Web Content Cartography

Bernhard Ager[†] Wolfgang Mühlbauer[‡]
Georgios Smaragdakis[†] Steve Uhlig[†]

[†]Technische Universtität Berlin / T-Labs
[‡]ETH Zürich

Internet Measurement Conference 2011

Motivation


Content is King

- Web traffic currently dominates: $\sim 60\,\%$
- Hosting infrastructures are the work-horse of content delivery
- But: "The only constant is change": Hyper-giants, Meta CDNs, CDNi, virtualization, applications

How is the content landscape evolving?

We need tools for

- Researchers: Understand the content eco-system better
- CPs: discover choice of available infrastructures
- ISPs: perform strategic decisions: Peering, CDN infrastructure
- Infrastructures: understand the utility of deploying infrastructure

How we complement existing work

Earlier approaches to characterize infrastructures

Hyper-giants, Google [La10]; Hosting models [Le09]; Rapidshare [An09], Akamai and Limelight [Hu08]; Akamai [Su06]; Akamai, Digital Island, and 12 more [Kr01]; ...

... and how our approach is different

- No a-priori signatures
- Aiming at the broad picture
- Automatable, lightweight

[La10] C. Labovitz, S. Lekel-Johnson, D. McPherson, J. Oberheide, and F. Jahanian. Internet Inter-Domain Traffic. In Proc. ACM SIGCOMM, 2010.

[Le09] T. Leighton. Improving Performance on the Internet. Commun. ACM, 2009.

[An09] D. Antoniades, E. Markatos, and C. Dovrolis. One-click Hosting Services: A File-Sharing Hideout. In Proc. ACM IMC, 2009.

[Hu08] C. Huang, A. Wang, J. Li, and K. Ross. Measuring and Evaluating Large-scale CDNs. In Proc. ACM IMC, 2008.

[Su06] A. Su, D. Choffnes, A. Kuzmanovic, and F. Bustamante. Drafting Behind Akamai: Inferring Network Conditions Based on CDN Redirections. IEEE/ACM Trans. Netw., 2009.

[Kr01] B. Krishnamurthy, C. Wills, and Y. Zhang. On the Use and Performance of Content Distribution Networks. In Proc. ACM IMW, 2001.

1. Motivation
2. Approach
3. Data
4. Results
5. Conclusion

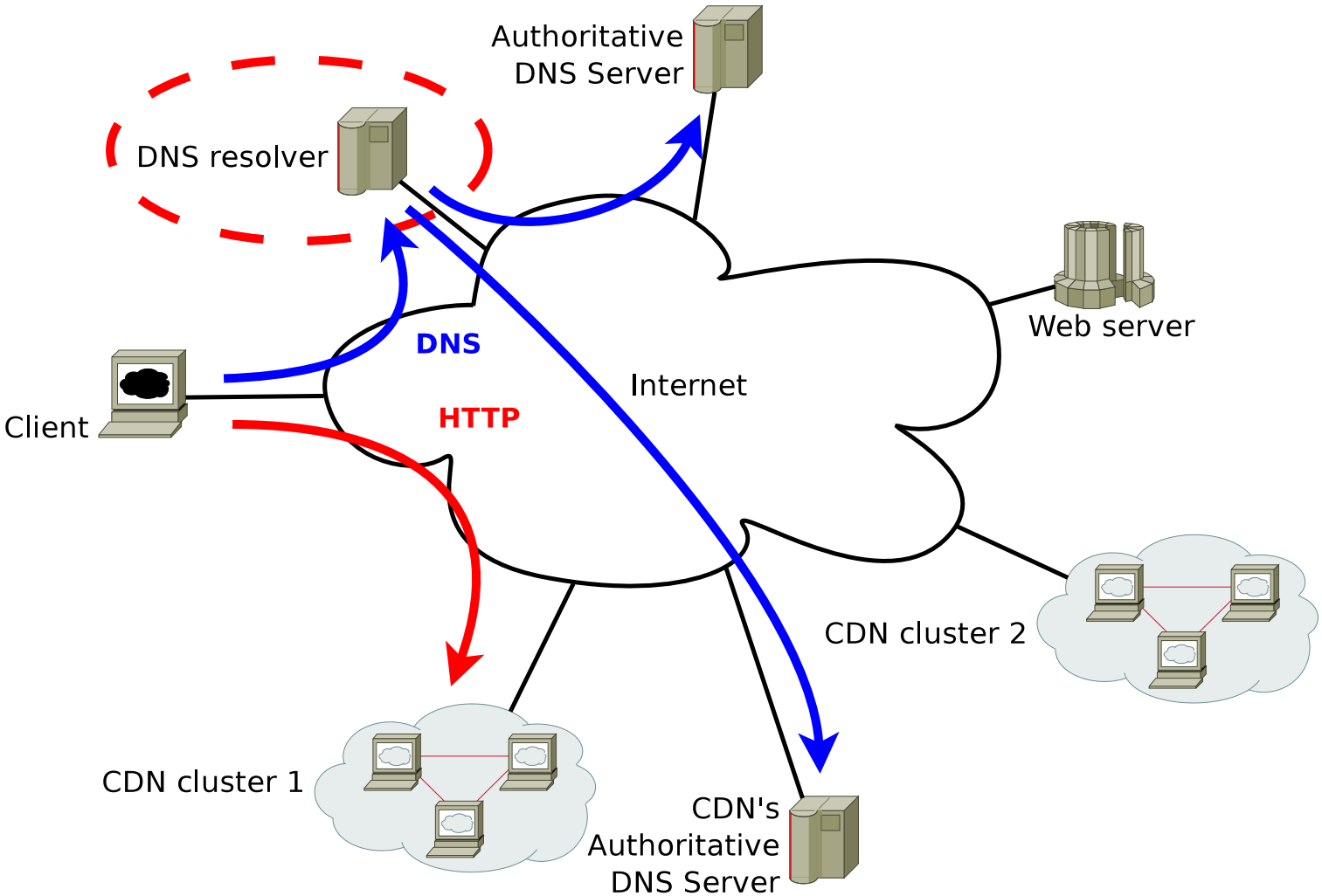
What are the characteristics of content hosting?

