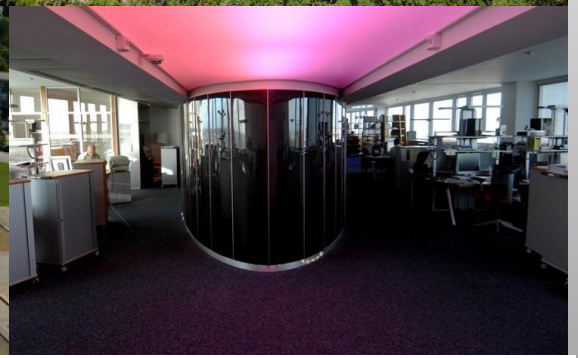




Web Content Cartography

Georgios Smaragdakis
Joint work with Bernhard Ager, Wolfgang Mühlbauer, and Steve Uhlig

Cartography

Cartography (from Greek Χάρτης, *chartes* or *charax* = sheet of papyrus (paper) and *graphein* = to write) is the study and practice of *making maps*.

Ptolemy's map, published in Strasbourg in 1513

I would also add:

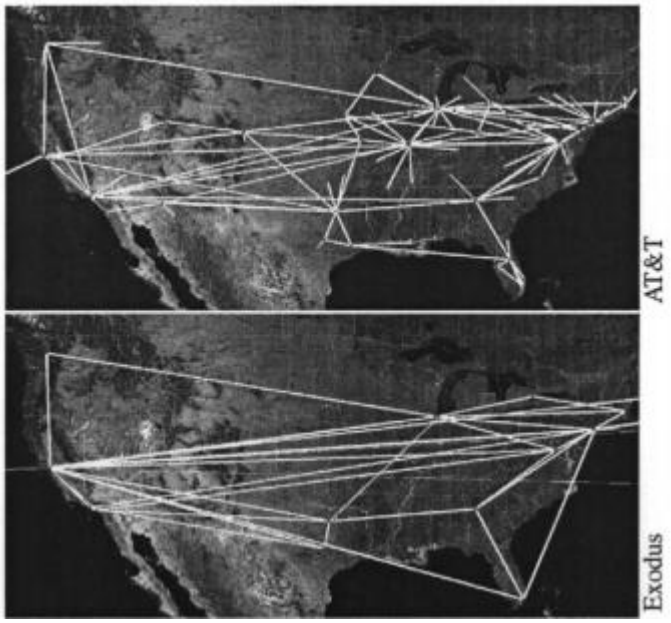
– *Annotated* maps

– Based on *Inference*

Cartography and the Internet

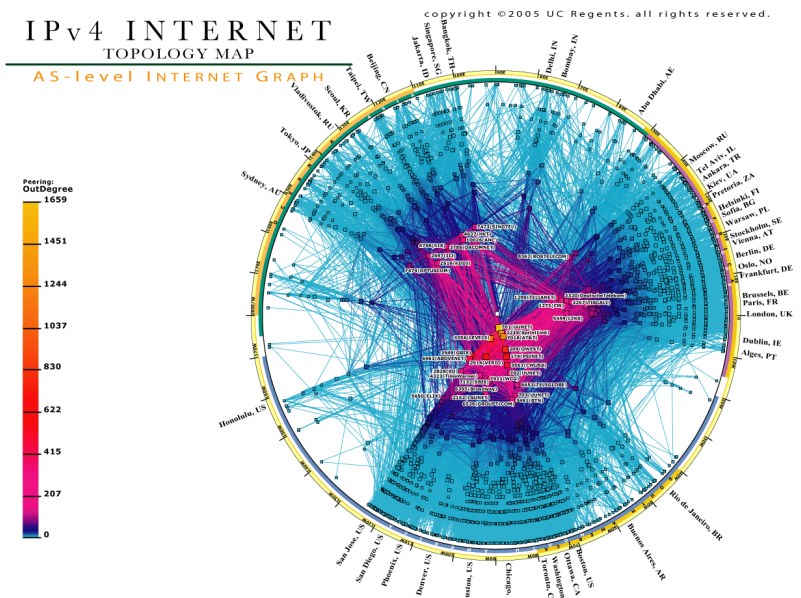
In the recent years the focus was on:

Inference of topology via active or passive measurements



→ In the router-level

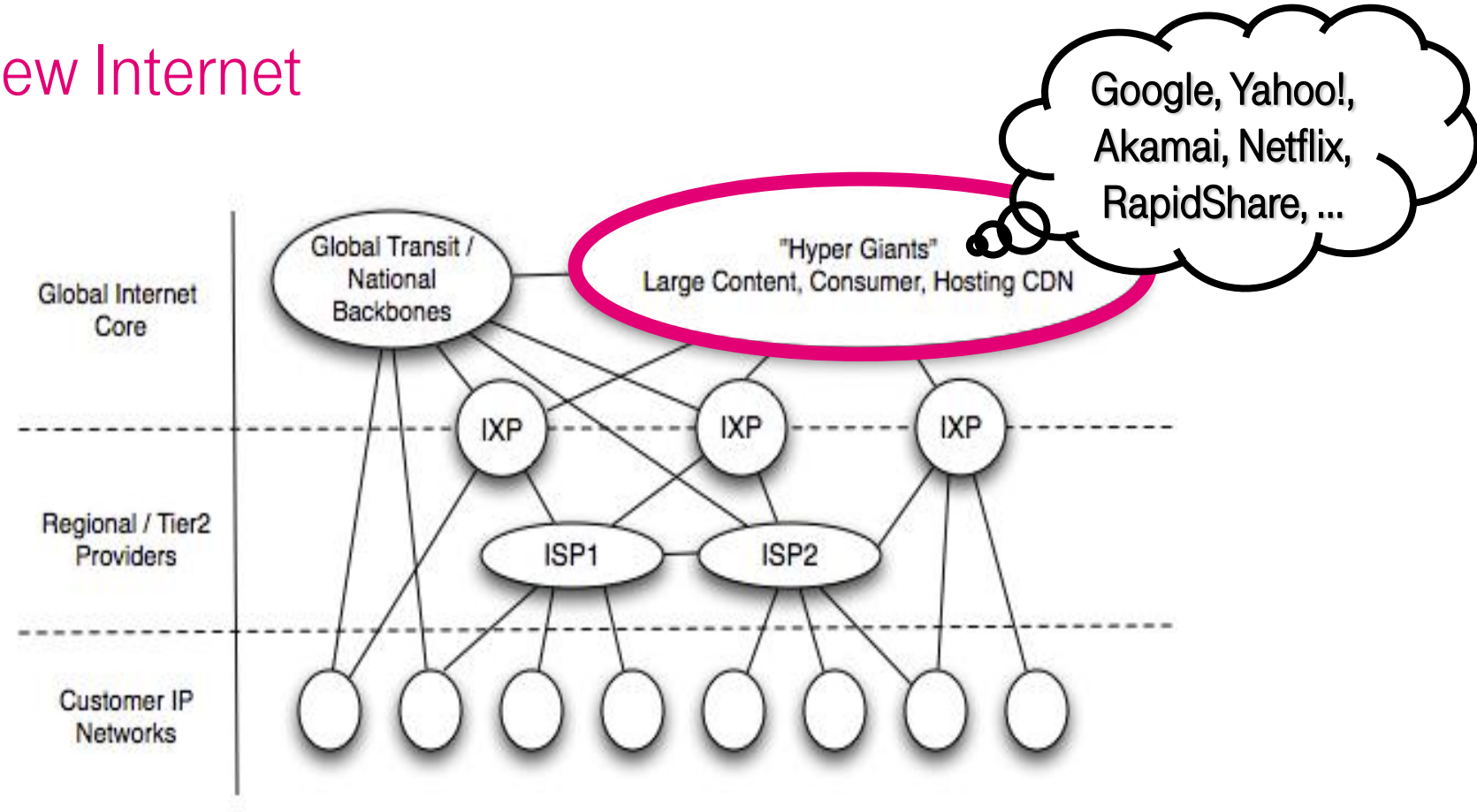
Source: Rocketfuel, UW



→ In the AS-level

Source: CAIDA

The New Internet



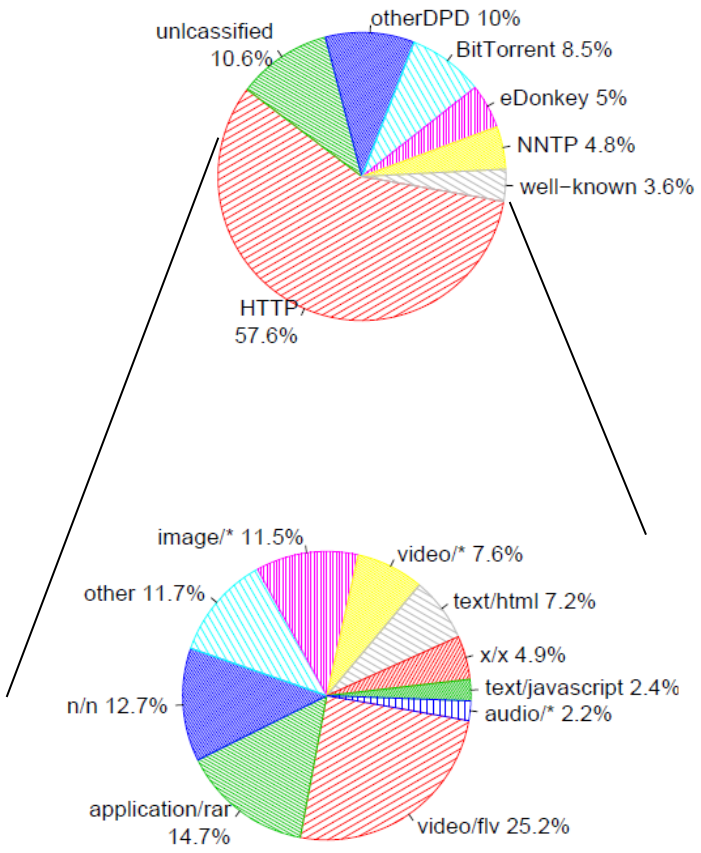
Source: Arbor Networks 2010

→ New core of interconnected content and consumer networks [1]

→ *It is easy to deploy a massive and distributed deployment of servers*

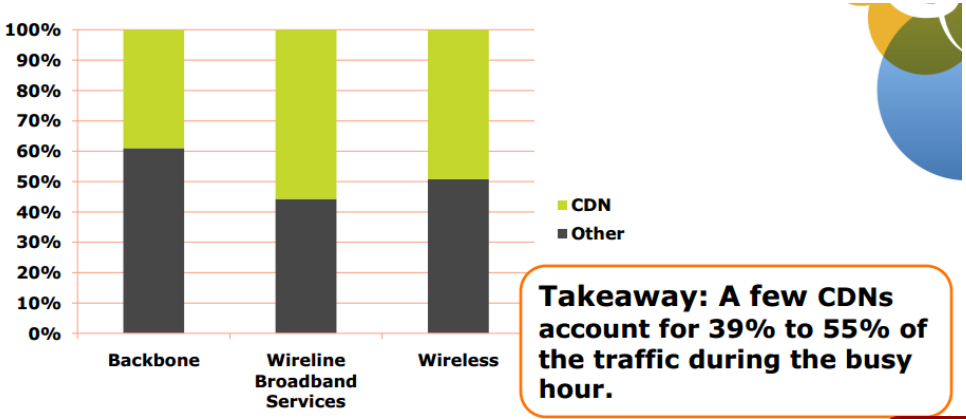
[1] "Internet Interdomain Traffic", Labovicz, Lekel-Johnson, McPherson, Oberheide, Jahanian, SIGCOMM 2010

Content is King



Source: "On Dominant Characteristics of Residential Broadband Internet Traffic
Maier, Feldmann, Paxson, Allman, IMC'09

- HTTP is responsible for around 60% of the total traffic
- This trend will continue (flash video, cloud applications, datacenters, smartphones)
- 50% of the Traffic in a Tier-1 US carrier is due a small number of CDNs
- Netflix in 2011 was responsible for more then 30% of peak hour traffic



Takeaway: A few CDNs account for 39% to 55% of the traffic during the busy hour.

