University of Amsterdam

Research Project

Data Map

Authors:
S.A. Gieske
T. Houtenbos

Supervisors:
Dr. J.J. van der Ham
B. van Kampen

Master System and Network Engineering

February 2014

UNIVERSITY OF AMSTERDAM

Abstract

Graduate School of Informatics

Master System and Network Engineering

Data Map

by S.A. Gieske
T. Houtenbos

The research described in this paper gives an analysis of the scope of data sharing of the top visited websites with third parties. Third parties for global top 10,000 visited websites and national top 1,000 websites are identified through several methods. An analysis is done on the security of their connections. In addition, the geographical distribution of these third parties is identified.

Results show DNS records introduce third parties through MX and NS records. A significant amount of MX records point to the same organisations, which shows a dominant presence in this field. Third parties introduced via NS records are mostly localized in range of the third party. HTTP requests also show a significant high introduction of third parties in which a first party introduced 13 third parties on average. These requests are classified and results show a large number of HTTP requests for advertisement goals. Traceroute also introduce several third parties since the data is routed through several parties. A low number of secure connections is found for HTTP-traffic and e-mail traffic which suggests data can potentially be accessed by (unauthorized) parties due to security vulnerabilities. The analysis of the geographical distribution of data sharing shows only 8.4% of domains host all their services in the first-party country. Roughly the same amount of domains domains (8.44%) geographically distribute data to parties in 10 to 20 different countries. Most domains (83.2%) distribute data across 2 to 9 different countries, on average to 5.79 different countries.

This research shows a wide scope of data sharing with third parties. It also illustrates difference in data sharing with third parties between global and national domains.

Contents

Abstract **i**

Contents **ii**

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Research Question . . . 2
- 1.3 Related Work . . . 2

2 Methodology **4**
- 2.1 Identification of third parties . . . 5
  - 2.1.1 Third parties through DNS records . . . 5
  - 2.1.2 Third parties through HTTP requests . . . 5
  - 2.1.3 Third parties through data routes . . . 6
    - 2.1.3.1 HTTP-traffic . . . 6
    - 2.1.3.2 E-mail traffic . . . 7
- 2.2 Secure Connections . . . 7
  - 2.2.1 First-party website . . . 8
    - 2.2.1.1 DNSSEC . . . 8
    - 2.2.1.2 HTTPS . . . 8
  - 2.2.2 E-mail traffic . . . 8
    - 2.2.2.1 DKIM . . . 9
    - 2.2.2.2 TLS . . . 9
- 2.3 Geographical Distribution . . . 9
- 2.4 Global and National Comparison . . . 10

3 Results **11**
- 3.1 Identification of third parties . . . 11
  - 3.1.1 Third parties through DNS records . . . 11
  - 3.1.2 Third parties through HTTP-requests . . . 14
  - 3.1.3 Third parties through data routes . . . 15
    - 3.1.3.1 Traceroute . . . 15
    - 3.1.3.2 E-mail trace . . . 16
- 3.2 Secure connections . . . 16
  - 3.2.1 HTTP traffic . . . 16
    - 3.2.1.1 HTTPS . . . 17
    - 3.2.1.2 DNSSEC . . . 17
  - 3.2.2 E-mail traffic . . . 17

3.2.2.1 TLS 18
3.2.2.2 DKIM 18
3.3 Geographical Distribution 18
3.3.1 Autonomous Systems 20
3.3.2 Global and National comparison 20

4 Conclusions **23**
4.1 Conclusions 23
4.2 Future Work 25

A E-mail template used **27**

Bibliography **28**

Chapter 1

Introduction

1.1 Motivation

A big evolvement on the web is the composition of "first-party" web pages from content of multiple "third-party" websites. These third parties enable first-party websites to trivially implement several web services ranging from advertising to social network integration and more. These contributions have brought tremendous value to the web.

However, they also give rise to privacy concerns. Security vulnerabilities are introduced via several design choices. The composed web pages often make use of HTML, JavaScript and CSS in which no restrictions are made for the inclusion of elements from or complete control delegation to unrelated third-party websites. These security vulnerabilities, such as cross-site scripting[1] and cross-site request forgery[2], can enable the unauthorized third parties to retrieve information from the first-party websites in which the user voluntarily participated.

In addition, recent revelations about the scope of privacy invading data mining by the NSA and other secret services[3] have intensified the debate on the privacy of user data and data sharing. Several big companies have also been accused of collaboration with the NSA[4] on data sharing. According to the survey conducted by Annalect[5] in 2013 on online privacy concerns, the percentage of concerned Internet users has increased in July by 19% to 57% after seven weeks of daily coverage of these revelations. As a response, nearly one-third (31%) of the respondents said to have taken actions to protect their online privacy. However, almost half of the respondents (48%) feel they do not know enough about how their information is collected and 61% feel they do not have control over how their personal information is used.

This research strives to define the scope of privacy infringing data sharing to increase awareness among Internet users of the status of their online privacy.

1.2 Research Question

The research question on which is focused is set as: *'What is the scope of (privacy) infringing data sharing of the top visited websites with third parties?'* and will be applied on the Alexa's top 10000 websites. [1]

In addition, international and national data sharing will be analysed. Experiments on the Alexa's top 1000 websites of three country-specific domains will be conducted from within their associated country. The NL-domain, the CN-domain and the US-domain are chosen, each associated with the Netherlands, the United States and China respectively. The focus on these countries is made because they differ in privacy laws in comparison with one another.

In order to answer the research question, this research will focus on the following four subquestions:

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?*

1.3 Related Work

Third-party data sharing is still a novel research topic in the field of data privacy. Mayer and Mitchell [6] have identified 6 different third-party business models in their research: advertising companies, analytics services, social integration, content providers, front-end service and hosting platforms. Research on privacy diffusion by third parties is conducted by Krishnamrthy and Wills[7], which provides a longitudinal study on third-party penetration showing an increase of nearly 60% for first-party servers which set cookies to be used by third-party JavaScript. These studies show a diverse and widely spread integration of third-party data sharing.

The sharing of data with third parties can be executed via different techniques. A popular technique is fingerprinting, which is the collection of information for the purpose of identification. Eckersley[8] demonstrated in 2010 the successful combining of benign characteristics of a browser's environment to create unique device-specific fingerprints. Research by Deyer et al.[9] shows a negative picture of the usefulness of efficient, low-level countermeasures against website-fingerprinting attacks. In research by Herrman et

[1] `http://www.alexa.com/topsites`

al. [10] experiments are conducted using a Multinomial Naive-Bayes classifier to show the ability to successfully create a website fingerprint from websites downloaded over an encrypted connection. Research by Acar et al.[11] analysed Alexa's top million websites and uncovered 13 different fingerprinting providers among 404 websites. The researchers also created a framework (FPDetective) in order to uncover fingerprinting done by third parties.

