. 7. Fraction of websites located inside/outside India.

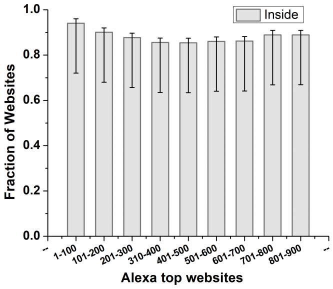

Fig. 8. Fraction of websites located inside/outside Saudi Arabia.

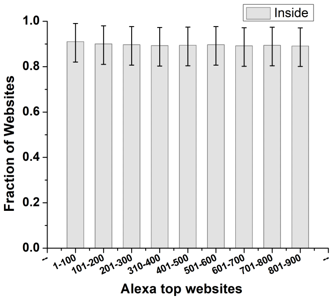

Fig. 9. Fraction of websites located inside/outside Brazil.

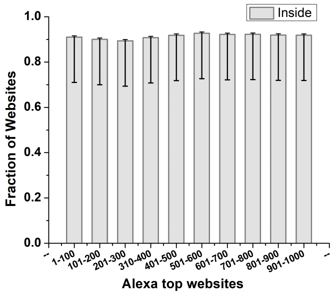

Fig. 10. Fraction of websites located inside/outside United States.

Countries			Predicted		Accuracy (%)
			IN	OUT	
India (IN)	Actual	IN	19	2	89.73
		OUT	17	147	
Iran (IR)		IN	27	1	94.12
		OUT	5	69	
Saudi (SA)		IN	47	1	89.87
		OUT	7	24	
Brazil (BR)		IN	33	3	90.71
		OUT	10	94	
USA (US)		IN	617	4	92.41
		OUT	45	213	

Fig. 11. Confusion Matrix for different countries.

than 89% of these targets, were correctly classified by R-CBG as inside (or outside) the chosen country. *E.g.,* corresponding to US, we identified total of 879 RIPE nodes (responsive to pings, and within 5000 kms), that were selected as targets. Of these, 621 were hosted inside and 258 were outside. R-CBG correctly classified 617/621 nodes as inside, whereas 213/258 as outside. About 45 nodes which were actually hosted outside, were miss-classified as inside. Most of these were located near the border of US (either Mexico or Canada). We tested the efficacy of R-CBG for each of the five countries under consideration and obtained similar trend for all. The Confusion Matrix is shown in Fig. 11.

Moreover, for a few university websites that we know were hosted within the university itself, we attempted to multilaterate them with R-CBG. *E.g.* when we selected `www.columbia.edu` as target and RIPE probes in US as reference nodes, R-CBG predicted the website to be hosted within the US itself. Later, when we changed the set of reference nodes to be RIPE nodes located in other countries (*e.g.,* Brazil), R-CBG estimated the website to be hosted outside. We repeated the same exercise for nine more university websites and R-CBG correctly classified them inside/outside depending on the location of reference nodes.

Caveat: For very small countries like Switzerland, R-CBG might not result in high accuracy. This is because, the intersection region formed by the PCRs would be so large that it might always subsume that entire country, irrespective of the location of target IP address (inside or outside the country). This might violate the rationale behind our four-point heuristics, for these small countries. Thus, before applying R-CBG to any new country, its accuracy needs to be tested again.

4.2 Multilaterating Alexa Websites

Having established the accuracy of R-CBG, we used it to test if a website is located inside the country or not. Fig. 6, 7, 8, 9 and 10 show the fraction of websites that are located inside and outside the country's boundary. *E.g.,* for Brazil, we observed that ≈ 89% of Alexa top-1k websites are located inside its geographic periphery. Similarly, for Iran ≈ 66%, India ≈ 61%, Saudi Arabia ≈ 86% and US ≈ 92% of the websites were found to be located inside. However it must be noted that, in all the five countries, there were significant number of websites (atleast 8% of Alexa top-1k websites) that were certainly located *outside* the country's boundary. Thus clients residing in these countries may use such sites with anti-censorship approaches like Decoy Routing.

Interestingly, we also observed that the sites hosted outside varied evenly in their popularity. *E.g.,* in SA 8% of the top-100 websites were found to be located outside. Similar trend was also observed for top 800-900 websites. Thus, selectively censoring the relatively less popular sites may not significantly impact circumvention systems that rely on websites positioned outside, however would hinder them.

4.3 Multilaterating Alexa Websites When Resolved From Outside the Country

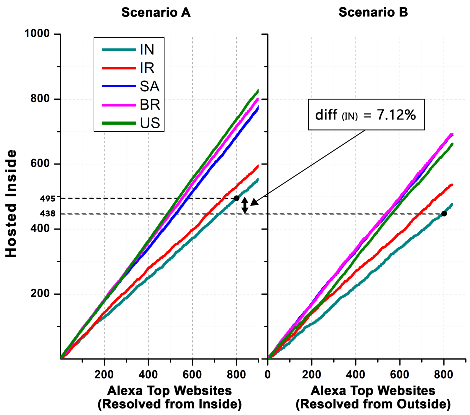

Fig. 12. Number of websites located inside when resolved from within and outside the country.

As already described, for majority of the websites a client obtains a corresponding IP address in its own country (likely due to CDNs). This might render anti-censorship approaches like Decoy Routing, Domain Fronting, CacheBrowser *etc.* ineffective. However, it can be argued that a client can still use these approaches, if it somehow obtains IP addresses corresponding to *foreign front-ends* of the same Alexa website. *E.g.,* if a client uses such IP addresses for Decoy Routing, the decoy routed packets may cross the country's network boundary, eventually being intercepted by the DR.

To obtain IP addresses outside the censor's boundary, we individually resolved Alexa top-1k websites from a host (we control) in an uncensored country (Ireland) and recorded the IP addresses. Ideally these IP addresses should be located outside the censor's boundary and may be used by these anti-censorship approaches. Thus, to test our hypothesis, we multilaterated the resulting IP addresses from each of these countries, choosing the original set of reference nodes positioned within the said countries (as already mentioned in §3.3).

For each country, Fig. 12 represents the number of country specific Alexa top-n websites, identified to be hosted inside, when domains were resolved from within the censorious country, and from a non-censorious foreign country. *Scenario A* is when websites were both resolved and multilaterated from the said country. Whereas, *Scenario B* corresponds to websites being resolved externally, but multilaterated using reference nodes inside the country.

Ideally, the websites which were earlier ascribed as inside the country (in §4.2), should have now been reported as outside. However, we observed no significant differences. The total number of websites hosted inside do not differ much in both the scenarios. *E.g.,* in India 495/800 websites were inside in *Scenario A*, whereas in *Scenario B* it was 438/800. This implies that either majority of the websites were anycasted or non-CDN hosted (positioned only at a single location) within the bounds of the censor. This small difference (7.12%) is likely due to DNS based CDNs (explained in detail in the next subsection). Thus we now describe our approach to identify which type of CDN a website is using.

4.4 Identifying Type of CDN

Anycasted IP addresses are announced at multiple locations across the globe. Geolocating such IPs, using SOL multilateration (ref. §2.3) would never yield an intersection of *all* PCRs. We use this observation to differentiate between anycast and other forms of hosting. This observation holds valid because probe packets of different reference hosts would be routed to their likely closest anycasting site. Thereafter, by employing SOL constraint one estimates maximum possible distance travelled by a packet in the observed RTT. Using these distances to multilaterate anycasted IP addresses would lead to zero (or very few) overlapping circles (ref. Fig. 4).

To identify the different types of CDN hosting (for the Alexa-1k websites), we selected 25 globally distributed RIPE nodes and resolved each of the websites from them. A website that resolved to the same IP address across all the probes, was multilaterated using the SOL principle. The presence of an intersection region of *all* PCRs, indicates that the website is positioned at only one location. The absence indicates anycast hosting. On the other hand, a website that resolved to multiple IP addresses from the different probes, very likely uses DNS-based CDN. However, anycasting *also* allows a site to have multiple IP addresses which may simultaneously be advertised from various geographic locations.