Source: "Traffic Types and Growth in Backbone Networks", Gerber, Doverspike, OFC/NFOEC 2011

The process of building maps of hosting infrastructures

Identification:

→ Which are those of infrastructures?

Classification:

→ Where are they located?

- → At the network level
- → Geographically

→ Who is operating them

→ Deployment schemes

→ Which role does each infrastructure play

Why Web Content Cartography?

Research Community:

→ To Better understand the *evolving* content ecosystem

Content Producers:

→ To be able to better express their *preference* on the location

Content Distributors:

→ To understand their position in the market; *strategic deployment*

ISPs:

→ To Support of *strategic decisions*: peering vs. datacenter deployment

New Architectures

→ To Increase *Content Awareness*

→ Check our work on “Content-Aware Traffic Engineering”, Sigmetrics’12

& “Improving Content Delivery with PaDIS”, IEEE Internet Computing, IMC’10

The “Expert Reviewer” Comment

Well, there are papers in the area that studied the deployment of, e.g.,

- Akamai [1,2]
- RapidShare [3]
- Google [4]

What is new here...?

[1] “Drafting Behind Akamai: Inferring Network Conditions based on CDN Redirections”, Su, Choffnes, Kuzmanovic, Bustamante, IEEE/ACM ToN 2009.

[2] “Measuring and Evaluating Large-Scale CDNs”, Huang, Wang, Li, Ross, IMC 2008

[3] “One-Click Hosting Services: A File-Sharing Hideout”, Antoniadis, Markatos,Dovrolis, IMC 2009

[4] “Answering “What-If” Deployment and Configuration Questions with WISE”, Bin Tariq, Zeitoun, Valancius, Feamster, Ammar, SIGCOMM 2008

The “Expert Reviewer” Comm

Well, there are p rs in the a
deployment

- Akamai [1,2]
- RapidShar
- Google [4]

No A-priori Signatures

Automated Identification & Classification

Lightweight

Redirections ; nes, Kuzmanovic,

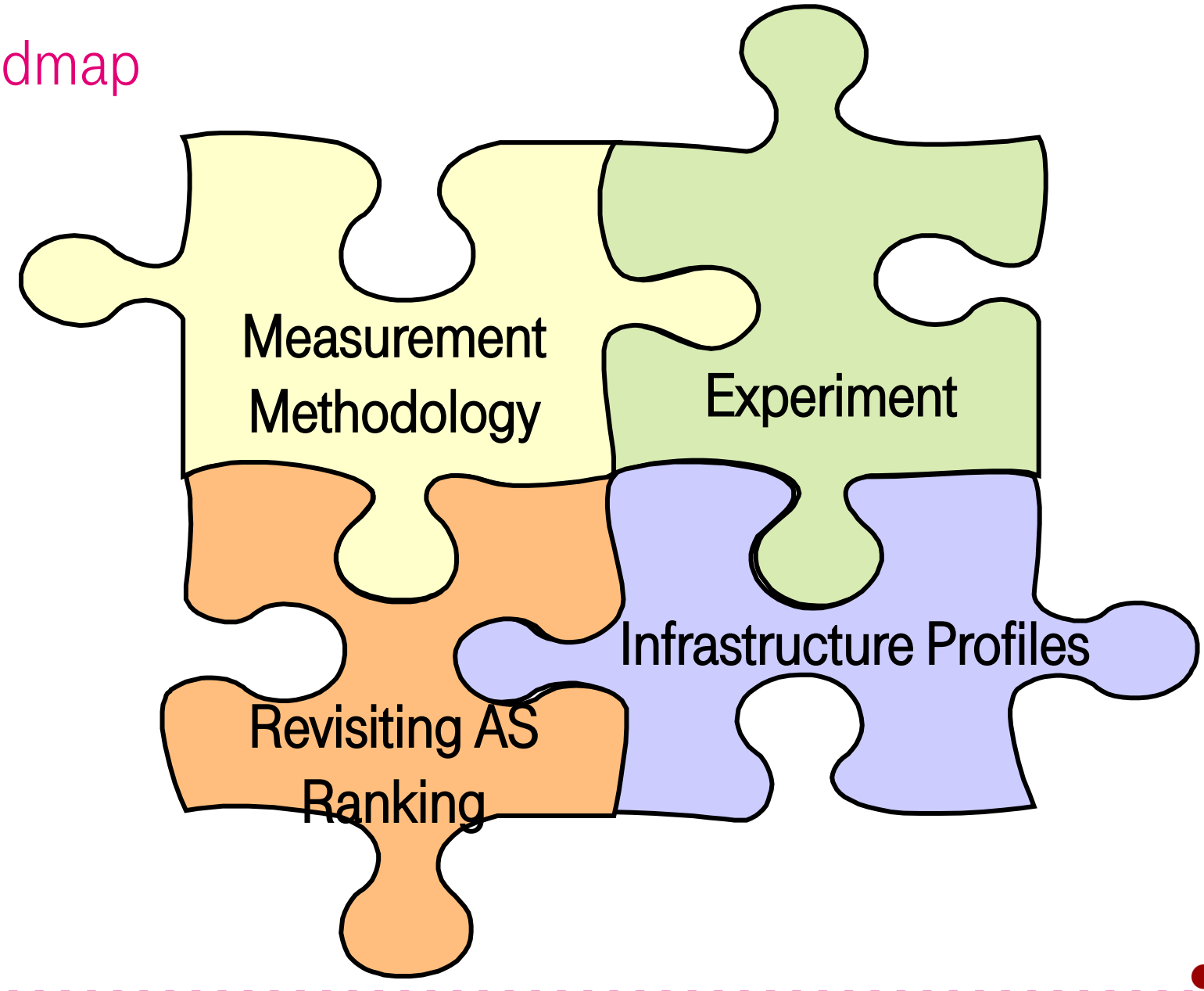
oss, IMC 2008

Markatos,Dovrolis, IMC 2009

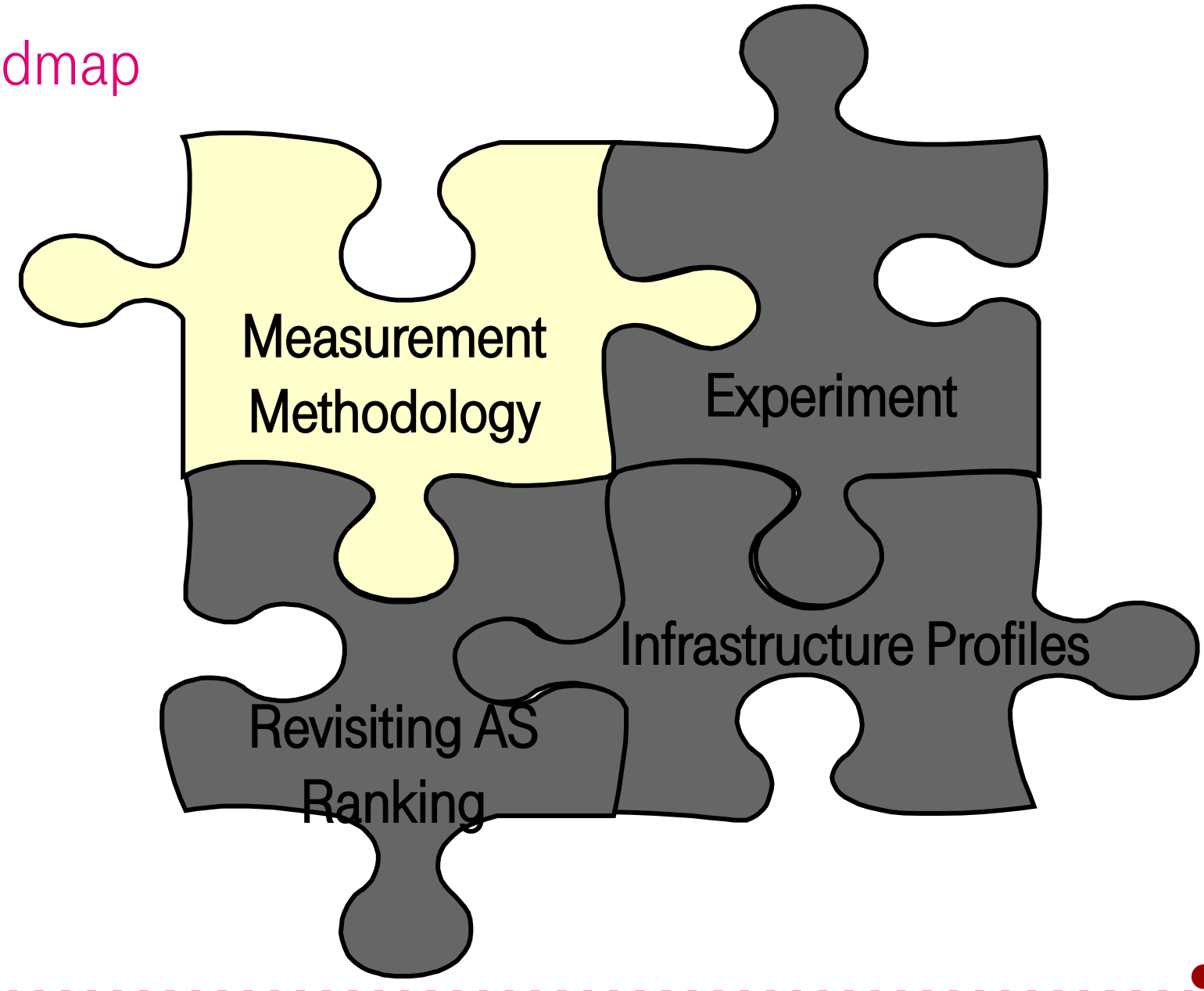
s with WISE”, Bin Tariq, Zeitoun, Valancius, Feamster,

Am

Roadmap


Measurement Methodology
Experiment
Infrastructure Profiles
Revisiting AS Ranking

Roadmap


Measurement Methodology
Experiment
Infrastructure Profiles
Revisiting AS Ranking

High-level Idea

The process of building maps of hosting infrastructures

Identification:

→ Which are those of infrastructures?

Classification:

→ Where are they located?

- → At the network level
- → Geographically

→ Who is operating them

→ Deployment schemes

→ Which role does each infrastructure plays

High-level Idea

The process of building maps of hosting infrastructures

Identification:

→ Which are those of infras

Classification:

→ Where are they located?