Other third-party data sharing research has resulted from a more practical point-of-view. Mozilla has developed the plug-in LightBeam [2] for FireFox which enables users to see the first- and third-party sites which are interacted with on the Web. Another browser tool developed in order to uncover third-party sites is Ghostery [3], which is available in several web browsers. This tool scans the page for trackers and enables the user to block its tracking. Another project named *'Where is my Data?'*[4] is carried out by hackerspace RandomData, whom created a website to increase awareness on the location of user data concerning e-mail storage location.

This research will contribute by combining the identification of the data sharing third parties (whether personal user data or data obtained via fingerprinting) with the geographical location of these parties in order to evaluate the extent of the privacy of Internet user data. This research will also compare differences between national and global data sharing of parties between different countries taking into account the different privacy laws that are in effect.

[2] `https://addons.mozilla.org/en-US/firefox/addon/lightbeam/`

[3] `http://www.ghostery.com/`

[4] `http://whereismydata.nl/`

Chapter 2

Methodology

The research question is to be answered through four subquestions of which each requires a different approach. The approach for each subquestion is described below and their methodology is described in the proceeding sections.

1. **Which third parties are involved when visiting a website?**
   This research question requires the identification of third parties through different connections. The identification methods are described in Section 2.1.

2. **Can data potentially be accessed by third parties?**
   The involved parties can be contacted through secure or non-secure connections which influence the type of data a third party can access. Non-secure connections can introduce security vulnerabilities where more (unauthorized) third parties can be introduced. In Section 2.2 the identification of secure connections is described.

3. **What is the geographical distribution of your data?**
   Every third party has a geographical location. These locations are identified through the identification of IP address(es) of the third parties and the associated country. The geographical distribution is based on country-level locations since data sharing regulation are for a significant part influenced by the government of a country. The methods applied to uncover these locations are described in Section 2.3.

4. **Which differences in data sharing can be found between countries for national and global first parties?**
   The global first parties are identified as the Alexa's top 10,000 domains and the national first parties as the Alexa's top 1,000 country-specific top-level domains. The third parties and their locations for the national and global first parties are compared in order to answer this subquestion. The methods for this comparison are described in Section 2.4

2.1 Identification of third parties

Third parties can be associated with a first-party domain via several different connections. For example, in the outsourcing of services, such as e-mail or name servers, third parties are introduced. This can also occur through the integration of third-party web services directly on the first-party website. Third parties can also be found in transit and are identified as hops in a data transmission route.

In this research, the identification for third parties is done by third-party domain extraction from Domain Name System (DNS) records and from HTTP requests logs including JavaScript integration code. In addition, third parties are also identified in data transmission routes of HTTP-traffic and e-mail traffic. These identification methods are described in the following sections.

2.1.1 Third parties through DNS records

The Domain Name System (DNS) is a hierarchical distributed naming system for resources connected to the internet. This system associates various information with domain names assigned to these resources. The information associated with a domain is stored in DNS records in zones. Several of these records can contain third parties when these services are outsourced to other companies.

The identification of third parties through DNS records is conducted by retrieval of the following three records: name server (NS) records, mail server (MX) records and canonical name (CNAME) records. NS records delegate a DNS zone to use the given authoritative name servers. MX records maps domain names to a mail transfer agent which is a service that is often outsourced. High availability for these services is required and therefore desire fault tolerant systems. The implementation and maintenance of these services is fairly complex and are therefore often outsourced to third parties. The CNAME records are an alias for another domain name and may also reveal another third party.

The experiments are conducted with the use of the simple ‘dig’ command to query these records and their results are stored in the database for further analysis.

2.1.2 Third parties through HTTP requests

When an Internet user visits a website several HTTP requests are sent between the browser and web server in order to correctly display the website. If this website makes use of incorporated third-party web services on the website, HTTP requests are also sent to these parties. These requests will contain the third-party domain and are used for the identification of the third parties.

The HTTP requests are logged with the use of the FPDetective framework[1]. The FPDetective uses a crawler to drive a browsers to websites and navigate through the pages. The output of the FPDetective process are parsed for unique domain names of third parties. These third parties are then stored in the database for further analysis.

As a result of the many different web services that can be integrated on a website, the third parties will also be classified with the use of a classification database for third-party trackers. In this experiment the Ghostery database is used, which contains an extensive classification for the categories *Ad, Analytics, Privacy, Tracker* and *Widget* and uses regular expression patterns to match classes to URLs. The global domains will be classified intensively by finding all matches of patterns in all fully qualified domain names. The national domains are classified with a lighter classification method in which patterns are matched to host names. This method is unable to retrieve all matches of the intensive classification. However, the matches that occur in a domain are able to occur in other domains and it will show a rough overview of HTTP request classifications in order to compare the national domains.

2.1.3 Third parties through data routes

Data is gathered from the parties involved for the creation of a website when it is visited. This data is passed through the network over several hops. These hops can be identified as third parties as they are involved in storing data for a short time and passing it to the next hop. Parties in transit can also be an access point for performing man-in-the-middle attacks and are therefore an important part of the third-party scope.
In this section a differentiation is made between HTTP-traffic and e-mail traffic as a result of their different routing methods.

2.1.3.1 HTTP-traffic

In order to analyse the route of the HTTP-traffic, traceroutes can be performed to all IP addresses of found first- and third-party domains.

Traceroutes can differ due to policy based routing in which network administrators can determine and implement routing policies on packet information such as protocols. In addition, firewalls and routers often block the ICMP protocol completely or disallow the ICMP echo requests (ping requests), and/or block various UDP ports. This also results in different traceroutes due to different protocols. In this research multiple protocols are examined in order to gain a more complete picture of the routes. The protocols analysed in this research are UDP, TCP, and ICMP with the use of the tool scamper[2] for bulk measurements on traceroutes. The TCP traceroute was performed with destination port

[1] https://github.com/fpdetective/fpdetective/

[2] http://www.caida.org/tools/measurement/scamper/

80 (HTTP) to properly traverse and detect HTTP loadbalancers. For UDP traceroutes the default destination port (33435) was used.

2.1.3.2 E-mail traffic

Regular traceroutes (UDP, TCP, and ICMP) are performed to the mail server of the domain via the DNS MX records. However, there are some limitations to the number of information gathered using this mechanism for the analysis of e-mail routing. Firstly, this traceroute only shows the e-mail route in one direction. However, in contrast to other used protocols, this e-mail response can come from a different mail server, which may take a different route across the internet. Secondly, there is also an (internal) e-mail route between mail servers before the route from the final mail server to the receiving mail server is traversed across the internet. Thirdly, some e-mail is scanned by a third party anti-spam service after reception by the mail server which is also not shown in a regular traceroute.