Web content cartography

- What are those hosting infrastructures?
- Where are they located?
  - At the network level
  - Geographically
- Who is operating them?
- Which role does each infrastructure play?

We propose web content cartography: building maps of hosting infrastructures

A sketch of HTTP content delivery



Identifying hosting infrastructures

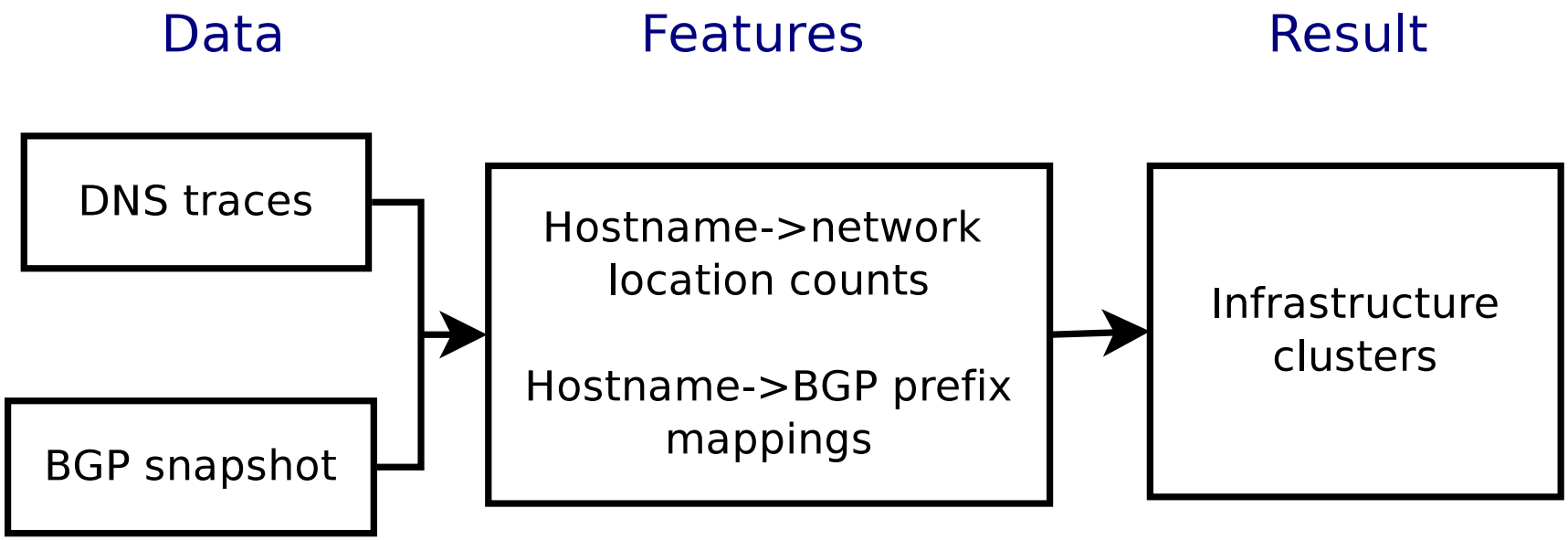
Key observation

DNS exposes the network footprint of hosting infrastructures

Goals

1. Find out geographic locations
2. Find out network locations
3. Identify hosting infrastructures

Identifying infrastructures



Two-level clustering process

- First phase: k-means
- Second phase: based on address space

Collecting data

Hostnames

Requirement: Good coverage of hosting infrastructures

- Extracted from Alexa top 1 Mio. list
- 2000 TOP, 2000 TAIL, $\sim$ 3000 EMBEDDED, $\sim$ 850 CNAMES

