Towards Measuring Content Locality

James I. Madeley
Loughborough University
j.i.madeley@lboro.ac.uk

Amreesh Phokeer
Internet Society
phokeer@isoc.org

Aftab Siddiqui
Internet Society
siddiqui@isoc.org

Theophilus A. Benson
Carnegie Mellon University
theophilus@cmu.edu

ABSTRACT

Local access to Internet content is generally known to be cheaper and faster, but a lack of local content affects users' ability to access other resources, for example in cases where utility bills are paid online. However, despite this importance, we lack a comprehensive framework to perform such content locality analysis. In this work, we present a framework and provide preliminary evidence to support its effectiveness. To quantify the extent to which Internet traffic stays local in a country, we use the list of the Top 1000 most popular websites from Google's Chrome User Experience Report (CrUX) as a proxy to measure traffic locality. Our solution removes censored sites per country and determines whether the remaining websites are hosted natively on a server or on a Content Delivery Network (CDN) cache. We then find the location of the server or CDN cache using a mix of techniques, including geo-hints extraction, IP geolocation, and website location sourcing.

ACM Reference Format:
James I. Madeley, Amreesh Phokeer, Aftab Siddiqui, and Theophilus A. Benson. 2024. Towards Measuring Content Locality. In *Applied Networking Research Workshop (ANRW '24), July 23, 2024, Vancouver, AA, Canada.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3673422.3674895

1 INTRODUCTION

Traffic locality, i.e., placement and delivery of content within the same region as the users, is crucial from a performance, cost, and availability standpoint. In fact, a recent outage in West Africa [1] illustrated the importance of content locality because the lack of local services made citizens unable to use local utilities (e.g., citizens were unable to pay for electricity and were left without power). Similarly, businesses were unable to operate due to the lack of online services. Regarding the performance implications of traffic locality, previous work has shown that reducing the distance between CDN servers and end-users improves the performance of the Web [2]. Despite the importance of keeping local traffic local and the many efforts to address this issue, in places like Sub-Saharan Africa, the most popular content is still being fetched from Europe or America [3].

Measuring traffic locality is challenging for several reasons: first, content providers and CDNs employ complex mechanisms to redirect users to the closest content, thus we need a local geographic footprint to accurately determine location. Second, even with a local vantage point, we need to infer the location of a server using a combination of techniques, with varying degrees of accuracy. Finally, while the content may be local, routing may be non-local if it traverses an out-of-country region due to a lack of peering between operators and CDN networks [4].

In this paper, we propose a methodology for determining traffic locality, introduce an initial prototype of this methodology, and present a preliminary measurement method of traffic locality.

2 PLATFORM DESIGN

While there has been prior work on analyzing traffic location [5–7] and a myriad of existing tools for IP Geolocation [8–10], such techniques are susceptible to geographic bias due to the limited number of vantage points used to perform their scans. We therefore look at a novel framework that combines multiple techniques to fulfill the following requirements:

Requirement #1 The system should use vantage points that are local to the end-users to perform scans from the network/country of interest.

Requirement #2 The system should perform latency measurements from local vantage points towards the target server (i.e. where content is located).

Requirement #3 The system should be able to detect the hops on the path (e.g. transit providers, peers or IXPs) from a local vantage point towards the target server.

ANRW '24, July 23, 2024, Vancouver, AA, Canada

ACM ISBN 979-8-4007-0723-0/24/07.
https://doi.org/10.1145/3673422.3674895

Region	Website Locality
Africa	13%
Americas	21%
Asia	42%
Europe	40%
Oceania	16%

Table 1: Average Website Locality by Region

As discussed in Section 1, the location of seemingly 'local' content is not necessarily local to the end-user. Local networks either peer directly with the externally hosted content providers or use transit links to provide their customers access to the content.

3 METHODOLOGY

Our methodology consists of the following four steps:

Step 1: First, our framework captures a list of the top 1000 most popular websites by country from the Chrome User Experience Report (CrUX) [11]. There are many such curated databases of popular domains, including Cloudflare Radar [12], Tranco [13], and Similarweb [14]. While our framework is sufficiently modular to support any such database, we use CrUX as it is known to be the most accurate toplist [15].

Step 2: Next, we sanitize the list, again by country, to remove any censored websites. In particular, we pay careful attention to ethics because we plan to leverage residential proxies (RESIP) (e.g., Oxylabs [16] and Bright Data [17]) to run measurements. RESIP platforms allow you to run measurements by selecting a local vantage point within a network in a country (where probes are present). This provides a quasi-real-world experience from the end-user perspective. We use the CitizenLab [18] list to identify censored sites.

Step 3: Then, for each domain we determine which CDN provider is hosting the website. To achieve this, we extract the IP address and perform a WHOIS lookup, then a DNS CNAME lookup, and finally, we extract the "server" tag from HTTPS response header which sometimes contains the name of the CDN. In the case a website is not hosted using a CDN, we categorize it as "natively hosted". Our CDN detection methods are inspired by the FindCDN tool [19].

Step 4: Finally, we geolocate the CDN or application cache of a domain by using a mix of techniques. A big majority of CDNs would provide geo-hints in their HTTP Response header. We leverage this technique to determine location. Huang *et al.* [20] used a similar technique in their demo. Some applications such as online streaming services or social media would use their own caching mechanism but still provide geo-hints in either the HTTP Response header or within the body itself. In cases where no geo-hints are available, we curate the location information directly from the CDN websites.

Our core methodology provides baseline measurements of whether when accessing a website, the traffic remains local or goes through external links. As part of future work, we plan to extend the framework to use BGP control-plane information to improve the accuracy and precision of our tool.