In order to properly overcome these limitation of the traditional traceroute to the mail server, the e-mail server can be probed with an e-mail requesting a response. The response will show the source IP of the final mail server transmitting the e-mail, which allows to perform a regular traceroute to this server. Furthermore, the e-mail contains headers showing the (internal) e-mail route between mail servers. Lastly, the e-mail should contain tracking images that will be downloaded by anti-spam servers for optical scanning, these downloads are logged to reveal the location of the third party anti-spam servers.

In this research several e-mails were sent to each of the Alexa's top 10,000 domain and each e-mail contained a unique link to the personal page and profile picture of a fictional persona. Each domain is linked to a generated unique fictional persona with a matching e-mail address in order to create a comprehensive analysis.

2.2 Secure Connections

The extent to which third parties can possibly access more data than an Internet user authorizes can be analysed by identifying secure connections between the user and the first party for authentication. In this section a distinction in approach is made between communication to the first-party website or to the third-party e-mail service. These two approaches are described in the following section.

2.2.1 First-party website

In this research two methods will be conducted for the identification of security for first-party websites: DNSSEC and HTTPS. As seen in section 2.1.1, third parties can easily be introduced via DNS records and are a point in which unauthorized parties can be inserted. In order to authenticate the domain and the integrity of its records DNS Security Extensions (DNSSEC) can be used. Another security issue entails that the communication between the Internet user and the first party can contain sensitive user data and can be subject to injection of eavesdropping or man-in-the-middle attacks where parties can see the data. In order to prevent this data sharing with (unauthorized) third parties connections can be made using HTTPS. The two security approaches are described below.

2.2.1.1 DNSSEC

DNSSEC uses public-key cryptography to sign DNS resource records. The signed resource records can be authenticated with a DNSKEY, which is present at the domain. The DNSKEY itself is authenticated via a chain of trust. The experiments are conducted with the use of the simple 'dig' command to query for the presence of DNSKEY records which indicates whether DNSSEC is implemented.

2.2.1.2 HTTPS

The Hypertext Transfer Protocol (HTTP) is the foundation of data communication on the World Wide Web. HTTP Secure (HTTPS) layers HTTP over the SSL/TLS protocol for secure communication over a network and is widely employed on the World Wide Web. It is used for authentication websites and encryption of communication. Communication to first parties and their third parties run over a HTTP or HTTPS connection. To identify secure connections a parser runs over the FPDetective log files mentioned in subsection 2.1.2 and matches each URL for HTTP or HTTPS connections.

2.2.2 E-mail traffic

All the e-mails received as response to the e-mail probe messages sent (as described in 2.1.3.2) are checked for several headers. These headers indicate the mail servers support of multiple cryptographic capabilities that secure the e-mail transmission between mail servers.

2.2.2.1 DKIM

Domains can implement DomainKeys Identified Mail (DKIM) Signatures to increase security. DKIM does not provide encryption of the message content, but consists of a cryptographic signature that guarantees the e-mail source server is a valid origin for the domain. DKIM support is still fairly limited and early adopters have started implementing DKIM before it became a standard in 2011. Currently not all implementations are valid because several changes have been made since the original proposal in 2007 [12][13][14]. Although most DKIM signatures are valid merely checking for the presence of DKIM headers is not sufficient and the DKIM hash should be validated to check if the server is properly configured according to the standard [15].

2.2.2.2 TLS

E-mail servers are able to encrypt messages using TLS to another mail sever and in doing so protect the message content in transit between these e-mail servers. However, not all e-mail servers support TLS and in order to use TLS, the destination e-mail server must advertise its support for TLS and the sending server must be configured to use TLS connections when possible. Since 1995 SMTP Service Extensions are supported and this led to the implementation of the StartTLS command in 2002, TLS capability should be actively reported since the updated ESMTP standard from 2008 [16][17][18]. ESMTP is currently the standard method for sending e-mail and TLS is generally well supported.

The TLS connections that are made in transit appear in the e-mail headers of the e-mail. The headers are parsed for all servers that support TLS. Support for encrypted connections should be checked for each mail server since e-mail transits several intermediate mail servers which all need to individually enable TLS support. When encryption is not enabled on all hops of the route there exists a point where the message transits the internet unencrypted.

2.3 Geographical Distribution

The geographical distribution of third parties is measured by the distribution over countries in which third party IP addresses are located. The retrieval of country-level locations are conducted with the use of a GeoIP database[3] which contain many mappings from IP addresses to countries.

In addition, bulk retrievals of WhoIs records are performed to identify the Autonomous Systems (AS) which are under the control of one or more network operators. This retrieval method can identify important AS areas and their locations.

[3] free country level database provided by MaxMind dated 2013-08-06 (GEO-106FREE Build 1)

2.4 Global and National Comparison

International domains are compared with national domains by selecting the top 10,000 websites, from the top 1 million websites as defined by Alexa[4]. In order to make a proper comparison, national websites the top 1,000 domains with the country top level domain are selected per server location. For the Netherlands these are the .NL domains, for China the .CN domains, and for the United States the .US, .MIL, and .GOV domains. The extra domains associated with the United States were added for a good comparison of the official government websites that are hosted on the countries, because the United States uses the .GOV and .MIL extensions exclusively for national websites.

The list of top domains of top level domains of a country is used, instead of the most popular websites per country, because the most popular domains are fairly similar with the top 10.000 websites and don't reflect the behaviour of websites exclusive to the country. Since the domains are national most are presumably hosted in the same country, it is thus expected that the national domains are associated with less other countries than the international domains. The hypotheses is made that the national domains will also have differences in third parties due to their geographical location and differences in privacy regulations.

[4] referred to on 2014-01-08

Chapter 3

Results

In this section the results of the subquestions found in section 1.2 are described. In each section the results of the proposed methods are described for the Alexa top 10,000 domains after which the comparison between the global and national domains is made.

A focus is put on second-level domains which commonly refer to the organisation that registered the domain at the domain name registrar. These organisations can hold different domain names and these domain names are therefore seen as belonging to the same third party. Organisations are said to be hosting services in-house when a referral is made, for DNS records or HTTP requests, which holds the same second level domain as the organisation.

3.1 Identification of third parties

A multitude of methods is used to attain a domain or an IP address of a third party based on the domain of the first party. Some third parties are fully integrated as primary service provider for the first-party (like DNS- or mailservers), other third parties provide content, functionality or services indirectly by linking from the website and had to be obtained by visiting the website using an automated process. Lastly, to attempt identification of even remotely related third parties which cannot be identified directly, network probes like traceroute can identify an additional scala of third parties related to the first-party domain.