Traces

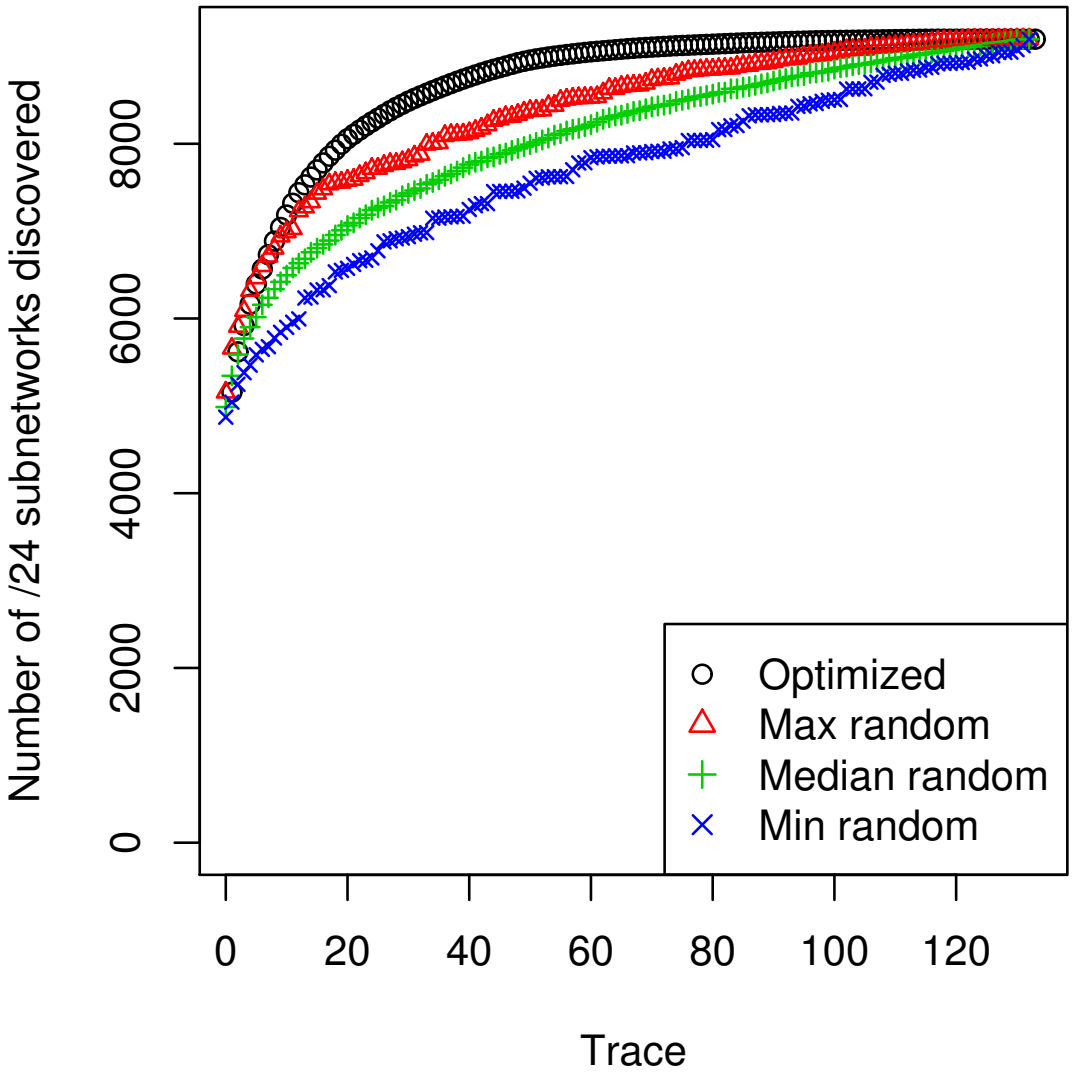
Requirement: Sampling a large enough network footprint

- Script
- Run by volunteers
- Trace collection via website

Traces	133
ASN	78
Countries	27
Continents	6

Estimating coverage

How should you choose vantage points?

