Florian Haemmerling
TU Berlin WS 08/09

Unconstrained Endpoint Profiling (Googling the Internet)

Network Architectures: Internet Measurement

Unconstrained Endpoint Profiling
Ein Überblick

Unconstrained Endpoint Profling
~ Uneingeschränkte Profilbildung von Endpunkten

- *Was ist das Profil von Endpunkten?*
- *Was für Einschränkungen wurden beseitigt?*

Was ist das Profil von Endpunkten?

überregionale Analyse der Internetnutzung

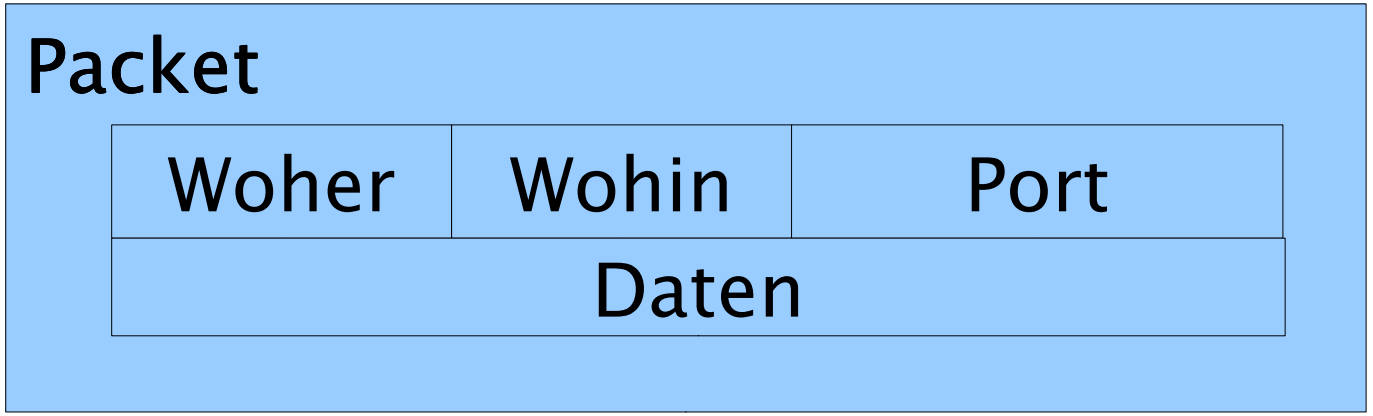
- Beobachtung regionaler Trends
  - ➔ besuchte Webseiten
  - ➔ genutze Dienste
  - ➔ verwendete Betriebssysteme

- Lokalität des Internetverkehrs

Was für Einschränkungen wurden beseitigt?

Wie wird Netz-Verkehr analysiert?

→ z.B. Packetanalyse



Einschränkungen von Paketanalyse:

- Doppelbelegung von Ports
- Dynamische Port Belegung
- Port 80 als „Universal-Tunnel“
- Benutzung von Proxys

→ ähnliche Einschränkungen für Flow-Traces

Was für Einschränkungen wurden beseitigt?

Einschränkungen von Packetanalyse:

- Doppelbelegung von Ports
- Dynamische Port Belegung
- Port 80 als „Universal-Tunnel“
- Benutzung von Proxys

Weitere Einschränkungen:

- Paket-traces werden benötigt
- Hohes Datenaufkommen
- kleiner Ausschnitt

Wie umgeht UEP diese Einschränkungen?

Anstatt Packetanalyse

Aktivität zu IP xyz

Google Such-Anfrage

IP abc

Google

Merkmal zu IP abc

Welche IPs werden bei Google angefragt?

Welche IPs werden bei Google angefragt?

zur Erinnerung:

überregionale Analyse der Internetnutzung

IPs sind regional vergeben!