4 PRELIMINARY RESULTS AND EXTENSIONS

Our methodology is reliant on the availability of residential proxy endpoints in the country we are measuring. Figure 1 shows the average locality results across the five regions. Our results can be used to give an indication of which countries, CDNs, and website categories provide the most local content, as well as information about changing locality as measurements are repeated over time.

Website complexity: Currently, we do not consider the size and origin of web objects that constitute a website. Prior work shows that $\tilde{3}5\%$ of bytes downloaded come from different sources across more than 60% of websites [21]. We plan to extend our methodology to consider embedded data by conducting recursive measurements on links within websites, giving a more accurate picture of the composition of a website.

Network latency: Once the user request reaches a CDN provider, the content may not be fetched from a local cache (i.e. cache-miss). We will complement our use of geo-hints with RTT-based geolocation, e.g., work by Patel *et al.* [22]. Using RTT measurements, we can more confidently determine whether a local destination is reached.

Network Paths: We can also explore the use of AS paths to infer locality. We can analyze the AS path used to reach a location from a given starting point and then geolocate each AS on the path to determine whether traffic is likely to leave that country to reach a destination.

5 CONCLUSION

We are creating a measurement tool that can be used periodically to determine how many of the most popular domains within a country are hosted and served locally. The results can be visualized and compared with previous results to see how the landscape of traffic locality changes with time. Traffic samples can then be used to estimate how much traffic per country is accessed locally. The work can be extended by using different measurement techniques. We also intend to create a publicly available platform for viewing and visualizing the results.

REFERENCES

[1] Internet Society. Major internet outages across western and southern africa today, 2024.

[2] Adnan Ahmed, Zubair Shafiq, Harkeerat Bedi, and Amir Khakpour. Peering vs. transit: Performance comparison of peering and transit interconnections. In *2017 IEEE 25th International Conference on Network Protocols (ICNP)*, pages 1–10. IEEE, 2017.

[3] Enrico Calandro, Josiah Chavula, and Amreesh Phokeer. Internet development in africa: a content use, hosting and distribution perspective. In *e-Infrastructure and e-Services for Developing Countries: 10th EAI International Conference, AFRICOMM 2018, Dakar, Senegal, November 29-30, 2019, Proceedings 10*, pages 131–141. Springer, 2019.

[4] Josiah Chavula, Nick Feamster, Antoine Bagula, and Hussein Suleman. Quantifying the effects of circuitous routes on the latency of intra-africa internet traffic: a study of research and education networks. In *e-Infrastructure and e-Services for Developing Countries: 6th International Conference, AFRICOMM 2014, Kampala, Uganda, November 24-25, 2014, Revised Selected Papers 6*, pages 64–73. Springer, 2015.

[5] Bernhard Ager, Wolfgang Mühlbauer, Georgios Smaragdakis, and Steve Uhlig. Web content cartography. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '11, page 585–600, New York, NY, USA, 2011. Association for Computing Machinery.

[6] Mingkui Wei, Khaled Rabieh, and Faisal Kaleem. Cacheloc: Leveraging cdn edge servers for user geolocation. In *International Conference on Security and Privacy in Communication Systems*, pages 22–40. Springer, 2020.

[7] Kevin Vermeulen, Loqman Salamatian, Sang Hoon Kim, Matt Calder, and Ethan Katz-Bassett. The central problem with distributed content: Common cdn deployments centralize traffic in a risky way. In *Proceedings of the 22nd ACM Workshop on Hot Topics in Networks*, pages 70–78, 2023.

[8] Eris Networks S.A.S. Ip geolocation api & free address database | db-ip, 2024.

[9] Geobytes. Ip address locator | geobytes, 2019?

[10] IP2Location. Ip address to ip location and proxy information | ip2location, 2024.

[11] Google. Overview of crux, 2024.

[12] Inc Cloudflare. Cloudflare radar, 2024.

[13] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczynski, and Wouter Joosen. Tranco: A research-oriented top sites ranking hardened against manipulation. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS)*, NDSS 2019, Reston, VA, USA, 2019. Internet Society.

[14] Similarweb LTD. Similarweb rank - top websites by country and category, 2024.

[15] Kimberly Ruth, Deepak Kumar, Brandon Wang, Luke Valenta, and Zakir Durumeric. Toppling top lists: evaluating the accuracy of popular website lists. In *Proceedings of the 22nd ACM Internet Measurement Conference*, IMC '22, page 374–387, New York, NY, USA, 2022. Association for Computing Machinery.

[16] oxylabs.io. Oxylabs - premium proxy service to gather data at scale, 2024.

[17] Bright Data Ltd. Bright data - all in one platform for proxies and web data, 2024.

[18] Citizen Lab and Others. Url testing lists intended for discovering website censorship, 2014. https://github.com/citizenlab/test-lists.

[19] FindCDN GitHub Contributors. Findcdn, 2024.

[20] Run Huang, Mengying Zhou, Tiancheng Guo, and Yang Chen. Locating cdn edge servers with http responses. In *Proceedings of the SIGCOMM '22 Poster and Demo Sessions*, SIGCOMM '22, page 19–21, New York, NY, USA, 2022. Association for Computing Machinery.

[21] Michael Butkiewicz, Harsha V. Madhyastha, and Vyas Sekar. Understanding website complexity: measurements, metrics, and implications. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '11, page 313–328, New York, NY, USA, 2011. Association for Computing Machinery.

[22] Kishan B. Patel, Nadine Moukdad, and S. Anand. Geolocation of ip hosts in large computer networks with congestion. In *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, pages 1–6, Operations Center, Building 1, 445 Hoes Lane, Piscataway, NJ, USA, 2020. IEEE.