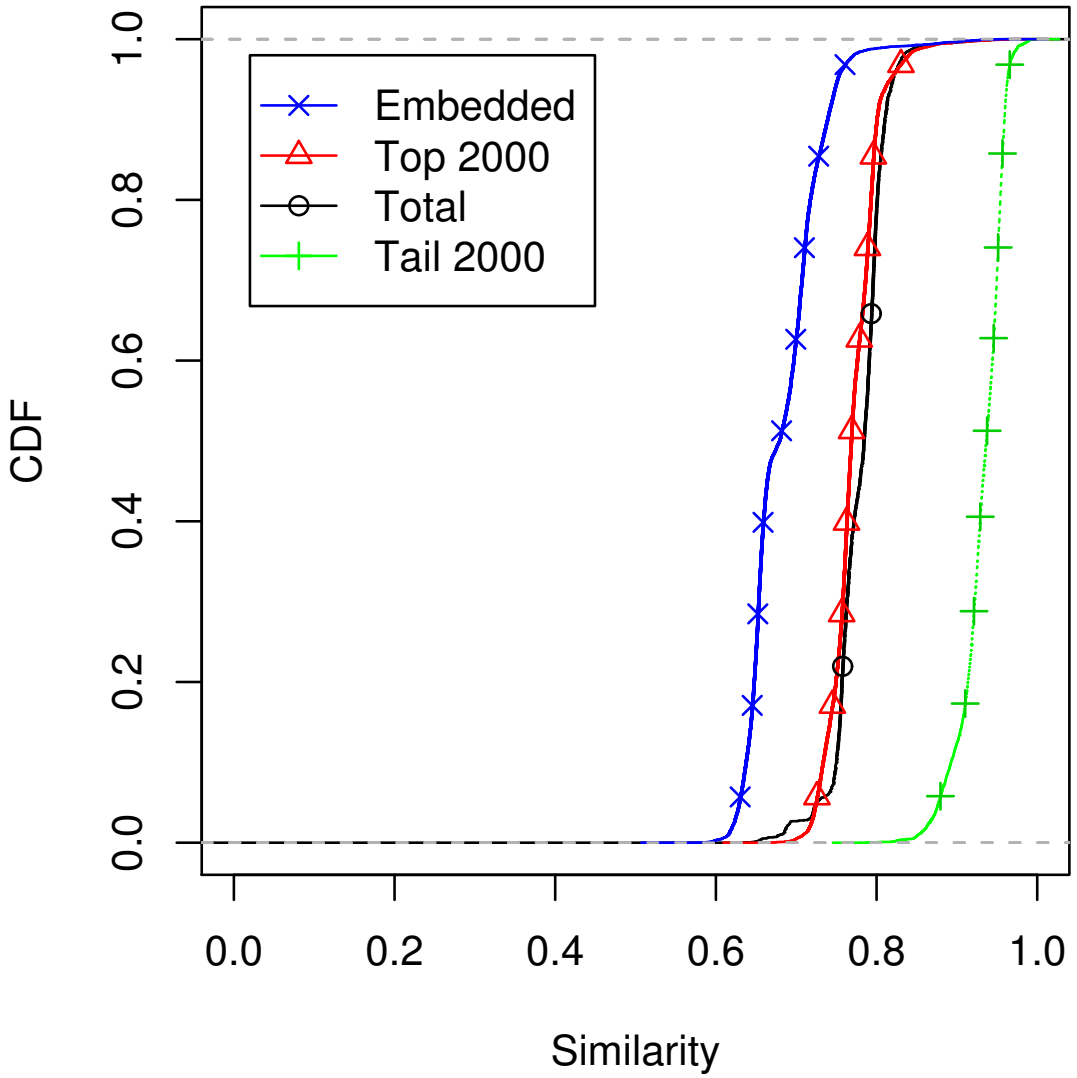


Insights

- Optimized: sampling diversity comes from geographic and network diversity: first 30 traces from 30 ASs in 24 countries
- Median: tail traces yield 20 /24s per trace ⇒ limited utility even when adding more traces

Estimating coverage

How should you choose hostnames?



Insights

- TAIL: similarity high ⇒ mostly centralized
- EMBEDDED: similarity low ⇒ better distributed

Characterizing infrastructures

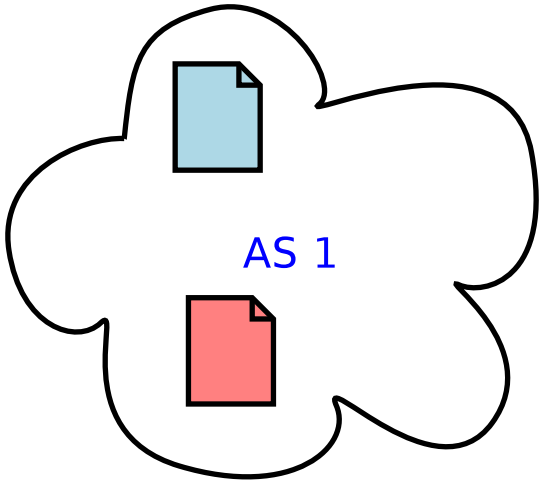
Rank	# hostnames	owner	content mix
1	476	Akamai	
3	108	Google	
4	70	Akamai	
5	70	Google	
6	57	Limelight	
7	57	ThePlanet	
12	28	Wordpress	

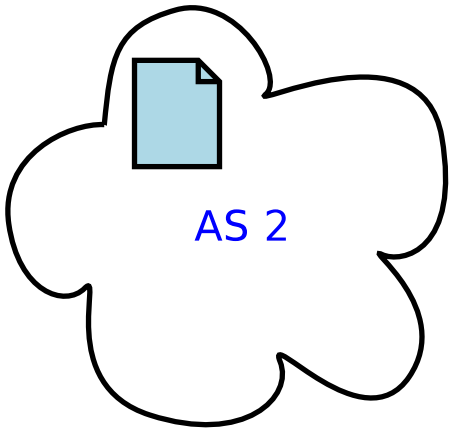
only on TOP, both on TOP and EMBEDDED, only on EMBEDDED, TAIL.

Main findings in Top 20

- Some companies run multiple infrastructures
- EMBEDDED often dominating
- TAIL content is important: consolidation

Content potential and monopoly



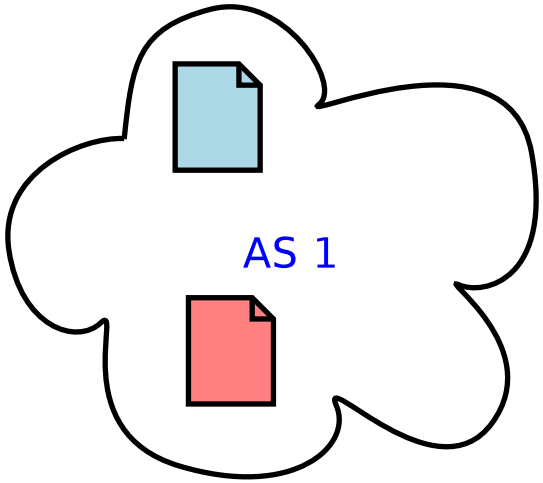


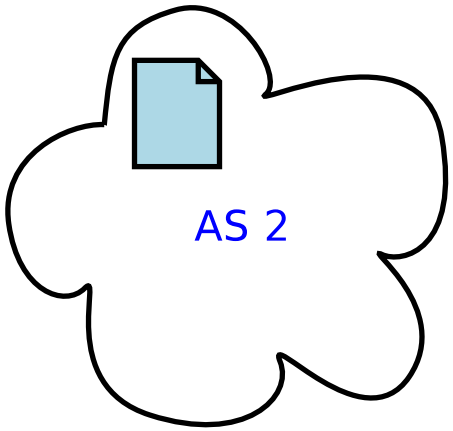
Location	CP
AS 1	1
AS 2	0.5

Content Potential (CP)

Fraction of content available from a location.