3.1.1 Third parties through DNS records

In total 57,771 DNS records are found which contain 7,611 unique second level domains. However, the services can also be hosted in-house and as a result 7,310 third parties are found. In table 3.1 an overview is given on the distribution of third-party introduction among records and show the mean of third parties introduced via the different resource records.

	CNAME	MX	NS
Total records	115	25,468	32,188
Third parties[1]	62	4,567	4,536
Mean	1	2.8	3.2
Stdev	0	2.1	1.7

Table 3.1: Overview of identification of third parties through resource records

Third parties are introduced mostly via MX and NS records, whereas only a small number of CNAME records are present. Among these third parties, significantly many first parties outsource their services to only a handful of third parties. This illustrates a presence of dominant organisations in this field. Detailed results of identification through the different records and comparisons between global and national domains are described below.

CNAME records

There are only 115 CNAME records found which contain 62 unique second level domains. There is only one in-house CNAME direction found. A maximal of 13 CNAME directions to one specific third party is found in the records of the top 10,000 websites. The previously mentioned third party is the DNS hosting service provider 'DNSPod' and this result shows that several first parties use this service for the DNS hosting.

Similar results remain among the top 1,000 of national domains. There are only 5 US-domain CNAME records and 1 NL-domain CNAME record found. However, 65 CN-domain CNAME records are retrieved. After analysis of these CNAME records the same explanation is found as described above, there are 18 CNAME directions to the DNS hosting service provider 'DNSPod'. This organisation is situated in China and can explain the bigger presence among the CN-domains.

The found results for CNAME records show that these records do not introduce a significant number of third parties. The global and national domains share this finding; however, the CN-domain shows a small popularity for a specific DNS hosting service provider.

MX records

There are a total of 25,468 MX records found of which 5,213 MX records point to in-house hosting. There are on average 2.8 ± 2.1 MX records per first party (n = 9,936).

[1] Summed more third parties due to overlap in second level domains of different records

These results are explainable since it is common for domains to run multiple mail servers to reach a high availability in e-mail services. There are 4,567 third parties introduced; however, there is a significant number of MX records that point to two third parties: 12,325 out of 25,468 MX records, and are referred to by 4,272 first parties. These mail server domains are '*GOOGLE.com*' and '*GOOGLEMAIL.com*' and both belong to Google mail servers.

	TOP	**NL**	**US**	**CN**
Total	25,468	2,428	1,693	1,051
Third party	GOOGLEMAIL.com + GOOGLE.com	GOOGLEMAIL.com + GOOGLE.com	GOOGLEMAIL.com + GOOGLE.com	qq.com
Records to third party	12,325	694	464	404

Table 3.2: global and national MX records & biggest third party comparison

Among NL- and US-domains the two third parties mentioned above are also significantly present: approximately 25% of the found MX records points to these parties. However, the CN-domains hold other third parties at the top such as 'qq.com' to which approximately 40% of the found MX records point to.

The results for the identification of third parties via MX records show a significant retrieval of third parties. In the global and national domains many first parties have MX records pointing to the same third parties, which shows a dominant presence of these third parties. The CN-domain stands out as its results show another third party mail service as most implemented through MX records.

NS records

In total there are 32,188 NS records found in which 61 name server services are hosted in-house. There are on average 3.2 ± 1.7 NS records per first party (n = 9,817). A restriction for NS records is that a domain requires at least 2 NS records. In order to maintain high availability for the site, multiple NS records are often introduced. There are 4,536 third parties introduced in total and among these parties there are 4 DNS services that are referred to significantly more: '*akam.net*', '*dynect.net*', '*dnsmadeeasy.com*' and '*cloudflare.com*' and are together referred to 6,320 times. These domains host DNS services on multiple location across different continents which explains the dominant position in the NS records for the Alexa top 10,000 domains.

On average there are 2 to 3 NS records found per first party in national domains. Similar to the global domains, each national domain also contains DNS services to which significantly more NS records are referred to. In contrast, the biggest third parties for the global and national domains mostly differ. This result follows as name server services which are relatively close to the domain are often deployed. The countries analysed in this research are on different continents and therefore show a different top name server services.

There are many third parties introduced through NS records. Third party name server services are fairly country specific due to the requirement of having relatively close name servers for domains.

3.1.2 Third parties through HTTP-requests

There are in total 203,076 HTTP-requests found for the visits of the top 10,000 domains, of which 27,302 requests are in-house directions. There are 17,216 third parties introduced with an average of 13 ± 13.2 third parties introduced via HTTP-requests per first party. These results show that websites introduce many different third parties via the code integration.

#	Third party	# First parties
1	google-analytics.com	6282
2	doubleclick.net	4532
3	google.com	4327
4	facebook.com	3875
5	googleapis.com	3162

Table 3.3: Number of first parties sending HTTP-request to specified third party

In table 3.3 an overview is given for the top 5 most requested third parties per first party. This table shows a significant integration of popular third parties in first-party websites.

#	Class	# HTTP-requests
1	Ad	35,614
2	Analytics	14,800
3	Widget	13,691
4	Tracker	11,663
5	Privacy	220

Table 3.4: Number of HTTP-requests classified in specified class

An extensive ghostery classification is done on the HTTP-request for the Alexa top 10,000 domains in order to analyse the distribution of the different classes among the HTTP-requests. In total 75,988 HTTP-request fit in the tracker classification model of Ghostery. Table 3.4 shows the number of HTTP-requests belonging to each class. These results show a strong incorporation of HTTP-requests for ads on websites.

In table 3.5 an overview is given for classification of HTTP-request using host names, as described in section 2.1.2. This classification gives an overview of the interrelated classifications since the same set of matches is possible. As seen in table 3.5 the distribution of HTTP-requests among the different classes is similar between the NL- and

Class	NL	US	CN
Ad	4036	2285	887
Analytics	745	732	1326
Widget	689	688	360
Tracker	567	415	100
Privacy	18	18	0
Total	6,055	4,138	2,673

Table 3.5: # HTTP-requests classified in specified class per national domain

US-domains and show that most classified HTTP-requests belong to the Ad class. However, the CN-domains have a different class distribution and show that most classified HTTP-requests for CN-domains belong to the Analytics class

The results for the identification of third parties through HTTP-requests show a significant number of third parties introduced. The classification results indicate that a significant part of the classified HTTP-requests belong to the class Ad, which indicates advertisements. Global and national results of NL- and US-domains are very similar. In contrast, most of the HTTP-requests within CN-domains are classified as Analytics.

3.1.3 Third parties through data routes

The data routes detected can be divided between two types, the traceroutes and the e-mail trace via headers (see table 3.6).