To differentiate the two, we randomly chose a single IP address for all such websites, and multilaterated it using SOL principle. Again, the presence of an intersection of *all* PCRs indicates that the IP is positioned only at one location, contraindicating anycasting, confirming DNS-based hosting. Fig. 13 schematically describes the approach. We further validate our CDN classification scheme in Appendix A.2.

Nations	CDN				Non-CDN	
	DNS		ANYCAST			
	Hosted In (%)	Hosted Out (%)	Hosted In (%)	Hosted Out (%)	Hosted In(%)	Hosted Out (%)
IN	8.12	6.77	35.92	11.81	19.68	17.71
IR	1.31	2.42	11.78	15.01	52.87	16.62
SA	8.41	1.19	57.44	3.66	20.80	8.51
BR	9.36	2.07	52.88	2.72	27.97	5.01
US	14.97	3.49	56.31	2.67	20.62	1.95

Table 1. Type of CDNs used by Alexa top-1k websites.

Table 1 represents the type of CDNs a website is using. Columns 1, 3 and 5 represent fraction of websites located inside the country, whereas remaining represent websites hosted outside (coloured as green). Anti-censorship approaches (like Decoy Routing, Cache-Browser) require websites to be located outside the censor's boundary. However, websites using *DNS-based CDNs* which are located *inside* (column 1) can also be used for such purposes. These websites very likely have front-ends spread across the globe. If a client (somehow) obtains an IP address of a foreign front-end, its request will cross the country's boundary. Thus, websites using DNS-based CDNs can also be considered as a viable option for such anti-censorship approaches.

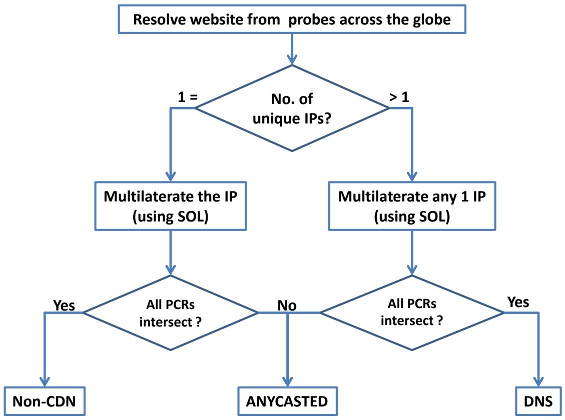
Resolve website from probes across the globe
No. of unique IPs?
1 =
> 1
Multilaterate the IP (using SOL)
Multilaterate any 1 IP (using SOL)
All PCRs intersect ?
All PCRs intersect ?
Yes
No
Yes
Non-CDN
ANYCASTED
DNS

Fig. 13. Identifying type of CDN used.

Additionally, the Table 1 also explains the reason for minor differences (*e.g.,* 7.12% for IN) reported in Fig. 12. It corresponds to the fraction of websites using DNS based CDNs, hosted inside. When websites are resolved from a foreign host would likely map to their respective foreign front-end IP address, while others (anycast and non-CDN based) would not. *E.g.*, the IP addresses for 61.88% websites for India, when resolved internally, were identified to be inside. This number dropped to 54.8% when these sites were resolved externally. This difference is close to the fraction of websites using DNS based CDNs (*i.e.,* 8.12%), hosted inside.

5 Inferences From Results

5.1 On Internet Traffic Locality

Our results reveal that for five countries under consideration, 60%-90% of Alexa top-1k websites are located within the country. Moreover, this trend remains same for Alexa top-5k websites also (ref. §6.1). This traffic localization could hinder powerful countries to surveil transitory (or terminal) traffic of "underserved" nations.

For instance, for BR we found that 90% of Alexa top-100 websites (ref. Fig. 9) are hosted within the country itself. However, Edmundson et al. [33] reported that only less than 17% of Alexa top-100 websites are located within BR. This can be explained as follows. Edmundson *et al.,* rented VPSes in developing nations – executed `traceroutes` to top Alexa websites – converted IP paths to country level paths (using Maxmind) – simply reported the domicile of Alexa websites as the country hosting the last IP addresses in the `traceroute` path.

Unfortunately, IP to country mapping is not trivial. In general these databases rely on Internet routing registries and map the IP addresses to their corresponding ASes and report the country of the IP addresses' as the country where the AS is headquartered. Further, IP to AS assignment and geolocation databases are erroneous and could lead to incorrect inferences [76], as acknowledged by authors themselves. Moreover, it is already known that converting traceroute paths to country level paths is not straightforward and could also lead to incorrect inferences [54, 55, 58, 59].

Additionally, the majority of these websites are hosted on CDNs. Wohlfart *et al.,* [78] recently revealed that CDNs like Akamai have complex infrastructural deployments. Their information is neither available in publicly accessible BGP data nor could be captured by active measurements like `traceroutes`. They reported that Akamai have about 6.1k "explicit" peerings (standard peerings where Akamai is one of the two involved peers) and about 28.5k "implicit" peerings (where neither of the involved peers is Akamai). These implicit peerings of CDNs with different ISPs across the globe further complicates the *IP to ASN (or country) mappings*, as it becomes non-trivial to identify whether the IP address (assigned to a CDN front-end) is owned by the CDN or the ISP. Thus, we relied on R-CBG (a multilateration technique) to identify whether a particular IP is located within (or outside) the given country. Since this approach does not require information from any of the third-party sources, the possible errors associated with these sources would not impact our results.

As already mentioned, R-CBG classified majority of the Alexa top-1k websites as hosted within the nation. However, it can be argued that even if websites are located inside the nation, web requests may follow *tromboning paths* — paths that originate and end in the same country, but transit a foreign country [33]. Our research indicates that, even if tromboning paths exist, they are rare. A packet following a tromboning path (relatively longer path) would very likely result in higher RTT, in comparison to a non-tromboning path [63]. Such high RTTs would have resulted in wider PCRs, resulting in large intersection regions. In such cases, R-CBG would have incorrectly classified inside nodes as outside (ref. point 4 of §3.1). But, this is not the case as R-CBG has high accuracy ($> 89\%$). *E.g.,* in US, R-CBG correctly identified 617 (out of 621) internal RIPE nodes. Such observations likely indicate the absence of tromboning paths. Moreover, they also convey that the client's traffic is majorly localized in its own country, and rarely follows a hierarchical path [39]. Thus, we believe that CDNs have resulted in "Internet flattening", also reported by others [32, 56].

5.2 Hindrances to Anti-Censorship

Decoy Routing: It assumes that there are some unblocked "overt" websites, located outside the censored regimes. Thus, it is expected that when a client sends web requests (carrying special steganographic tags) to these "overt" websites, they cross the censors' nation boundary. *En-route* these requests are intercepted by Decoy Routers (DRs), also positioned outside. Based on the tags, the DR identifies the packets and diverts them towards the intended "covert" destination.

Existing DR placement schemes [41, 44, 46] assume that specially crafted web requests from DR clients destined to overt websites would *cross censor's boundary* and en-route intercepted by the DRs. However, traffic localization due to CDNs may inadvertently impact such schemes.

In 2011, Houmansadr *et al.* [44], showed that DRs placed in two tier-1 ASes could deliver DR service to all Internet hosts. Thereafter, Schuchard *et al.* [68] demonstrated how a sufficiently powerful adversary can change its routing policies and simply route around ASes where DRs are positioned. Later, Houmansadr *et al.* [46] reverted that such a move could be prohibitively expensive for an adversary. If DRs are placed in enough foreign ASes they could completely intercept all network paths originating from a censored country. For instance, cooperation of such 900 friendly foreign ASes could provide DR services to all Chinese clients. Further, Gosain *et al.* [41] proposed an improved placement strategy that involves placing DRs in about 30 ASes of the globe. The authors argued that these strategically chosen ASes intercept more than 90% of AS level paths from almost all countries of the world.

