Towards Measuring Content Locality

James I. Madeley, Amreesh Phokeer, Theophilus A. Benson, Aftab Siddiqui

ANRW’24
23rd July 2024
Local Traffic

User → Local Server → Local Server → Local Government Website

International Network
External Traffic

User → Local Server → Local Server → Local Government Website

International Network
Previous Works on CDN Deployment

Web Content Cartography

Bernhard Ager
T-Labs/TU Berlin
bernhard@net.t-labs.tu-berlin.de

Georgios Smaragdakis
T-Labs/TU Berlin
georgios@net.t-labs.tu-berlin.de

The Central Problem with Distributed Content
Common CDN Deployments Centralize Traffic In A Risky Way

Kevin Vermeulen
Loqman Salamatian
Sang Hoon Kim
Columbia University
Columbia University
Columbia University

CacheLoc: Leveraging CDN Edge Servers for User Geolocation

Wolfgang Mühlbauer


1 Cyber Forensics Intelligent Center, Computer Science, Sam Houston State University, Huntsville, TX
2 Computer Science and Cybersecurity, Metropolitan State University, Saint Paul, MN

Internet development in hosting and distribution

Enrico Calandro1, Josiah Chavula2, and Amreesh Phoiker

1 Research ICT Africa, Cape Town, South Africa
ecalandro@researchictafrica.net
2 University of Cape Town, Cape Town, South Africa
jchavula@cs.uct.ac.za
3 AFRINIC, Ebene, Mauritius
amreesh@afринic.net

Peering vs. Transit: Performance Comparison of Peering and Transit Interconnections

Adnan Ahmed and Zubair Shafiq
The University of Iowa

Harkeerat Bedi and Amir Khakpour
Verizon Digital Media Services
Previous Works on Geolocation

Locating CDN Edge Servers with HTTP Responses
Run Huang, Mengying Zhou, Tiar
Shanghai Key Lab of Intelligent Information Processing, Schc
{runhuang19,myzhou19,tcguo20,ch

Geolocation of IP Hosts in Large Computer Networks with Congestion
Kishan R. Patel
Nadine Moukdad
Computer Science University,
NY 10458
mukdad@gmail.com
S. Anand
Department of ECE
New York Institute of Technology
New York, NY 10023
Email: asanthan@nyit.edu

Towards Geolocation of Millions of IP Addresses
Zi Hu John Heidemann Yuri Pradkin
USC/Information Sciences Institute {zihu, johnh, yuri}@isi.edu
Motivation

Policy | Performance | Persistence
       |             | (availability)
Motivation

Policy
- South Africa’s POPI Act / EU’s GDPR

Performance

Persistence
(availability)

Internet Society’s 50/50 Vision
Motivation

Policy

South Africa’s POPI Act / EU’s GDPR

Internet Society’s 50/50 Vision

Performance

Live streaming favours local content

Persistence

(availability)
Motivation

Policy

South Africa’s POPI Act / EU’s GDPR

Internet Society’s 50/50 Vision

Performance

Live streaming favours local content

Persistence (availability)

Resilience against faults e.g. cable cuts
Methodology
Methodology

1. Get top 1,000 websites per country
Methodology

1. Get top 1,000 websites per country

Google Chrome Report

2. Remove censored websites

Censorship Database
Methodology

1. Get top 1,000 websites per country

2. Remove censored websites

3. Determine website hosting

DNS, WHOIS, Headers
Methodology

1. Get top 1,000 websites per country

2. Remove censored websites

3. Determine website hosting

4. Geolocate CDNs/domains

Google Chrome Report

Censorship Database

RESIPs, CDN DBs, Market Share

DNS, WHOIS, Headers
### Methodology – Determine Website Hosting

**www.studysmarter.co.uk, NL (Netherlands)**

<table>
<thead>
<tr>
<th>Service</th>
<th>Returned</th>
<th>Parsed</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHOIS</td>
<td>CDN77 _, GB</td>
<td>CDN77</td>
</tr>
<tr>
<td>DNS CNAME</td>
<td>prod-web-uk.b-cdn.net.</td>
<td>BunnyCDN</td>
</tr>
<tr>
<td><strong>Result</strong></td>
<td><strong>Tie</strong></td>
<td></td>
</tr>
</tbody>
</table>
Methodology – Determine Website Hosting

www.studysmarter.co.uk, NL (Netherlands)

<table>
<thead>
<tr>
<th>Service</th>
<th>Returned</th>
<th>Parsed</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHOIS</td>
<td>CDN77 _, GB</td>
<td>CDN77</td>
</tr>
<tr>
<td>DNS CNAME</td>
<td>prod-web-uk.b-cdn.net.</td>
<td>BunnyCDN</td>
</tr>
</tbody>
</table>

Result: Tie

<table>
<thead>
<tr>
<th>Headers</th>
<th>BunnyCDN-DE1-1080</th>
<th>BunnyCDN</th>
</tr>
</thead>
</table>

Result: BunnyCDN
Methodology – Geolocate Domains without CDNs

wikipedia.org, NL (Netherlands)
Methodology – Geolocate Domains without CDNs

wikipedia.org, NL (Netherlands)
Methodology – Geolocate Domains without CDNs

wikipedia.org, NL (Netherlands)

DNS
Cloudflare
Google
Quad9

Domain
wikipedia.org

IP Address
185.15.59.224
Methodology – Geolocate Domains without CDNs

wikipedia.org, NL (Netherlands)
Methodology – Geolocate Domains without CDNs

wikipedia.org, NL (Netherlands)
Methodology – Geolocate CDNs
Methodology – Geolocate CDNs