→ IPs einer Region ergooglen

- Asien (China)
- S. Amerika (Brazilien)
- N. Amerika (USA)
- Europa (Frankreich)

UEP im Detail

Server Listen
- serverlists.com
- dienste-liste.de

Block Listen
- blockedsites.org
- blacklists.com

Foren
- kleinesforum.de
- funforum.org

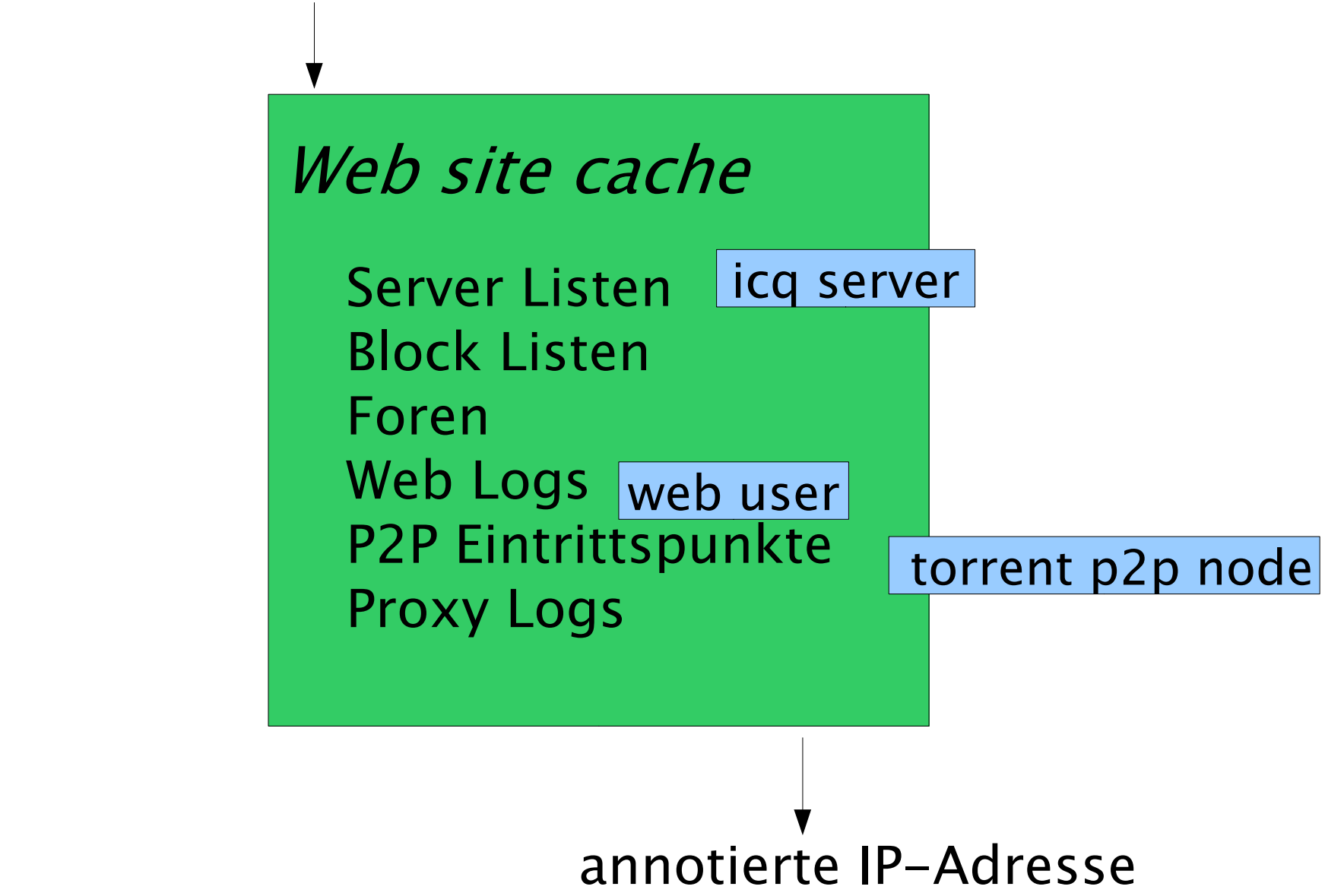
Web Logs
- weblogs.com
- mein-log.de

P2P Eintrittspunkte
- esel-files.de
- p2p-netzwork.com

Proxy Logs
- proxys.de
- proxylogs.com

UEP im Detail

unbekannte IP-Adresse



Beobachtete Trends

zur Erinnerung:

- besuchte Webseiten
- genutze Dienste
- verwendete Betriebssysteme

- Asien (China)
- S. Amerika (Brazilien)
- N. Amerika (USA)
- Europa (Frankreich)

→ Am häufigsten aufgerufene Websites
→ Ist Windows überall dominant?