Content potential and monopoly





Location	CP	NCP
AS 1	1	0.75
AS 2	0.5	0.25

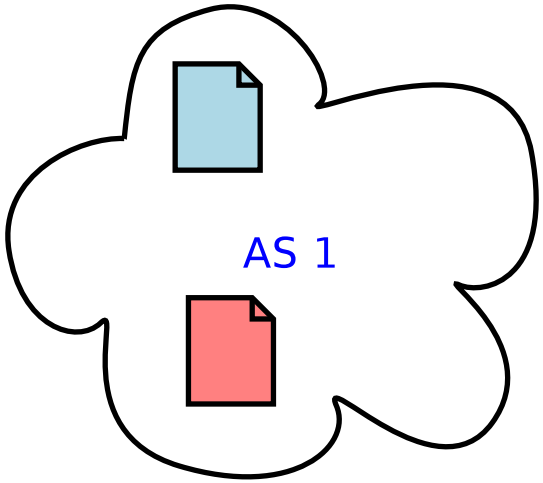
Content Potential (CP)

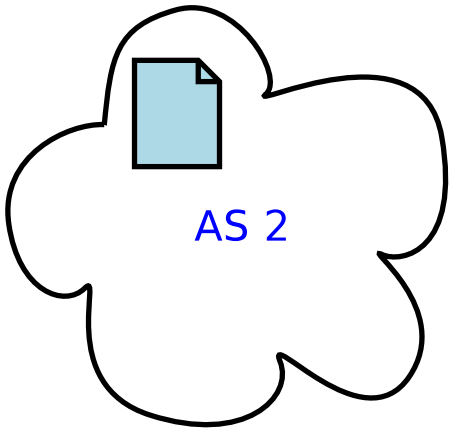
Fraction of content available from a location.

Normalized Content Potential (NCP)

CP weighted by distributedness.

Content potential and monopoly





Location	CP	NCP	CMI
AS 1	1	0.75	0.75
AS 2	0.5	0.25	0.5

Content Potential (CP)

Fraction of content available from a location.

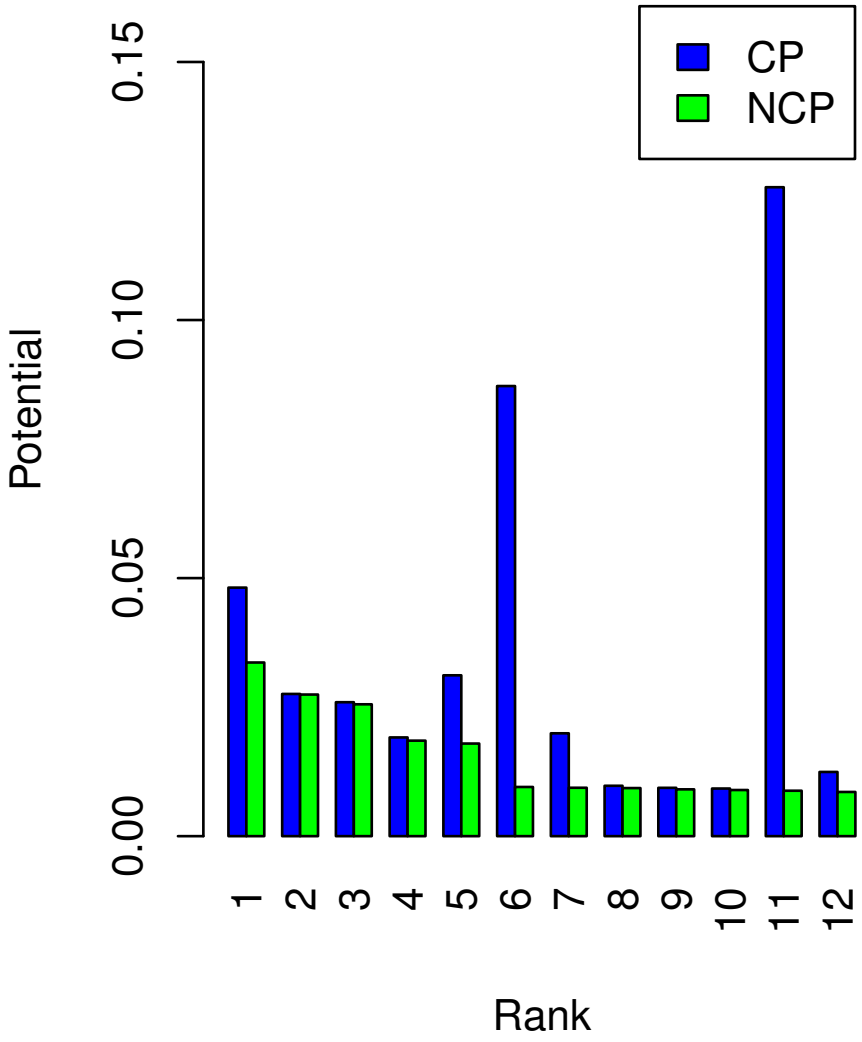
Normalized Content Potential (NCP)

CP weighted by distributedness.