	Total IPs	Mean	STD	Top
Traceroute	60,646	14.50	6.81	46
E-mail trace	9,325	2.56	2.34	44

Table 3.6: Unique third-party IPs identified

3.1.3.1 Traceroute

A regular traceroute is a fairly straightforward route to the public IP of a domain and will on average go trough 14.5 (visible) hops but the maximum number unique IPs is much higher, up to 46. This value 46 is even higher than the maximum number of hops of 34 based on TTL of the packet. These positive discrepencies occur when packets sometimes take different routes because of loadbalancing or different handling of TCP, UDP, and ICMP traffc. On the other hand sometimes negative discrepencies occur because there often are some hops that do not return a packet when the TTL reaches zero so the total number of unique IP addresses is limited. This results in an average number of hops (based on TTL) of 12.1, but an average number of unique IPs is 14.5 (slightly more than the number of hops if only one mechanism would be used). The traceroutes are performed using TCP, UDP, and ICMP protocls, but since the resulting IPs overlap almost completely it these different protocols are not specified in table 3.6.

3.1.3.2 E-mail trace

The e-mail trace is a bit different, it can go trough multiple internal hosts (with IPs on a private subnet) as well as trough externally accessible mailservers. The total number of unique IPs encountered in the e-mail headers is 13,121 of which 3,796 are internal addresses (from the 10.x.x.x, 127.x.x.x, 192.168.x.x, and 172.16.x.x IP ranges). The other 9,325 IP addresses are of publicly accessible mailservers, the maximum number of hops used by one domain is 44 (This is an extreme outlier that used multiple different routes for different mails. Most routes are much shorter, especially since not all intermediate mailservers are always reported). A lot of the IP addresses found in e-mail headers are used by multiple domains or are also known as the public mailserver from the MX record, resulting in a total of only 2,087 unique additional IP addresses used for internal e-mail routing. Beside these 2,087 hosts, and additionally 3,796 unique internal IPs, are only found trough the e-mail traceroute. The internal IPs are not listed in table 3.6.

3.2 Secure connections

Several different types of secure connections are checked. Lots of domains implement one or more secure connections, but since there still are large numbers of insecure connections (see sections below) used the vast majority of all domains has at least one third-party connection that is vulnerable, potentially leaking personal data. An attacker will generally target the weakest link, so the security is as weak as the weakest link of all third parties. Because of this the number of domains that have near 100% secure connections most likely approach zero.

3.2.1 HTTP traffic

In this section results for the identification of secure connections in HTTP traffic are described. The results show a low implementation of HTTPS and DNSSEC. Less than 60% of the domains have implemented HTTPS despite it being a widely implemented protocol. This result is present in global and national domains, where deployment for HTTPS among CN-domains is significantly low. DNSSEC implementations are shown to be even lower with a implementation of less than 2% for global domains. The NL-domains show an implementation of 20%, opposite to lower implementation of DNSSEC in the other national domains. DNSSEC is however complex and therefore not yet implemented in many domains, which the results support.

3.2.1.1 HTTPS

A total of 5,739 domains are secured with HTTPS in the Alexa top 10,000 domains. Although HTTPS is a widely implemented protocol, almost half of the domains do not implement this. In comparison to the national domains, a similar percentage of the NL-domains have HTTPS incorporated (589 domains). The US-domains contain 393 domains which have HTTPS implemented. A significant low number of CN-domains implement HTTPS, which are only 72 domains. This result is surprising since the Chinese government has implemented the China's Great Firewall and HTTPS URLs are a way to circumvent the inspection of content by the firewall. A hypotheses for this result is that China recently has been blocking HTTPS versions of these websites.

The low implementation of HTTPS can increase the vulnerability of man-in-the-middle attacks on the Internet.

3.2.1.2 DNSSEC

A total of 117 domains have DNSSEC implemented for the Alexa top 10,000 domains. The national domains also show a low implementation of DNSSEC; 184 of the NL-domains and 200 of the US-domains have DNSSEC implemented. A significant low number of DNSSEC implementations is found in the CN-domains: only 22 domains show an implementation of DNSSEC.

Although DNSSEC show advantages for the security of domains, as described in section 2.2, implementation of DNSSEC is complex in large-scale networks and therefore deployment of this protocol is low. The results found in this research on DNSSEC implementations among top visited domains support this statement.

3.2.2 E-mail traffic

A total of 64,032 e-mails have been sent (6 per domain) to the top 10,000 global domains and the op 1,000 .NL domains. From the majority of domains we received an (automated) response, the total is listed in table 3.7. A significant low number of DKIM deployments of 7.3% is found.

	Domains	Percentage
At least one response e-mail received	8,917	81.4%
No response e-mails received	2,035	18.6%
Total domains e-mail sent	10,952	100%

Table 3.7: Number of domains by e-mail status

3.2.2.1 TLS

Of the number of domains that received e-mail 10% has at least one mailserver that has ESMTP TLS enabled, as shown in table 3.8. Although some servers have TLS enabled only 6.2% protects the entire e-mail route by encrypting every (known) hop. This means that of all domains that implement TLS well over half (56.2%) implements TLS correctly across the entire route. All TLS implementations seem to be of sufficient strength to be cryptographically effective so no further distinction can be made as to the effectiveness of the TLS implementation.

	Domains	Percentage of total	Percentage of valid
TLS disabled	7,934	90.0%	90.0%
TLS enabled	982	10.0%	10.0%
TLS complete	552	6.2%	56.2%
TLS partial	430	4.8%	43.8%

Table 3.8: Number of domains by TLS encryption status

3.2.2.2 DKIM

Table 3.9 shows an overview of the deployment of DKIM in the domains. Only a small number of domains has DKIM enabled (7.3%) of which most have a correct implementation of the DKIM standard. The percentage of observed valid DKIM signatures is 91%, which is fairly close to the previous result of 92.3% as described in the RFC 4871 Implementation Report from 2011 [15]. This small percentage of invalid DKIM signatures can occur because of misconfiguration of the mail server, expired keys, or changes made to the message content by intermediate mail servers.

There is a significantly low deployment of DKIM signatures among the top 10,000 global domains, which lowers security in the authentication of the e-mail source server for belonging to a valid origin for the domain.

	Domains	Percentage of total	Percentage of valid
No DKIM present	8262	75.4%	92.7%
DKIM enabled	655	6.0%	7.3%
DKIM valid	596	5.4%	91%
DKIM invalid	59	0.5%	9%

Table 3.9: Number of domains by DKIM header status

3.3 Geographical Distribution

The first- and third-party domains are distributed across 119 different countries or regions. The top 10 countries listed in table 3.10 account for 77.5% of all IP addresses encountered and therefore are the most relevant countries for the most popular websites. The top 10 countries by number of unique domains (including sub domains) are the same as the top 10 countries by number of unique IP addresses, only the ordering is slightly different, for example China is 5th by number of domains but second by number of IPs.

