Institutionen för datavetenskap
Department of Computer and Information Science

Final thesis

Swedes Online: You Are More Tracked Than You Think

by

Joel Purra

LIU-IDA/LITH-EX-A--15/007--SE

2015-02-19

Linköpings universitet
SE-581 83 Linköping, Sweden

Linköpings universitet
581 83 Linköping

This page intentionally left blank.
Almost.

Linköping University
Department of Computer and Information Science

Final Thesis

Swedes Online: You Are More Tracked Than You Think

by

Joel Purra

LIU-IDA/LITH-EX-A--15/007--SE

2015-02-19

Supervisor:	Patrik Wallström The Internet Infrastructure Foundation (.SE)
Supervisor:	Staffan Hagnell The Internet Infrastructure Foundation (.SE)
Examiner:	Niklas Carlsson Division for Database and Information Techniques (ADIT)

This page intentionally left blank.
Almost.

Master's thesis
Swedes Online: You Are More Tracked Than You Think

Joel Purra
mig@joelpurra.se, joepu444
http://joelpurra.com/
+46 70 352 1212

v1.0.0

Abstract

When you are browsing websites, third-party resources record your online habits; such *tracking* can be considered an invasion of privacy. It was previously unknown how many third-party resources, trackers and tracker companies are present in the different classes of websites chosen: globally popular websites, random samples of .se/.dk/.com/.net domains and curated lists of websites of public interest in Sweden. The in-browser HTTP/HTTPS traffic was recorded while downloading over 150,000 websites, allowing comparison of HTTPS adoption and third-party tracking within and across the different classes of websites.

The data shows that known third-party resources including known trackers are present on *over 90%* of most classes, that third-party hosted *content* such as video, scripts and fonts make up a large portion of the known trackers seen on a typical website and that tracking is just as prevalent on *secure* as insecure sites.

Observations include that Google is the most widespread tracker organization *by far*, that content is being served by known trackers may suggest that trackers are moving to providing services to the end user to *avoid being blocked* by privacy tools and ad blockers, and that the small difference in tracking between using HTTP and HTTPS connections may suggest that users are given a *false sense of privacy* when using HTTPS.

Acknowledgments

The Internet Infrastructure Foundation[1] (.SE)

.SE is also known as Stiftelsen för internetinfrastruktur (IIS).

This thesis was written in the office of – and in collaboration with – .SE, who graciously supported me with domain data and internet knowledge. Part of .SE's research efforts include continuously analyzing internet infrastructure and usage in Sweden. .SE is an independent organization, responsible for the Swedish top level domain, and working for the benefit of the public that promotes the positive development of the internet in Sweden.

Thesis supervision

Niklas Carlsson Associate Professor (Swedish: docent and universitetslektor) at Division for Database and Information Techniques (ADIT), Department of Computer and Information Science (IDA), Linköping University, Sweden. Thank you for being my examiner!

Patrik Wallström Project Manager within R&D, .SE (The Internet Infrastructure Foundation), Sweden. Thank you for being my technical supervisor!

Staffan Hagnell Head of New Businesses, .SE (The Internet Infrastructure Foundation), Sweden. Thank you for being my company supervisor!

Anton Nilsson Master student in Information Technology, Linköping University, Sweden. Thank you for being my opponent!

Domains, data and software

.SE (Richard Isberg, Tobbe Carlsson, Anne-Marie Eklund-Löwinder, Erika Lund), DK Hostmaster A/S (Steen Vincentz Jensen, Lise Fuhr), Reach50/Webmie (Mika Wenell, Jyry Suvilehto), Alexa, Verisign. Disconnect.me, Mozilla. PhantomJS, jq, GNU Parallel, LyX. Thank you!

Tips, feedback, inspiration and help

Dwight Hunter, Peter Forsman, Linus Nordberg, Pamela Davidsson, Lennart Bonnevier, Isabelle Edlund, Amar Andersson, Per-Ola Mjömark, Elisabeth Nilsson, Mats Dufberg, Ana Rodriguez Garcia, Stanley Greenstein, Markus Bylund. Thank you!

And of course everyone I forgot to mention – sorry and thank you!

[1] https://www.iis.se/

Contents

List of Tables

List of Figures

Nomenclature

.com A generic top level domain. It has the greatest number of registered domain of all TLDs.

.dk The country code top level domain name for Denmark.

.net A generic top level domain.

.SE The Internet Infrastructure Foundation. An independent organization for the benefit of the public that promotes the positive development of the internet in Sweden. .SE is responsible for the .se country code top level domain.

.se The country code top level domain name for Sweden.

Alexa A web traffic statistic service, owned by Amazon.

ccSLD Country-code second-level domain. A SLD that belongs to a country code TLD. A ccSLD is not for public use, which are required to register their domains on the third domain level.

ccTLD A top level domain based on a country code, such as .se or .dk.

CDF Cumulative distribution function.

CDN Content delivery network

Content Information and data that is presented to the user. Includes text, images, video and sound.

Content delivery network (CDN) The speed at which data can be delivered is dependant on distance between the user and the server. To reduce latency and download times, a content delivery network places multiple servers with the same content in strategic locations, both geographic and network toplolgy wise, closer to groups of users.

For example, a CDN could deploy servers in Europe, the US and Australia, and reduce loading speed by setting up the system to automatically use the closest location.

Cumulative distribution function (CDF) In this thesis usually a graph which shows the ratio of a property as seen per domain on the x axis, with the cumulative ratio of domains which show this property on the y axis. The steeper the curve is above an x value range, the higher the ratio of domains which fall within the range.

DNT Do Not Track

Do Not Track (DNT) A HTTP header used to indicate that the server should not record and track the client's traffic and other data.

Domain name A human-readable way to navigate to a service on the internet: example.com. Often implicitly meaning FQDN. Domains are also used, for example, as logical entities in regards to security and privacy scopes on the web, often implemented as same-origin policies. As an example, HTTP cookies are bound to domain that set them.

External resource A resource downloaded from a domain other than the page that requested it was served from.

External service A third party service that delives some kind of resource to the user's browser. The service itself can vary from showing additional information and content, to ads and hidden trackers.

External services include file hosting services, CDNs, advertisting networks, statistics and analytics collectors, and third party content.

FQDN Fully qualified domain name

Fully qualified domain name (FQDN) A domain name specific enough to be used on the internet. Has at least a TLD and a second-level domain name - but oftentimes more depending on TLD rules and organizational units.

GOCS Government-owned corporations

Government-owned corporations (GOCS) State-owned corporations.

gTLD Generic top level domain such as .com or .net.

HAR HTTP Archive (HAR) format, used to store recorded HTTP metadata from a web page visit. See the software chapter.

HTTP Hypertext Transfer Protocol

HTTPS Secure HTTP, where data is transfered encrypted.

Hypertext Transfer Protocol (HTTP) A protocol to transfer HTML and other web page resources across the internet.

JavaScript Object Notation (JSON) A data format based on Javascript objects. Often used on the internet for data transfer. Used in this thesis as the basis for all data transformation.

jq A tool and domain specific programming language to read and transform JSON data. See the software chapter.

JSON JavaScript Object Notation

P3P Platform for Privacy Preferences (P3P) Project

Parked domain A domain that has been purchased from a domain name retailer, but only shows a placeholder message – usually an advertisement for the domain name retailer itself.

phantomjs Browser software used for automated web site browsing. See the software chapter.

Platform for Privacy Preferences Project (P3P) A W3C standard for HTTP where server responses are annotated with an encoded privacy policy, so the client can display it to the user. Work has been discontinued since 2006.

Primary domain For the thesis, the first non-public suffix part of a domain name has been labeled the primary domain. For example example.com.br has been labeled the primary domain for www.company-abc.com.br, as .com.br is a public suffix.

Public suffix The part of a domain name that is unavilable for registrations, used for grouping. All TLDs are public suffixes, but some have one or more levels of public suffixes, such as .com.br for commercial domains in Brazil or .pp.se for privately owned personal domains (a public suffix which has been deprecated, but still exists).

