Datamap

Sharon Gieske & Thijs Houtenbos

Supervisors: Jeroen van der Ham & Barry van Kampen

Master System and Network Engineering
University of Amsterdam

2014-02-05

Table of Contents

1. Introduction
   - Motivation
   - Related research
   - Focus of Research
2. Approach
3. Methods & Results
   - Identification & Classification
   - Secure connections
   - Localization
4. Conclusions

Motivation

When visiting a website (first party), often more parties involved (third parties)

Concerns:

- Privacy laws differ between countries
- NSA revelations
- Survey by Annalect[1] in 2013 on online privacy concerns:
  - Lack of knowledge on collection of their information (48%)
  - Lack of control over how personal information is used (61%)

Related research

Academic research (novel)

- Third-party type distribution
- Third-party penetration on JavaScript cookies [2]
- Fingerprinting

Projects:

- Waarismijndata.nl
- Mozilla Lightbeam
- Ghostery

Focus of Research

Research question:
What is the scope of (privacy) infringing data sharing of the top visited websites with third parties?

Subquestion:

1. *Which third parties are involved when visiting a website?*
2. *Can data potentially be accessed by third parties?*
3. *What is the geographical distribution of your data?*
4. *Which differences in data sharing can be found between countries for national and global first-parties?*

Approach

First parties:

- Alexa's top 10,000 websites
- Alexa's top 1,000 websites of NL, CN & US domains

Approach for third parties:

- Identification & classification
- Identification of secure connections
- Localization

Overview approach

How to find third parties?

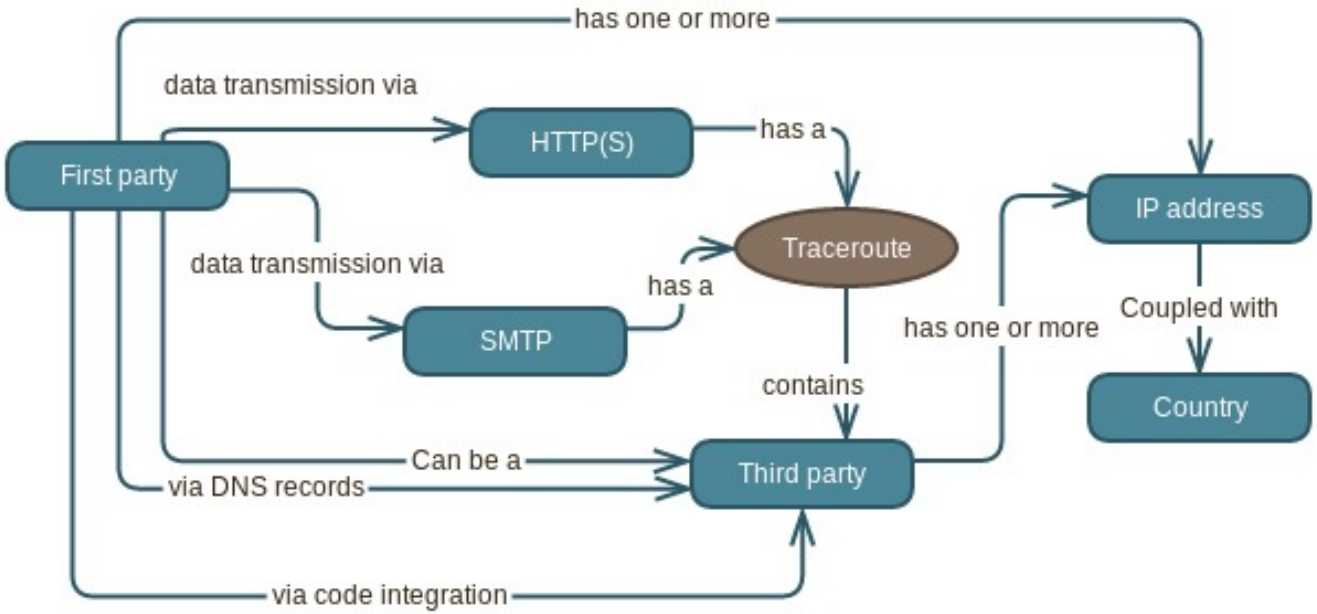


Figure: Relation overview

Identification & Classification

Third parties through:

- DNS resource records
- JavaScript Objects
  - Classification via Ghostery:
    Analytics, Widget, Tracker, Ad, Privacy
- Routes of data
  - Traceroutes for websites (ICMP, UDP, TCP)
  - E-mail routes via header analysis

Results: Identification & Classification

Total:

- Third parties: 23,420 third parties (84,647 subdomains)
- Traceroutes: 30,165 routes (46,668 hosts discovered)
- E-mails: 37,122 e-mail replies (13,287 hosts discovered)

Third parties overview:

	Total	Mean	STD	Top
DNS	9,164	2	1	8
JavaScript	17,215	13	16.2	133
Traceroutes	40,286	12.8	5.9	43
Email trace	13,121	29.8	34.3	99

Table: Third parties identified

Results: Identification & Classification (1/3)

Observations resource records:

- 2 significantly bigger MX third parties:
  - GOOGLE.com and googlemail.com (4,272 of 8,968 first parties)
  - also in US & NL domains. In CN domains: qq.com
- 4 big DNS name servers, differ on country-level
- No significant CNAME directions

Results: Identification & Classification (2/3)

top domain	# http-requests
doubleclick.net	10,573
facebook.com	9,541
google.com	7,904
google-analytics.com	7,024
twitter.com	5,997

Table: *Top domains in JavaScript code integration*

class	name	count
ad	DoubleClick	10,718
ad	AppNexus	3,278
widget	Facebook Connect	2,419
ad	Rubicon	2,363
ad	Quantcast	2,190

Table: *Top classifications in HTTP*

Observation

Code integration:

- analytics - CNZZ (CN)
- ad - Baidu Ads (CN)

Observations classification:

- $\pm$ 60% ad (Top, US, NL)
- $\pm$ 50% analytics classification (CN)

Results: Identification & Classification (3/3)

Observations on email headers:

- A total of 5,690 addresses was obtained (mostly internal)
- 2,033 are externally accessible unique IPs that were not found in other records

Figure: Nanoniem website

Secure connections

Secure connections:

- DNSSEC:
  - 117 first party domains with dnskeys
- HTTPS:
  - 5,811 first party domains secured
- TLS:
  - 2,749 of 13,669 (20% of distinct IPs)
- DKIM:
  - 584 of 8,905 (6.6% of total domains)

Observations:

HTTPS in top 1,000	594 (NL)	393 (US)	71 (CN)
DNSSEC in top 1,000	200 (US)	184 (NL)	17 (CN)

Localization

Localization:

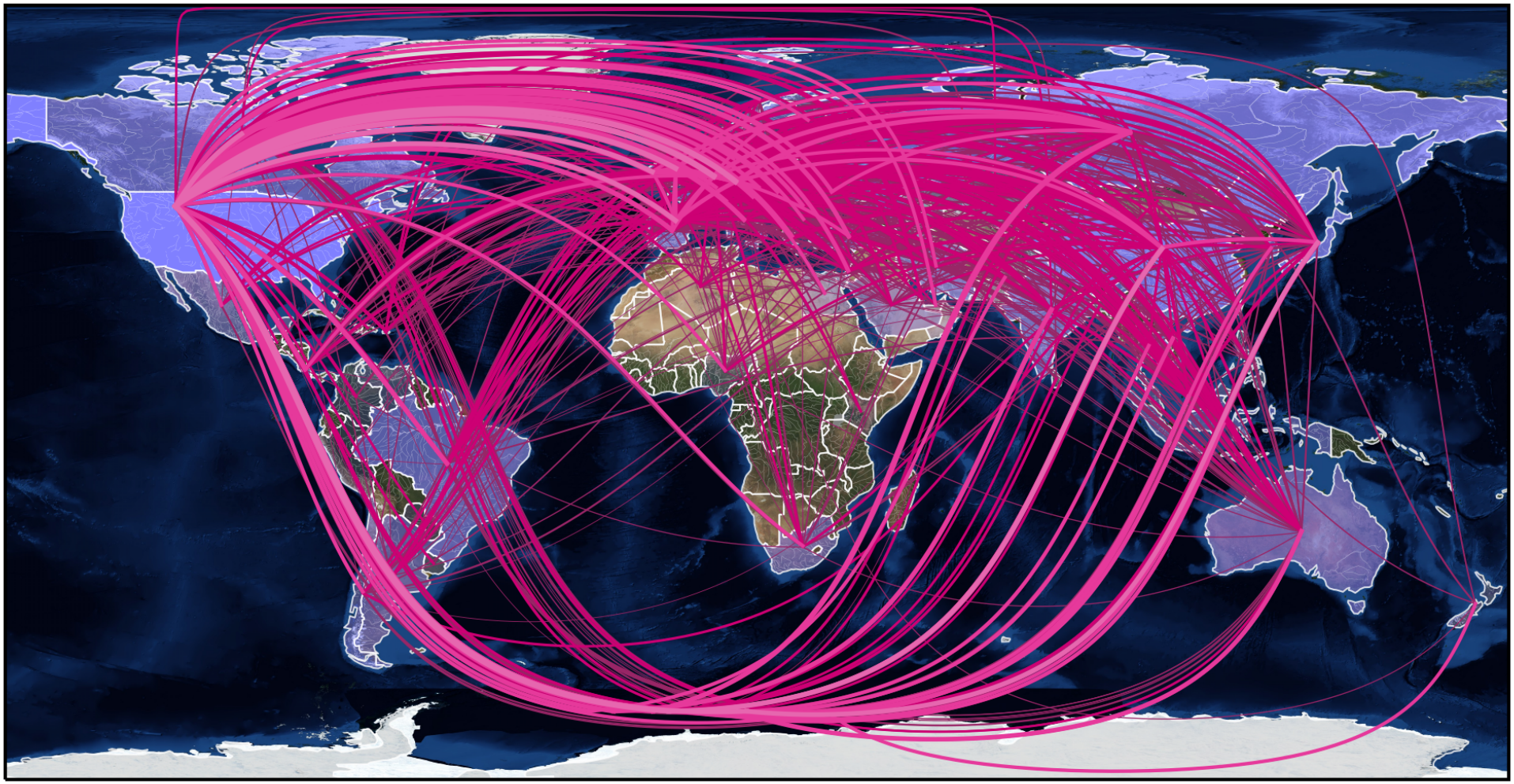
- IPv4 addresses via A records
- Additional IPv4 addresses from email
- Country via GeoIP database
- AS via WhoIs lookup

Third parties countries overview: The number of countries per domain

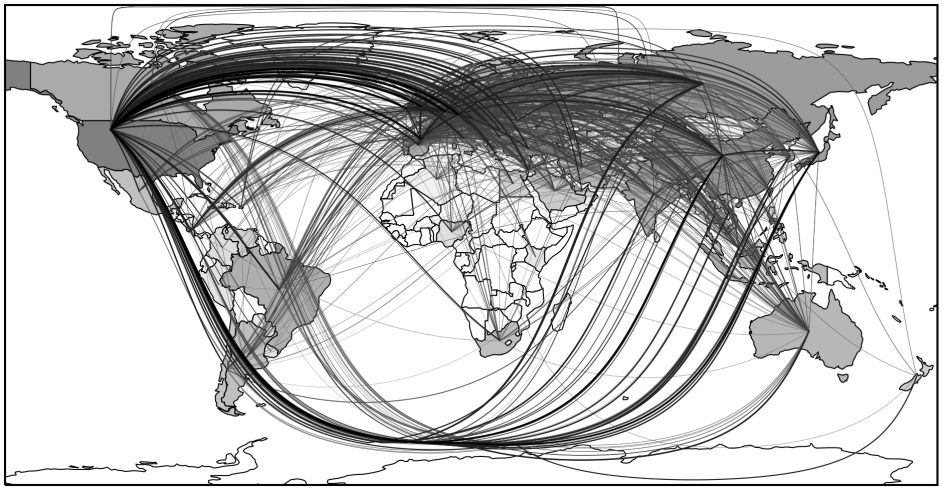
	Mean	STD	Top
Top 10,000	5.79	2.87	20 (mazika2day.com)
Top 1,000 NL	4.70	2.66	16 (sony.nl)
Top 1,000 US	3.37	2.13	13 (breakz.us)
Top 1,000 CH	-	-	-