→ At the network level

→ Geographically

→ Who is operating them

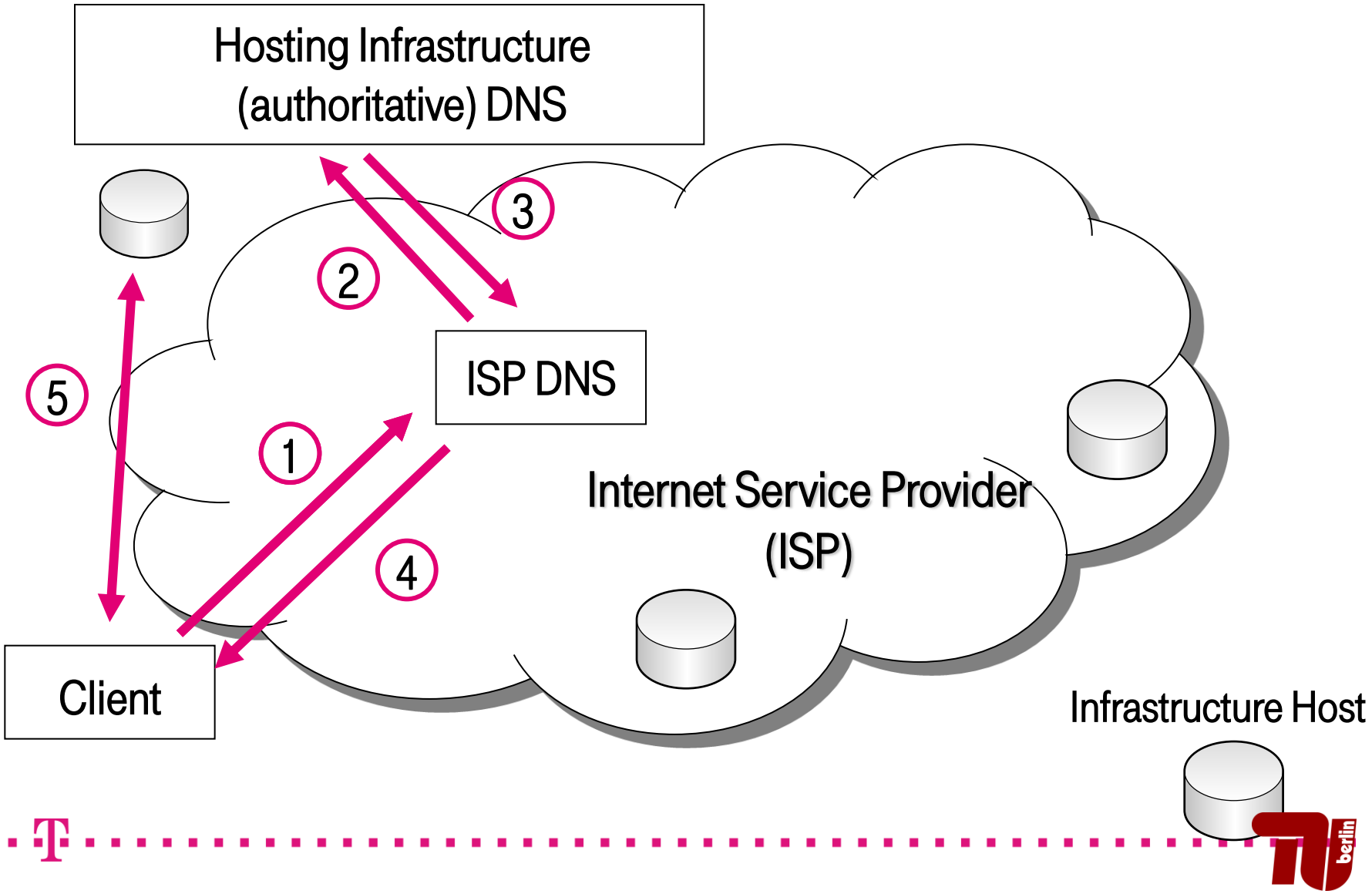
→ Deployment schemes

→ Which role does each inf

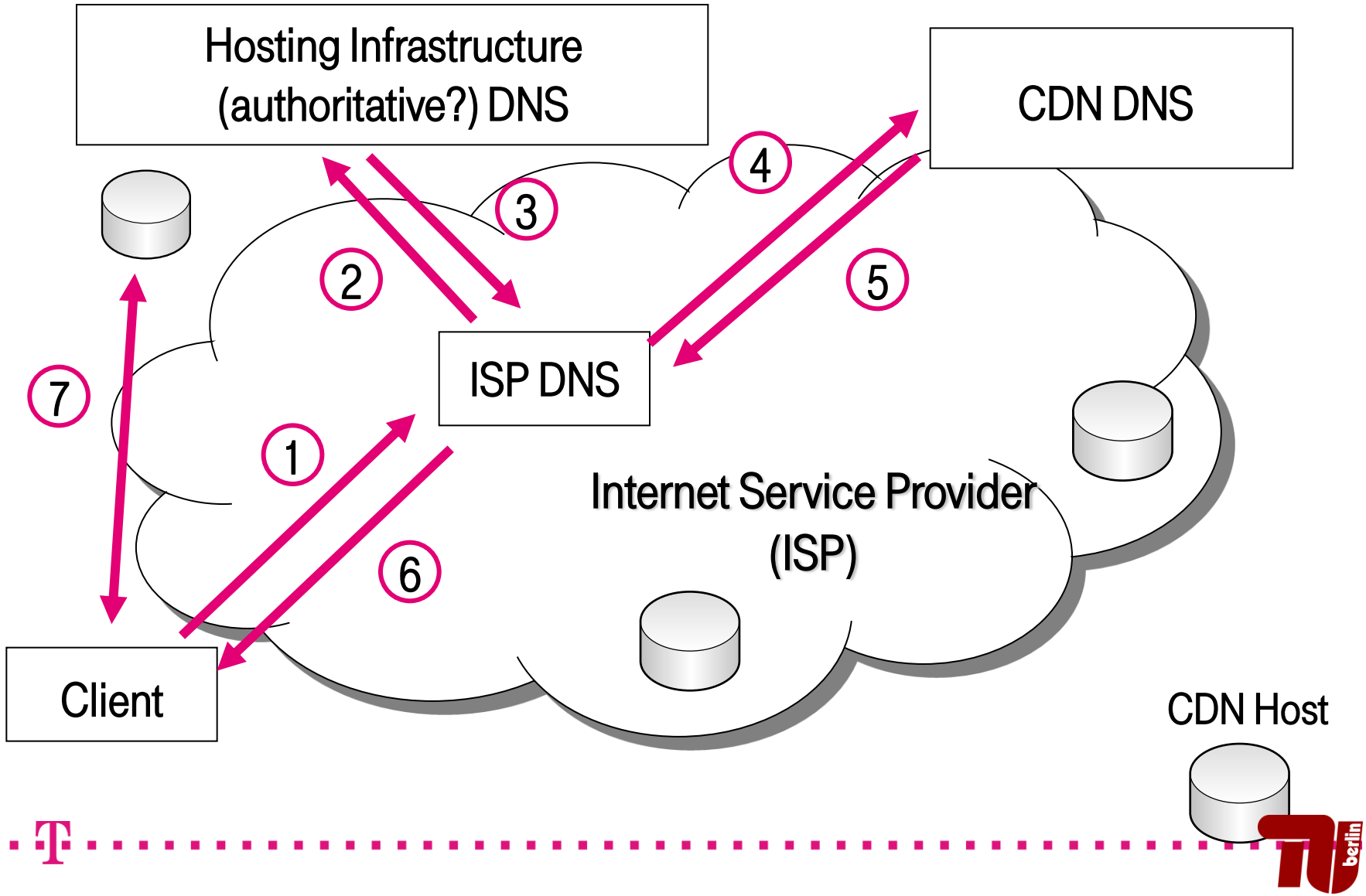
They expose their identity by redirecting you to a server!

Utilize the network footprint! Publicly available Information!

Content Distribution Prime



Content Distribution Prime



Welcome to Facebook - Log

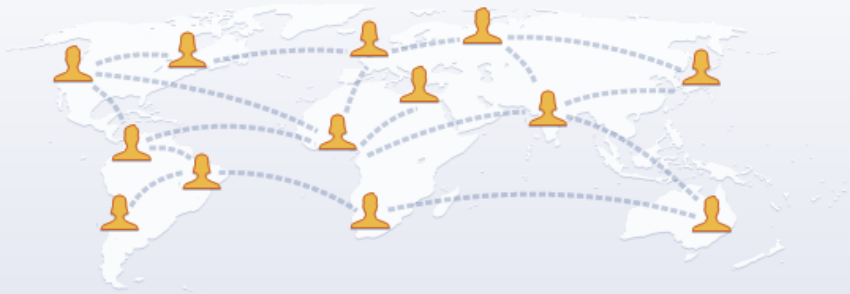
www.facebook.com/index.php?stype=lo&lh=Ac_ZX3VH_I5jQejP

facebook

Email | Password | Log In

Keep me logged in | Forgot your password?

Facebook helps you connect and share with the people in your life.



Sign Up

It's free and always will be.

First Name:

Last Name:

Your Email:

Re-enter Email:

New Password:

I am: Select Sex:

Birthday: Month: Day: Year:

Why do I need to provide my birthday?

By clicking Sign Up, you agree to our Terms and that you have read and understand our Data Use Policy.

Sign Up

Create a Page for a celebrity, band or business.

Deutsch English (US) Español Português (Brasil) Français (France) Italiano العربية हिन्दी 中文(简体) 日本語 ...

Facebook © 2012 · English (US)

Mobile · Find Friends · Badges · People · Pages · About · Advertising · Create a Page · Developers · Careers · Privacy · Terms · Help

Georgios Smaragdakis

www.facebook.com/georgios.smaragdakis

facebook Search Georgios Smaragdakis | Home

Welcome to Your Timeline — Preview

Start Tour | Publish Now

Right now, only you can see your timeline. This gives you a chance to:

1. Review what's on your timeline now, and add or hide whatever you want.
2. Click Publish Now or wait until **your timeline goes live on March 16.** Learn more.

Now
February
2012
2011
2010
2009
2008
2007
2003
2002
Born

Embedded objects

Advertisements

Georgios Smaragdakis

Update Info | Activity Log

Works at T-Labs

Studied Computer Science at Boston University

Lives in Berlin, Germany

About


Friends 541


Photos 210

Map 12

Likes 67

2

Status | Photo | Place | Life Event

What's on your mind?

Georgios Smaragdakis
Yesterday

Friends See All


Vijay Erramilli
91 mutual friends


Nikolaos Laoutaris
71 mutual friends


Nagia Chanaki
7 mutual friends


Dimitris Skoulaxinos
25 mutual friends

Sponsored

% Aktion: iPhone GRATIS
handyservice.de

Deutschlands größtes Mobilfunkversandhaus. Top-Handys, günstige Tarife und Gratis Prämien oder Ratenzahlungen sichern und sparen!

KATIE MELUA - Berlin
ticketmaster.de


Am 9. Juli live auf dem Classic Open Air Gendarmenmarkt. Katie Melua auf Secret Symphony Tour 2012. www.ticketmaster.de

Dein Auslandsjahr
schueleraustausch.net


Hol dir jetzt kostenlos die Kataloge aller Organisationen und vergleiche die Angebote mit unserem Vergleichstool

DNS Answer Anatomy

→ Requesting a photo from Facebook

More than 60% of websites (>30% of traffic) redirect to at least 5 non-original servers [1]

```
$ dig photos-h.ak.fbcdn.net

; <<>> DiG 9.7.0-P1 <<>> photos-h.ak.fbcdn.net

;; QUESTION SECTION:
;photos-h.ak.fbcdn.net.                IN      A

;; ANSWER SECTION:
photos-h.ak.fbcdn.net.        6099    IN      CNAME   photos-
d.ak.facebook.com.edgesuite.net.
photos-d.ak.facebook.com.edgesuite.net.       20492 IN CNAME a998.mm1.akamai.net.
a998.mm1.akamai.net.  7       IN      A       62.41.85.74
a998.mm1.akamai.net.  7       IN      A       62.41.85.90
...
```

2nd Level Domain → Application

facebook

Redirection → Content Provider

Akamai

[1] "Understanding Web Complexity," Butkiewicz, Madhyastha, Sekar, IMC 2011

High-level Idea

The process of building maps of hosting infrastructures

Identification:

→ Which are those of inf

Classification:

→ Where are they locate

→ At the network level

→ Geographically

→ Who is operating then

→ Deployment schemes

→ Which role does each

They expose their identity by redirecting you to a server!

Utilize the network footprint!
Publicly available Information!

Features

There are three different deployment schemes for Hosting Infrastructures [1]:

- → Centralized Hosting: Single server or Single Datacenter, e.g., Rackspace.
- → Datacenter-based Hosting: A small number of datacenters, e.g., Limelight.
- → Cache-based: Highly distributed infrastructures with many peerings or deep inside the network, e.g., Google and Akamai.