Resource An entity external to the HTML page that requested it. Types of resources include images, video, audio, CSS, javascript and flash animations.

Second-level domain (SLD) A domain that is directly below a TLD. Can be a domain registerable to the public, or a ccSLD.

SLD Second-level domain

Subdomain A domain name that belongs to another domain name zone. For example service.example.net is a subdomain to example.net.

Superdomain For the thesis, domains in parent zones have been labeled superdomains to their subdomains, such as such as example.se being a superdomain to www.example.se.

Third-party content Content served by another organization than the organization serving the explicitly requested web page. Also see external resource.

Third-party service A service provided by an organization other than the explicitly requested service. Also see external service.

TLD Top level domain.

Top level domain (TLD) The last part of a domain name, such as .se or .com. Registration of TLDs is handled by ICANN.

Tracker A resource external to the visited page, which upon access receives information about the user's system and the page that requested it.

Basic information in the HTTP request to the resource URL includes user agent (browser vendor, type and version down to the patch level, operating system, sometimes hardware type) referer (the full URL of page that requested the resource), an etag (unique string identifying the data from a previous request to the same resource URL) and cookies (previously set by the same tracker).

Uniform Resource Locator (URL) A standard to define the address to resources, mostly on the internet, for example http://joelpurra.com/projects/masters-thesis/

URL Uniform Resource Locator

Web browser Or browser. Software a user utilizes to retrieve, present and traverse information from the web.

Web service A function performed on the internet, and in this document specifically web sites with a specific purpose directed towards human users. This includes search engines, social networks, online messaging and email as well as content sites such as news sites and blogs.

Web site A collection of web pages under the same organization or topic. Often all web pages on a domain is considered a site, but a single domain can also contain multiple sites.

Zone A technical as well as administrative part of DNS. Each dot in a domain name represents another zone, from the implicit root zone to TLDs and privately owned zones – which in turn can contain more privately controlled zones.

Chapter 1

Introduction

How many companies are recording your online trail, and how much information does the average Swede leak while using popular .se websites? Many, and a lot – more than you may think. Large organizations like Google, Facebook and Amazon are able to connect the dots you leave behind during everyday usage, and construct a *persona* that reflects you from their perspective. Have you told your family, friends or colleagues about your gambling addiction, your sex toy purchases, or your alcoholism? Even if you did not tell anyone your deepest secrets, these companies might conclude that they can put labels on you by looking at everything you do online. And now they are selling it as hard facts behind the scenes.

While browsing the web users are both actively and passively being *tracked* by multiple companies, for the purpose of building a persona for targeted advertising. Sometimes the data collection is visible, as in social network sites and questionnaires, but it is most common in the form of different kinds of external resources which may or may not serve a purpose other than keeping track of your every click. Secure connections between server and client help against passive data collection along the network path, but not against site owners allowing in-page trackers. Tracking code is installed on web pages that have adverts as well as those that do not – the spread and reach of tracking across web pages and domains of different kinds increases the quality of the user data collected and inferred, making it more valuable for advertising purposes. With the extent of the use of trackers and other external resources largely unknown and ever evolving, what is already known raises privacy concerns – data considered personal leak without the user's knowledge or explicit permission and end up in privately owned databases for further distribution. Data collection is the new wild west, and you are the new cattle.

This thesis uses large-scale measurements to characterize how different kinds of domains in Sweden and internationally use website resources. Front pages of approximately 150,000 random .se, .dk, .com, .net domains and Swedish, Danish and Alexa's top domains were visited and their resources, including those dynamically loaded, recorded. Each domain was accessed both with insecure HTTP and secure HTTPS connections to provide a comparison. Resources were grouped by mime type, URL protocol, domain, if it matches the domain the request originated from and compared to lists of known trackers and organizations. The thesis makes three primary contributions:

1. Software for automated, repeatable retrieval and analysis of large amounts of websites has been developed, and released as open source (see Appendix B). Datasets based on publicly available domain lists have been released for scientific scrutinization[1]. The data allows

[1] http://joelpurra.com/projects/masters-thesis/

analysis of websites' HTTP/HTTPS requests including the use of resources internal versus external to the entry domain, which the most common confirmed tracker organizations are, what spread they have and how much the average internet user can expect to be tracked by visiting some of the most important and popular sites in Sweden, Denmark and worldwide. Downloading and analyzing additional/custom datasets is very easy.

2. HTTPS usage for different domains has been characterized from a Swedish perspective; adoption rates are compared between classes of domains within Sweden as well as against popular international domains (see Section 4.1). HTTPS adoption among globally popular websites (10-30%, 50% for the very top) and curated lists of Swedish websites (15-50%) is much higher than for random domains (less than 1%). This means that most websites in the world are susceptible to passive eavesdropping anywhere along the network path between the client and the server. But even with HTTPS enabled, traffic data and personally identifiable information is leaked through external resources and third-party trackers, which are just as prevalent on insecure HTTP as secure HTTPS enabled websites (see Section 4.2 and 4.3). This means that a secure, encrypted connection protecting against eavesdropping doesn't automatically lead to *privacy* – something which users might be lead to believe when it is called a "secure connection" as well as through the use of "security symbols" such as padlocks.

3. The use of *known* or *recognized* third-party trackers and other third-party (external) services for different classes of domains, has been analyzed. Using public lists of *recognized* tracker domains, we analyzed and compared the widespread adoption of these services across domains within Sweden, as well as internationally. The use of external resources is high among all classes of domains (see Section 4.2). Websites using strictly internal resources are relatively few; less than 7% of top sites, even less in most categories of curated lists of Swedish websites, but more common among random domains at 10-30%. This means most websites around the world have made an active choice to install external resources from third-party services, which means that users' traffic data and personal information is leaked (see Section 4.3). Most websites also have at least one *known* tracker present; 53-72% of random domains, 88-98% of top websites and 78-100% of websites in the Swedish curated lists.
   The number of known tracker organizations present is interesting to look at, as a higher number means users have less control over where leaked data ends up (4.3.2). Around 55% of random Swedish domains have 1-3 trackers, and about 5% have more than 3. Nearly 50% of global top sites load resources from 3 or more tracker organizations, while about 5% load from more than 20 organizations. Half of the Swedish media websites load more than 6 known trackers; a single visit to the front page of each of the 27 investigated sites would leak information in over 3,800 external requests (C.5) to at least 57 organizations (C.11.1). This means that any guesswork in what types of articles individuals read would read in a printed newspaper is gone – and with that probably the guesswork in exactly what kind of *personal opinions* these individuals hold.
   It is clear that Google has the widest coverage by far – Google trackers alone are present on *over 90%* of websites in over half of the datasets (4.3.3). That being said, it is also hard to tell how many trackers are missed – Disconnect's blocking list *only detects 10%* of external primary domains as trackers for top website datasets (4.3.4).

Chapter 2

Background

In everyday web browsing, browsers routinely access a lot of material from other domains or services than the one visited [11]. These external resources vary from content that the user explicitly want to obtain, to implicitly loaded third-party services, ads, and non-visible resources with the sole purpose of collecting user data and statistical material [22]. All are downloaded on behalf of the user with no or few limitations, and oftentimes without the user's need, understanding and explicit consent. These external resources can all be seen as browsing habit trackers, whose knowledge and power increase with any additional visits to other domains or services loading the same resources [35]. While privacy is both hard to define as well as relative to perspective and context, there is a correlation between trackers and online privacy; more trackers means it becomes harder to control the flow of personal information and get an overview of where data ends up [41, 7].

2.1 Trackers are a commercial choice

While online privacy has been in the spotlight due to recently uncovered mass surveillance operations, the focus has been on national government intelligence agencies collecting information around the globe. Public worry regarding surveillance in Sweden is low. Only 9% of adult Swedish internet users say they worry to some degree about government surveillance, but at 20% twice as many worry about companies' surveillance – a number that has been steadily rising from 11% in 2011 [2, 14]. Governments are able to intercept traffic data and metadata by, among several techniques, *covertly* hooking into the internet infrastructure and passively listening. Basic connection metadata can always be collected, but without secure connections between client and server, any detail in the contents of each request can be extracted.