Nevertheless, *all these schemes assume that the client–OD traffic crosses the censor's network boundaries. Our results show that in majority of the cases they do not.* Front-ends of $61\% - 92\%$ of the country specific popular websites are located within the client's own nation. This poses challenges for the placement of DRs *i.e.,* where to place DRs on the Internet.

It could be further argued that a local front-end would eventually contact its backend CDN server [64], requesting the content on behalf of the DR client. Thus, the web request would cross the censor's boundary and might be intercepted by DRs. Indeed it is possible; however, for DRs to identify a legitimate DR request, it should bear an embedded stenographic tag. But the web request generated by the local front-end does not bear this covert signature, as the front-end is agnostic to DR.

Furthermore, it must be noted that the fraction of country specific Alexa websites (10%–40%), that are hosted outside, could be directly used as overt sites for DR. However, with the proliferation of CDNs within ISPs and with relatively cheaper cloud infrastructure, more websites are being hosted on such platforms [29, 71]. Thus, following the current trend, the websites that are presently hosted outside, might migrate to CDNs in near future. These websites could be then hosted on front-ends within the censor's boundary, thus becoming unsuitable as overt sites for DR.

Future directions and challenges involved: We believe that in future, DR approaches like Slitheen [23], Waterfall of Liberty [61] and Conjure [38] could be further explored for circumvention. Slitheen and Waterfall of Liberty assume that a client would send a DR request to an overt website (*e.g.,* Alexa websites), hosted outside the censor's regime for signalling the decoy router. The usual response from these overt websites bear URLs to additional (leaf) content such as images, videos and ads *etc.* The DR station then replaces the leaf responses (corresponding to the requests to these embedded URLs) with the censored content. In other words, these schemes *do not send specially crafted DR requests to these embedded URLs.*

As already mentioned, if the overt websites are hosted on front-ends within censor's jurisdiction, DR request would not be intercepted by the decoy router, and it would cease to function. However, if the IP addresses associated with the embedded URLs are hosted outside the country's boundary, these schemes in future could be tailored to use them for sending the DR request as well. Thus, DR request to these IP addresses may cross the censor's boundary and might not be impacted by CDN traffic localization.

But, still there exist some additional concerns that DR solutions (like Slitheen) need to address — *i.e.,* what are the Alexa websites whose parallel connections could be used for DR by the DR clients? Exactly how many of these parallel connections terminate in front-ends located outside the censor's boundary *etc.*? We attempted to answer these questions and present our preliminary findings in §6.2.

Conjure [38], another novel DR scheme, involves clients connecting to an IP address in the unused IP address space of the friendly ISP (positioned outside the censor's boundary). The ISP's network infrastructure mirrors this traffic to the DR station.

However, the initial registration step in Conjure requires the client to connect to an overt website hosted outside the censor's jurisdiction. It is only after this step that the client derives an unused IP address to be used for actual DR. Further, these websites must also be popular in the censored country else they could be easily blocked by the censor. Moreover, if these overt websites are hosted on CDNs, the traffic localization problem may still persist.

Domain Fronting: A Domain Fronting (DF) client engages in an apparent HTTPS communication with the front-end of a popular CDN [37]. It actually bears a request to the blocked domain in the HTTP Host field, hidden from censor due to HTTPS encryption. The front-end (inside the censor's boundary) would forward the requests to a proxy (hosted on such services) that fetches the censored content for the client. However, since front-ends are themselves located within the censor's jurisdiction, censor can coerce the CDN provider to halt the use of such circumvention techniques.

E.g., Telegram (a popular IM app) has been subjected to a stern censorship from Russian authorities [65]. Thus, operators of Telegram resorted to use DF as a circumvention service, by relying on cloud services provided by Google, Amazon and Microsoft. However, it has been reported that Roskomnadzor (the Russian authority managing censorship) has coerced them to halt the use of DF [1, 6, 15]. Such instances back our claim that censors can coerce CDN providers to adhere to their specified policies regarding censorship, if they want to expand their business within censor's jurisdiction, hindering circumvention schemes like DF.

CacheBrowser: When clients access CDN-based web content, the requests are generally directed to the closest front-ends (often located in clients' country). However, Zolfaghari *et al.* [85] reported that powerful censors coerce CDN providers to filter content on front-ends located within their boundary. CacheBrowser (and CDNReaper) aims to disrupt this dynamic. It involves direct communication with IP addresses of foreign front-ends, rather than relying on regular DNS resolutions. However, anycast CDNs assign the same IP address to all their front-ends. Thus, for anycasted websites, it cannot be used; the request would never leave the clients' countries. *Thus, CacheBrowser can only be used to access websites that use DNS based CDNs.* Our results reveal that only a small fraction of Alexa top-1k websites rely on DNS based CDNs *i.e.,* ≈ 11% (ref. Tab. 1).

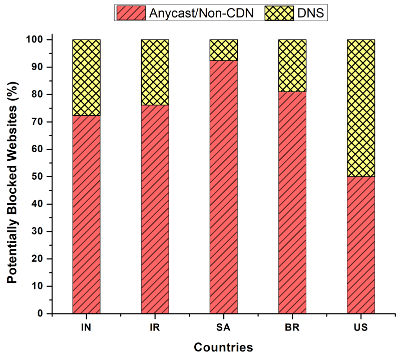

Fig. 14. Type of CDNs used by potentially blocked websites.

But, it can be argued that a censor might avoid blocking Alexa top-1k websites due to their popularity. Thus, we also tested if CacheBrowser can be employed for accessing potentially blocked websites (Citizen Lab's [5] country specific lists). It is evident from Fig. 14, that majority of such websites (for all five countries) were not using DNS-based CDNs[2], and thus may be inaccessible using CacheBrowser. *E.g.,* in BR, out of 2769 potentially blocked URLs, CacheBrowser can only be used to access about 525 of them (*i.e.,* ≈ 18%).

[2] The type of CDN was identified using the approach in §4.4.

CovertCast: It relies on popular live video streaming services (*e.g.,* YouTube) to secretly transport the content of blocked websites to clients residing in censored regimes [60]. Popular streaming video services like YouTube use CDNs and operate via front-ends (in censor's boundary). Thus, the adversary might coerce these services to stop support for CovertCast, particularly via front-ends positioned within its control. In our results, we observed popular live stream supported sites (*e.g.,* YouTube, Facebook, Instagram) have front-ends hosted within the countries under test.

6 Discussion

6.1 Analysis of Alexa top-5K Websites

In previous sections, we identified that majority of Alexa top-1k sites were located within the nations under test. However, one can question if our analysis holds good for other Alexa websites as well. Thus we further attempt to identify the locations of Alexa top-5k sites w.r.t the individual countries.

The ideal case would have been to use R-CBG to multilaterate each of these websites. However, the multilateration process is *costly* in terms of RIPE credits and time consumption. In total we required multilaterating $\approx 25K$ websites. In the absence of sufficient credits, we abandoned this idea and relied on RTT as a gross-metric to identify the location of these websites. It must be noted that measuring RTT using `ping` reduces the credit requirement by more than 15 times in comparison to running R-CBG for a single target.

Website Locality Using RTT Profiling: As already described in §2.3, RTT by and large correlates to distance. For most of the Alexa top-1k sites, the ones hosted inside have relatively smaller RTTs than those hosted outside. *E.g.,* Fig. 15 shows the distribution of these RTTs, recorded using a single RIPE probe in Iran. Evident from the figure, there are *two* distinguishable categories corresponding to websites located inside and outside. Similar trends were also observed for other countries (ref. Appendix A.1).