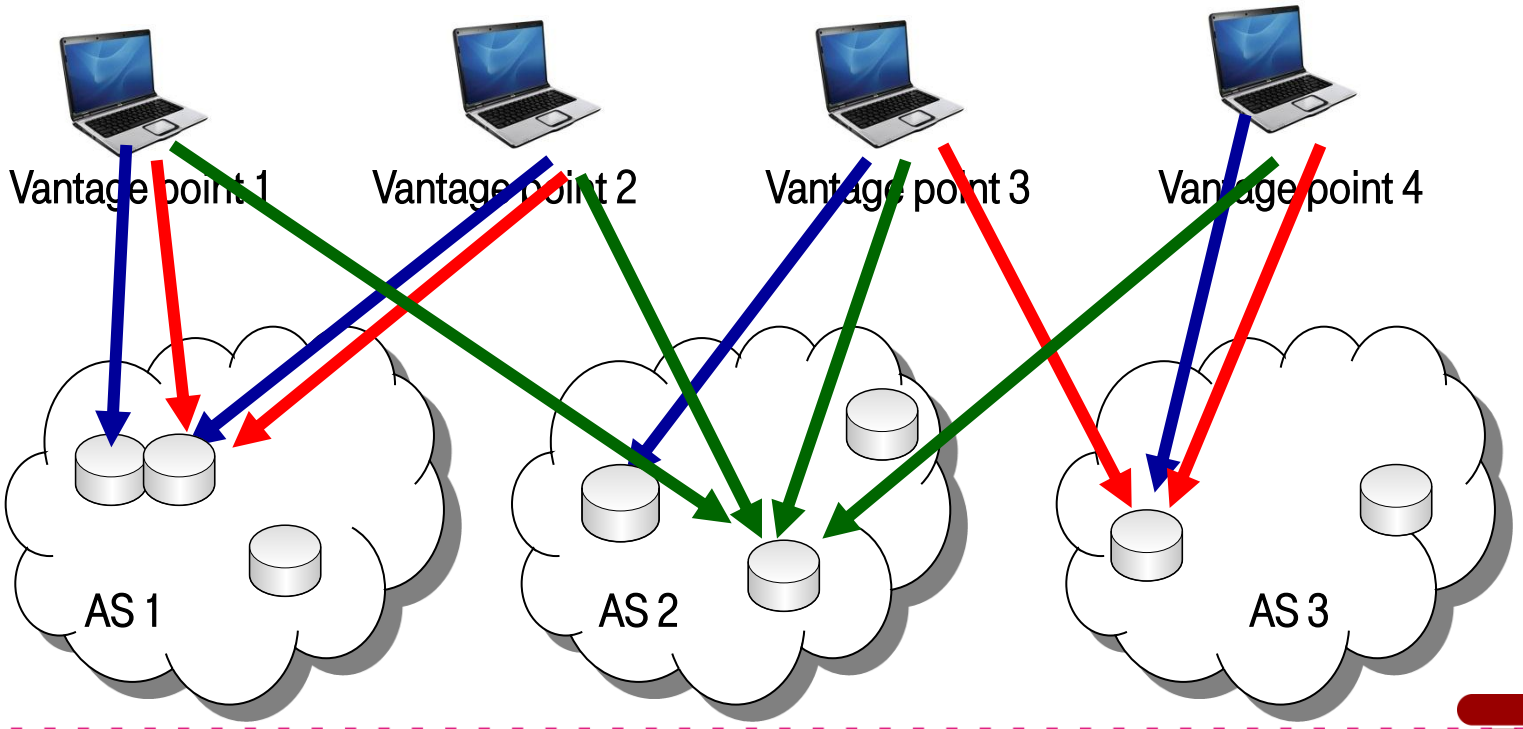
Features:

1. Number of IP addresses
2. Number of subnets, we use \24 to take into account the scale of deployment
3. Number of ASes hosted in
4. Topological characteristics of prefix, AS

[1] "Improving Performance on the Internet" Leighton CACM 2009

Feature Extraction via DNS

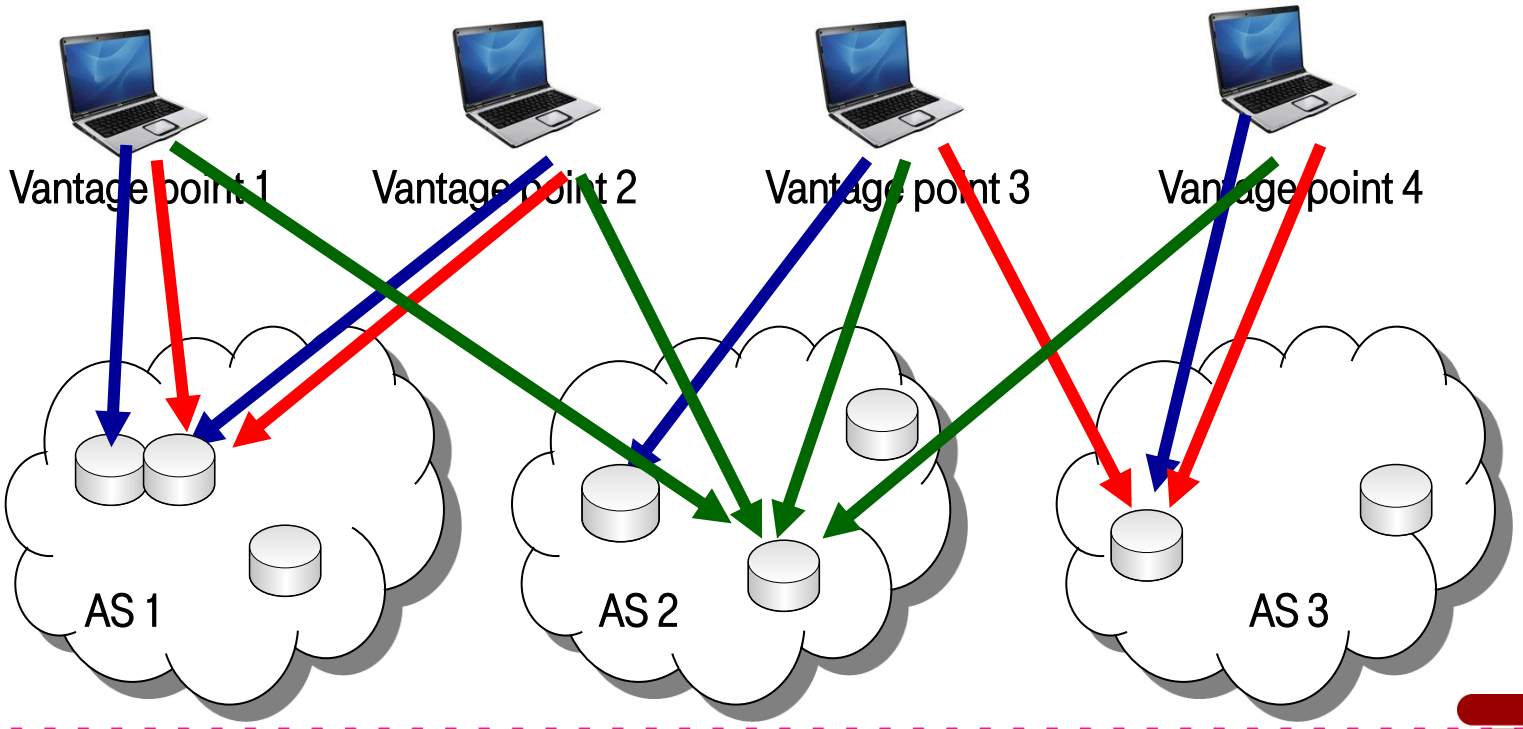
Features	*(#IPs ,*	*#prefixes ,*	*#ASes)*
Hostname 1	(4,	3,	2)
Hostname 2	(2,	2,	2)
Hostname 3	(1,	1,	1)



Feature Extraction via DNS

Features	*(#IPs ,*	*#prefixes ,*	*#ASes)*
Hostname 1	(4,	3,	2)
Hostname 2	(2,	2,	2)
Hostname 3	(1,	1,	1)

Be aware of User Mis-location [1]



[1] "Comparing DNS Resolvers in the Wild", Ager, Muehlbauer, Smaragdakis, Uhlig, IMC 2010

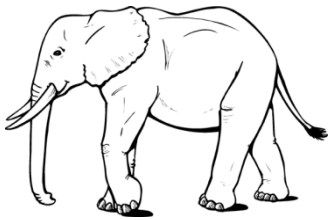
Step 1: Separating Large Hosting Infrastructures

Features	*(#IPs*	*#prefixes*	*#ASes)*
Hostname 1	(4,	3,	2)
Hostname 2	(2,	2,	2)
Hostname 3	(1,	1,	1)

k-means

→ Use k-means algorithm to partition the *hostnames* in clusters in the feature space (# of IPs, prefixes, ASes).

→ Clusters whose features have high values correspond to massively deployed infrastructures. (top-k clusters, k relatively small)

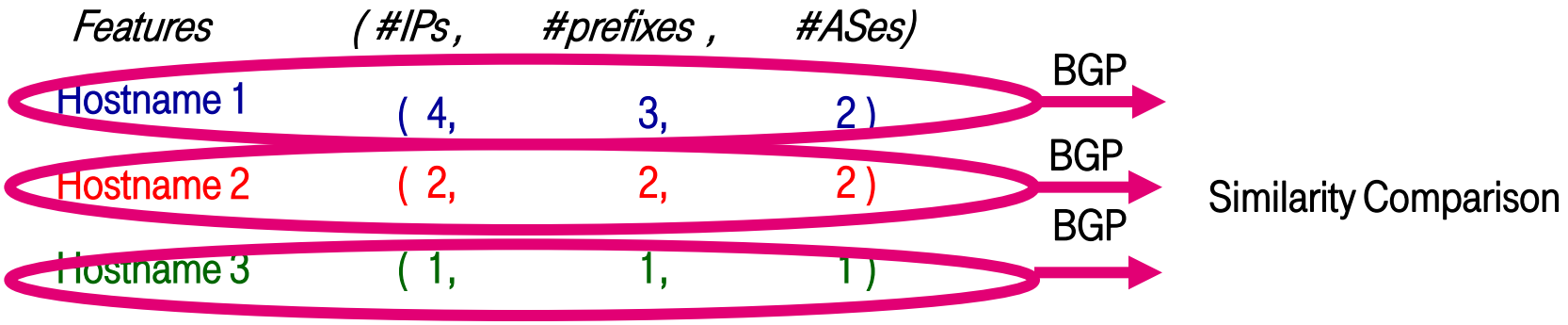


→ Small infrastructures are grouped together (they have similar small values for the features)

Step 2: Distinguishing Small Hosting Infrastructures

Features	*(#IPs ,*	*#prefixes ,*	*#ASes)*
Hostname 1	(4,	3,	2)
Hostname 2	(2,	2,	2)
Hostname 3	(1,	1,	1)

Step 2: Distinguishing Small Hosting Infrastructures

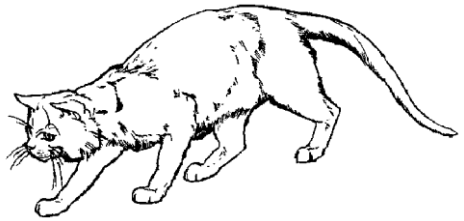


For the smaller clusters:

→ Build sub-clusters for each *hostname* and use network characteristics (*via BGP*), not only the features that correspond to size.