Content Monopoly Index (CMI)

CMI = NCP / CP

Normalized content potential: Top 12 ASs



Rank	AS name	CMI
1	Chinanet	0.699
2	Google	0.996
3	ThePlanet.com	0.985
4	SoftLayer	0.967
5	China169 BB	0.576
6	Level 3	0.109
7	China Telecom	0.470
8	Rackspace	0.954
9	1&1 Internet	0.969
10	OVH	0.969
11	NTT America	0.070
12	EdgeCast	0.688

Comparing AS rankings

CAIDA-cone [CAIDA]

- Number of customer ASs

Arbor [La10]

- Inter-AS traffic volume

Normalized potential

- Weighted content availability

Rank	CAIDA-cone	Arbor	Normalized potential
1	Level 3	Level 3	Chinanet
2	AT&T	Global Crossing	Google
3	MCI	Google	ThePlanet
4	Cogent/PSI	*	SoftLayer
5	Global Crossing	*	China169 backbone
6	Sprint	Comcast	Level 3
7	Qwest	*	Rackspace
8	Hurricane Electric	*	China Telecom
9	tw telecom	*	1&1 Internet
10	TeliaNet	*	OVH

[La10] C. Labovitz, S. Lekel-Johnson, D. McPherson, J. Oberheide, and F. Jahanian. Internet Inter-Domain Traffic. In Proc. ACM SIGCOMM, 2010.

[CAIDA] `http://as-rank.caida.org/`

Conclusion

Summary

- Lightweight discovery of hosting infrastructures
- Characterization of hosting infrastructures
  - We can detect the imhomogenous use of infrastructures
  - Tail content is important due to consolidation
- Content-centric AS rankings
  - "Content monopolies": Google, Chinese ISPs
  - Complementary to traditional rankings

Future work

- Relate with other metrics: traffic volume, finances, ...
- Explore the interactions of content delivery with the topology
- Break-down content by other categories: type, language, ...
- Follow-up work: increase coverage

Backup slides

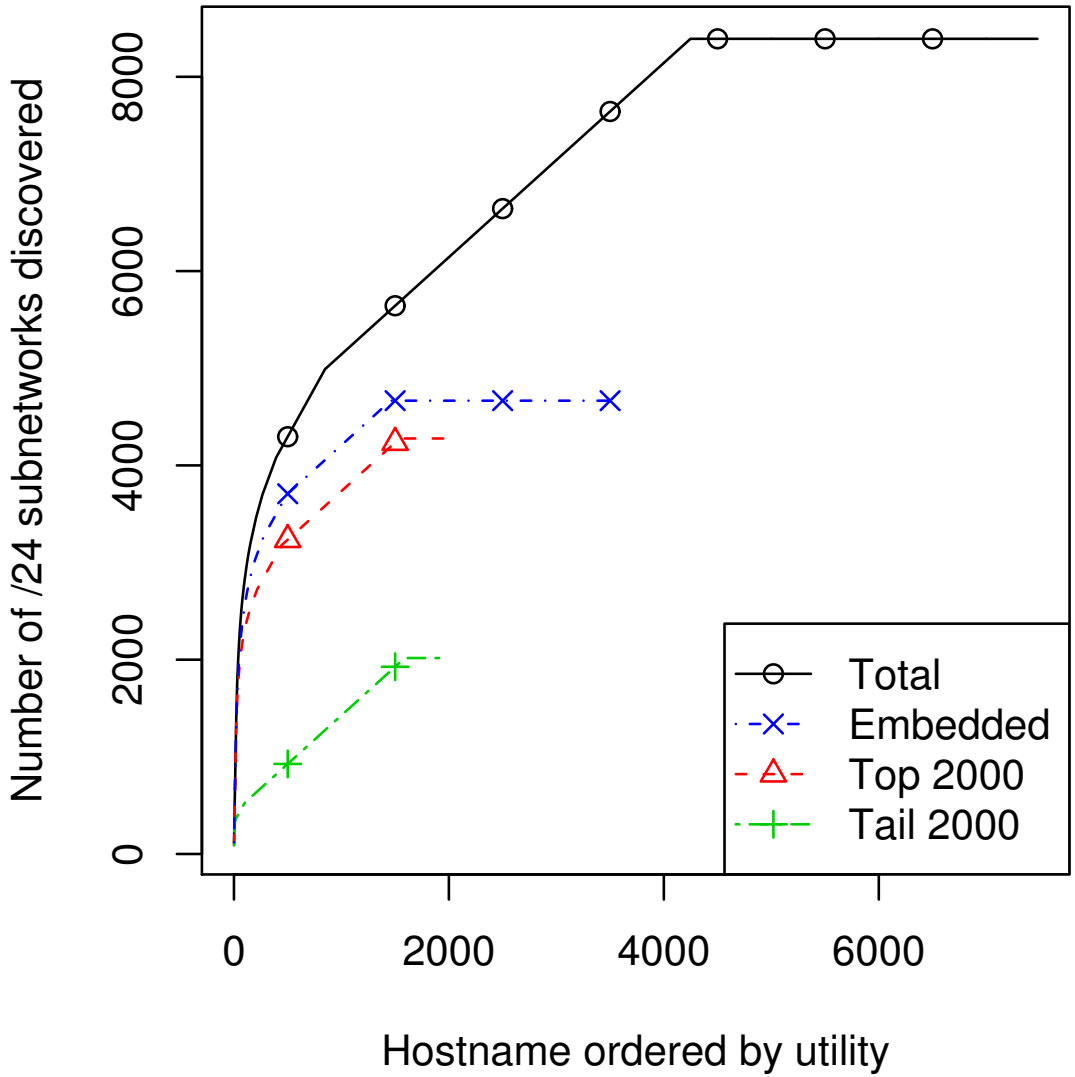
Backup slides

Top 20 content clusters by hostname count

Rank	# hostnames	# ASes	# prefixes	owner	content mix
1	476	79	294	Akamai	
2	161	70	216	Akamai	
3	108	1	45	Google	
4	70	35	137	Akamai	
5	70	1	45	Google	
6	57	6	15	Limelight	
7	57	1	1	ThePlanet	
8	53	1	1	ThePlanet	
9	49	34	123	Akamai	
10	34	1	2	Skyrock	
11	29	6	17	Cotendo	
12	28	4	5	Wordpress	
13	27	6	21	Footprint	
14	26	1	1	Ravand	
15	23	1	1	Xanga	
16	22	1	4	Edgecast	
17	22	1	1	ThePlanet	
18	21	1	1	ivwbox.de	
19	21	1	5	AOL	
20	20	1	1	Leaseweb	