From such observations, we believe that RTT *alone* can be used to determine whether websites are internal (or external) to the nation. To correctly disambiguate the two categories, we used Support Vector Machine (SVM) classifier [72]. We provide a set of labelled data points to the SVM classifier. The set contains Alexa top-1k website labelled as inside or outside (outcomes of R-CBG in §4.2) and their corresponding RTT values obtained from the reference node. SVM then creates the best possible line that divides the set of RTT into two distinct categories — inside and outside. To test the effectiveness of SVM we used K-fold cross-validation for a reference node [51]. It was then repeated for all reference nodes in a country. K-fold cross validation (K=5)[3], divides the samples into five equal sized partitions. Of the five partitions, four are used for training, the remaining one is used for testing the classifier. The entire process is repeated five times, selecting a different (non-repetitive) testing partition in every iteration. Then, the average classification accuracy (and corresponding standard deviation) of the five iterations is computed. For these nodes the cross-validation process observed an accuracy of $85 - 99\%$, barring a few outliers[4].

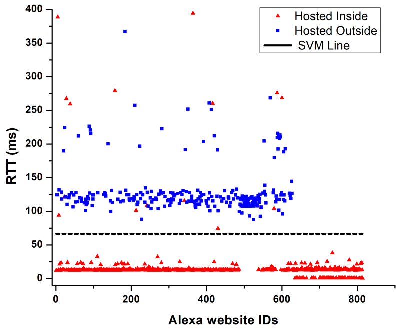

Fig. 15. RTT scatter plot for a probe in Iran.

Next, we computed the Coefficient of Variation (CoV) for all the reference nodes. The reference node with least CoV was considered as the most consistent one. Finally, for every country we selected the most consistently accurate reference node for further analyzing the locality of Alexa top-5k websites. Fig. 15 shows the SVM line (threshold) for such a reference node in Iran. Similar thresholds were also established for other nations (ref. Appendix A.1).

Fraction of Alexa Top-5k Websites Hosted Inside (or Outside) the Country: We measured the RTT for Alexa top-5k sites from the reference nodes with highest consistency. Based on the previously trained SVM classifiers, we predicted what fraction of 5k websites reside inside or outside for each country (ref. Fig. 16). By and large the fractions of websites hosted inside remain the same for both Alexa top-5k and top-1k. This yet again dispels the notion of nation state hegemony (over transitory network flows) even for a larger set of websites.

[3] K is generally selected as 5 (or 10), as these values have already been empirically shown to yield smaller bias [2, 16].

[4] For each country, 2-3 reference nodes, yielded lower accuracy ($\approx 50 - 60\%$). This could be due to variable congestion experienced by these nodes, resulting in dramatic RTT variations.

Additionally, this shows that most of these sites cannot be used by anti-censorship systems like Decoy Routing.

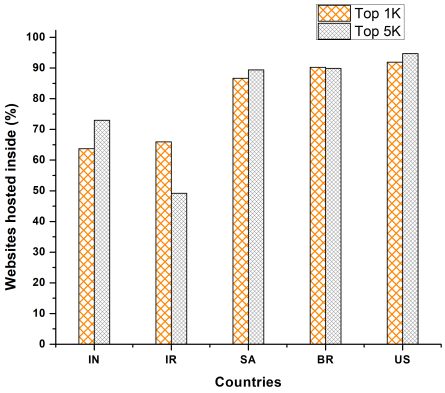

Fig. 16. Alexa top-5k websites hosted inside the country.

Interestingly for Iran, a large fraction of the more popular websites (ranked below 1k) were hosted inside, while the less popular ones (ranked above 1k) were hosted outside (ref. Fig.16). This behaviour can be explained as follows. From table 1, we observe that (1) majority (52.87%) of the Alexa top-1k websites used non-CDN infrastructure and were located inside, (2) only a small fraction (≈ 13%) of Alexa top-1k websites used CDNs and were also found inside. This indicates low CDN presence inside Iran. Overall it implies that, more than 65% of Iran's popular top-1k websites websites are hosted inside and a small fraction of these inside hosted websites use CDNs. As a consequence, for less popular websites (ranked above 1k) running on CDNs, the lack of internal front-ends may force requests to exit the nation boundary. This is likely why we observe an inversion in the trend.

6.2 Decoy Routing via Parallel (Leaf) Web Connections

For all the countries we observed that a majority of the web request (to Alexa websites) never crossed the nation boundaries. As already discussed, this may hinder the functioning of anti-censorship systems like Decoy Routing (ref. §5.2). Response for typical web requests bear HTML code embedded with URLs for content like CSS and images (required to render the pages). Web browsers establish individual (parallel) connections corresponding to each of these URLs. *Some of these parallel connections may be utilized as overt destinations for Decoy Routing*, if they terminate at IP addresses located outside the censors' control.

Most parallel connections do not cross the country's boundary. For example, when we analyzed the web transactions corresponding to Alexa top-100 websites - for India, we found that for 23 websites, the parallel connections terminated outside, even when the original web requests did not. We observed that most of these websites resulted in very few parallel connection terminating outside the nation (mode 1, median 3). However, for one particular website, this value was as high as 32. In future, existing Decoy Routing systems can evolve to make use of such embedded URLs (parallel connections) as overt destinations.

6.3 Comparison With Popular Geolocation Databases

It is known that popular geolocation databases are prone to errors at city level. However, it is believed that at country level they are relatively error free [47]. *E.g.*, very recently Edmundson *et al.* [33], used Maxmind for IP-to-country mapping. Using this they reported that majority of the Internet paths (to Alexa top-1k websites) crossed the national boundary. We thus compared the accuracy of Maxmind GeoIP2 database [12] with R-CBG. To do so, we compared the country level information for Alexa top-1k sites (for each of the countries) derived from the database, against the results obtained by multilaterating with R-CBG. Our results show (ref. Fig. 17) that geolocating IP addresses of popular websites using Maxmind results in large errors. For instance, in SA, Maxmind report less than 15% of the websites to be located inside the country itself, whereas our algorithm ascribed 90% of them to be inside. This is because, often these databases rely on static annotations [28], which map an IP address to the country where parent AS is headquartered. Thus, relying on such information [33], to determine the country through which traffic transits, seems inaccurate. Hence the notions that "powerful" countries intercept large fraction of traffic originating from underserved nations is unfounded.

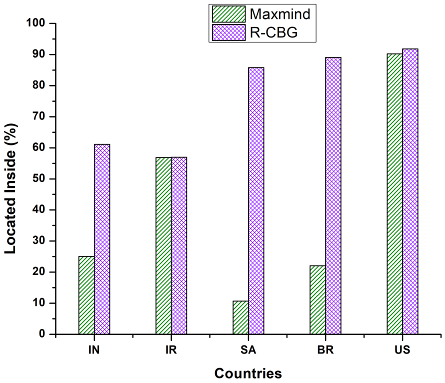

Fig. 17. Comparison of Maxmind with R-CBG.

6.4 Selection of Reference Nodes

The accuracy of R-CBG depends on the number of reference nodes and their geographic proximity to the target. We chose the reference nodes within (or close) to a country. Thus, selecting enough nodes ensured that the targets were close to a considerable fraction. We empirically observed that atleast 15 nodes were required in each of the countries for R-CBG to correctly multilaterate the target. Previous authors [42] however reported that about 30 reference nodes were required to geo-locate (using CBG) the targets. But, they chose the reference nodes at the continent level while we chose within the country (closer to the target).

6.5 Selection of Target Websites

We selected country specific Alexa top-n websites (n=100, 200...1000) for two reasons. Firstly, several anti-censorship approaches like Decoy Routing rely on these popular unblocked sites and require them to be positioned beyond the censors' control. Our analysis aids assessing the feasibility of using such systems in the nations considered. Secondly, these websites are a representation of the actual web traffic of users. Others such as Cisco *Umbrella* [4] and *Majestic Million* [11] are derived from indirect sources like DNS queries and URLs embedded in website ads, often rendered or accessed without the users' control. Our choices are in accordance to recommendations by Scheite *et al.* [67].