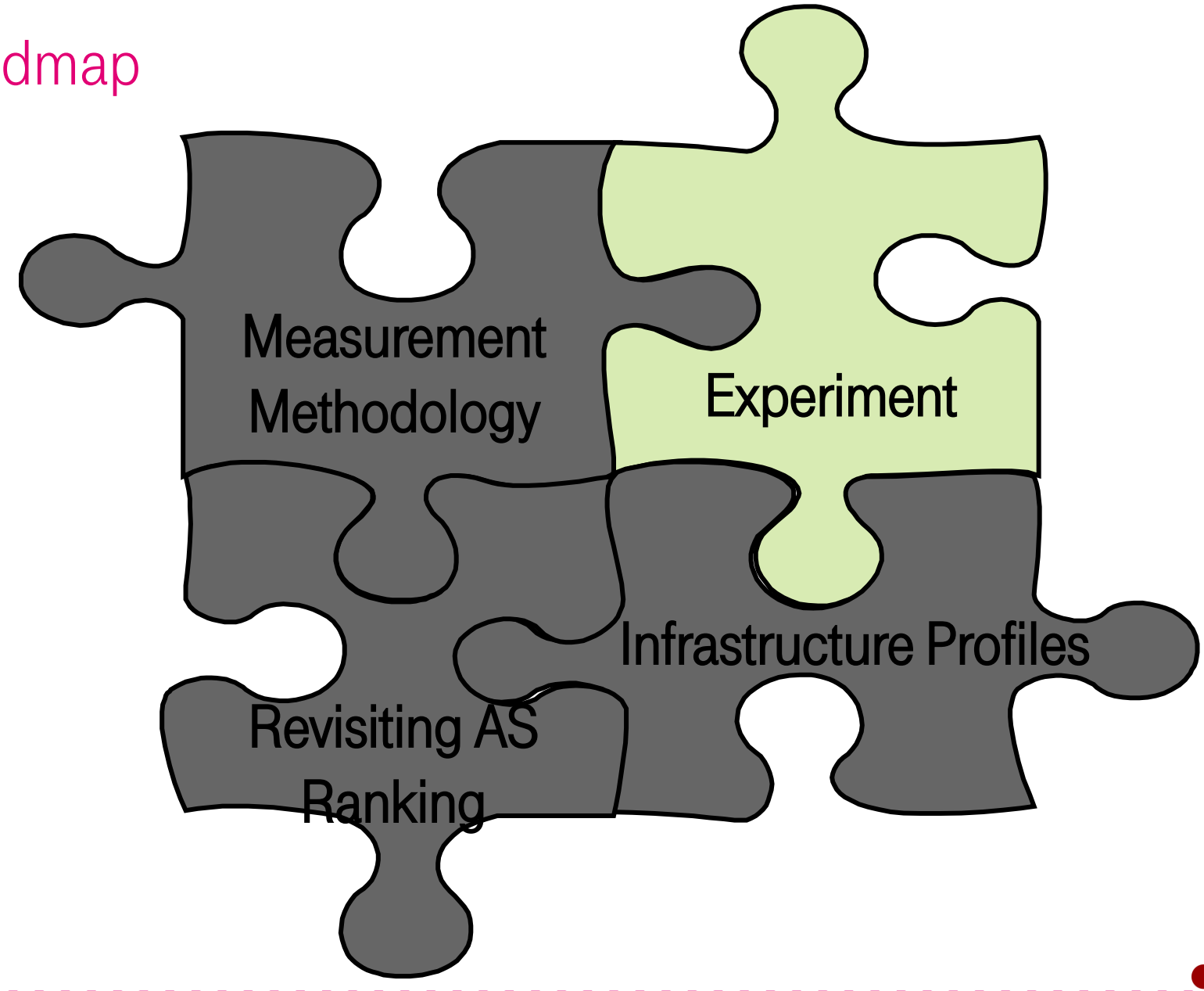
→ Test the similarity of any two sub-clusters,

if they have a high similarity score then merge them.



→ The final set of clusters contains all the hosting infrastructures. (Meta-CDNs are appear as separate clusters)

Roadmap


Measurement
Methodology
Experiment
Infrastructure Profiles
Revisiting AS
Ranking

Experiment & Collection of Traces

Selection of Hostnames.

Requirement: Achieve a good coverage of Hosting Infrastructures.

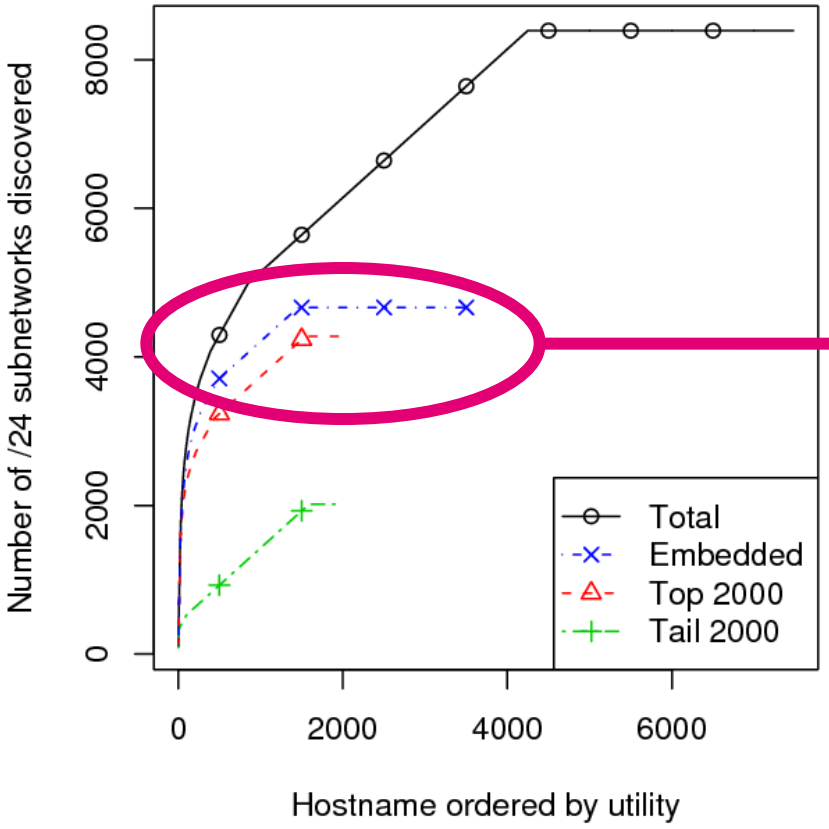
- → Top 2,000 hostnames from Alexa
- → Tail 2,000 hostnames from Alexa
- → Around 3,000 embedded hostnames from the top 2,000

Utilization of Vantage Points.

Requirement: Sampling a large network footprint.

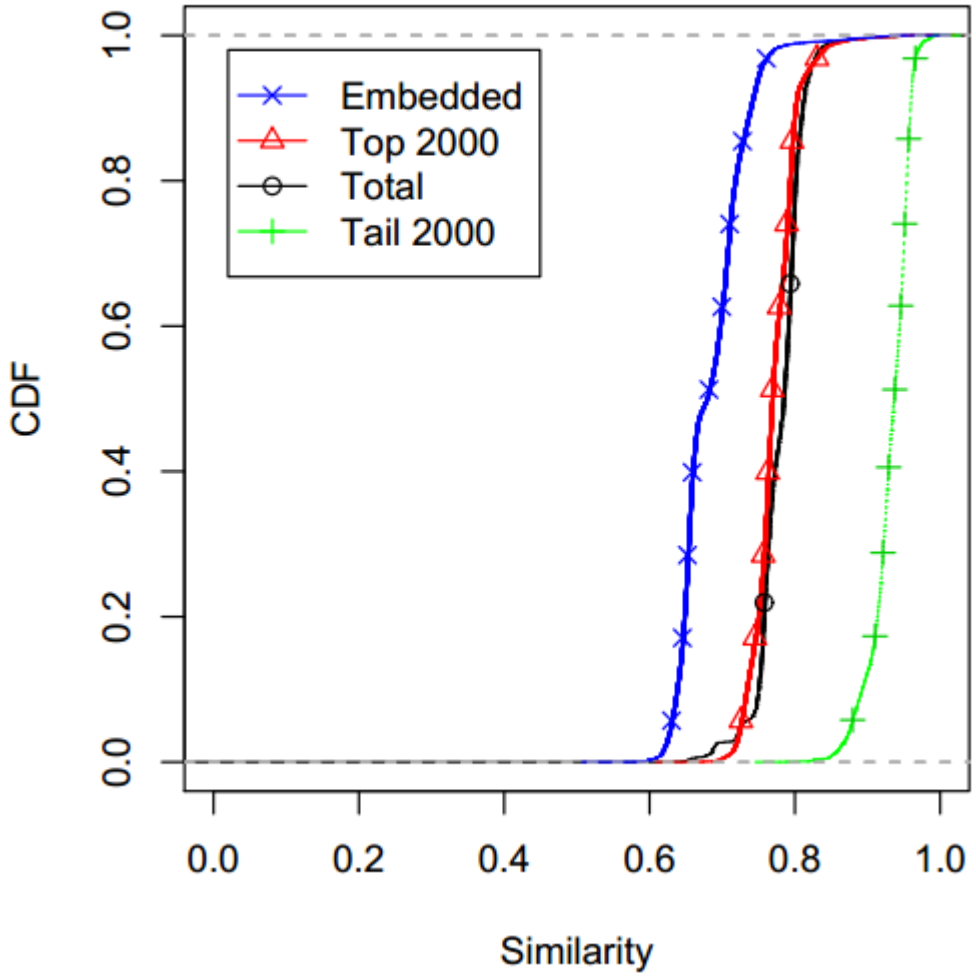
- → Friends & Family network, IMC 2010, IETF mailing lists
- → 133 Traces
- → 78 ASes, many commercial ISPs
- → 27 countries
- → 6 continents

Estimated Coverage: Hostnames



Top and Embedded objects are well distributed

Estimated Coverage: Hostnames

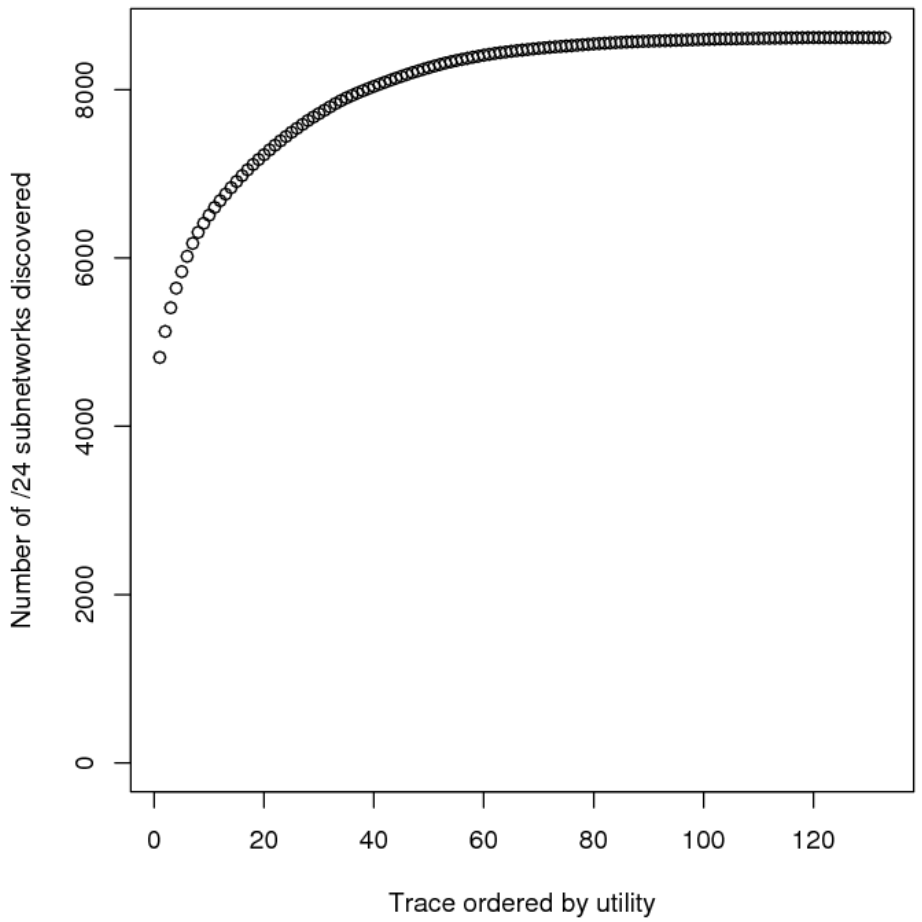


The coverage is complementary

→Tail: consolidation

→Top and Embedded: distributed and closer to the end-user

Estimated Coverage: Traces



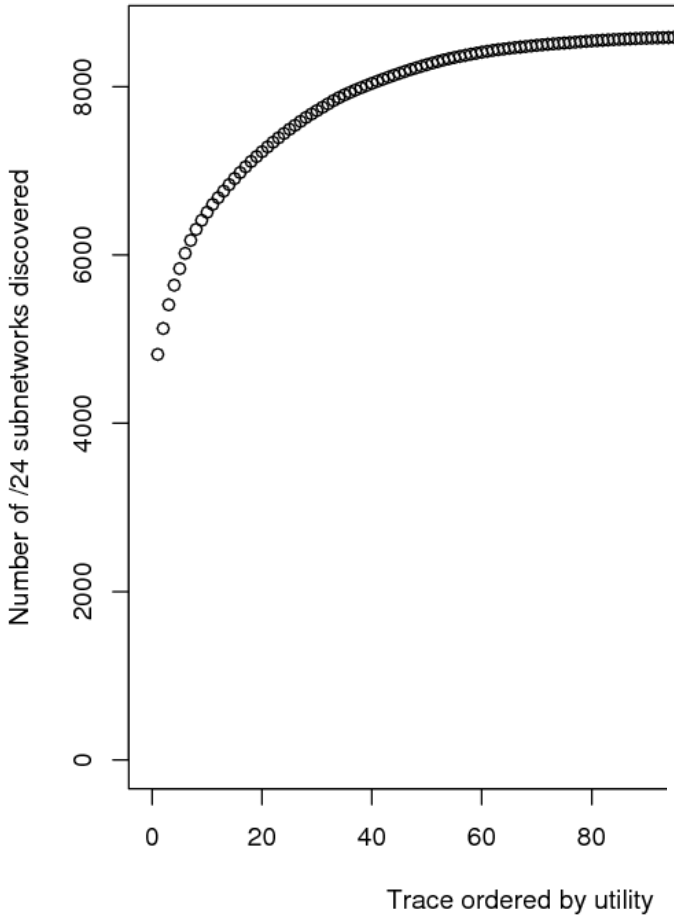
→ A single trace finds more than 4,000 subnets!