only on TOP, both on TOP and EMBEDDED, only on EMBEDDED, TAIL.

Marginal utility: hostnames



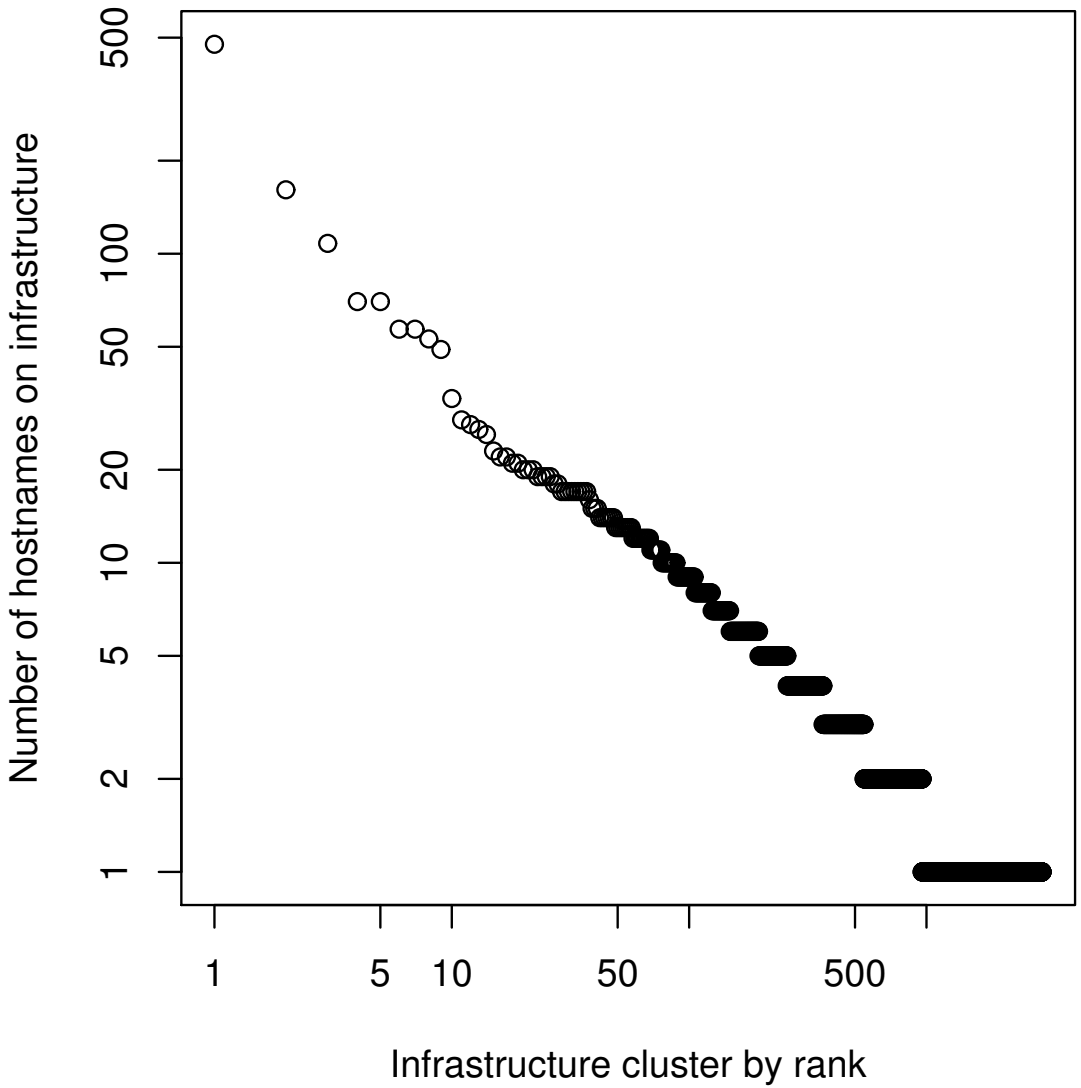
Content exchange matrix: TOP

Requested from	Served from					
	Africa	Asia	Europe	N. America	Oceania	S. America
Africa	0.3	18.6	32.0	46.7	0.3	0.8
Asia	0.3	26.0	20.7	49.8	0.3	0.8
Europe	0.3	18.6	32.2	46.6	0.2	0.8
N. America	0.3	18.6	20.7	58.2	0.2	0.8
Oceania	0.3	20.8	20.5	49.2	5.9	0.8
S. America	0.2	18.7	20.6	49.3	0.2	10.1