6.6 Stability of our Results

Scheite *et al.* [67] report that list of popular websites vary over time. Our measurements may be subject to such variations. Thus, we repeated our experiments for five consecutive months by selecting monthly snapshots of Alexa top-1k websites. As evident from Fig. 18, the fraction of websites hosted inside roughly remained same over the said period. On an average, the Alexa ranks of about 812 websites consistently remained under 1k (across the five monthly snapshots).

7 Limitations and Future Work

We studied only a few countries which are known to censor (or surveil) network traffic [7]. Further, R-CBG requires about 15 probes in a geographic vicinity of a country under study. Unfortunately, the RIPE nodes are concentrated in EU and North America [14, 76]. We thus chose countries where atleast 15 stable nodes were available. We restricted our study to five countries as our objective was to judiciously use the limited RIPE credits[5]. These countries represent a diverse distribution of Internet users — IN (≈ 560 M), US (≈ 292 M), BR (≈ 150 M), IR (≈ 62 M), SA (> 20 M) [9]. Moreover, we also believe that these countries represent a good representative set in terms of geo-political power, Internet infrastructure and open communication.

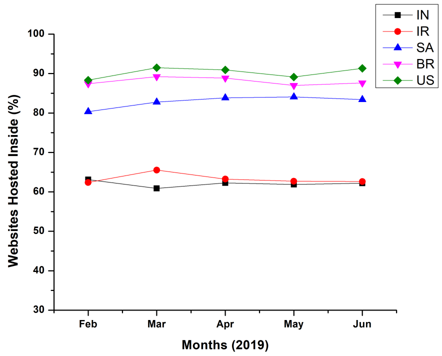

Fig. 18. Location stability for Alexa top-1k websites.

It must be noted that we wanted to study other well-known censors like China, Russia and Egypt etc. But, due to the unavailability of stable nodes and limited RIPE credits, we excluded them from our analysis. Thus, validating our claims and observations about other countries become an important part of our future work. Furthermore, R-CBG can be used for studying these countries as well; although its accuracy needs to be calculated with the new set of reference nodes.

Lastly, we used country-specific Alexa websites for our analysis as they represent the popular (unfiltered) websites for a given country. However, the "actual" popular websites in repressive countries could be different from country-specific Alexa websites; they could be blocked, and citizens would access them using VPNs *etc.* Thus, they would not be listed in the top Alexa websites. But, since the circumvention schemes considered in our study (*e.g.,* Decoy Routing) assume unhindered access to *popular unblocked* websites, we used Alexa websites in our study. However, in future, other lists of popular websites [4, 11], could also be used for further validating our claims.

8 Concluding Remarks

The proliferation of CDNs on the Internet, have brought web content closer to the end-user. On the positive side, it has improved users' web experience. But on the negative side, this content "closeness" may enhance nation

[5] The details on how RIPE credit system works can be found at [13].

states' ability to coerce content providers for regulating access within their own boundaries. To identify if a host is inside or outside a nation, we re-engineered Constraint Based Geolocation (CBG), a multilateration technique and call it as Region Specific CBG (R-CBG). From our tests (involving R-CBG) repeated for five months, we identified that majority (61% - 92%) of popular web content (Alexa top-1k websites) is located within the nation states. *E.g.,* in SA, ≈ 89% of SA's Alexa top-1k websites are hosted within the country itself.

The common circumvention solution involves proxy based systems. Sadly, these often bear easily identifiable traffic signatures (*e.g.,* IP address). Newer alternatives, like Decoy Routing, require access to popular website hosted outside the censors' control. Unfortunately, CDNs may hinder easy access to such websites. Our heuristics classified majority of the Alexa top-1k websites as hosted on CDNs, within the clients' domicile. Interestingly, this trend persists for Alexa top-5k websites, when tested using a novel RTT based heuristic. Thus, neither conventional (proxy based), nor heterodox (relying on web traffic) approaches alleviate the predicament. However, a small, yet significant set of websites (*E.g.,* ≈ 20% in SA), are hosted outside the censors' boundaries and may be used by such systems.

9 Acknowledgements

We would like to thank our shepherd Milad Nasr and other anonymous reviewers for their valuable inputs. A special mention to Piyush Kumar Sharma, who provided essential comments and feedback throughout the project. Further, this research received no specific grant from any funding agency in the public, commercial, or not-for profit sectors.

References

[1] Amazon discontinue support to domain fronting. https://signal.org/blog/looking-back-on-the-front/.

[2] Applied predictive modeling. https://bit.ly/3eUZ4bh.

[3] Caida as ranking. https://asrank.caida.org/.

[4] Cisco umbrella popular websites. https://umbrella.cisco.com/blog/2016/12/14/cisco-umbrella-1-million/.

[5] Citizen lab list of blocked websites. https://github.com/citizenlab/test-lists.

[6] Domain fronting discontinued by google and amazon. https://www.bloomberg.com/opinion/articles/2018-05-03/telegram-block-gets-help-from-google-and-amazon.

[7] Freedom house. https://freedomhouse.org/.

[8] Google global cache. https://peering.google.com/#/.

[9] Internet statistics. https://www.internetworldstats.com/top20.htm.

[10] Ip to asn using team cymru. https://asn.cymru.com/.

[11] majestic million popular websites. https://majestic.com/reports/majestic-million.

[12] Maxmind geoip2 database. https://www.maxmind.com/en/geoip2-services-and-databases.

[13] Ripe atlas credits. https://atlas.ripe.net/docs/credits/.

[14] Ripe atlas project. https://atlas.ripe.net/.

[15] Russia blocks telegram. https://tass.ru/pmef-2018/articles/5231399.

[16] Selecting value of k in k-fold cross validation. https://machinelearningmastery.com/k-fold-cross-validation/.

[17] Tor pluggable transport. https://2019.www.torproject.org/docs/pluggable-transports.html.en.

[18] HB Acharya, Sambuddho Chakravarty, and Devashish Gosain. Few throats to choke: On the current structure of the internet. In *2017 IEEE 42nd Conference on Local Computer Networks (LCN)*, pages 339–346. IEEE, 2017.

[19] Bernhard Ager, Wolfgang Mühlbauer, Georgios Smaragdakis, and Steve Uhlig. Web content cartography. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 585–600. ACM, 2011.

[20] Simurgh Aryan, Homa Aryan, and J Alex Halderman. Internet censorship in iran: A first look. In *Presented as part of the 3rd $USENIX$ Workshop on Free and Open Communications on the Internet*, 2013.

[21] Diogo Barradas, Nuno Santos, and Luís Rodrigues. Deltashaper: Enabling unobservable censorship-resistant tcp tunneling over videoconferencing streams. *Proceedings on Privacy Enhancing Technologies*, 2017(4):5–22, 2017.

[22] Armon Barton and Matthew Wright. Denasa: Destination-naive as-awareness in anonymous communications. *Proceedings on Privacy Enhancing Technologies*, 2016(4):356–372.

[23] Cecylia Bocovich and Ian Goldberg. Slitheen: Perfectly imitated decoy routing through traffic replacement. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1702–1714, 2016.

[24] Timm Böttger, Felix Cuadrado, Gareth Tyson, Ignacio Castro, and Steve Uhlig. Open connect everywhere: A glimpse at the internet ecosystem through the lens of the netflix cdn. *ACM SIGCOMM Computer Communication Review*, 48(1):28–34, 2018.

[25] Matt Calder, Xun Fan, Zi Hu, Ethan Katz-Bassett, John Heidemann, and Ramesh Govindan. Mapping the expansion of google's serving infrastructure. In *Proceedings of the 2013 conference on Internet measurement conference*, pages 313–326. ACM, 2013.

[26] Matt Calder, Ashley Flavel, Ethan Katz-Bassett, Ratul Mahajan, and Jitendra Padhye. Analyzing the performance of an anycast cdn. In *Proceedings of the 2015 Internet Measurement Conference*, pages 531–537. ACM, 2015.