#	Country	Number of domains	Number of IPs
1	United States	10036	30572
2	Netherlands	2587	5889
3	Germany	1684	4527
4	France	1661	2635
5	China	1153	7186
6	United Kingdom	1023	3158
7	Europe	1005	2334
8	Japan	739	3144
9	Russian Federation	698	2949
10	Canada	448	1127

Table 3.10: Number of unique domains and IPs per country

#	Country	Number of first-party domains
1	United States	2328
2	China	368
3	Netherlands	318
4	Germany	291
5	United Kingdom	182
6	France	170
7	Russian Federation	157
8	Japan	118
9	Iran, Islamic Republic of	100
10	Ireland	87

Table 3.11: Number of first-party domains country

For comparison the list of the top 10 countries where the first party domains are located is shown in table 3.11. Most countries listed in the top 10 countries with most third parties are also listed in the top 10 countries with most first parties, showing a clear relation between the geographical distribution of the country where the first party is located and the countries where associated third parties are located.

In table 3.12 the number domains per number of countries is listed. Some domains are only hosted in their country of origin, these 8.4% of domains are only associated with 1 country. Almost the same number (just sligtly larger percentage at 8.44%) is associated with 10 or more countries. The total number of domains is lower than 10,000 because for some domains no data is available because the GeoIP country information for that domain is not available.

Countries	Domains	Percentage of domains
1	832	8.4%
2	1,017	10.2%
3	1,374	13.8%
4	1,231	12.4%
5	1,166	11.7%
6	1,064	10.7%
7	966	9.7%
8	806	8.1%
9	652	6.6%
10+	840	8.44%
Total	9,948	100%

Table 3.12: Number domains per number of countries

3.3.1 Autonomous Systems

The IPs of first parties and the associated third parties are associated with 4,586 unique Autonomous Systems. The top 10 Autonomous Systems by number of unique IPs listed in table 3.13 account for 22.6% of all the IPs encountered of which the ASN was known. By far the largest Autonomous System encountered is *Amazon* (comprising the top 2 of largest Autonomous Systems), namely *AMAZON-02* and *AMAZON-AES*) which is used mostly for their cloud services. The list also contains Content Distribution Networks (CDNs) like *CloudFlare* and *Akamai*. More traditional hosting providers follow, namely *Softlayer*, *Leaseweb*, *Hetzner*, and *OVH* . The last entries in the top 10 are backbones that handle international traffic.

#	Autonomous System	Number of IPs
1	AMAZON-02/-AES - Amazon.com, Inc.	4844
2	CHINANET-BACKBONE No.31,Jin-rong Street	1469
3	CLOUDFLARENET - CloudFlare, Inc.	1407
4	AKAMAI-ASN1 Akamai International B.V.	1325
5	CHINA169-BACKBONE CNCGROUP China169 Backbone	1157
6	SOFTLAYER - SoftLayer Technologies Inc.	1026
7	LEVEL3 Level 3 Communications	864
8	HETZNER-AS Hetzner Online AG	796
9	LEASEWEB LeaseWeb B.V.	735
10	OVH OVH Systems	681

Table 3.13: Number of unique IPs for the top 10 Autonomous System

3.3.2 Global and National comparison

There are major differences in the number of countries associated with a single domain, this ranges from as little as 1 to as many as 20 different countries. The top 10.000 international domains are associated with the most countries as can be seen in figure 3.1, the top 1.000 country domains have slightly less associated countries as expected.

The top 1.000 Dutch domains (as seen in figure 3.2) are fairly similar in geographical distribution to the United States domains (as seen in figure 3.3), but are associated with slightly more countries. The Chinese domains (as seen in figure 3.4) are associated with remarkably fewer countries, and the domains that have the most associated countries are generally the more internationally oriented websites.

The maps below are a visual representation of all the connections between the countries of first parties and countries of associated third parties and clearly illustrate the difference in geographical distribution with additional shading per country that indicates the number of IPs located in that country:

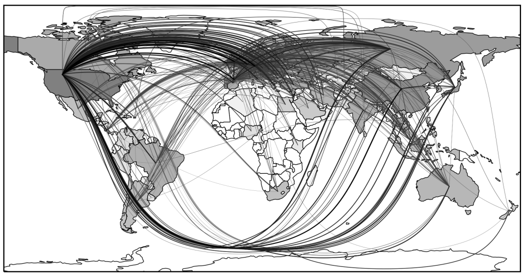

FIGURE 3.1: Top 10.000 domains

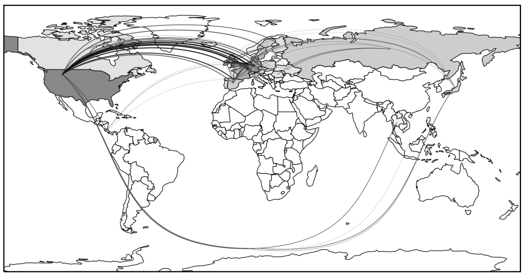

FIGURE 3.2: Top 1.000 .NL domains

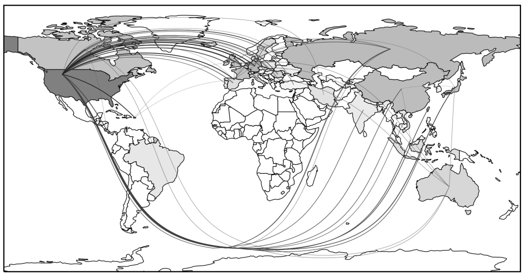

FIGURE 3.3: Top 1.000 .US domains

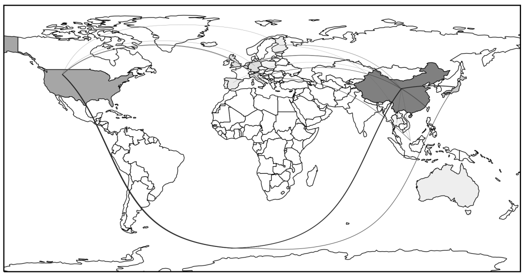

FIGURE 3.4: Top 1.000 .CN domains

	Top 10.000	Top .NL	Top .US	Top .CN
Mean number	3.90	3.75	1.97	1.36
Standard deviation	2.27	2.06	1.52	0.70
Maximum number	17	13	11	11
Top domain	mazika2day.com	sony.nl	luislauro.us	napolimagazine.com.cn

TABLE 3.14: Number of counties identified for all first and second party domains. Note: .US domains includes .MIL and .GOV domains

Note that the .CN domain with the most associated countries - *napolimagazine.com.cn* listed in table 3.14 with 11 countries - is in fact an international (Italian) website and clearly an outlier. The .CN domain with the second most associated countries is *huihui.cn* with 6 countries. This domain much better represents the typical maximum number of countries associated with a domain in China.

