Measuring Content Locality: An Ensemble of Methods

Martin Thodi[1], Marihno Barcellos[2], Amreesh Phokeer

[1]Net4D Lab, University Of Cape Town

[2]University of Waikato, New Zealand

Content locality

- extent to which Internet content/traffic is served from within a user's country/region.
- Includes
  - locally hosted content
  - in-country CDN delivery.

ISOC 50/50 Vision

keep at least half of all Internet traffic in selected economies local

Why it matters?

- Better performance through reduced latency.
- Lower ISP costs by reducing international transit.
- Improved resilience (e.g., 2024 West Africa cable outage).
- Supports digital economic development.)

Approach

Measurement Challenge

- AS-level Traffic Volumes are confidential
- From outside: CDN routing mechanisms e.g., Anycast obscure

This work

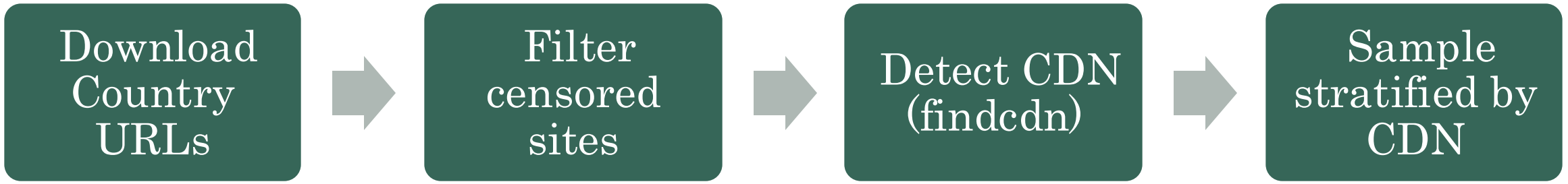
- Use web content as proxy for Internet Research
  - **How hosted?** Native, Content Delivery Networks (CDNs), Mix of both
- Ensemble approach
  - DNS localization, CDN geo-hints,
  - Streaming services handled differently
- Measure from end user vantage points (residential proxies)

Research Questions

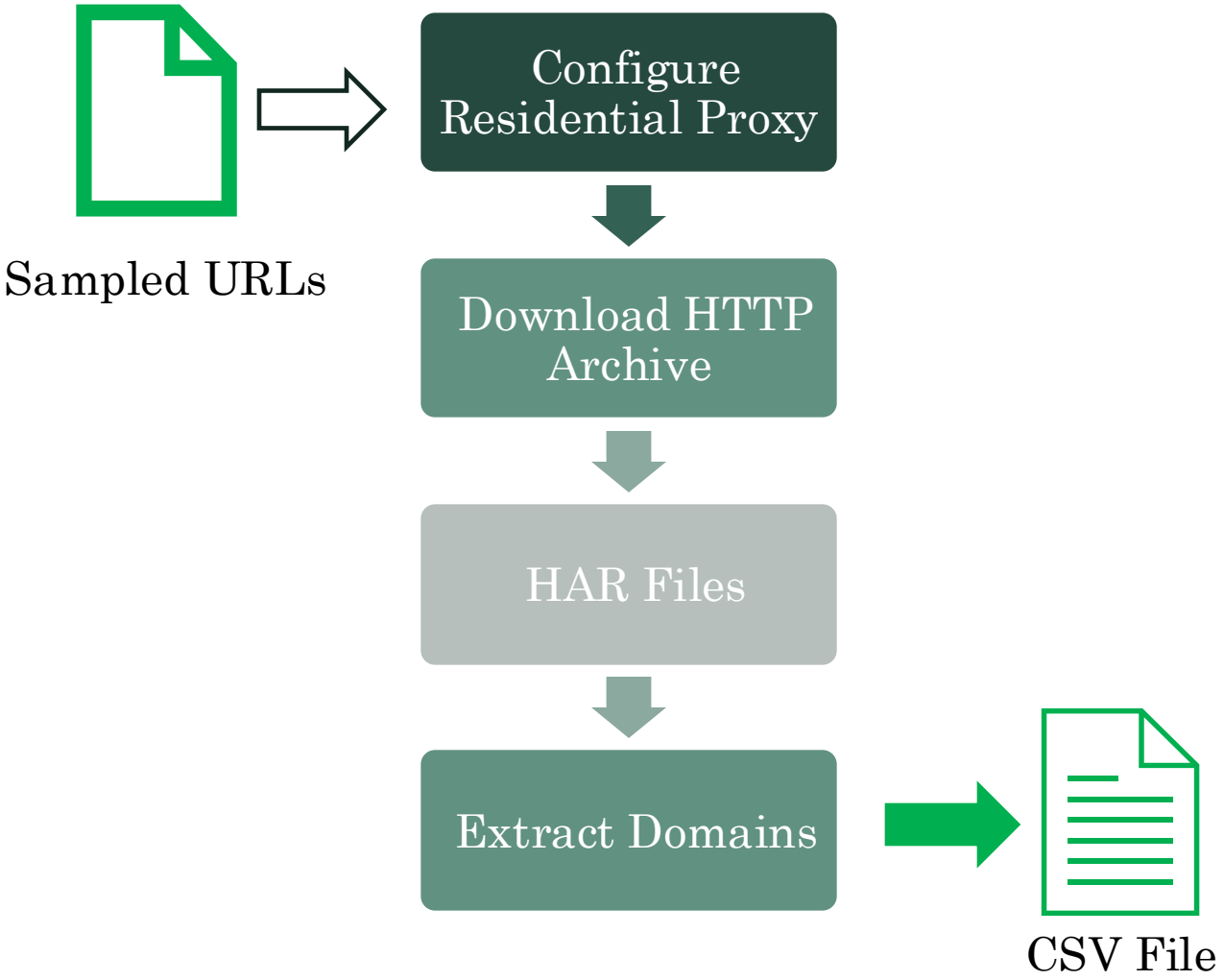
- **RQ1**: To what extent do content delivery mechanisms localize traffic for users in different countries and autonomous systems, and how does locality vary across CDN providers, content types, and geographic regions?
- **RQ2**: How do geo-hints-based and DNS-based methods compare in measuring content locality, and what systematic differences exist in their locality estimates across different CDN architectures and deployment strategies?