[27] Abdelberi Chaabane, Terence Chen, Mathieu Cunche, Emiliano De Cristofaro, Arik Friedman, and Mohamed Ali Kaafar. Censorship in the wild: Analyzing internet filtering in syria. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, pages 285–298. ACM, 2014.

[28] Balakrishnan Chandrasekaran, Mingru Bai, Michael Schoenfield, Arthur Berger, Nicole Caruso, George Economou, Stephen Gilliss, Bruce Maggs, Kyle Moses, David Duff, et al.

Alidade: Ip geolocation without active probing. *Department of Computer Science, Duke University, Tech. Rep. CS-TR-2015.001*, 2015.

[29] David Choffnes, Jilong Wang, et al. Cdns meet cn an empirical study of cdn deployments in china. *IEEE Access*, 5:5292–5305, 2017.

[30] Danilo Cicalese, Danilo Giordano, Alessandro Finamore, Marco Mellia, Maurizio Munafò, Dario Rossi, and Diana Joumblatt. A first look at anycast cdn traffic. *arXiv preprint arXiv:1505.00946*, 2015.

[31] Ronald Deibert and Rafal Rohozinski. Liberation vs. control: The future of cyberspace. *Journal of Democracy*, 21(4):43–57, 2010.

[32] Amogh Dhamdhere and Constantine Dovrolis. The internet is flat: modeling the transition from a transit hierarchy to a peering mesh. In *Proceedings of the 6th International COnference*, page 21. ACM, 2010.

[33] Anne Edmundson, Roya Ensafi, Nick Feamster, and Jennifer Rexford. Nation-state hegemony in internet routing. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, COMPASS '18, pages 17:1–17:11, New York, NY, USA, 2018. ACM.

[34] Daniel Ellard, Christine Jones, Victoria Manfredi, W Timothy Strayer, Bishal Thapa, Megan Van Welie, and Alden Jackson. Rebound: Decoy routing on asymmetric routes via error messages. In *2015 IEEE 40th Conference on Local Computer Networks (LCN)*, pages 91–99. IEEE, 2015.

[35] Roya Ensafi, David Fifield, Philipp Winter, Nick Feamster, Nicholas Weaver, and Vern Paxson. Examining how the Great Firewall discovers hidden circumvention servers. In *Internet Measurement Conference*. ACM, 2015.

[36] Xun Fan, Ethan Katz-Bassett, and John Heidemann. Assessing affinity between users and cdn sites. In *International Workshop on Traffic Monitoring and Analysis*, pages 95–110. Springer, 2015.

[37] David Fifield, Chang Lan, Rod Hynes, Percy Wegmann, and Vern Paxson. Blocking-resistant communication through domain fronting. *Proceedings on Privacy Enhancing Technologies*, 2015(2):46–64, 2015.

[38] Sergey Frolov, Jack Wampler, Sze Chuen Tan, J Alex Halderman, Nikita Borisov, and Eric Wustrow. Conjure: Summoning proxies from unused address space. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 2215–2229, 2019.

[39] Lixin Gao. On inferring autonomous system relationships in the internet. *IEEE/ACM Transactions on Networking (ToN)*, 9(6):733–745, 2001.

[40] Manaf Gharaibeh, Anant Shah, Bradley Huffaker, Han Zhang, Roya Ensafi, and Christos Papadopoulos. A look at router geolocation in public and commercial databases. In *Proceedings of the 2017 Internet Measurement Conference*, pages 463–469. ACM, 2017.

[41] Devashish Gosain, Anshika Agarwal, Sambuddho Chakravarty, and HB Acharya. The devil's in the details: Placing decoy routers in the internet. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, pages 577–589. ACM, 2017.

[42] Bamba Gueye, Artur Ziviani, Mark Crovella, and Serge Fdida. Constraint-based geolocation of internet hosts. *IEEE/ACM Transactions on Networking (TON)*, 14(6):1219–1232, 2006.

[43] John Holowczak and Amir Houmansadr. Cachebrowser: Bypassing chinese censorship without proxies using cached content. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 70–83. ACM, 2015.

[44] Amir Houmansadr, Giang TK Nguyen, Matthew Caesar, and Nikita Borisov. Cirripede: Circumvention infrastructure using router redirection with plausible deniability. In *Proceedings of the 18th ACM conference on Computer and communications security*, pages 187–200, 2011.

[45] Amir Houmansadr, Thomas J Riedl, Nikita Borisov, and Andrew C Singer. I want my voice to be heard: Ip over voice-over-ip for unobservable censorship circumvention. In *NDSS*, 2013.

[46] Amir Houmansadr, Edmund L Wong, and Vitaly Shmatikov. No direction home: The true cost of routing around decoys. In *NDSS*, 2014.

[47] Bradley Huffaker, Marina Fomenkov, et al. Geocompare: a comparison of public and commercial geolocation databases-technical report. Technical report, Cooperative Association for Internet Data Analysis (CAIDA), 2011.

[48] Josh Karlin, Daniel Ellard, Alden W Jackson, Christine E Jones, Greg Lauer, David Mankins, and W Timothy Strayer. Decoy routing: Toward unblockable internet communication. In *FOCI*, 2011.

[49] Ethan Katz-Bassett, John P John, Arvind Krishnamurthy, David Wetherall, Thomas Anderson, and Yatin Chawathe. Towards ip geolocation using delay and topology measurements. In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, pages 71–84. ACM, 2006.

[50] Sheharbano Khattak, Tariq Elahi, Laurent Simon, Colleen M Swanson, Steven J Murdoch, and Ian Goldberg. Sok: Making sense of censorship resistance systems. *Proceedings on Privacy Enhancing Technologies*, 2016(4):37–61, 2016.

[51] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[52] Sándor Laki, Péter Mátray, Péter Hága, Tamás Sebők, István Csabai, and Gábor Vattay. Spotter: A model based active geolocation service. In *2011 Proceedings IEEE INFOCOM*, pages 3173–3181. IEEE, 2011.

[53] Philip Levis. The collateral damage of internet censorship by dns injection. *ACM SIGCOMM CCR*, 42(3), 2012.

[54] Matthew Luckie, Amogh Dhamdhere, Bradley Huffaker, David Clark, and KC Claffy. Bdrmap: Inference of borders between ip networks. In *Proceedings of the 2016 Internet Measurement Conference*, pages 381–396, 2016.

[55] Matthew Luckie et al. A second look at detecting third-party addresses in traceroute traces with the ip timestamp option. In *International Conference on Passive and Active Network Measurement*, pages 46–55. Springer, 2014.

[56] Matthew Luckie, Bradley Huffaker, Amogh Dhamdhere, Vasileios Giotsas, et al. As relationships, customer cones, and validation. In *Proceedings of the 2013 conference on Internet measurement conference*, pages 243–256. ACM.

[57] Kyle MacMillan, Jordan Holland, and Prateek Mittal. Evaluating snowflake as an indistinguishable censorship circumvention tool. *arXiv preprint arXiv:2008.03254*, 2020.

[58] Alexander Marder, Matthew Luckie, Amogh Dhamdhere, Bradley Huffaker, kc claffy, and Jonathan M Smith. Pushing the boundaries with bdrmapit: Mapping router ownership at internet scale. In *Proceedings of the Internet Measurement Conference 2018*, pages 56–69, 2018.

[59] Alexander Marder and Jonathan M Smith. Map-it: Multipass accurate passive inferences from traceroute. In *Proceedings of the 2016 Internet Measurement Conference*, pages 397–411, 2016.

[60] Richard McPherson, Amir Houmansadr, and Vitaly Shmatikov. Covertcast: Using live streaming to evade internet censorship. *Proceedings on Privacy Enhancing Technologies*, 2016(3):212–225, 2016.

[61] Milad Nasr, Hadi Zolfaghari, and Amir Houmansadr. The waterfall of liberty: Decoy routing circumvention that resists routing attacks. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 2037–2052, 2017.