In contrast, external resources are approved by and actively installed by site and service owners, and presented openly to users with basic technical skills and tools. Reasons can be technical, for example because distributing resources among systems improves performance [22, 21]. Other times it is because there are positive network effects in using a third-party online social network (OSN) to promote content and products. Ads are installed as a source of income. More and more commonly, allowing a non-visible tracker to be installed can also become a source of income – data aggregation companies pay for access to users' data on the right site with the right quantity and quality of visitors. Because these external resources are used on behalf of the service, they are also loaded when end-to-end encryption with HTTPS is enabled for enhanced privacy and security. This encryption bypass gives these private trackers more information than possible with large-scale passive traffic interception, even when there is a security nullifying mixture of encrypted and unencrypted connections.

2.2 What is known by trackers?

Depending on what activities a user performs online, different things can be inferred by trackers on sites where they are installed. For example, a tracker on a news site can draw conclusions about interests from content a user reads (or choses *not* to) by tagging articles with refined keywords and creating an interest graph [24]. The range of taggable interests of course depend on the content of the news site. Private and sensitive information leaked to third-party sites during typical interaction with some of the most popular sites in the world include personal identification (full name, date of birth, email, ip address, geolocation) and sensitive information (sexual orientation, religious beliefs, health issues) [35].

Social buttons, allowing users to share links with a simple click, are tracking users whether they are registered or not, logged in or not [39]. They are especially powerful when the user is registered and logged in, combining the full self-provided details of the user with their browsing habits – all within the bounds of the services' privacy policies agreed to by the user. Once a user has provided their personal information, it is no longer just the individual browser or device being tracked, but the actual *person* using it – even after logging out [19, 23]. This direct association, as opposed to inferred, to the person also allows for tracking across devices where there is an overlap of services used.

2.3 What is the information used for?

Publishers reserve areas of their web pages for displaying different kinds and sizes of advertisements alongside content. Ads chosen for the site may be aligned with the content but it is more valuable the more is known about the visitors. Combining and aggregating information from past visitors means that more information can be assumed about future visitors, on a statistical basis, which will define the general audience of the site. To generate even more revenue per displayed ad, individual users are *targeted* with *personalized* ads depending on their specific personal data and browsing history [16].

Indicators such as geographic location, hardware platform/browser combinations have been shown to result in *price steering* and *price discrimination* on some e-commerce websites [18, 36]. While the effects of a web-wide user tracking have not been broadly measured with regards to pricing in e-commerce, using a larger and broader portion of a user's internet history and contributions would be a logical step for online shopping, as it has been used to personalize web search results and social network update feeds [9, 38].

Social networks can use website tracking data about their users' to increase per-user advertising incomes by personalization, but they will try to keep most of the information to themselves [40, 3, 46]. There are also companies that only collect information for resale – *data brokers* or *data aggregators* – which thrive on combining data sources and package them as targeted information for other companies to consume[1]. The market for tracking data resale is expected to grow, as the amount of data increases and quality improves. The Wall Street Journal investigated some of these companies and their offerings:

> Some brokers categorize consumers as "Getting By," "Compulsive Online Gamblers" and "Zero Mobility" and advertise the lists as ideal leads for banks, credit-card issuers

[1] CBS 60 Minutes – *The Data Brokers: Selling your personal information* http://www.cbsnews.com/news/the-data-brokers-selling-your-personal-information/

> and payday and subprime lenders, according to a review of data brokers’ websites. One company offers lists of ”Underbanked Prime Prospects” broken down by race. Others include ”Kaching! Let it Ride Compulsive Online Gamblers” and ”Speedy Dinero,” described as Hispanics in need of fast cash receptive to subprime credit offers.[2]

[2] Wall Street Journal – *Data Brokers Come Under Fresh Scrutiny; Consumer Profiles Marketed to Lenders* http://online.wsj.com/news/articles/SB10001424052702303874504579377164099831516

Chapter 3

Methodology

Emphasis for the thesis is on a technical analysis, producing aggregate numbers regarding domains and external resources. Social aspects and privacy concerns are considered out of scope.

3.1 High level overview

Based on a list of domains, the front page of each domain is downloaded and parsed the way a user's browser would. The URL of each requested resource is extracted, and associated with the domain it was loaded from. This data is then classified in a number of ways, before being boiled down to statistics about the entire dataset. Lastly, these aggregates are compared between datasets. In the following sections we describe each of these steps in more detail. For yet more details of the methodology, we refer to Appendix A[1]. The software developed is described in Appendix B and the details of the results are presented in Appendix C.

The thesis is primarily written from a Swedish perspective. This is in part because .SE[2] has access to the full list of Swedish .se domains, and in part because of their previous work with the *.SE Health Status* reports (6.1). The reports focus on analyzing government, media, financial institutions and other nation-wide publicly relevant organization groups' domains, as they have been deemed important to Sweden and Swedes. This thesis incorporates those lists, but focus on only the associated websites.

3.2 Domain categories

Curated lists The *.SE Health Status* reports use lists of approximately 1,000 domains in the categories counties, domain registrars, financial services, government-owned corporations (GOCS), higher education, ISPs, media, municipalities, and public authorities (A.1.1). The domains are deemed important to Swedes and internet operations/usage in Sweden.

Top lists Alexa's Top 1,000,000 sites (A.1.5) and Reach50 (A.1.6) are compiled from internet usage, internationally and in Sweden respectively. The Alexa top list is freely available and used in other research; four selections of the 1,000,000 domains were used – top 10,000, random 10,000, all .se and all .dk domains.

[1] To help the reader, explicit references of the form A.1 is used to refer to Section A.1 of Appendix A.

[2] This thesis was written in the office of The Internet Infrastructure Foundation (.SE), the .se TLD registry.

Random zone lists To get snapshot of the status of general sites on the web, random selections directly from the .se (A.1.2), .dk (A.1.3), .com and .net (A.1.4) TLD zones were used. The largest set was 100,000 .se domains; 10,000 domains each from .dk, .com and .net were also used.

Table 3.1 summarizes the domain lists and samples from each of theses lists used in the thesis. More details on each sublist is provided in Appendix A. However, at a high level we categorize the lists in three main categories. In total there are more than 156,000 domains considered.

We note that it is incorrect to assume that domain ownership is always based on second-level domains, such as iis.se or joelpurra.com. Not all TLDs' second-level domains are open for registration to the public; examples include the Brazilian top level domain `.br`, which only allows commercial registrations under `.com.br`. There is a set of such *public suffixes* used by browser vendors to implement domain-dependent security measures, such as preventing super-cookies (A.2). The list has been incorporated into this thesis as a way to classify (A.5.4) and group domains such as `company-abc.com.br` and `def-company.com.br` as separate entities, instead of incorrectly seeing them as simple subdomains of the public suffix `.com.br` – technically a second-level domain.

For the thesis, the shortest non-public suffix part of a domain has been labeled the primary domain. The domain `example.com.br` is the primary domain for `machine100.services.example.com.br`, as `.com.br` is a public suffix. The term superdomain has also been used for the opposite of subdomain; `example.org` is a superdomain of `www.example.org`.

3.3 Capturing tracker requests

One assumption is that all resources external to the initially requested (origin) domain can act as trackers, even for static (non-script, non-executable) resources with no capabilities to dynamically survey the user's browser, collecting data and tracking users across domains using for example the `referer` (sic) HTTP header [22]. While there are lists of known trackers, used by browser privacy tools, they are not 100% effective [35, 22] due to not being complete, always up to date or accurate. Lists are instead used to emphasize those external resources as *confirmed* and *recognized* trackers.

Resources have not been blocked in the browser during website retrieval, but have been matched by URL against a third-party list in the classification step (A.5.4) of the data analysis. This way trackers dynamically triggering additional requests have also been recored, which can make a difference if they access another domain or another organization's trackers in the process.