Step #1 – Preparing URLS

- Use web content as a proxy for Internet Traffic
  - Measured from user vantage points (using Residential Proxies)
- Google's Chrome User Experience (CrUX) Report
  - Get top 1000 sites in a country



Step #2 – Capturing & Processing HTTP Content



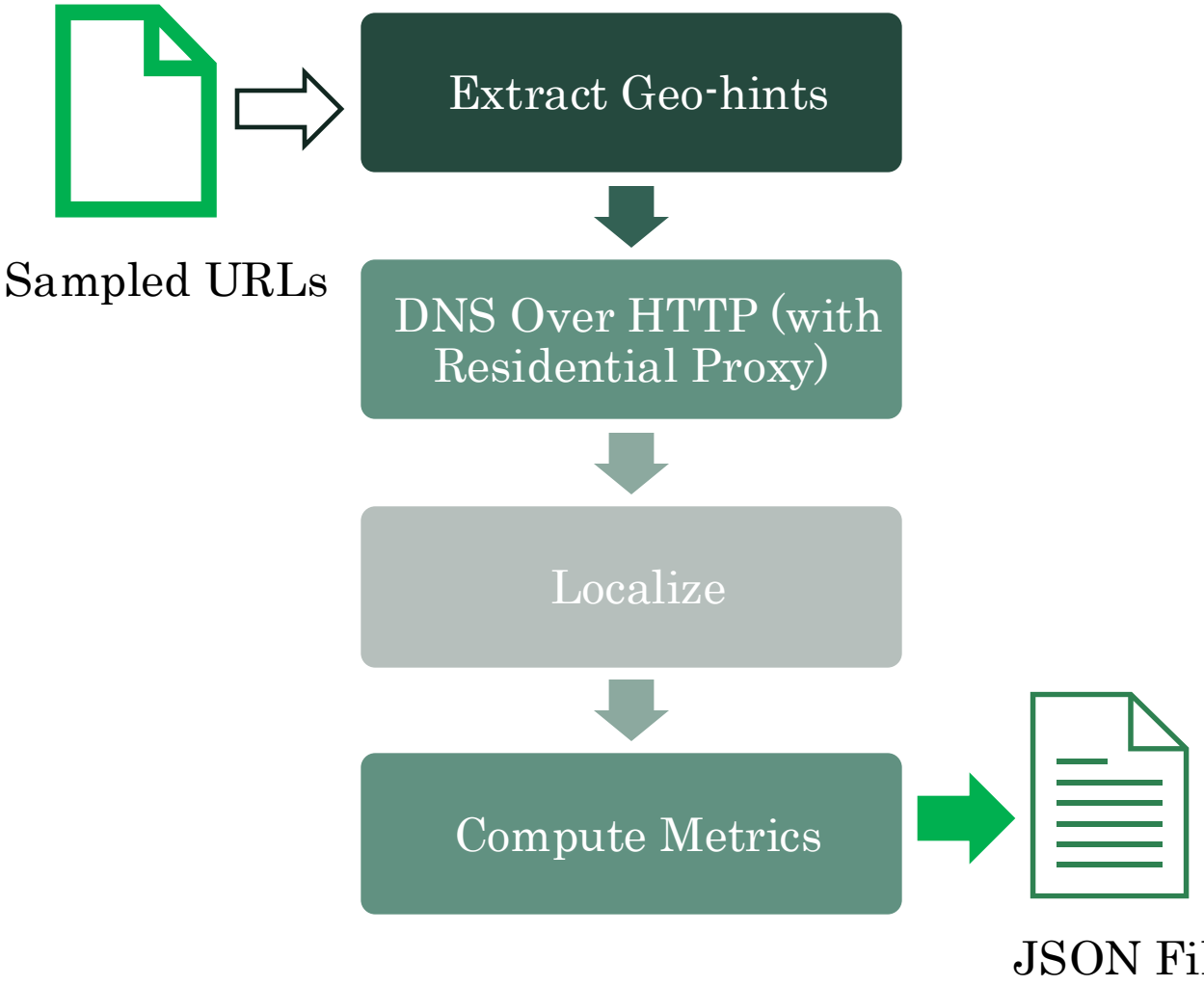
HTTP Archive Downloading

- Country or ASN vantage points
  - APNIC's ASN User Population Dataset for ASN targeting
- Browser automation (Playwright)
- Recursive Crawl
  - **3** levels deep, **5** endpoints per Level
  - No object saving (images/style sheets etc)