[62] Venkata N Padmanabhan and Lakshminarayanan Subramanian. Determining the geographic location of internet hosts. In *SIGMETRICS/Performance*, pages 324–325, 2001.

[63] Roberto Percacci and Alessandro Vespignani. Scale-free behavior of the internet global performance. *The European Physical Journal B-Condensed Matter and Complex Systems*, 32(4):411–414, 2003.

[64] Enric Pujol, Philipp Richter, Balakrishnan Chandrasekaran, Georgios Smaragdakis, Anja Feldmann, Bruce MacDowell Maggs, and Keung-Chi Ng. Back-office web traffic on the internet. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, pages 257–270. ACM, 2014.

[65] Reethika Ramesh, Ram Sundara Ram, Matthew Bernhard, Victor Ongkowijaya, Leonid Evdokimov, Anne Edmundson, Steven Sprecher, Muhammad Ikram, and Roya Ensafi. Decentralized control: A case study of russia. In *Proceedings of the 2020 Conference on Network and Distributed System Security Symposium (NDSS)*. San Diego, Calafornia, 2020.

[66] Reethika Ramesh, Ram Sundara Raman, Matthew Bernhard, Victor Ongkowijaya, Leonid Evdokimov, Anne Edmundson, Steven Sprecher, Muhammad Ikram, and Roya Ensafi. Decentralized control: A case study of russia. In *Network and Distributed Systems Security Symposium (NDSS)*, 2020.

[67] Quirin Scheitle, Oliver Hohlfeld, Julien Gamba, Jonas Jelten, Torsten Zimmermann, Stephen D Strowes, and Narseo Vallina-Rodriguez. A long way to the top: significance, structure, and stability of internet top lists. In *Proceedings of the Internet Measurement Conference 2018*, pages 478–493. ACM, 2018.

[68] Max Schuchard, John Geddes, Christopher Thompson, and Nicholas Hopper. Routing around decoys. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 85–96. ACM, 2012.

[69] Will Scott, Thomas Anderson, Tadayoshi Kohno, and Arvind Krishnamurthy. Satellite: Joint analysis of cdns and network-level interference. In *2016 {USENIX} Annual Technical Conference* 16, pages 195–208, 2016.

[70] Piyush Kumar Sharma, Devashish Gosain, Himanshu Sagar, Chaitanya Kumar, Aneesh Dogra, Vinayak Naik, HB Acharya, and Sambuddho Chakravarty. Siegebreaker: An sdn based practical decoy routing system. *Proceedings on Privacy Enhancing Technologies*, 3:243–263, 2020.

[71] Volker Stocker, Georgios Smaragdakis, William Lehr, and Steven Bauer. Content may be king, but (peering) location matters: A progress report on the evolution of content delivery in the internet. 2016.

[72] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[73] Paul Syverson, R Dingledine, and N Mathewson. Tor: The second generation onion router. In *Usenix Security*, 2004.

[74] Michael Carl Tschantz, Sadia Afroz, Vern Paxson, et al. Sok: Towards grounding censorship circumvention in empiricism. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 914–933. IEEE, 2016.

[75] Yong Wang, Daniel Burgener, Marcel Flores, Aleksandar Kuzmanovic, and Cheng Huang. Towards street-level client-independent ip geolocation. In *NSDI*, volume 11, pages 27–27, 2011.

[76] Zachary Weinberg, Shinyoung Cho, Nicolas Christin, Vyas Sekar, and Phillipa Gill. How to catch when proxies lie: Verifying the physical locations of network proxies with active geolocation. In *Proceedings of the Internet Measurement Conference 2018*, pages 203–217. ACM, 2018.

[77] Florian Wohlfart, Nikolaos Chatzis, Caglar Dabanoglu, Georg Carle, and Walter Willinger. Leveraging interconnections for performance: the serving infrastructure of a large cdn. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, pages 206–220.

[78] Florian Wohlfart, Nikolaos Chatzis, Caglar Dabanoglu, Georg Carle, and Walter Willinger. Leveraging interconnections for performance: The serving infrastructure of a large cdn. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, SIGCOMM '18, pages 206–220, New York, NY, USA, 2018. ACM.

[79] Bernard Wong, Ivan Stoyanov, and Emin Gün Sirer. Octant: A comprehensive framework for the geolocalization of internet hosts. In *NSDI*, volume 7, pages 23–23, 2007.

[80] Eric Wustrow, Colleen M Swanson, and J Alex Halderman. Tapdance: End-to-middle anticensorship without flow blocking. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, pages 159–174, 2014.

[81] Eric Wustrow, Scott Wolchok, Ian Goldberg, and J Alex Halderman. Telex: Anticensorship in the network infrastructure. In *USENIX Security Symposium*, page 45, 2011.

[82] Jing'an Xue, David Choffnes, and Jilong Wang. Cdns meet cn an empirical study of cdn deployments in china. *IEEE Access*, 5:5292–5305, 2017.

[83] Tarun Kumar Yadav, Akshat Sinha, Devashish Gosain, Piyush Kumar Sharma, and Sambuddho Chakravarty. Where the light gets in: Analyzing web censorship mechanisms in india. In *Proceedings of the Internet Measurement Conference 2018*, pages 252–264. ACM, 2018.

[84] Bahador Yeganeh, Reza Rejaie, and Walter Willinger. A view from the edge: A stub-as perspective of traffic localization and its implications. In *Network Traffic Measurement and Analysis Conference (TMA)*, pages 1–9. IEEE, 2017.

[85] Hadi Zolfaghari and Amir Houmansadr. Practical censorship evasion leveraging content delivery networks. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1715–1726. ACM, 2016.

A Appendix

A.1 RTT Profiles for RIPE Probes in Various Countries

Figures 19, 20, 21, 22 represent the SVM line (threshold) for most consistent reference nodes in India, Saudi Arabia, Brazil and Unites States.

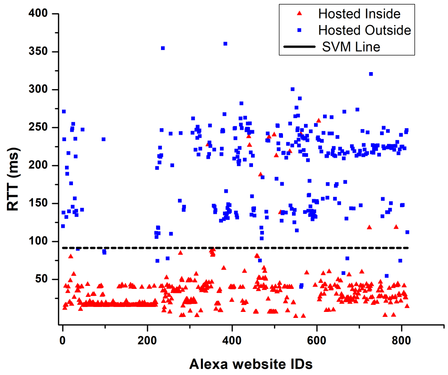

Fig. 19. RTT scatter plot for a probe in India.

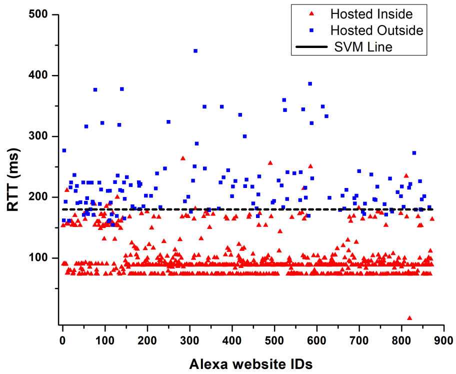

Fig. 20. RTT scatter plot for a probe in Saudi Arabia.

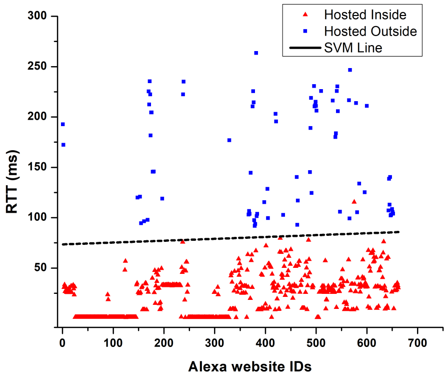

Fig. 21. RTT scatter plot for a probe in Brazil.