The tracker list of choice is the one used in the privacy tool Disconnect.me, where it is used to block external requests to (most) known tracker domains (A.3). It consists of 2,149 domains, each belonging to one of 980 organizations and five categories – see Table 3.2 for the number of domains and organizations per category. The domain level blocking fits well with the thesis' internal versus external resource reasoning. Because domains are linked to organizations as well as broadly categorized, blocking aggregate counts and coverage can form a bigger picture.

Not all domains in the list are treated the same by Disconnect.me; despite being listed as known trackers, the content category (A.3.6) is not blocked by default in order to not disturb the normal user experience too much. Most organizations are only associated with one domain, but some organizations have more than one domain (A.3.3). Figure 3.1 shows the number of organizations (out of the 980 organizations) that have a certain number of tracker domains (x axis). We see that 47% (459 of 980) have at least two domains listed by Disconnect.me. Google (rightmost point) alone has 271 domains and Yahoo has 71. Some organizations have their domains categorized in more than one category, as shown in detail in Table 3.3. Due to the

Name	Date	Total size	Selection	Selection size	Unique
.SE health status	2014-03-27	980	curated		915
.se zone	2014-07-10	1 318 000	random	100 000	100 000
.dk zone	2014-07-23	1 260 000	random	10 000	10 000
.com zone	2014-08-27	114 178 000	random	10 000	10 000
.net zone	2014-08-27	15 096 000	random	10 000	10 000
reach50.com	2014-09-01	50	top		50
Alexa Top 1M	2014-09-01	1 000 000	top	10 000	9 986
			random	10 000	9 959
			.se		3 364
			.dk		2 637
Total		132 852 050		156 907	156 045

Table 3.1: Domain lists in use

Category	Domains	Organizations
Advertising	1 326	732
Analytics	230	145
Content	513	111
Disconnect	38	3
Social	43	14

Table 3.2: Disconnect's categories

relaxed blocking of the content category this can provide a way to track users despite being labeled a tracker organization.

While cookies used for tracking have been a concern for many, they are not necessary in order to identify most users upon return, even uniquely on a global level [10]. Cookies have not been considered to be an indicator of tracking, as it can be assumed that a combination of other server and client side techniques can achieve the same goal as a normal tracking cookie [1].

3.4 Data collection

The lists of domains have been used as input to har-heedless, a tool specifically written for this thesis (B.2.2). Using the headless[3] browser phantomjs, the front page of each domain has been accessed and processed the way a normal browser would (A.4.3). HTTP/HTTPS traffic metadata such as requested URLs and their HTTP request/response headers have been recorded in the HTTP Archive (HAR) data format (B.1.1).

In order to make comparisons between insecure HTTP and secure HTTPS, domains have been accessed using both protocols. As websites traditionally have been hosted on the www subdomain, not all domains have been configured to respond to HTTP requests to the primary domain – thus both the added www prefix and no added prefix have been accessed. This means four variations for each domain in the domain lists, quadrupling the number of accesses (A.4.4) to over 600,000. List variations have been kept separate; downloaded and analyzed as different datasets (3.6).

Multiple domains have been retrieved in parallel (A.4.3), with adjustable parallelism to fit the computer machine's capacity (A.4.5). To reduce the risk of intermittent errors – either in software, on the network or in the remote system – each failed access has been retried up to two times (A.4.6).

Details about website and resource retrieval can be found in A.4.

3.5 Data analysis and validation

With HAR data in place, each domain list variation is analyzed as a single dataset by the purpose-built har-dulcify (B.2.3). It uses the command line JSON processor jq (B.1.3) to transform the JSON-based HAR data to formats tailored to analyze specific parts of each domain and their HTTP requests/responses.

Data extracted includes URL, HTTP status, mime-type, referer and redirect values – both for the origin domain's front page and any resources requests by it (A.5.2). Each piece of data is then expanded, to simplify further classification and extraction of individual bits of information; URLs are split into components such as scheme (protocol) and host (domain), the status is labeled by status group and the mime-type split into type and encoding (A.5.3).

Once data has been extracted and expanded, there are three classification steps. The first loads the public suffix list and matches domains against it, in order to separate the FQDN into public suffixes, private prefixes and extract the primary domain (A.5.4). The primary domain, which is the first non-public suffix match, or the shortest private suffix, is used as one of the basic classifications; is an HTTP request made to a domain with the same primary domain as the origin domain's primary domain? Other basic classifications (A.5.4) compare the origin domain with each requested resource's URL, to see if they are made to the same domain, a subdomain or a superdomain. Same domain, subdomain, superdomain and same primary domain requests

[3] The browser does not have a visible window, as it is built for automation.

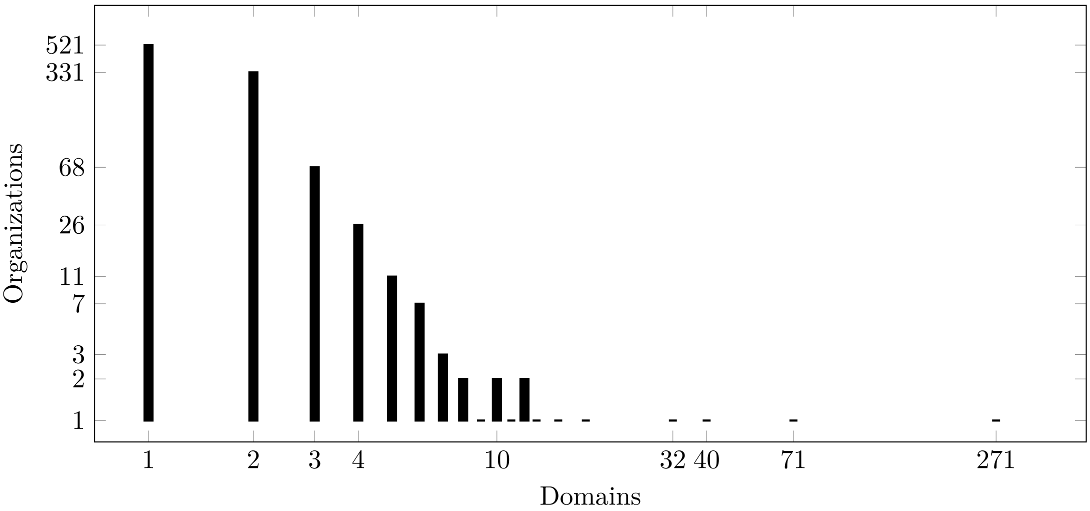

Figure 3.1: Domains per organization

Name	Count	Advertising	Analytics	Content	Disconnect	Social
Yahoo!	4	x	x	x		x
Amazon.com	3	x	x	x		
33Across	2	x		x		
Adobe	2	x		x		
Akamai	2	x		x		
AOL	2	x		x		
AT Internet	2	x	x			
Automattic	2		x	x		
comScore	2	x	x			
Facebook	2			x	x	
Google	2			x	x	
Hearst	2	x	x			
IBM	2	x	x			
LivePerson	2		x	x		
Microsoft	2	x		x		
Nielsen	2	x	x			
Oracle	2	x		x		
QuinStreet	2	x	x			
TrackingSoft	2	x	x			
WPP	2	x	x			

Table 3.3: Organizations in more than one category

often overlap in their classification – collectively they are called internal requests. Any request not considered an internal request is an external request – which is one of the fundamental ideas behind the thesis' result grouping (C.4). Mime-types are counted and grouped, to show differences in resource usage (C.9). To get an overview of domain groups, their primary domains and public suffixes (C.10) are also kept. Another fundamental distinction is also wether a request is secure – using the HTTPS protocol – or insecure. Finally, Disconnect's blocking list (3.3) is mixed in, to be able to potentially classify each requests' domain as a known tracker (A.5.4), which includes a mapping to categories and organizations (C.11).

After classification has completed, numbers are collected across the dataset's domains (A.5.5). Counts are summed up per domain (C.5), but also reduced to boolean values indicating if a request matches a certain classification property or primary/tracker domain, so that a single domain making an excessive number of requests would not skew numbers aggregated across domains. This allows domain coverage calculations, meaning on what proportion of domains a certain value is present.