Beobachtete Trends – aufgerufene Websites

Top 1: Google

insgesamt deutlich über 50 %
regional jeweils ~ 40 – 70 %

Top 2: 

~ 13 % ohne China
~ 8 % mit China
In China stehen
47.584 Google Aufrufe
gegen
170 Wikipedia Aufrufe

Warum ist das so?

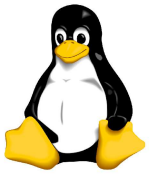
Zensur und Sperrung von Wikipedia durch die chinesische Regierung!

Danach: regional sehr unterschiedlich

Beobachtete Trends – Dominanz von Windows

Meist Benutztes Betriebssystem:

- China
- USA



- Frankreich
- Brasilien

(vielleicht) unerwartet aber erklärbar:

Behörden, Ämter, Schulen setzen vermehrt auf Linux

Ganz genau:

Windows

Ch: > 77 %
US: ~ 48 %
Br: < 40 %
Fr: ~ 21 %

Debian

Ch: < 1 %
US: < 3 %
Br: ~ 37 %
Fr: > 31 %

Ubuntu

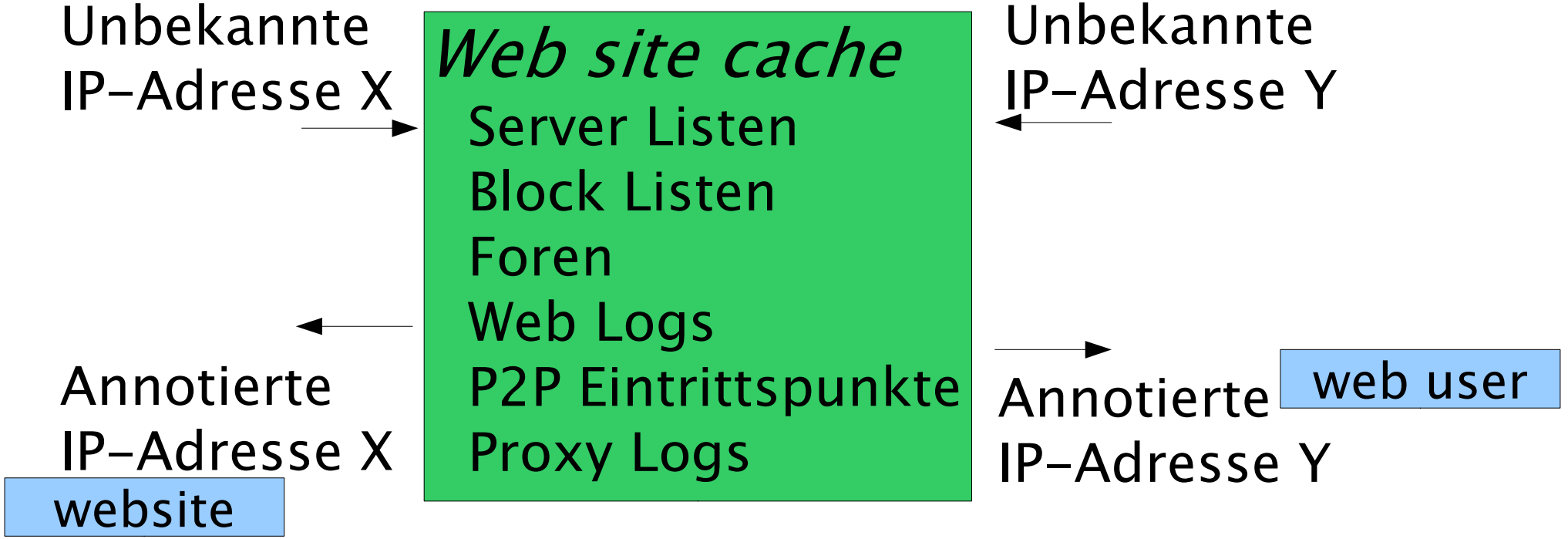
Ch: < 7 %
US: ~ 5 %
Br: > 15 %
Fr: ~ 15 %