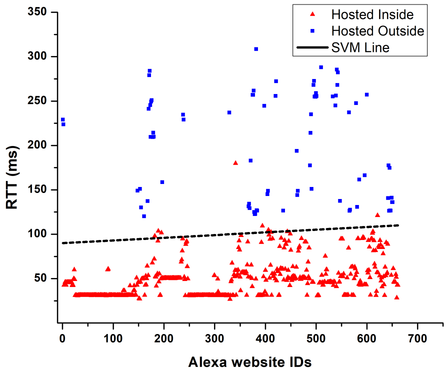

Fig. 22. RTT scatter plot for a probe in United States.

A.2 Validating our CDN Classification Scheme

In §4.4, we explained our approach to identify different types of hosting infrastructure being used by popular Alexa websites (*i.e.,* CDN or non-CDN).

We now describe how we additionally validate our CDN classification scheme. It is already known that Akamai uses DNS based CDNs [30], whereas Cloudflare uses anycast based CDNs [77, 78]. We leverage this information to gauge the accuracy of our approach to identify different type of hosting infrastructures. Our classification approach would yield 100% accuracy, if all the websites that are hosted on Akamai would be classified as, hosted on "DNS based CDN", whereas websites hosted on Cloudflare would be classified as hosted on "anycast based CDN".

To begin with, we obtained the names of the organizations of the hosting infrastructure (*e.g.,* Cloudflare) for Alexa top-1k websites, using Team Cymru's [10] IP to ASN mapping.

From ASNs we obtained their organization names using Caida dataset [3]. Figures 23, 24, 25, 26, 27 depict the hosting infrastructure of country specific Alexa top-1k websites. It is evident that popular CDNs like Cloudflare, Akamai, Amazon, Google and Microsoft *etc.* host a large fraction of these websites. *E.g.,* Akamai hosts ≈ 1.6% (*i.e.* 16) of SA's Alexa top-1k websites, while Cloudflare alone hosts more than 33% of them (*i.e.* 332). Using our method (ref. Subsec. 4.4), we classified the 16 (out of 16) websites (that were hosted on Akamai) to be using DNS based infrastructure. Similarly, among 332 websites hosted on Cloudflare, we correctly identified 328 to be using anycast based CDN and only

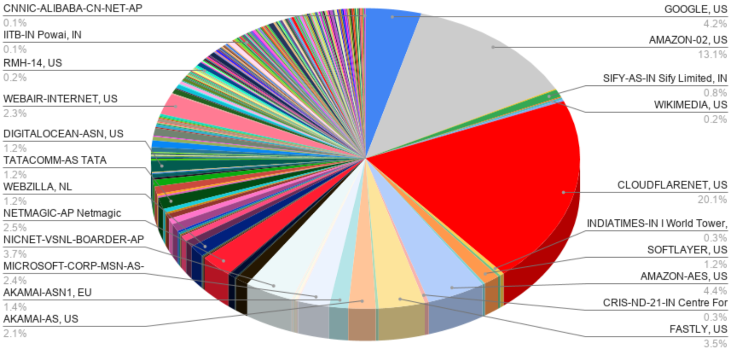

Fig. 23. Hosting Infrastructure of IN specific Alexa top-1k websites.

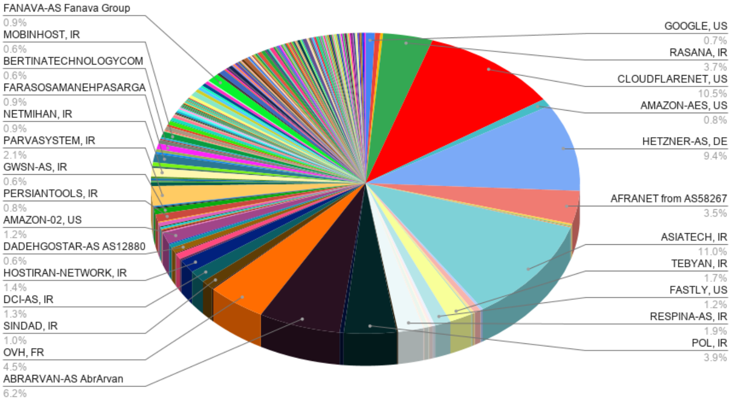

Fig. 24. Hosting Infrastructure of Alexa top-1k websites specific to IR.

Country	Cloudflare Websites predicted as anycast based (in %)	Akamai Websites predicted as DNS based (in %)
IN	100	97.23
IR	100	*
BR	99.01	100
SA	99.2	100
US	100	97.5

Table 2. Percentages of websites hosted on Cloudflare as anycast based and those which are hosted on Akamai as DNS based. (*) represents that for IR, none of the Alexa top-1k websites are hosted on Akamai, and is thus not applicable in this validation tests.

four to be using non-CDN infrastructure. We obtained such promising results for other countries as well (summarized in Table 2). Thus, by leveraging the knowledge that Akamai relies on DNS based CDN, and Cloudflare uses anycast based CDN, we validate our CDN classification approach with more than 97% accuracy for all cases.

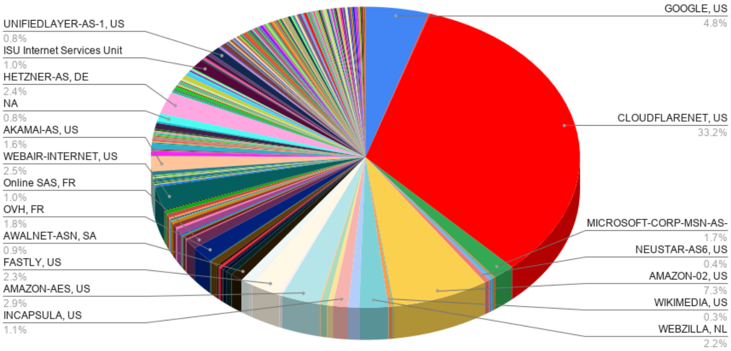

Fig. 25. Hosting Infrastructure of Alexa top-1k websites specific to Saudi Arabia.

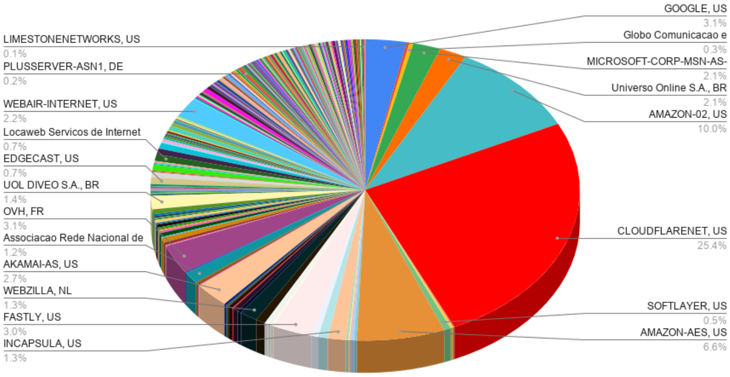

Fig. 26. Hosting Infrastructure of BR specific Alexa top-1k websites.

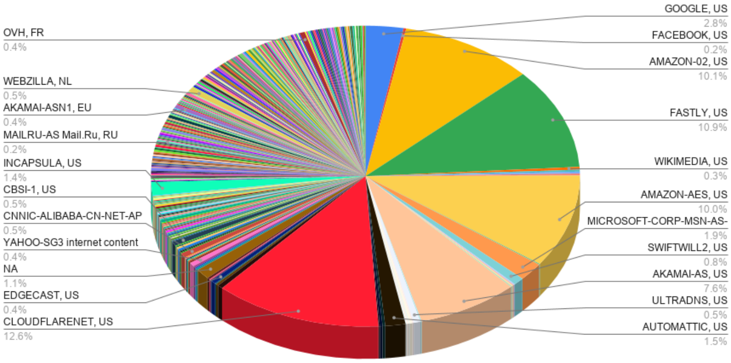

Fig. 27. Hosting Infrastructure of US specific Alexa top-1k websites.