The difference between the number of countries listed in table 3.14 and 3.15 is explained by the intermediate countries as revealed by the traditional traceroute, or alternate mail

	Top 10.000	Top .NL	Top .US	Top .CN
Mean number	5.79	4.70	3.37	2.43
Standard deviation	2.87	2.66	2.13	1.03
Maximum number	20	16	13	16
Top domain	mazika2day.com	sony.nl	breakz.us	napolimagazine.com.cn

Table 3.15: Number of counties identified when considering all traceroutes to the first and third parties, including any country that traffic passes through.
Note: .US domains includes .MIL and .GOV domains

servers of the e-mail return path trough those countries. These differences are often zero for national domains, but for most domains there are one or more additional countries that the HTTP, DNS, or e-mail traffic passes trough.

To illustrate the difference in the number of countries involved between first- to third-party connections, and the traceroutes between them, the known .NL domain - *alternate.nl* - is taken as an example and is illustrated below. In figure 3.5 the connections to the third-party endpoints are drawn to 7 countries, but in figure 3.6 the connections trough the intermediate hosts from the traceroute are drawn to 12 countries. This difference of 5 countries is well above the average difference of almost 1 country for .NL domains, but this domain is already an exception with the number of countries associated with it almost thrice the standard deviation. Another observation that can be made from figure 3.6 is that the additional countries from the traceroute are the surrounding countries, this is true for the vast majority of all additional countries found by traceroutes. When checking the corporate website referenced trough emphalternate.nl it is clearly stated that Alternate in fact has local office branches in the surrounding countries[2].

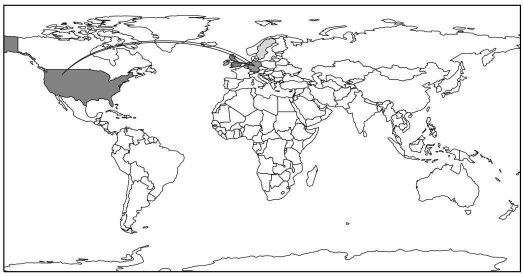

Figure 3.5: First- to third-party connections of alternate.nl

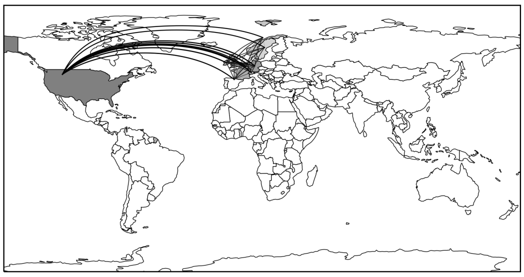

Figure 3.6: Complete traceroute connections of alternate.nl

[2] Alternate international EU map: `http://www.alternate.com/?n=alternateEu`

Chapter 4

Conclusions

4.1 Conclusions

In order to find the scope of data sharing of first parties a large database, containing information on third parties and geographical locations, is created from which results are extracted and conclusions are drawn. Four subquestions are set to get insight on the research question which is set as: *'What is the scope of (privacy) infringing data sharing of the top visited websites with third parties?*. The conclusions on these four subquestions are described below.

Which third parties are involved when visiting a website?

Third parties are identified through different channels. Results show third parties are introduced through DNS resource records. For 10,000 global domains the MX records introduce 4,567 different third parties, where NS records introduce 4,536 different third parties and CNAME records introduce a low number of 115 third parties. There are organisations to which a significant amount of MX records point, for national as well as global domains, which show a dominant position for these organisations. Dominant organisations are also present in the introduction of third parties through NS records, however these organisations are located in proximity to the location of the first party and therefore differ between national domains.

Turning to identification of third parties via HTTP-requests, these results show a significantly higher introduction of third parties. A total of 17,216 third parties are identified and on average first-party web sites introduce 13 third parties. A significant integration of popular third parties in first-party websites is also shown. The results of classification of these HTTP-request show many third parties are introduced with the goal of advertisement or analytics.

With regard to traceroutes these add additional intermediate third parties. A total of 69,971 third parties are identified as either part of the traceroute to the first and third party servers (HTTP, DNS, mailserver), or an e-mail trace that maps the return path of an e-mail to our server. These third parties add a significant amount of extra geographical distribution between the first- and third-party endpoints. On average they add almost 1 country to the list of associated countries, but for some domains the difference is much higher, like for the example domain *alternate.nl* these traceroutes reveal an additional 5 associated countries.

Can data potentially be accessed by third parties?

In order to identify if data can be potentially accessed by third parties secure connections are identified. Among HTTP(S)-traffic less than 60% of the domains are secured with HTTPS. In addition, a significant low deployment of DNSSEC is present of which the results correlate with the complex implementation of DNSSEC in large-scale networks. These results show a presence of security vulnerabilities among the domains which can potentially result in data access by (unauthorized) third parties.

E-mail traffic is considerably unsecure, with only 6.2% of the domains implementing a secure TLS connection between all intermediate mailservers, although it should be noted that up to 10% have some hops protected with TLS so at least attempt to provide a partially secure connection. Beside TLS e-mails can be protected with DKIM. Although even a correct DKIM implementation will do nothing to protect yout e-mails from being read by a man in the middle attack you will be able to cryptographically verify the source of the message and if the message has been tampered with. Even if DKIM might offer a security benefit it is only correctly implemented by 5.4% of the domains.

What is the geographical distribution of your data?

The most conservative geographical distribution of your data is to only host all services (and presumably your data) in the country of the first-party. But only a small minority (8.4%) of domains use this minimal geographical distribution by hosting all their services in the first-party country. Just a little bit more domains (8.44%) geographically distribute your data to a lot of countries (at least 10, and up to a maximum of 20 different countries). The other 83.2% of the domains share the data with services in 2 to 9 different countries, on average your data will be geographically distributed across 5.79 different countries. The most common country your data is distributed to is the United States, this is by far the biggest location for third-party services, even more than Netherlands, Germany, France, China, and the United Kingdom combined. But this distribution is largely due to the number of first-party websites being located in the United States. There is a clear relation between the geographical distribution of the country where the first party is located and the countries where associated third parties are located so these third-party results are to be expected.

Which differences in data sharing can be found between countries for national and global first-parties?

The results in this research show a big similarity between global domains, NL-domains and US-domains with regard to the identification of third parties. The third parties introduced via MX records and HTTP-requests comprise on average of the same domains. Classification of HTTP-requests also show similar results in classes of requests and HTTP-requests are often classified as advertisements. Third parties introduced via NS records differ, which is explained by the requirements for proximity of the DNS server to the domain. Implementation of security protocols in HTTP-traffic also show similar deployments in domains. In contrast, the CN-domains show several differences with the domains mentioned above. With regard to MX records and HTTP-requests, CN-domains show a similar distribution of third parties, but involve different third parties. These differences are explained due to the integration of many services within America and Europe and the use of the Latin alphabet which is present in the NL-domains and US-domains, whereas China is more isolated within the same continent and mostly employ Mandarin Hanzi (although sometimes also Cyrillic) alphabet.