Ch: China; US: USA; Br: Brasilien; Fr: Frankreich

Lokalität des Internetverkehrs

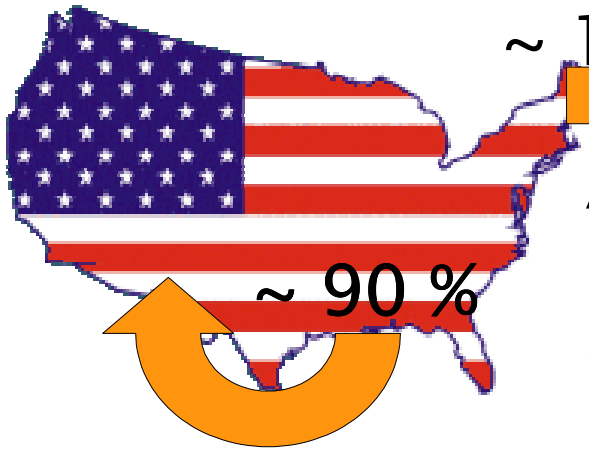
Beantwortung der Frage:
Von wo werden Anfragen einer Region beantwortet?

zur Erinnerung:



- Heraussuchen von Kommunikationspartnern
- Sender- und Empfängernetz auflösen

Lokalität des Internetverkehrs



~ 10 %

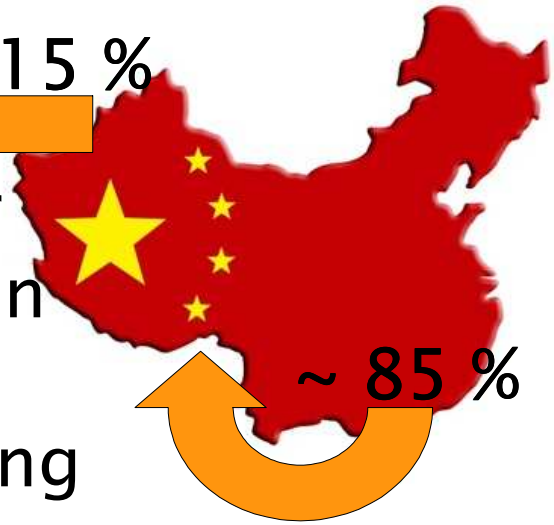
Asien
Europa
Canada

~ 90 %

USA

~ 15 %

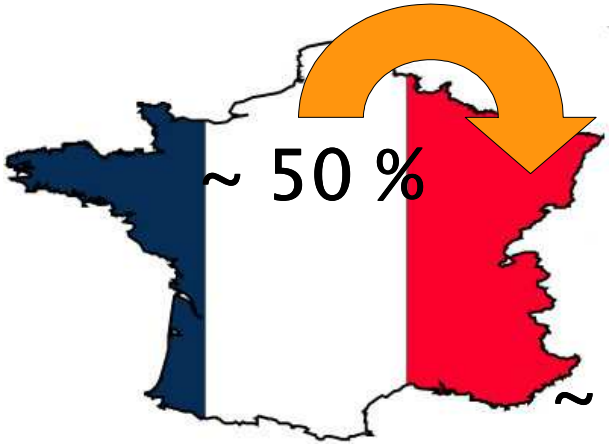
Singapur
Australien
Taiwan
Hong Kong
Japan
USA