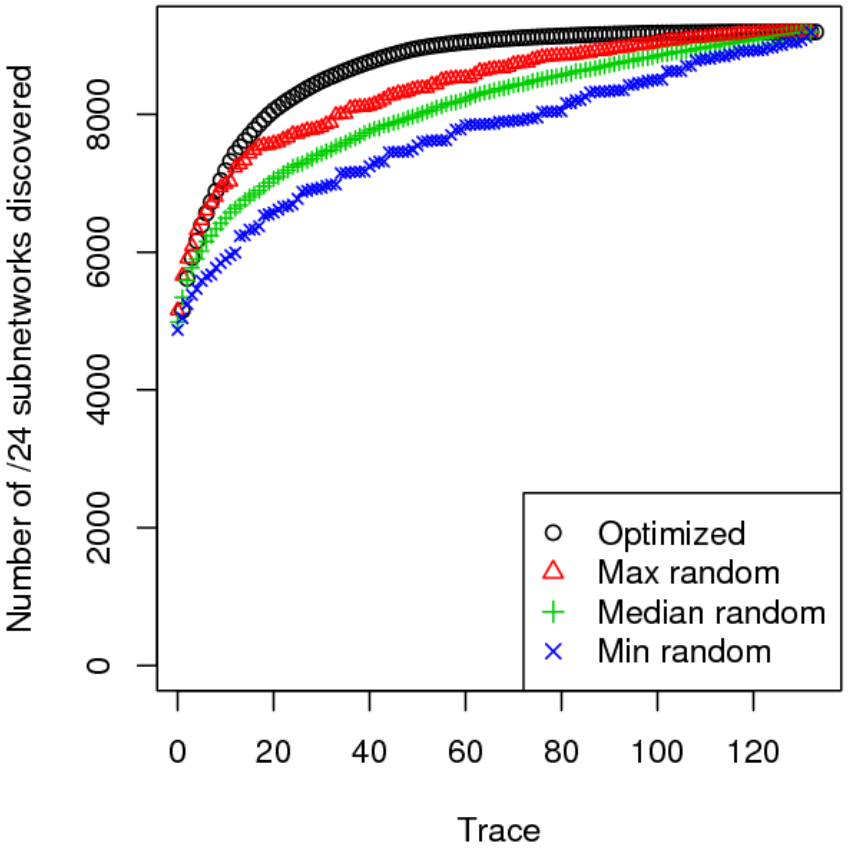
→ Best 30 traces in 30 different ASes in 24 countries.

→ Best 80 traces belong in 67 ASes and 26 countries.

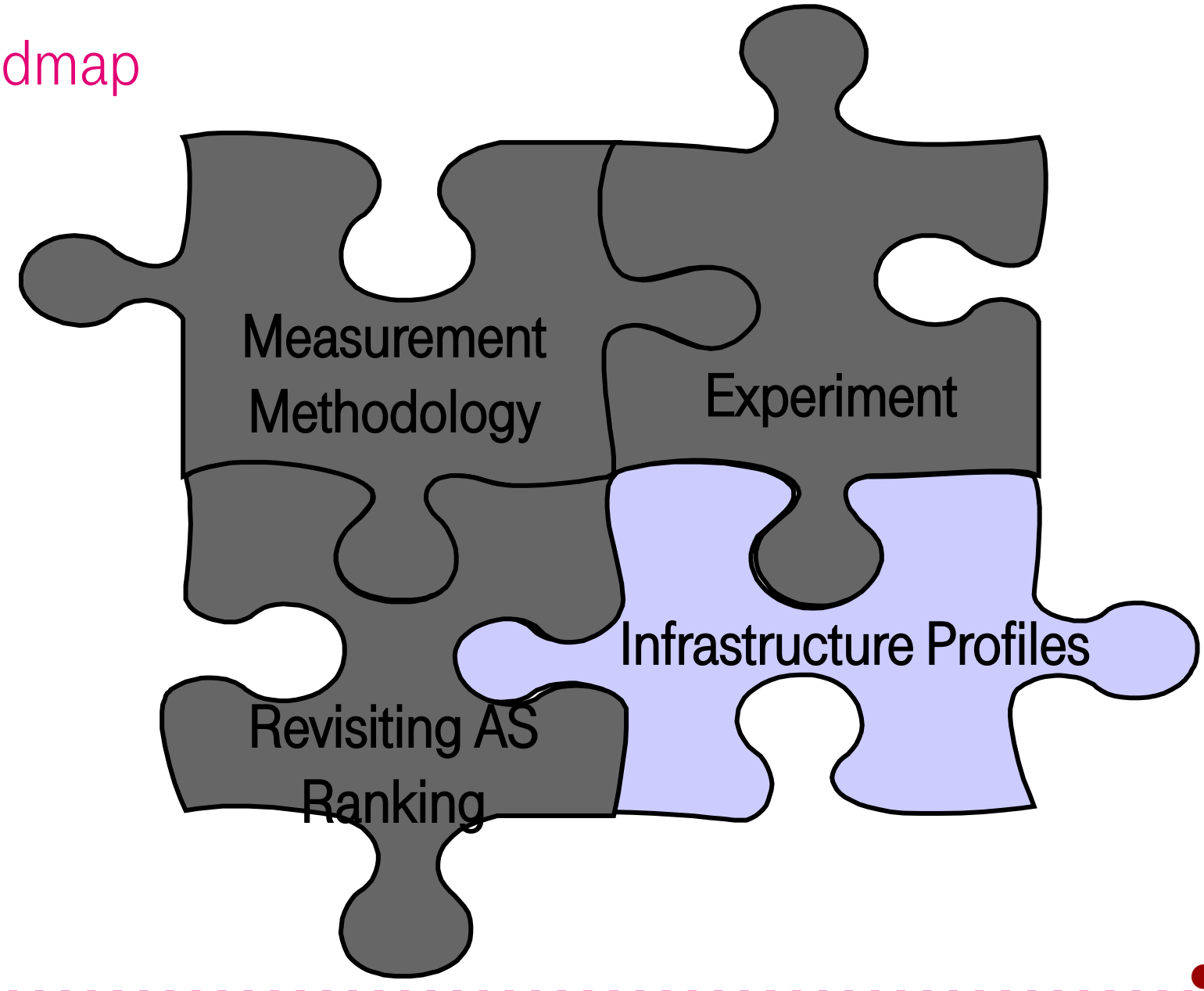
→ Marginal utility to add more traces

Estimated Coverage: Traces

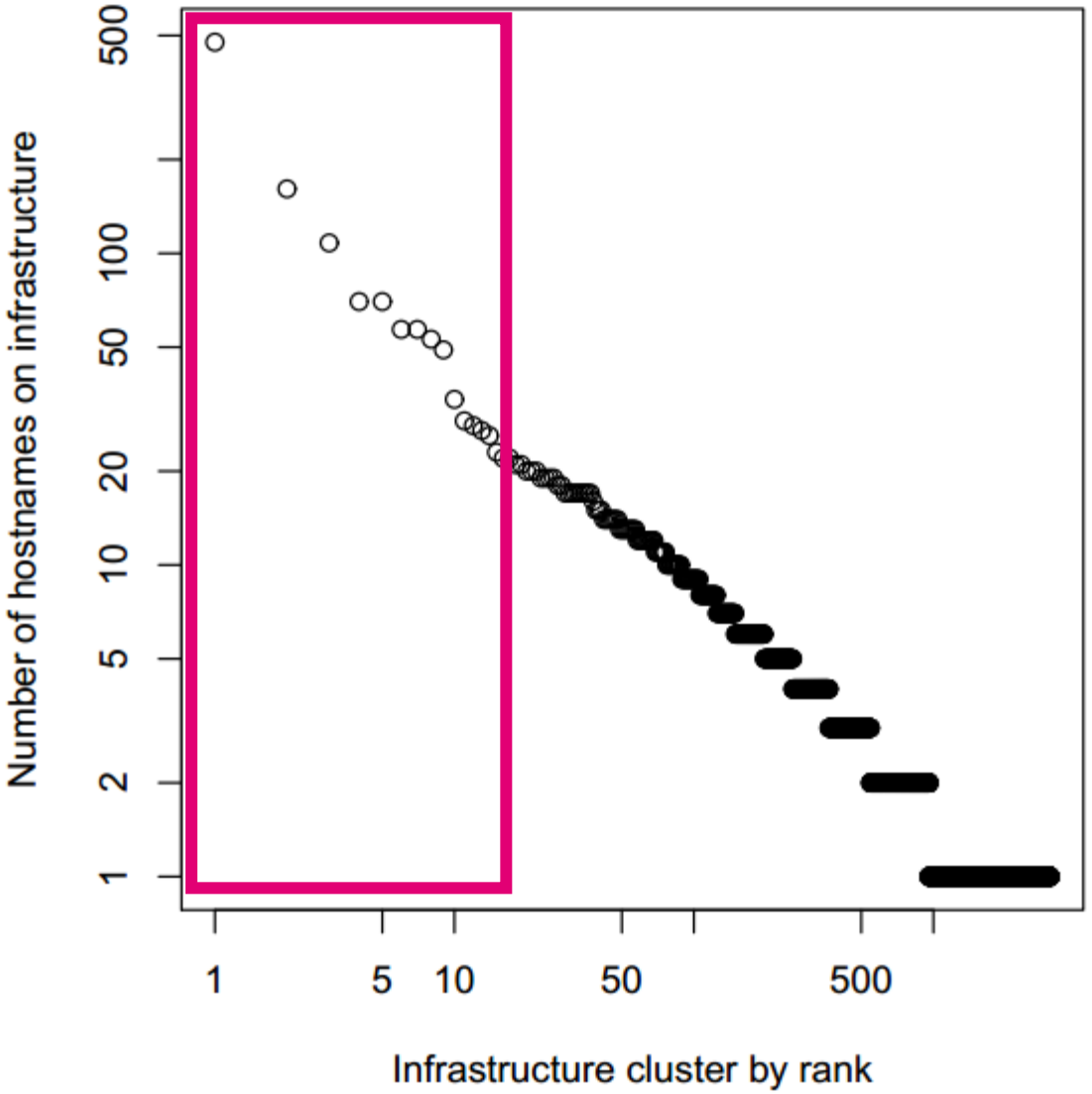




Roadmap


Measurement Methodology
Experiment
Infrastructure Profiles
Revisiting AS Ranking

Cluster Results: The Elephants



Cluster Results: The Elephants

Different Usage of the same Hosting Infrastructure

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 OSN	
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	