Most of the results presented in the thesis report are based on non-failed origin domains. Non-failed means that the initial HTTP request to the domain's front page returned a proper HTTP status code, even if it was not indicative of a complete success (C.2). Subsequent requests made while loading the front pages were grouped into unfiltered, only internal and only external requests (C.4). The analysis is therefore split into six versions (B.2.3), although not all of them are interesting for the end results.

Apart from these general aggregates, har-dulcify allows specific questions/queries to be executed against any of the steps from HAR data to the complete aggregates. This way very specific questions (A.5.6), including Google Tag Manager implications (A.4.3) and redirect chain statistics (C.8), can be answered based using the input which fits best. There are also multiset scripts, collecting values from several or all 72 datasets at once. Their output is the basis for most of the detailed results' tables and graphs; see Appendix C.

See also analysis methodology details in Appendix A.5.

3.6 High level summary of datasets

Domains lists chosen for this thesis come in three major categories – top lists, curated lists and random selection from zone files (Section 3.2 and Table 3.1, Section A.1). While the top lists and curated lists are assumed to primarily contain sites with staff or enthusiasts to take care of them and make sure they are available and functioning, the domain lists randomly extracted from TLD zones might not. Results (Chapter 4, Appendix C) seem to fall into groups of non-random and randomly selected domains – and result discussions often group them as such.

Table 3.4 shows the top TLDs in the list of unique domains; while random TLD samples of course come from a single TLD, top lists are mixed. Looking at the complete dataset selection, the gTLD .org, ccTLDs .ru and .de are about the same size. This list can be compared to the per-TLD (or technically public suffix) results in Table C.10, which shows the coverage of TLDs for external requests per dataset.

The curated *.SE Health Status* domain categories in Table 3.5 show that the number of domains per category is significantly lower than the top lists and random domains. This puts a limit on the certainty with which conclusions can be drawn, but still serves a purpose in that the categories often show different characteristics.

The most interesting category is the media category, as it is the most extreme example in terms of requests per domain and tracking (C.6). While the thesis is limited to the front pages of each domain (3.7), it would be interesting to see if users are still tracked after logging in to financial websites (C.4). It is also interesting to see how public authorities, government, county and municipality websites include trackers from foreign countries (C.11.1).

3.7 Limitations

With lists of domains as input, this thesis only looks at the front page of domains. While others have spidered entire websites from the root to find, for example, a specific group of external services [44], this is an overview of all resources. The front page is assumed to contain many, if not most, of the different types used on a domain. Analysis has mostly been performed on each predefined domain list as a whole, but dynamic – and perhaps iterative – re-grouping of domain based on results could improve accuracy, understanding and crystallize details. It would also be of interest to build a graph of domains and interconnected services, to visualize potential information sharing between them [1].

Rank	Count	TLD
1	103 645	.se
2	21 203	.com
3	12 610	.dk
4	11 012	.net
5	650	.ru
6	639	.org
7	619	.de
8	441	.jp
9	334	.br
10	316	.uk

Table 3.4: TLDs in dataset in use

Category	Domains	Unique	Description
Counties	21	21	Sweden's counties.
Domain registrars	146	146	Registrars selling .se domains; most are based in Sweden.
Financial services	79	79	Banks, insurance and others registered with the authorities.
GOCS	60	60	Swedish government-owned corporations.
Higher education	49	49	Universities and colleges.
ISPs	20	20	Internet service providers registered with the authorities.
Media	33	32	Companies and organizations in radio, TV and print media.
Municipalities	290	290	Sweden's municipalities.
Public authorities	282	226	Swedish public authorities.

Table 3.5: *.SE Health Status* domain categories

Chapter 4

Results

This chapter presents the main results, exemplified primarily by four datasets in their HTTP-www and HTTPS-www variations; Alexa's top 10k websites, Alexa's top .se websites, .se 100k random domains and Swedish municipalities. Some result highlights from Swedish top/curated datasets, random .se domains and findings from other domains in different categories are presented in the table below. Supplementary results and additional details are provided in Appendix C.

Swedish top/curated	Random .se	Other findings
	Internal vs external	
Over 90% of most categories' domains rely on external resources – external resources are considered trackers (C.4)	Uses more external resources than .dk, but less than .com and .net (C.4)	There are at least as many external resources, meaning as much tracking, on secure as insecure top domains (alexa.top.10k-hw and alexa.top.10k-sw in Figure 4.3(a))
	39% use *only* external resources (se.r.100k-hw in Figure 4.2(a))	94% of 5,959 HTTPS-www variation domains call external domains (C.5)
	Many random domains use only external resources due to being parked (4.2) or redirecting away from the origin domain (C.8)	78% of 123,000 HTTP-www variation domains call external domains (C.5)
	Secure vs insecure	
Only 13 of 290 municipalities have fully secure websites; no Swedish media sites are completely secure (C.7)	Only 0.3% respond to secure requests, in line with .dk and .net, while .com has 0.5-0.6% response rate (C.2)	

Swedish top/curated	Random .se	Other findings
25% of Swedish municipalities responding to secure requests load 90% of their resources securely – it's close, but still considered insecure (se.hs.municip-sw in Figure 4.3(b))		
Financial institutions redirect from secure *to insecure* sites for 20% of responding domains (C.8)		
	Known trackers	
A single visit to each media sites would leak information to at least 57 organizations (C.11.1)	Disconnect *only detects 3%* of external primary domains as trackers (4.3.4)	A few global top domains load more than 75 known trackers on their front page alone (C.11.1)
70% use content from known trackers (C.11.3)	58% use content from known trackers (C.11.3)	Disconnect's blocking list *only detects 10%* of external primary domains as trackers for top website datasets (4.3.4)
	Over 40% use Google Analytics or Google API (C.11.2)	
	Other	
Swedish media seems very social, with the highest Twitter and Facebook coverage (C.11.4)		Twitter has about half the coverage of Facebook (C.11.4)
		50% of top sites always redirect to the www subdomain, 13% always redirect to their primary domain (C.8)

Table 4.1: Results summary

4.1 HTTP, HTTPS and redirects

Figure 4.1(a) shows the ratio of domains' HTTP response status code on the x axis (C.3). There were few 1xx, 4xx and 5xx responses; the figure focuses on 2xx and 3xx; no response is shown as `null`. In general, HTTPS usage is very low at less than 0.6% among random domains – see random .se (`null`) responses for the se.r.100k-sw dataset. Reach50 top sites are leading the way with a 53% response rate (C.2).

Sites which implement HTTPS sometimes take advantage of redirects to direct the user from an insecure to a secure connection, for example when the user didn't type in `https://` into the browser's address bar. Surprisingly, this is not very common – while many redirect to a preferred variant of their domain name, usually the www subdomain, only a few percent elect to redirect

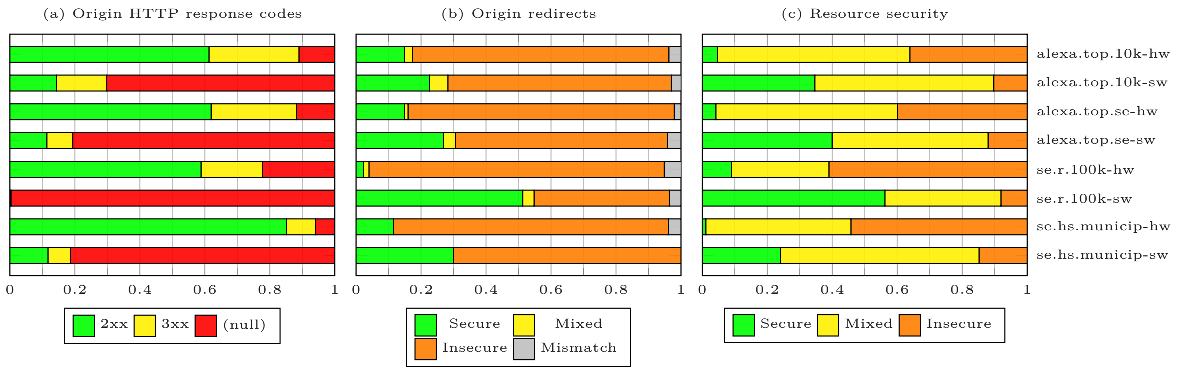

Figure 4.1: Selection of HTTP-www and HTTPS-www variations from Figure C.1, C.6 and C.4