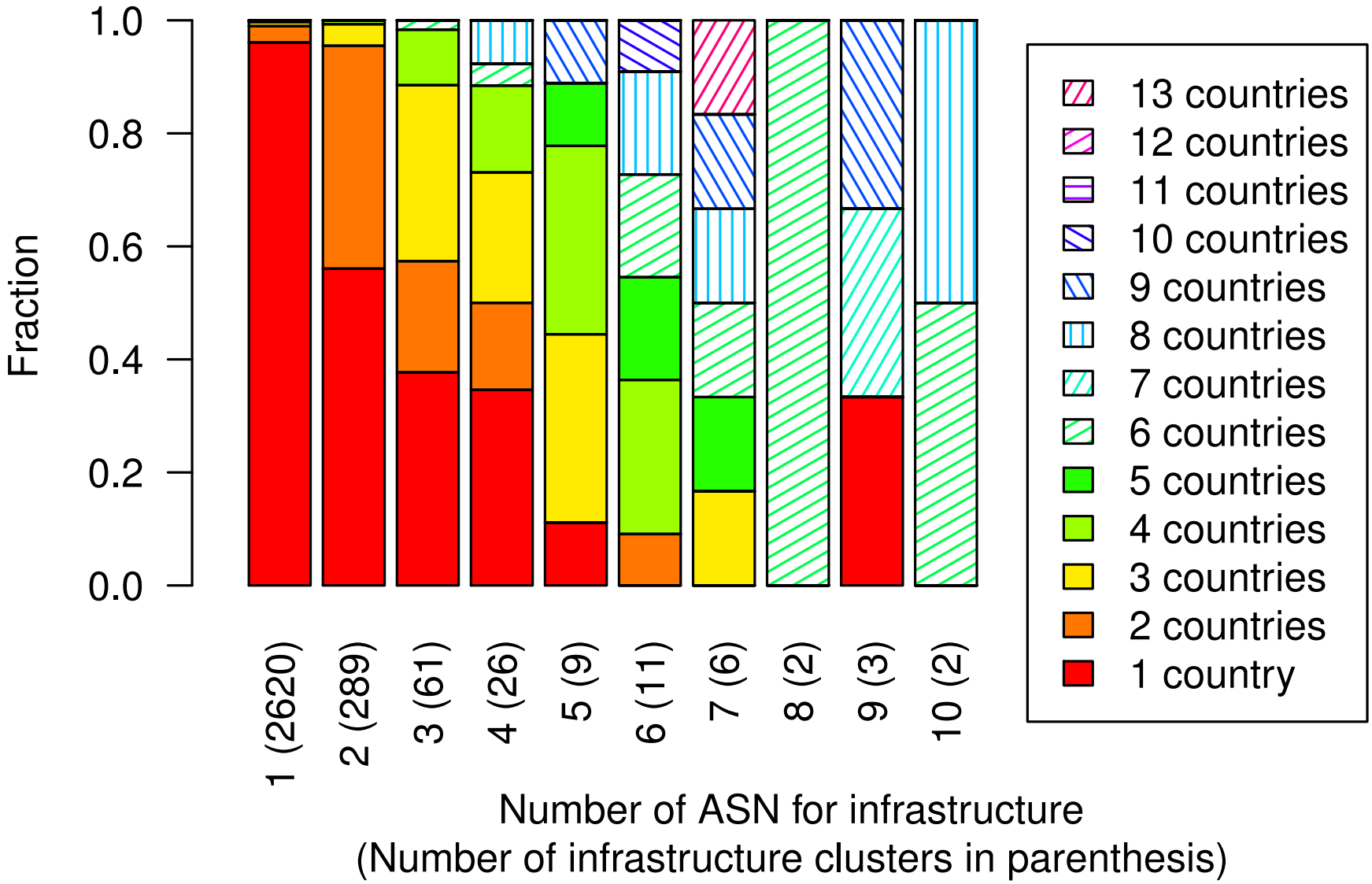
Content exchange matrix: EMBEDDED

Requested from	Served from					
	Africa	Asia	Europe	N. America	Oceania	S. America
Africa	0.3	26.9	35.5	35.8	0.3	0.6
Asia	0.3	37.9	18.3	40.1	1.1	0.6
Europe	0.3	26.8	35.6	35.6	0.4	0.6
N. America	0.3	26.5	18.4	52.9	0.3	0.6
Oceania	0.3	29.2	18.5	38.7	11.3	0.6
S. America	0.3	26.4	18.2	39.3	0.3	14.2

Sizes of similarity clusters



Hosting model of the less distributed infrastructures



Determining the hosting model

An example: Skyrock vs. Cotendo

Rank	# hostnames	# ASes	# prefixes	owner	content mix
10	34	1	2	Skyrock	
11	29	6	17	Cotendo	

only on TOP, both on TOP and EMBEDDED, only on EMBEDDED, TAIL.

Skyrock

- 4 /24-subnetworks
- Website offering blogs/OSN
- Single country: France

Cotendo

- 21 /24-subnetworks
- Website offers CDN service
- 8 countries on 4 continents

Determining the hosting model

An example: Skyrock vs. Cotendo

Rank	# hostnames	# ASes	# prefixes	owner	content mix
10	34	1	2	Skyrock	
11	29	6	17	Cotendo	

only on TOP, both on TOP and EMBEDDED, only on EMBEDDED, TAIL.

Skyrock

- 4 /24-subnetworks
- Website offering blogs/OSN
- Single country: France

⇒ Data center

Cotendo

- 21 /24-subnetworks
- Website offers CDN service
- 8 countries on 4 continents

⇒ Global scale CDN