Processing Output

- Embedded domain extraction
- Content size and type
- Timing information
- CDN Identification

Step #3 – Localization



Geo-hints Method

- Extract IATA codes from CDN responses
- CDN specific Rules + Regex parsing
- Very **reliable** but **low coverage** (~56%)

DNS Method

- Resolve domains via DNS over HTTP (DoH)
- Geo-locate IPs with IPinfo
- Great coverage but **low accuracy** for Anycast address

Prioritization

Geo Hints > DoH

Country	Total Domains	Locality Score (%)	Regional Locality (%)	Global Locality (%)
Algeria (DZ)	587	13.3	0	**86.7**
Brazil (BR)	692	**67.4**	2.3	30.3
Egypt (EG)	551	15.2	0	**84.8**
India (IN)	483	**61.4**	0	38.6
Japan (JP)	779	**69.5**	1.8	28.7
Malawi (MW)	966	3.0	**62.3**	34.7
South Africa (ZA)	799	**65.9**	0	34.1
Spain (ES)	816	**67.6**	0.2	32.2
Tanzania (TZ)	499	10.2	43.5	46.3
Thailand (TH)	626	43.7	25.1	31.2
United States (US)	905	**83.1**	0.2	16.7
Vietnam (VN)	597	20.3	12.7	67

Locality Overview

KTL Dashboard

Country	Domains (%)	Bytes (%)
South Africa (ZA)	52.9	62.21
United States (US)	19.15	18.35
France (FR)	5.49	5.22
Malawi (MW)	1.55	3.05
Belgium (BE)	0.93	2.54
Netherlands (NL)	4.87	2.53
United Kingdom (GB)	1.66	1.43
Germany (DE)	3.93	1.34
Others	9.52	3.33

Country Content Locality Matrix For Malawi (MW)

Geo-hints vs. DNS

Country	Total Domains	Geo-hints Coverage (%)	DNS-Geo Agreement (%)
Algeria (DZ)	587	54.3	50.5
Brazil (BR)	692	52.7	63
Egypt (EG)	551	52.1	56.6
India (IN)	483	48.4	62.9
Japan (JP)	779	48	**71.8**
Malawi (MW)	966	54.6	**70.3**
South Africa (ZA)	799	54.1	58.9
Spain (ES)	816	46.4	67.3
Tanzania (TZ)	499	53.7	53.2
Thailand (TH)	626	46.6	59.3
United States (US)	905	51.2	**91.8**
Vietnam (VN)	597	41.5	63.8

On-going work

Extended Country Coverage

Run Campaigns for more countries

Longitudinal analysis

Add traceroutes

Compensate for DoH-Anycast Inaccuracy

CDN Specific locality analysis

Better/Representative Sampling

ASN-level Locality variations

Outputs

- Measurement Pipeline
- API for Integration with ISOC Pulse
- Visualization Dashboard
  - Explore the preliminary results

Key Takeaways

- **Locality Varies Significantly By Country**
  - High locality: US (83%), Japan (70%), Spain (68%), Brazil (67%)
  - Low locality: Algeria (13%), Egypt (15%), Tanzania (10%), Malawi (3%)
  - Regional patterns matter: Malawi has 62% regional (mostly South Africa)
- **Regional Hubs Are Critical For Africa**
  - South Africa serves as primary regional hub for neighboring countries
  - Malawi: 62% regional (ZA), 3% local, 35% global
  - Tanzania: 44% regional locality despite only 10% local

Thank you!

REFERENCES

[1] Internet Society. "The 50/50 Vision for Internet Traffic: Methodology v2.0." January 2024. https://pulse.internetsociety.org

[2] Madeley, J.I., Phokeer, A., Siddiqui, A., and Benson, T.A. "Towards Measuring Content Locality." Applied Networking Research Workshop (ANRW '24), July 2024. DOI: 10.1145/3673422.3674895

[3] Ahmed, A., Shafiq, Z., Bedi, H., and Khakpour, A. "Peering vs. transit: Performance comparison of peering and transit interconnections." IEEE ICNP, 2017.

[4] Calandro, E., Chavula, J., and Phokeer, A. "Internet development in Africa: A content use, hosting and distribution perspective." AFRICOMM 2018.

[5] Internet Society. "50/50 Vision." https://www.internetsociety.org/issues/access/50-50-vision/