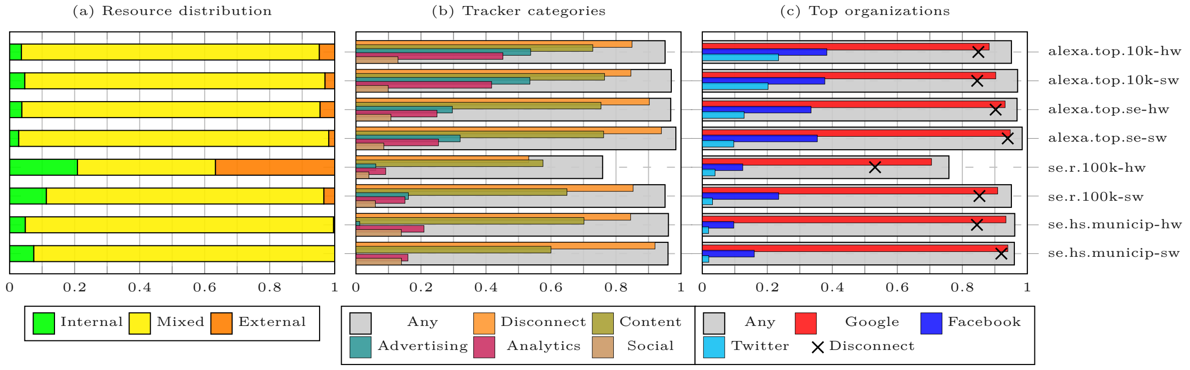

Figure 4.2: Selection of HTTP-www and HTTPS-www variations from Figure C.2, C.9 and C.10

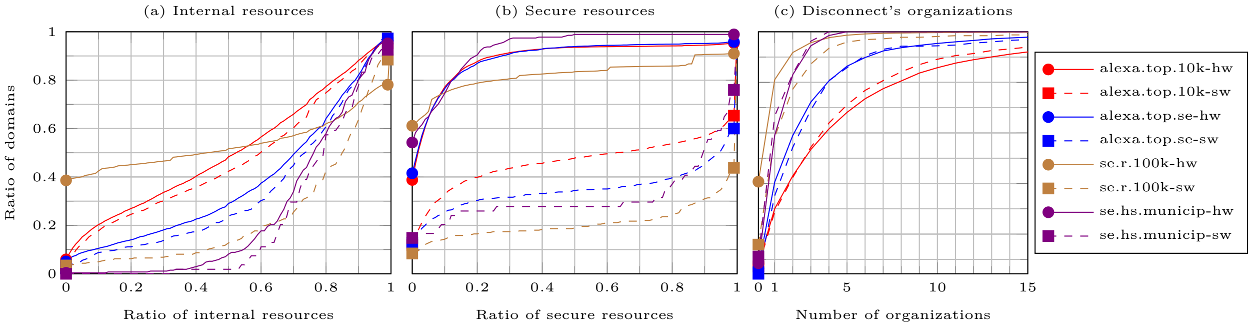

Figure 4.3: Small versions of Figure C.3, C.5 and C.8 showing a selection of HTTP-www and HTTPS-www variations

to a secure URL (C.8). The average number of redirects for domains with redirects is 1.23, but some domains have multiple, chained redirects; a few even to a *mixture* of HTTP and HTTPS URLs.

Figure 4.1(a) shows the ratio of domains responding with redirects (x axis' 3xx responses), and the effect of these redirects are detailed in Figure 4.1(b) as ratio of domains (x axis) which are strictly secure, have mixed HTTP/HTTPS redirects, are strictly insecure or which could not be determined because of recorded URL mismatches. It is surprising to see that redirects more often point to *insecure* than secure URLs – even if the origin request was made to a secure URL. The secure random .se domains (se.r.100k-sw) have a higher secure redirect ratio, but due to the very low response rate of 0.3% when using HTTPS – and even fewer which use redirects – it is hard to draw solid conclusions.

It seems that Swedish media shun secure connections – *not one* of them present a fully secured domain, serving mixed content in case of responding to secure requests. At the same time, they use the highest count of both internal and external resources – with numbers several times higher than other domain lists – and more than 20% of requests go to known trackers.

4.2 Internal and external requests

With each request classified as either internal or external to the originating domain, it is easy to see how sites divide their resources (C.4). Less than 10% of top sites (for example alexa.top.10k-hw) use strictly internal resources, meaning up to 93% of top sites are composed using at least a portion of external resources. See the percentage of domains (x axis) using strictly internal, mixed and strictly external resources in Figure 4.2(a) for a selection of datasets, and Figure C.2 for all datasets. This means external resources – in this thesis seen as trackers – have actively been installed, be it as a commercial choice or for technical reasons (2.1). The difference between HTTP and HTTPS datasets is generally small, showing that users are *as tracked* on secure as on insecure sites.

Figure 4.3(a) shows the cumulative distribution function (CDF) of the ratio of external resources used by each domain, with 0% and 99% internal resources marked. In particular, we show the ratio of domains (y axis) as a function of the ratio of internal resources seen by each domain (x axis). This maps to the bar graph in Figure 4.2(a); 0% is all external, over 99% is all internal – in between means mixed security.

Similar to the HTTPS adoption, we observe significant differences between randomly selected domains and the most popular (top ranked) domains. See how dataset HTTP/HTTPS variation lines follow each other for most datasets, again pointing towards "secure" HTTPS serving *as many trackers* as insecure HTTP. This means that a secure, encrypted connection protecting against eavesdropping doesn't automatically lead to *privacy* – something which users might be lead to believe when it is called a "secure connection" as well as through the use of "security symbols" such as padlocks.

For the HTTP variation of random .se domains (se.r.100k-hw) 40% use strictly external resources; this seems to be connected with the fact that many domains are *parked*[1] and load all their resources from an external domain which serves the domain name retailer's resources for all parked domains. The same domains seem to not have HTTPS enabled, as can be seen in 4.1(a), and the remaining HTTPS domains show the same internal resource ratio characteristics as top domains. There is a wide variety of parked page styles, as well as other front pages without actual content, but they have not yet been fully investigated and separately analyzed (7.5.4).

[1] A parked domain is one that has been purchased from a domain name retailer, but only shows a placeholder message – usually an advertisement for the domain name retailer itself.

4.3 Tracker detection

While looking at the number of requests made to trackers can give a hint towards how much tracking is installed on a website, it can be argued that one or two carefully composed requests can contain the same information as several request. The difference is merely technical, as not all types of resources can be efficiently bundled and transferred in a single requests, but require more than one – therefore it's more interesting to look at the number of organizations which resources are loaded from (C.11.1). Looking at categories can also be interesting – especially for the content category, which isn't blocked by default by Disconnect.me.

4.3.1 Categories

Figure 4.2(b) shows coverage of the five categories in Disconnect.me's blocking list (3.3, A.3.1), as well the grey "any" bar showing the union of known tracker coverage (x axis). The special Disconnect category (A.3.7) is predominant in most datasets, showing coverage almost as large as the union of all categories. The second largest category is content – which is *not blocked by default* by Disconnect.me, as these requests have been *deemed desirable* even to privacy-aware users. This means that even when running Disconnect.me's software, users are *still tracked* on 60-70% of websites (C.11.3).

4.3.2 Organizations per domain

Figure 4.3(c) shows the CDF of the ratio of domains (y axis) with the number of organizations detected per domain (x axis) for a selection of datasets. The random .se domain HTTP variation (se.r.100k-hw) has a different characteristic than others, with 40% of domains having no detected third party organizations; it can be due to domain registrars who serve parked domain page not being listed as trackers. Around 55% of random Swedish HTTP domains (se.r.100k-hw) have 1-3 trackers, and about 5% have more than 3.