Table: Third parties identified

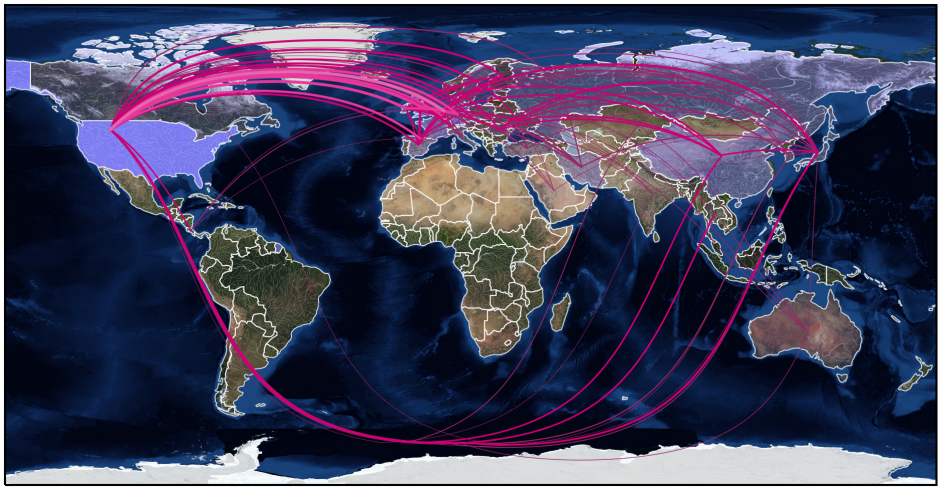
Localization of third parties



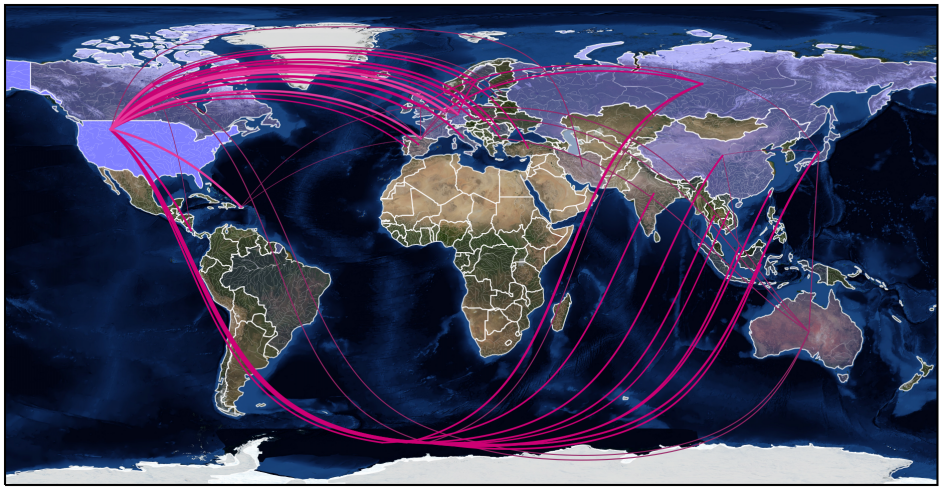
Localization of third parties



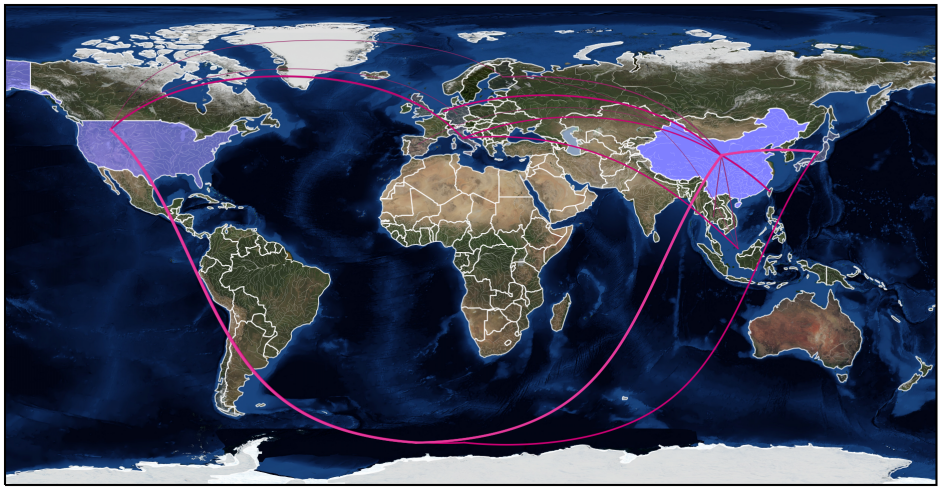
(a) Global overview



(b) NL top 1000 overview



(c) US top 1000 overview



(d) CN top 1000 overview

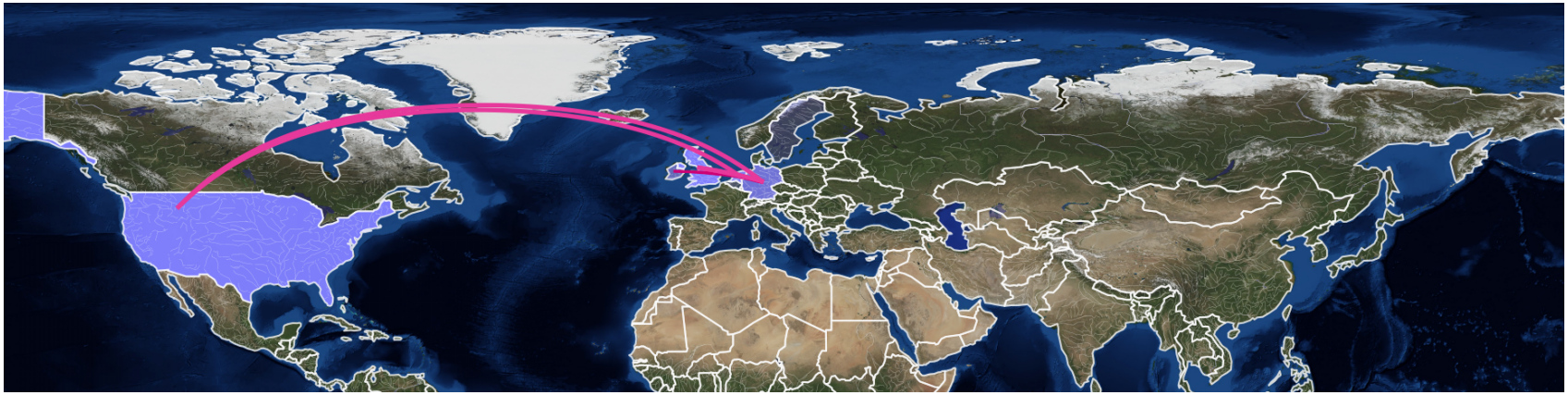


Figure: Third parties of alternate.nl

Figure: All intermediate routes of alternate.nl

Conclusions (1/2)

1. *Which third parties are involved when visiting a website?*
   - Many third parties per first party
   - Big domains in resource records stand out
   - US & NL very similar, CN different
   - Most third parties obtained via HTTP
   - Classification mostly advertisement

2. *Can data potentially be accessed by third parties?*
   - Mostly non-secure, differs per country

Conclusions (2/2)

3. *What is the geographical distribution of your data?*
   - The number of countries varies wildly (from 1 to 20)
   - Local websites use slightly less foreign servers
   - CN stands out: Most third parties are local, with US second

4. *Which differences in data sharing can be found between countries for national and global first-parties?*
   - US and NL are very similar
   - CN stands out: Firewall, less westernization

Future Work

- Other code integration methods (Flash Objects)
- Extensive classification
- Indexing of countries to privacy policies
- Analysis of more countries

To conclude

Significant wide scope of data sharing with third parties

$\rightarrow$ mostly via code integration

$\rightarrow$ big players in field

Differences between countries in secure connections and routing

Questions

References I

Annalect. *Annalect Q2 2013 Online Consumer Privacy Study*. 2013. Americans Concerns About the Privacy of Online Information Jump in the Wake of NSA Disclosures.

Balachander Krishnamurthy and Craig Wills. Privacy diffusion on the web: a longitudinal perspective. In *Proceedings of the 18th international conference on World wide web*, pages 541–550. ACM, 2009.