~ 85 %

China

Frankreich



~ 50 %

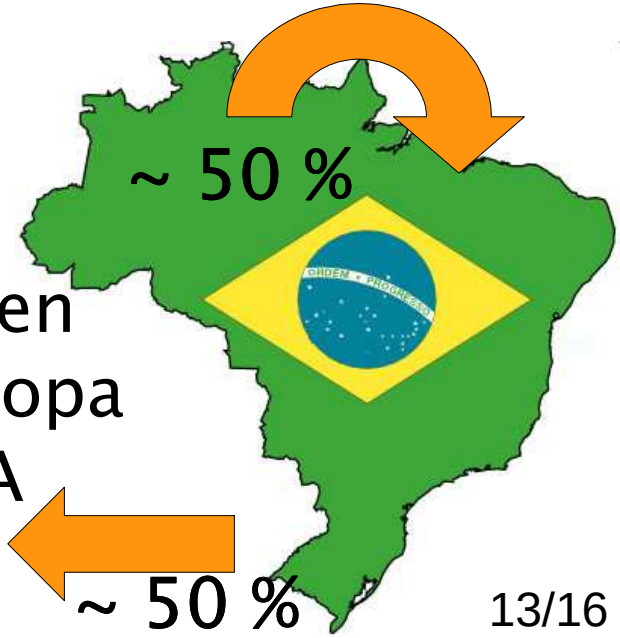
Süd Amerika
Asien
USA

~ 50 %

Brasilien

~ 50 %

Asien
Europa
USA



~ 50 %

Einschränkungen / Grenzen von UEP

Zur Erinnerung:

Unconstrained Endpoint Profling

~ Uneingeschränkte Profilbildung von Endpunkten

Ist UEP komplett uneingeschränkt?

Zur Erinnerung:

IP abc → Google → Merkmal zu IP abc

Google ist nicht gleich Google!

Einschränkungen von UEP durch Google:

- regionale Filterung von Google Ergebnissen
- Aktualität von Google Ergebnissen

Einschränkungen / Grenzen von UEP

Einschränkungen von UEP durch Google:

- regionale Filterung von Google Ergebnissen
- Aktualität von Google Ergebnissen

Gibt es weitere Einschränkungen?

Zur Erinnerung:

Web site cache aus

- Server Listen
- Block Listen
- Foren
- Web Logs
- P2P Eintrittspunkte
- Proxy Logs

– Wie öffentlich zugänglich sind diese Daten wirklich?

– Übergewichtung gefundener Einträge

Zusammenfassung

UEP ermöglicht :

- überregionale Analyse der Internetnutzung
- ohne Zugriff auf Paket-Traces

Trotz Einschränkungen durch:

- Abhängigkeit von Google
- Aufbau des Web site caches

Aussagekräftige Trends über:

- Webseitenzugriffe
- Nutzung von Diensten
- Betriebssystemverteilung
- und weitere

Literaturverzeichnis

UEP:

I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci. Unconstrained Endpoint Profiling (Googling the Internet). In Proceedings of ACM SIGCOMM 2008, Seattle, WA, August 2008.

BLINC:

T. Karagiannis, K. Papagiannaki, and M. Faloutsos. BLINC: Multilevel Traffic Classification in the Dark. In ACM SIGCOMM, Philadelphia, PA, August, 2005.

Earyl Applikation Identification:

L. Bernaille, R. Texeira, and K. Salamatian. Early Application Identification. In CONEXT, Lisboa, Portugal, December 2006.

Google.com Suchmaschine:

http://www.google.com

Standard Ports:

http://www.iana.org/assignments/port-numbers (2008-12-15)

Skype Port-Konflikt:

http://support.skype.com/en_US/faq/FA528/Conflicts-with-applications-such-as-Apache-or-IIS-working-on-port-80-443 (2008-12-15)

Linux in Brasilien:

http://www.brazzil.com/2004/html/articles/mar04/p107mar04.htm (2008-12-22)

Linux in Frankreich:

http://www.linuxinsider.com/story/35108.html (2008-12-22)
http://www.redhat.com/about/news/prarchive/2007/frenchministry.html (2008-12-22)

Wikipedia in China:

http://www.heise.de/newsticker/meldung/print/48343 (2008-12-15)
http://www.heise.de/newsticker/meldung/print/113650 (2008-12-15)