Once again it can be seen that the amount of tracking is the same in other HTTP-www variations as in their respective HTTPS-www variation – as the figure shows, the lines follow each other. Most websites also have at least one known tracker present; 53-72% of random domains have at least one tracker installed, while 88-98% of top websites have trackers and 78-100% of websites in the Swedish curated lists. In the larger Alexa global top 10,000 dataset (alexa.top.10k-hw and alexa.top.10k-sw), 70% of sites allow more than one external organization, 10% allow 13 or more and 1% even allow more than 48 trackers – and that is looking only on the front page of the domain.

Out of the Swedish media domains, 50% share information with more than seven tracker organizations – and one of them is sharing information with 38 organizations. Half of the Swedish media websites load more than 6 known trackers; a single visit to the front page of each of the 27 investigated sites would leak information in over 3,800 external requests (C.5) to at least 57 organizations (C.11.1). This means that any guesswork in what types of articles individuals read would read in a printed newspaper is gone – and with that probably the guesswork in exactly what kind of *personal opinions* these individuals hold. While it is already known that commercial media outlets makes their money through advertising, this level of tracking might be surprising – it seems to indicate that what news users read online is very well known.

4.3.3 Google's coverage is impressive

Figure 4.2(c) shows Google, Facebook and Twitter's coverage. It also shows the grey "any" bar showing the union of known tracker coverage and an `x` marking the coverage of the entire Disconnect category of Disconnect.me's blocking list, which they are part of (A.3.7). The organization with the most spread, *by far*, is Google. Google alone has higher coverage than the Disconnect category, meaning that a portion of websites use resources from Google domains in the content category (A.3.6).

The runner ups with broad domain class coverage are Facebook and Twitter, but in terms of domain coverage they are still far behind – see Section (C.11.4). Google is very popular with all top domains and most Swedish curated datasets have a coverage above 80% – and many closer to 90%. Random domains have a lower reliance on Google at 47-62% – still about half of all domains. Apart from the .SE Health Status list of Swedish media domains, Facebook doesn't reach 40% in top or curated domains. Facebook coverage on random zone domains is 6-10%, which is also much lower than Google's numbers. Twitter has even lower coverage, at about half of that of Facebook on average. As can be seen in Figure 4.2(c), Google alone oftentimes has a coverage *higher* than the domains in the Disconnect category – it shows that Google's content domains are in use (A.3.3). While Disconnect's blocking list contains very many Google domains (A.3.2), the coverage is not explained by the number of domains they own, but by the popularity of their services (C.11.2). In fact, at around 90% of the *total* tracker coverage, Google's coverage approaches that of the union of *all* known trackers.

4.3.4 Tracker detection effectiveness

While all external resources are considered trackers, parts of this thesis has concentrated on using Disconnect.me's blocking list for tracker verification. But how effective is that list of 2,149 *known* and *recognized* tracker domains across the datasets? See Section C.12 and Figure C.11 for the ratio of detected/undetected domains. While some of the domains which have not been matched by Disconnect are private/internal CDNs, the fact that less than 10% of external domains are blocked in top website HTTP datasets (such as alexa.top.10k-hw) is notable. The blocking results are also around 10% or lower for random domain HTTP datasets, but it seems it might be connected to the number of domains in the dataset. Only 3% of the 15,746 external primary domains in .se 100k random domain HTTP dataset (se.r.100k-hw) were detected. Smaller datasets, including HTTPS datasets with few reachable websites, have a higher detection rate at 30% and more.

Chapter 5

Discussion

Two previously investigated pieces of data this thesis' subject was based upon were .SE's statistics regarding the use of Google Analytics and the adoption rates for HTTPS on Swedish websites. Both have been thoroughly investigated and expanded upon. Below is a comparison to the *.SE Health Status* reports as well as a few other reports, in terms of results and methodology. After that, a summary of the software developed and open source contributions follow.

5.1 *.SE Health Status* comparison

5.1.1 Google Analytics

One of the reasons this thesis subject was chosen was the inclusion of a Google Analytics coverage analysis in previous reports. The reports shows overall Google Analytics usage in the curated dataset of 44% 2010, 58% in 2011 and 62% in 2012 [30, 31, 32].

Thesis data from filtered HTTP-www *.SE Health Status* domain lists shows usage in the category with the least coverage (financial services) is 59% while the rest are above 74% (C.11.2); the average is 81%. The highest coverage category (government owned corporations) is even above 94%. Since Google Analytics can now be used from the DoubleClick domain as well as Google offering several other services, looking only at the Google Analytics domain makes little sense – instead it might make more sense to look at the organization Google as a whole. The coverage jumps quite a bit, with most categories landing above 90% (C.11.4), which is also the *.SE Health Status* average.

This means that traffic data from at least 90% of web pages considered important to the Swedish general public end up in Google's hands. In a broader scope considering all known trackers, 95% of websites have them installed.

It is possible to extract the exact coverage for both Google Analytics and DoubleClick from the current dataset. Google Analytics already uses a domain of its own, and by writing a custom question differentiating DoubleClick's ad specific resource URLs from analytics specific resource URLs, analytics on doubleclick.net can be analyzed separately as well.

5.1.2 Reachability

The random zone domain lists (.se, .dk, .com, .net) have download failures for 22-28% of all domains when it comes to HTTP without www and HTTP-www variations, where www has fewer failures (C.2). The HTTP result for .se is consistent with results from the *.SE Health Status* reports, according to Patrik Wallström, where they only download www variations. Curated *.SE Health Status* lists have fewer failures for both HTTP, generally below 10% for the `http://www.` variation – perhaps explained by the thesis software and network setup (A.4.6). Several prominent media sites with the same parent company respond as expected when accessed with a normal desktop browser – but not automated requests, suggesting that they detect and block some types of traffic.

5.1.3 HTTPS usage

.SE have measured HTTPS coverage among curated health status domains since at least 2007 [27, 28, 30, 31, 32, 29]. The reports are a bit unclear about some numbers as measurement methodology and focus has shifted over the years, but the general results seem to line up with the results in this thesis. Quality of the HTTPS certificate is also considered by looking at for example expiry date, deeming them correct or not. Data comparable to this thesis wasn't published in the report from 2007 and HTTPS measurements were not performed in 2012. Also, measurements changed in 2009 so they might not be fully comparable with earlier reports.

Table 5.1 shows values extracted from reports 2008-2013 as well as numbers from this thesis. Thesis results show a 24% HTTPS response rate (C.2) while the report shows 28%. The 2013 report also finds that 24% of HTTPS sites redirect from HTTPS back to HTTP. In this thesis it is shown that 22% of *.SE Health Status* HTTPS domains have non-secure redirects (C.8) – meaning insecure or mixed security redirects – which is close to the report findings.

5.2 *Cat and Mouse*

Similar to the methodology used in this thesis, the *Cat and mouse* paper by Krishnamurthy and Wills [22] use the Firefox browser plugin AdBlock to detecting third-party resources – or in their case advertisements. The ad blocking list Filterset.G from 2005-10-10 contains 108 domains as well as 55 regular expressions. Matching was done after collecting requested URLs using a local proxy server, which means that HTTPS requests were lost.

As this thesis uses in-browser URL capturing, HTTPS requests have been captured – a definite improvement and increase in the reliability of result. On the other hand, not performing URL path matching (7.5.1) and instead only using domains (the way Disconnect does it) might lead to fewer detected trackers, as the paper shows that only 38% of their ad matches were domain matches, plus 10% which matched both domain and path rules. Their matching found 872 unique servers from 108 domain rules – the 2,149 domains (1,326 in the advertisement category) in the currently used Disconnect dataset might be enough, as subdomains are matched as well.