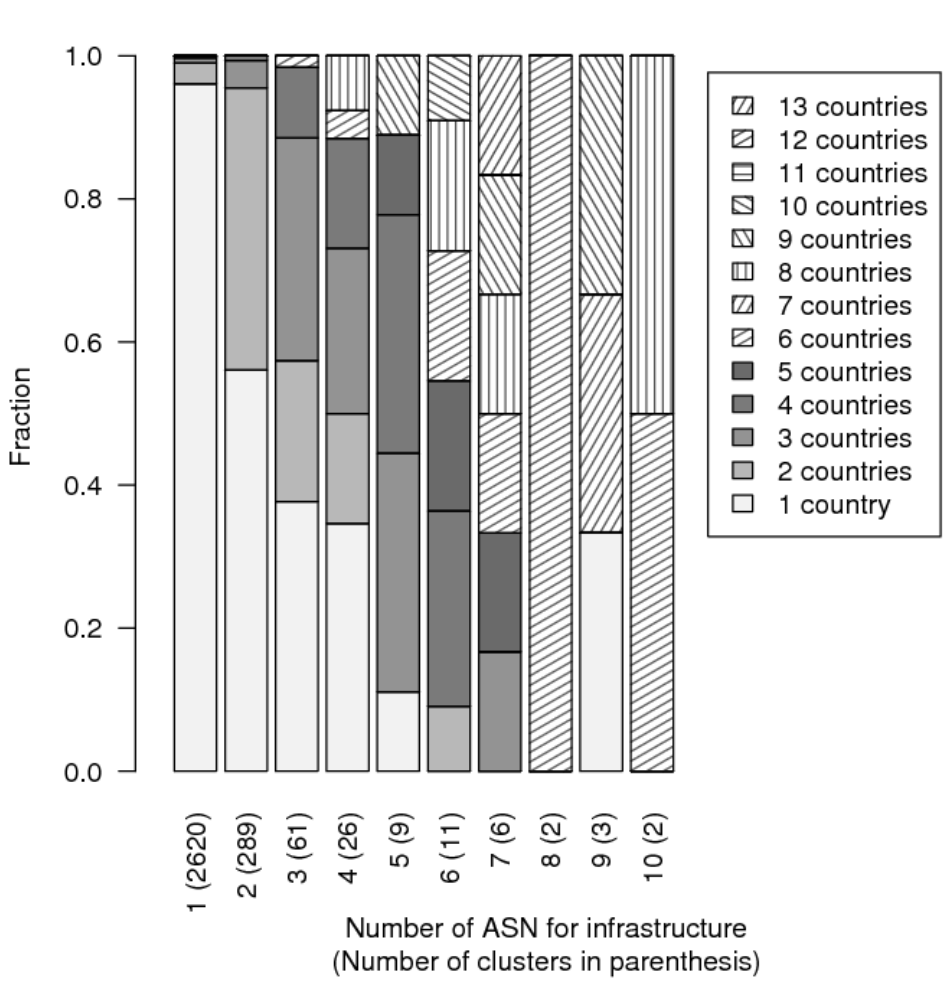
TOP

TOP & EMBEDDED

EMBEDDED

TAIL

Footprint of Hosting Infrastructures



Rank	Country
1	USA (CA)
2	China
3	USA (TX)
4	Germany
5	Japan
6	France
7	Great Britain
8	Netherlands
9	USA (WA)
10	USA (unknown)
11	Russia
12	USA (NY)
13	Italy
14	USA (NJ)
15	Canada
16	USA (IL)
17	Australia
18	Spain
19	USA (UT)
20	USA (CO)

A Continent-level View

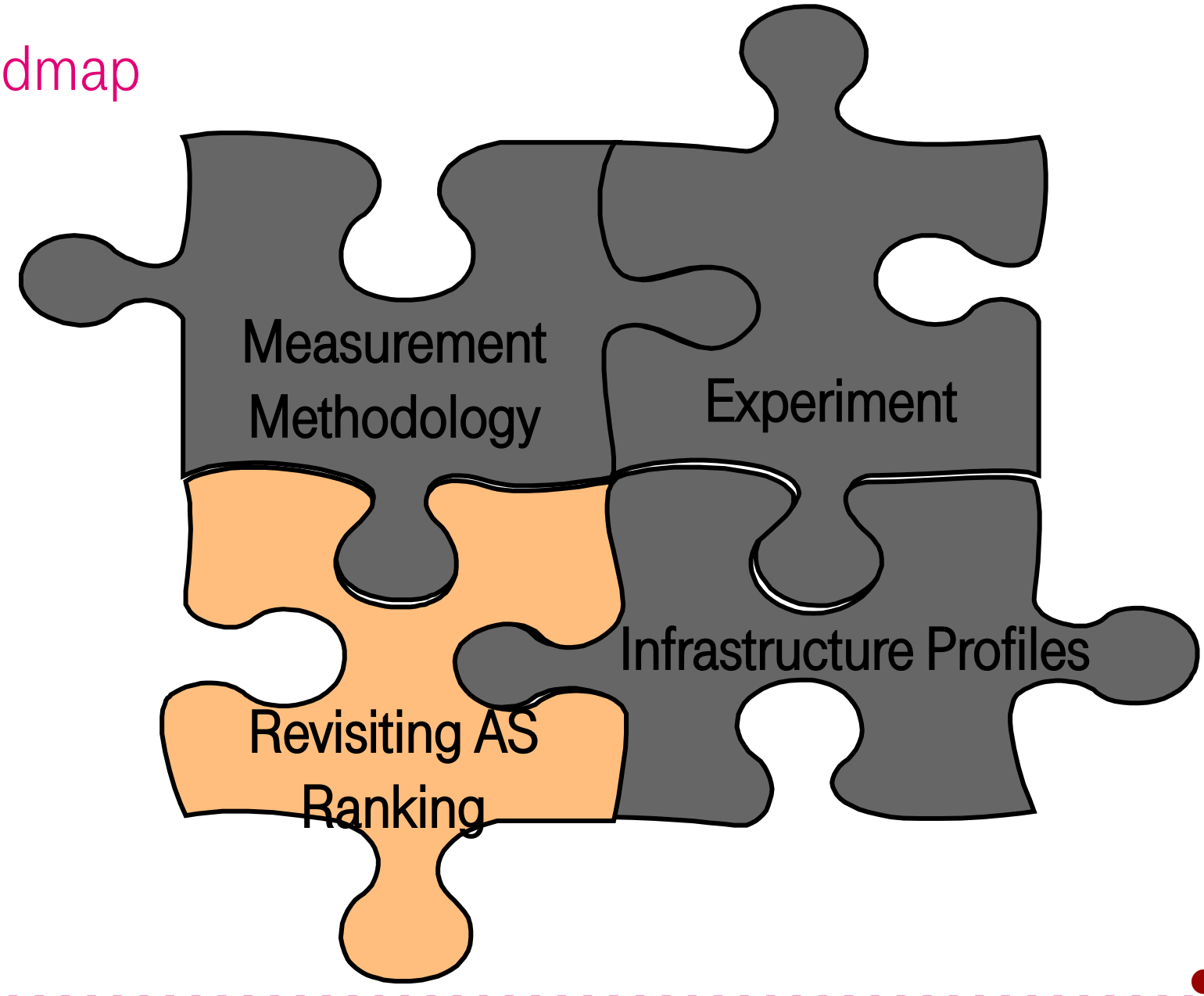
Top Alexa hostnames

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

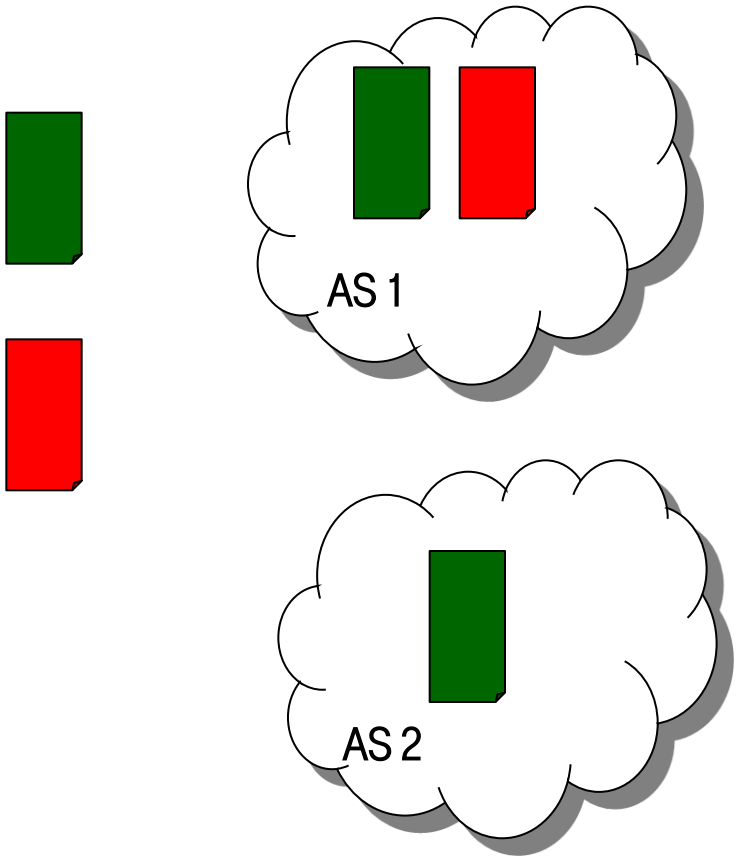
Embedded hostnames

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

Roadmap


Measurement Methodology
Experiment
Infrastructure Profiles
Revisiting AS Ranking

Metrics



Location	CP
AS1	1
AS2	0.5

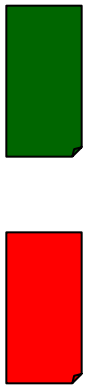
Content Potential (CP):

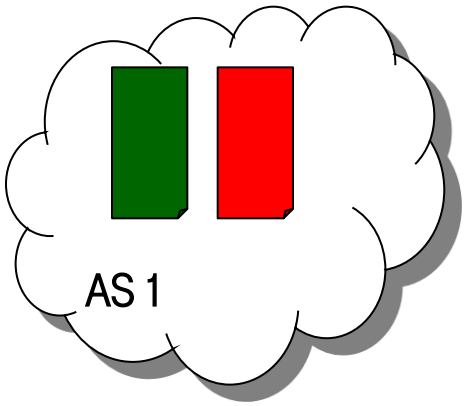
Fraction of content available in a location

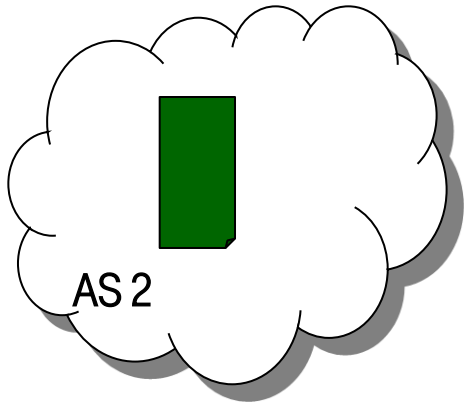
The content can be downloaded from all the instances of the Hosting Infrastructure [1]

[1] "Content Delivery Networks: Protection or Threat?", Triukose, Al-Qudah, Rabinovich , ESORICS 2009

Metrics





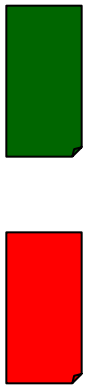


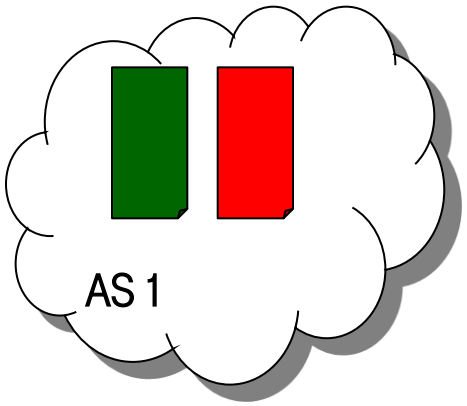
Location	CP	NCP
AS1	1	0.75
AS2	0.5	0.25

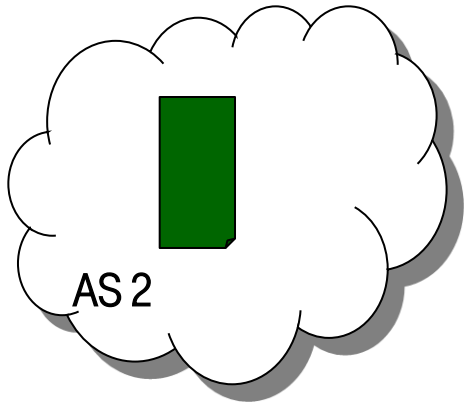
Normalized Content Potential (NCP):