4.2 Future Work

The following adjustments for the methods used in this research, as well as improvements using other techniques, are proposed:

- **Other code integration identification:** Third parties are identified, among other things, through their code integration on first-party websites. In this research third parties via code integration are retrieved from HTTP-requests which can include JavaScript Objects. There are however more types of code integrations possible, such as Flash Objects, Java Applets, Silverlight Objects, DirectX Object integration, or other types of add-ons and plug-ins. These different code integrations can hold different third parties. Experiments on the retrieval of different code integrated third parties is proposed for future work.
- **Intensive classification of HTTP-requests:** In this research HTTP-requests are classified using the Ghostery database. However, it was not possible to classify all requests with the use of this database. In future work the HTTP-requests can be analysis with the use of multiple databases in order to retrieve more insight in the goals of these HTTP-requests.

- **Validation checks for secure connections:** Secure connections are identified, however a full check of valid certificates regarding HTTPS and validation of DNSKEYs are not performed. In order to complete the scope of deployment of these secure connections these validations can be done.

- **Mailserver security validation:** Attempt to negotiate different (in)secure encryption methods with the remote mailserver to assess the strength of the cryptographic cyphers available.

- **Expansion on countries:** An analysis is done on three different countries which were chosen because of their different regimes. For future work an analysis of all countries will be interesting in order to completely map the scope of third party data sharing in the world.

- **Country index:** An overview is given of the geographical distribution of user data. An index can be assigned to these countries regarding, for example, their privacy regulations, freedom of press, or any other index relevant to the public.

Appendix A

E-mail template used

```
Dear sir/madam,

I am writing in regards to the privacy policy of your website DOMAINNAME.
We are currently researching the user privacy on the most popular websites.

Could you please provide answers to the following questions:
- Where is the privacy policy located on DOMAINNAME (it’s hard to find)?
- Does your privacy policy contain a synopsis for easy reading?
- Do your third-party affiliates strictly adhere to your privacy policy?
- Can you provide a list of your third-party affiliates that could possibly have (had)
  access to my data?
- Do you or any of your third-party affiliates perform fingerprinting and/or tracking?
  And can you list the types of tracking?

For a thorough research we need this information about the privacy policy of DOMAINNAME.

Thanks in advance! I am looking forward to your reply.

Kind regards,
FIRST AND LAST NAME

----------

LAST NAME, INITIALS uses Nanoniem (http://www.nanoniem.nl/)
Nanoniem is the new secure online non-anonymous email service.

Personal E-Mail:
USERNAME@nanoniem.nl

Personal Website:
http://www.nanoniem.nl/USERNAME
```

Bibliography

[1] Philipp Vogt, Florian Nentwich, Nenad Jovanovic, Engin Kirda, Christopher Kruegel, and Giovanni Vigna. Cross site scripting prevention with dynamic data tainting and static analysis. In *NDSS*, 2007.

[2] Adam Barth, Collin Jackson, and John C Mitchell. Robust defenses for cross-site request forgery. In *Proceedings of the 15th ACM conference on Computer and communications security*, pages 75–88. ACM, 2008.

[3] Glenn Greenwald. *No Place to Hide: Edward Snowden, the NSA, and the US Surveillance State.* Metropolitan Books, 2014.

[4] Barton Gellman and Laura Poitras. *U.S., British intelligence mining data from nine U.S. Internet companies in broad secret program.* Washington Post, 07-06-2013.

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

[6] J.R. Mayer and J.C. Mitchell. Third-party web tracking: Policy and technology. In *Security and Privacy (SP), 2012 IEEE Symposium on*, pages 413–427, 2012. doi: 10.1109/SP.2012.47.

[7] 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.

[8] Peter Eckersley. How unique is your web browser? In *Privacy Enhancing Technologies*, pages 1–18. Springer, 2010.

[9] Kevin P Dyer, Scott E Coull, Thomas Ristenpart, and Thomas Shrimpton. Peek-a-boo, i still see you: Why efficient traffic analysis countermeasures fail. In *Security and Privacy (SP), 2012 IEEE Symposium on*, pages 332–346. IEEE, 2012.

[10] Dominik Herrmann, Rolf Wendolsky, and Hannes Federrath. Website fingerprinting: attacking popular privacy enhancing technologies with the multinomial naïve-bayes classifier. In *Proceedings of the 2009 ACM workshop on Cloud computing security*, pages 31–42. ACM, 2009.

[11] Gunes Acar, Marc Juarez, Nick Nikiforakis, Claudia Diaz, Seda Gürses, Frank Piessens, and Bart Preneel. Fpdetective: dusting the web for fingerprinters. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 1129–1140. ACM, 2013.

[12] E. Allman, J. Callas, M. Delany, M. Libbey, J. Fenton, and M. Thomas. DomainKeys Identified Mail (DKIM) Signatures. RFC 4871 (Proposed Standard), May 2007. URL `http://www.ietf.org/rfc/rfc4871.txt`. Obsoleted by RFC 6376, updated by RFC 5672.

[13] D. Crocker. RFC 4871 DomainKeys Identified Mail (DKIM) Signatures – Update. RFC 5672 (Proposed Standard), August 2009. URL `http://www.ietf.org/rfc/rfc5672.txt`. Obsoleted by RFC 6376.

[14] D. Crocker, T. Hansen, and M. Kucherawy. DomainKeys Identified Mail (DKIM) Signatures. RFC 6376 (INTERNET STANDARD), September 2011. URL `http://www.ietf.org/rfc/rfc6376.txt`.

[15] M. Kucherawy. RFC4871 Implementation Report. RFC 4871 (Implementation Report), March 2011. URL `http://www.ietf.org/iesg/implementation/report-rfc4871.txt`.

[16] J. Klensin, N. Freed, M. Rose, E. Stefferud, and D. Crocker. SMTP Service Extensions. RFC 1869 (INTERNET STANDARD), November 1995. URL `http://www.ietf.org/rfc/rfc1869.txt`. Obsoleted by RFC 2821.

[17] P. Hoffman. SMTP Service Extension for Secure SMTP over Transport Layer Security. RFC 3207 (Proposed Standard), February 2002. URL `http://www.ietf.org/rfc/rfc3207.txt`.

[18] J. Klensin. Simple Mail Transfer Protocol. RFC 5321 (Draft Standard), October 2008. URL `http://www.ietf.org/rfc/rfc5321.txt`.