CDN → ASN → Residential Proxy → Search String
Methodology – Geolocate CDNs

CDN → ASN → Residential Proxy ← Search String

Residential Proxy → Headers, Body
Methodology – Geolocate CDNs

1. Residential Proxy
   - CDN
   - ASN
   - Search String

2. Headers, Body

3. REGEX

4. Geo-hints
Methodology – Geolocate CDNs

- Residential Proxy
- Search String
- CDN
- ASN
- Headers, Body
- REGEX
- Geo-hints
- Locality
Methodology – Geolocate CDNs

- Residential Proxy
- CDN
- ASN
- Headers, Body
- REGEX
- Geo-hints
- Locality
- Search String
- Results
Methodology – Geolocate CDNs

BunnyCDN, AS 200, NL (Netherlands)
Methodology – Geolocate CDNs

BunnyCDN, AS 200, NL (Netherlands)
Methodology – Geolocate CDNs

BunnyCDN, AS 200, NL (Netherlands)

![Diagram showing the process of geolocating CDNs]

- **Residential Proxy**
  - **Headers, Body**
  - **REGEX**
  - **Geo-hints**
  - **Locality**

**Search String**

Methodology – Geolocate CDNs

BunnyCDN, AS 200, NL (Netherlands)

BunnyCDN, 200
CDN   ASN

Residential Proxy

Headers, Body

REGEX

Geo-hints

Locality

Results

Search String

Methodology – Geolocate CDNs

BunnyCDN, AS 200, NL (Netherlands)

BunnyCDN 200
CDN ASN

Residential Proxy

Search String

BunnyCDN-AMS1-725

Headers, Body

REGEX

Geo-hints

Locality

Results
Methodology – Geolocate CDNs

BunnyCDN, AS 200, NL (Netherlands)

BunnyCDN
CDN
200
ASN

Residential Proxy

Headers, Body

REGEX

Geo-hints

Locality

Search String

Methodology – Geolocate CDNs

BunnyCDN, AS 200, NL (Netherlands)

BunnyCDN 200
CDN ASN

BunnyCDN-AMS1-725
Headers, Body

REGEX
Geo-hints
Locality

Search String

Results

Methodology – Geolocate CDNs

BunnyCDN, AS 200, NL (Netherlands)

BunnyCDN 200
CDN ASN

Residential Proxy

Search String

Headers, Body

REGEX

Geo-hints

Locality

Results

BunnyCDN, AS 200, Local

Methodology – Combining Results
Methodology – Combining Results

<table>
<thead>
<tr>
<th>CDN</th>
<th>ASN</th>
<th>Market Share</th>
<th>Local (Y/N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BunnyCDN</td>
<td>AS 200</td>
<td>50%</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>AS 250</td>
<td>50%</td>
<td>N</td>
</tr>
</tbody>
</table>
## Methodology – Combining Results

<table>
<thead>
<tr>
<th>CDN</th>
<th>ASN</th>
<th>Market Share</th>
<th>Local (Y/N)</th>
<th>Local Domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>BunnyCDN</td>
<td>AS 200</td>
<td>50%</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AS 250</td>
<td>50%</td>
<td>N</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hosting Method</th>
<th>Domain Count</th>
<th>Locality</th>
<th>Local Domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>BunnyCDN</td>
<td>10</td>
<td>50%</td>
<td>5</td>
</tr>
<tr>
<td>Native</td>
<td>5</td>
<td>20%</td>
<td>1</td>
</tr>
</tbody>
</table>
## Methodology – Combining Results

<table>
<thead>
<tr>
<th>CDN</th>
<th>ASN</th>
<th>Market Share</th>
<th>Local (Y/N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BunnyCDN</td>
<td>AS 200</td>
<td>50%</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>AS 250</td>
<td>50%</td>
<td>N</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hosting Method</th>
<th>Domain Count</th>
<th>Locality</th>
<th>Local Domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>BunnyCDN</td>
<td>10</td>
<td>50%</td>
<td>5</td>
</tr>
<tr>
<td>Native</td>
<td>5</td>
<td>20%</td>
<td>1</td>
</tr>
</tbody>
</table>

**Country Locality**: 40%
Limitations
Limitations

Website Complexity
Limitations

Website Complexity

Proxy Probes

Source: https://brightdata.com/locations
Limitations

Website Complexity
Proxy Probes
“Fog of Cloud”
Preliminary Results – Regional Locality

The chart above illustrates the website locality (% Local vs External) across different regions:

- **Africa**: 80% Local, 20% External
- **Americas**: 80% Local, 20% External
- **Asia**: 80% Local, 20% External
- **Europe**: 80% Local, 20% External
- **Oceania**: 80% Local, 20% External
Preliminary Results – CDN Locality (World)
Preliminary Results – Hosting Type (World)

![Hosting Type, World, 2024-06-07](image-url)

- **Africa**: CDN = 40%, Native = 60%
- **Americas**: CDN = 40%, Native = 60%
- **Asia**: CDN = 40%, Native = 60%
- **Europe**: CDN = 40%, Native = 60%
- **Oceania**: CDN = 40%, Native = 60%
Preliminary Results – 5 Most Popular CDNs (World)
Preliminary Results – Categories Locality (USA, 100 Sites)
Future Work
Future Work

Recursive Search

Traceroute/Latency
Future Work

Recursive Search

Reasons for Locality

Traceroute/Latency

Visualisation
Future Work

Recursive Search

Reasons for Locality

Categorization

Traceroute/Latency

Visualisation
Summary

Local traffic is (often) beneficial

Conducting global Internet traffic locality measurements

Expanding methodology

Any questions?

Check out the paper

Learn about the 50/50 Vision