CP normalized by #locations that host the same content, for each content

Metrics





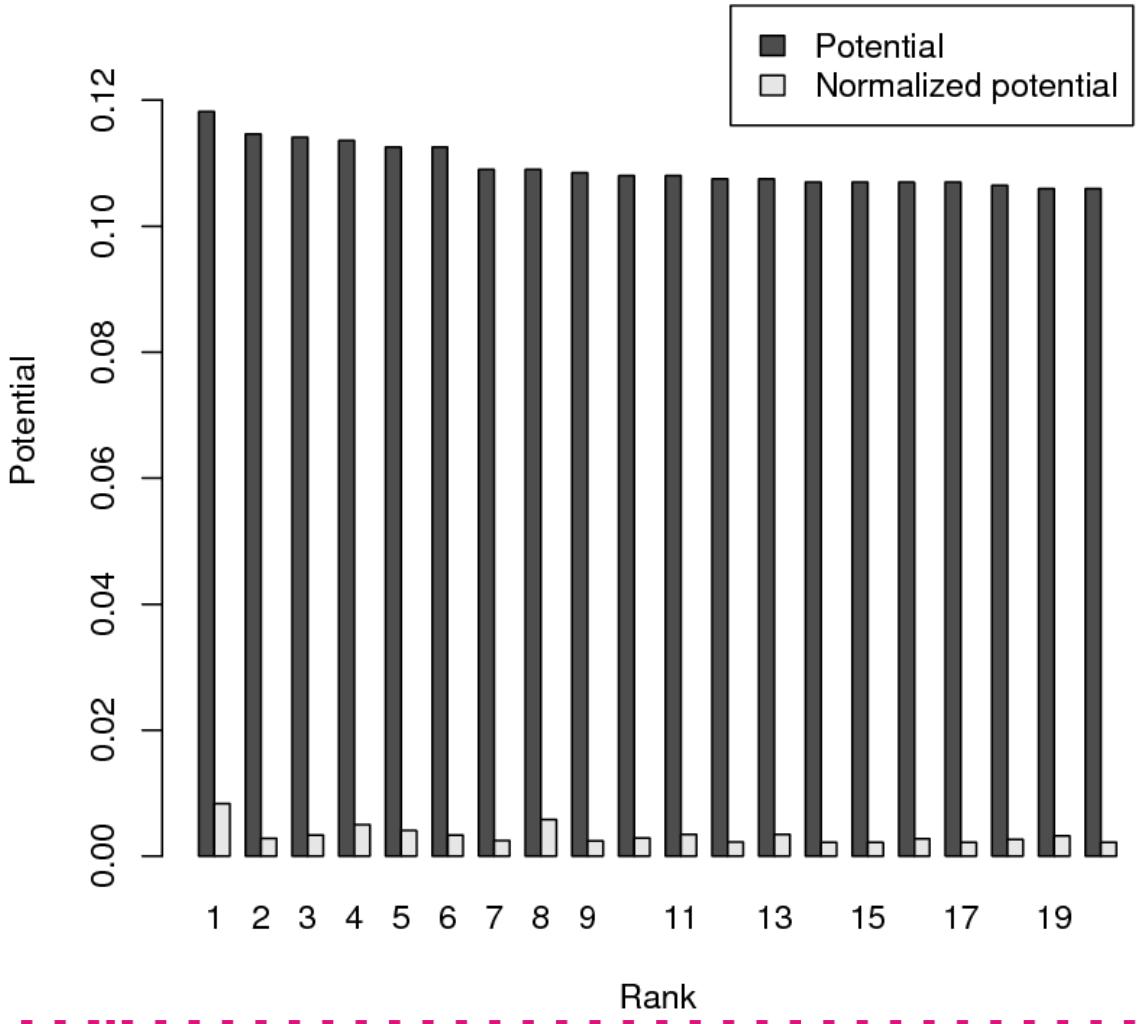


Location	CP	NCP	CMI
AS1	1	0.75	0.75
AS2	0.5	0.25	0.5

Content Monopoly Index (CMI):

CMI = NCP / CP

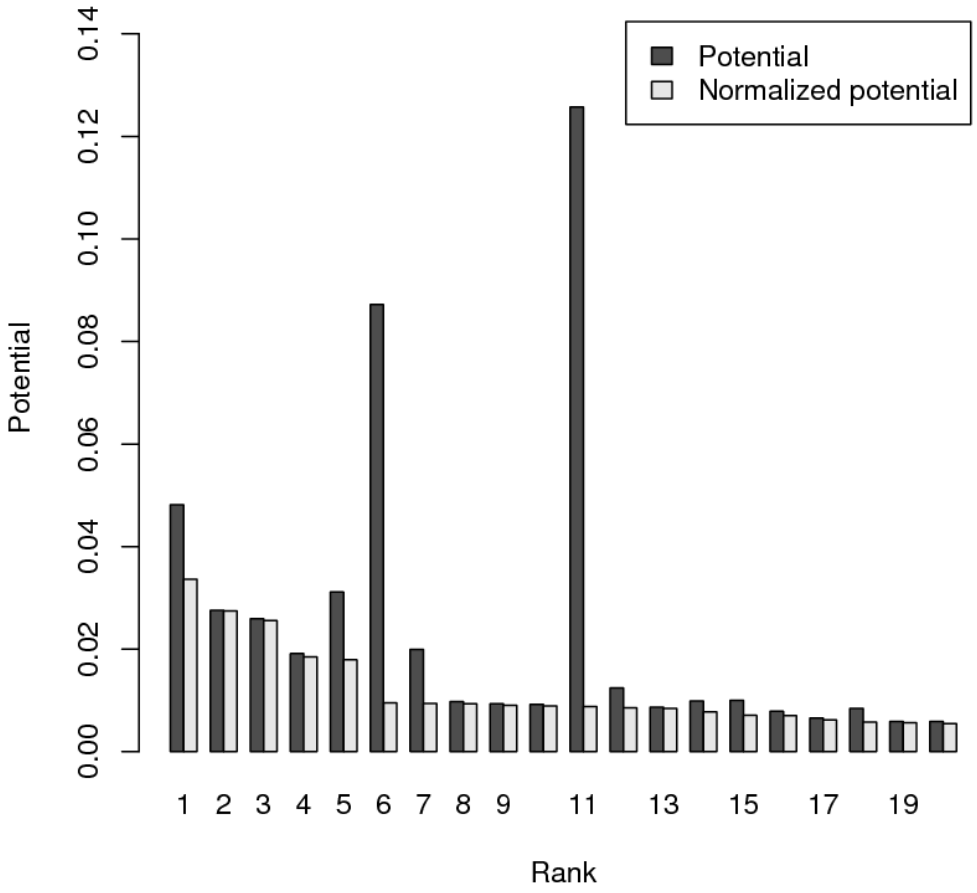
Content Delivery Potential



Tier-1 offer connectivity & CDN, and CDN hosting infrastructures

Rank	AS name	CMI
1	NTT America	0.070
2	Tinet	0.029
3	Global Crossing	0.034
4	KDDI	0.025
5	Akamai Europe	0.019
6	TeliaNet Global	0.027
7	Deutsche Telekom	0.033
8	Korea Telecom	0.030
9	Qwest	0.036
10	Bandcon	0.045
11	Cable and Wireless	0.021
12	SingTel Optus	0.019
13	Akamai	0.018
14	France Telecom - Orange	0.017
15	Internode	0.017
16	Comcast	0.017
17	StarHub	0.018
18	nLayer	0.020
19	Beyond The Network	0.018
20	TATA	0.023

Normalized Content Delivery Potential



High Content Delivery Potential leads to Monopoly

Rank	AS name	CMI
1	Chinanet	0.699
2	Google	0.996
3	ThePlanet.com	0.985
4	SoftLayer	0.967
5	China169 Backbone	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
13	GoDaddy.com	0.969
14	Savvis	0.785
15	China169 Beijing	0.706
16	Amazon.com	0.895
17	LEASEWEB	0.942
18	Cogent	0.687
19	Hetzner Online	0.962
20	AOL	0.932

Comparison of AS Rankings

Rank	CAIDA-degree	CAIDA-cone (#customers)	Arbor (iner-AS traffic)	Potential	Normalized Potential	Content Monopoly Index
1	Level 3	Level 3	Level 3	NTT	Chinanet	Google
2	Cogent/PSI	AT&T	Global Crossing	Tinet	Google	ThePlanet
3	AT&T	MCI	Google	Global Crossing	ThePlanet	GoDaddy
4	MCI	Cogent/PSI	*	Deutsche Telekom	SoftLayer	1&1 Internet
5	Hurricane	Global Crossing	*	KDDI	China169 Backbone	OVH
6	Qwest	Sprint	Comcast	Telia	Level 3	Softlayer
7	Sprint	Qwest	*	Akamai	Rackspace	Hetzner
8	Global Crossing	Hurricane	*	Bandcon	China Telekom	Rackspace
9	tw telekom	tw twlekom	*	Cable & Wireless	1&1 Internet	LEASEWEB
10	INIT7	TeliaNet	*	Qwest	OVH	AOL

Google, Akamai??

Summary

→ Signature-free Identification and Classification of Hosting Infrastructures

→ Lightweight Discovery of Hosting Infrastructure

→ Content-centric AS rankings

→ Tracking of trends in content delivery, e.g., Monopolized delivery and content hosted in Asia.

Current Agenda

→ Improvement of Coverage

→ Incorporate Passive Measurements and Information from Content Providers

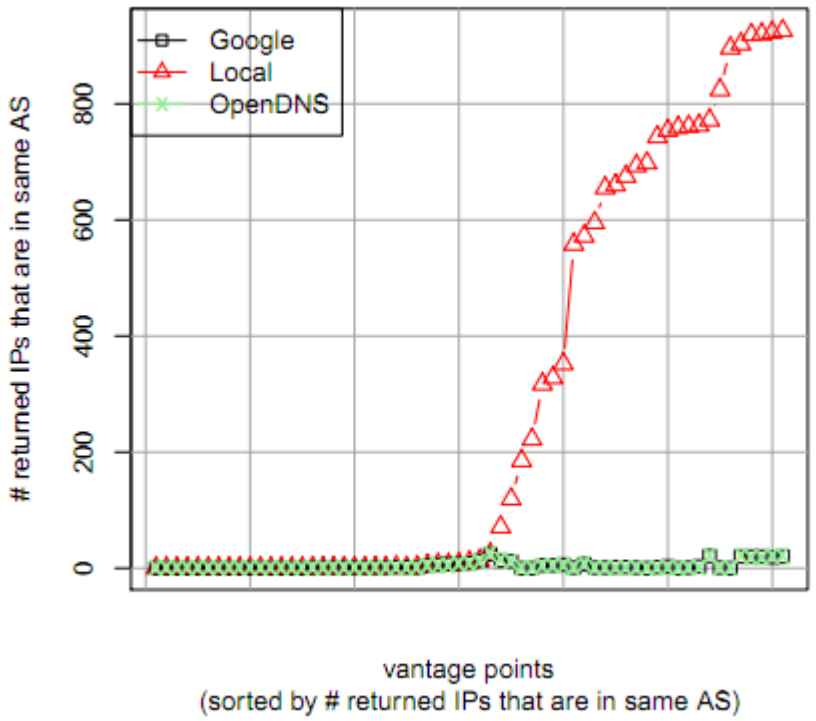
→ Shed light on Strategic Decisions on Content Deployment

→ Investigate further the Interplay between Topological Changes and Content Hosting

Thank you!

Back-up Slides

DNS Quality of Answer (ACM IMC 2010, Comparing DNS Resolvers in the Wild)



→ In many ISPs, the DNS deployment leads to high locality

→ Still, a high number of requests could have been served locally

→ GoogleDNS and OpenDNS mis-direct users (mis-location or sth. else?)