The paper also discusses serving content alongside advertisements as a way to avoid blocking of trackers (C.11.3), as well as obfuscating tracker URLs by changing domains or paths, perhaps by using CDNs (C.11.2). While this thesis has not confirmed that this is the case, it seems likely that some easily blockable tracking is being replaced with a less fragile business model where the tracker also adds value to the end user. There are two ways to look at this behavior – do service providers add tracking to an existing service, or do they build a service to support tracking? For Google, the most prevalent tracker organization, it might be a mixture of both. In the case of AddThis, a wide-spread (C.11.2) social sharing service (A.3.8), it seems the service is provided as a way to track users. The company is operated as a marketing firm selling audience information to online advertisers, targeting social influencers[1].

The report looks the top 100 English language sites from 12 categories in Alexa's top list, plus another 100 sites from a political list. These sites come from 1,116 domains. A total of 1,113 pages were downloaded from that set, plus 457 pages from Alexa's top 500 in a secondary list. The overlap was 180 pages. See Table 5.2 for ad coverage per dataset and Table 5.3 for ad domain match percentage.

The paper's top ad server (determined by counting URLs) was doubleclick.net at 8.8%. While thesis data hasn't been determined in the same way, it seems that Doubleclick has strengthened their position since then: comparing the coverage of doubleclick.net (C.11.2) to other advertisement domains, organizations or even the category seems to point to that Doubleclick alone has more coverage than the other advertisers together for several datasets. Advertisement coverage was 58% for Alexa's top 500, while this thesis detects 54% advertisement coverage *plus* an additional 52% doubleclick.net coverage – the union has unfortunately not been calculated.

5.3 *Follow the Money*

Gill and Erramilli *et al.* [16] have explored some of the economical motivations behind tracking users across the web. Using HTTP traffic recorded from different networks, the paper looks at the presence of different *aggregators* (trackers) on sites which are *publishers* of content. To distinguish publishers from aggregators, the domains in each session are grouped with regards to the originating request's domain's IP-address' network autonomous system (AS) number – requests to another AS number are counted as third parties/aggregators. In some cases looking at AS numbers leads to confusion, for example when multiple publishers are hosted on CDN service; separating publishers and aggregators by domain names is then used.

The largest dataset, a mobile network with 3,000,000 users and 1,500,000,000 sessions presumably excludes HTTPS traffic, as it would be unavailable to public network operators. The paper's Figure 3 shows that aggregator coverage is higher for the very top publishers; over 70% for the top 10.

The coverage of top aggregators on top publishers are shown in Table 5.4, alongside numbers from this thesis. This thesis doesn't use recorded HTTP traffic in the same way, and download each domain only once per dataset, but looking at publisher coverage should allow a comparison.

Google again shows a significantly greater coverage at 80%, compared to Facebook's 23% in second place. It looks like the paper has grouped AdMob under Global Crossing, which was a large network provider connecting 70 countries before being acquired in 2011. AdMob was acquired by Google already in 2009, so it's unclear why it's listed separately; one reason might be because the dataset is mobile-centric and that AS number is still labeled Global Crossing. Thesis results show even higher numbers for Disconnect's matching of Google and Facebook – 4 and 14 percentage points – even when looking at the Alexa's top 10,000 sites. Microsoft doesn't seem to have as much coverage in the top 100,000, but the other organizations show about the same coverage.

5.4 Trackers which deliver content

In Disconnect's blocking list, there is a category called content (A.3.6). While all other categories are blocked by default, this one is not as it represents external resources deemed *desirable* to

[1] https://en.wikipedia.org/wiki/AddThis

Source	Domains	Answering	Correct	All rate	Correct rate	Redirects to HTTP
2008	1 685	722	112	0.750	0.160	
2009	663	168	129	0.250	0.190	
2010	670	227	190	0.340	0.280	
2011	912	175	159	0.190	0.170	
2012	913					
2013	1 224	458	339	0.280	0.190	0.240
Thesis	911	218		0.240		0.220

Table 5.1: *.SE Health Status* HTTPS coverage 2008-2013

Category/TLD	Pages	With ads	Percent
all	1 113	622	56
reference	99	32	32
regional	98	59	60
science	95	31	33
shopping	97	54	56
arts	98	76	78
business	98	61	62
computers	88	47	53
games	95	61	64
health	99	40	40
home	97	60	62
news	98	83	85
political	88	61	69
recreation	95	46	48
com	832	532	64
org	65	24	37
gov	50	3	6
net	27	13	48
edu	43	5	12
global500	457	266	58

Table 5.2: *Cat and Mouse* ad coverage

Ad Server	Percentage
doubleclick.net	8.8
advertising.com	5.2
falkag.net	4.7
atdmt.com	4.4
blogads.com	3.7
akamai.net	3.2
zedo.com	3.1
2mdn.net	2.8
com.com	1.9
2o7.net	1.9

Table 5.3: *Cat and Mouse* ad server match distribution

Aggregator	Frac. Rev.	Frac. Users	Frac. Pubs.	D. Org.
Google	0.18	0.17	0.80	0.84
Facebook	0.06	0.09	0.23	0.37
GlobalCrossing (AdMob)	0.04	0.11	0.19	
AOL	0.03	0.04	0.07	0.05
Microsoft	0.03	0.04	0.17	0.02
Omniture	0.03	0.05	0.07	0.10
Yahoo! (AS42173)	0.03	0.04	0.07	0.04

Table 5.4: *Follow the Money* aggregators' revenue, users and publisher coverage fractions plus this thesis' Disconnect organization coverage

Disconnect's users. So while they are known tracker domains, they are allowed to pass "by popular demand" – similar to CDNs (A.2, 5.2). This brings an advantage to companies that can deliver content, as they can just as well use content usage data as pure web bug/tracker usage data when analyzing patterns.

Google has several popular embeddable services in the content category (coverage from large datasets in C.11.2 in parentheses), including Google Maps[2] (2-5%), Google Translate[3] and least but not least YouTube[4] (3-7%). Lesser known examples include Recaptcha[5] which is an embeddable service to block/disallow web crawlers/bots access to web page features. Those are visible examples, which users interact with; Google Fonts[6] which serves modern web fonts for easy embedding, is still visible but not branded. Google Hosted Libraries[7] is another very popular service yet unbranded service. It hosts popular javascript libraries from Google's extensive CDN network instead of the local server for site speed/performance gains, and is not visible as components – but they cannot be removed without affecting functionality. Especially the two latter, served from the googleapis.com (30-56%) domain, are prevalent in several of the datasets – and they are usually loaded on every single page on a website, and thus gain full insight on users' click paths and web history. The content tracking is passive and on the HTTP level, as opposed to scripts executing and collecting data such as Google Analytics (30-80%) and Doubleclick (11-53%).

As Disconnect's blocking blacklist is shown to cover only a fraction of the external domains in use, a whitelist could be an alternative. As Disconnect already has whitelisted the content category, it can be considered a preview of what shared whitelisting might look like. It is already the second largest category, in terms of coverage, with over 50% of domains in most datasets having matches (C.11.3). While the number of organizations being able to track users might be reduced by whitelisting, it seems the problem needs more research (7.5.1).

5.5 Automated, scalable data collection and repeatable analysis

One of the prerequisites for the type of analysis performed in this thesis was that all collection should be automated, repeatable and be able to handle tens of thousands of domains at a time. This goal has been achieved, and a specialized framework for analyzing web pages's HTTP requests has been built. While most of the code has been tailored to answer questions posed in this thesis, it is also built to be extendable, both in and between all data processing steps. More data can be included, additional datasets can be mixed in, separate questions can be written to query data from any stage in the data preparation or analysis. Tools have been written to easily download and compare separate lists of domains, and by default data is kept in its original downloaded form so that historical analysis can be performed.

It might be hard to convince other researchers to use code, as it might not fulfill all of their wishes at once on top of any "not invented here" mentality. Fortunately, the code is easy to run, and with proper documentation other groups should be able to at least test simple theories regarding websites. Some of the lists of domains used as input are publicly available, and thus results can also be shared. This should encourage other groups, as looking at example data might spark interest.

[2] https://developers.google.com/maps/documentation/embed/

[3] https://translate.google.com/manager/website/

[4] https://developers.google.com/youtube/player_parameters

[5] https://developers.google.com/recaptcha/

[6] https://www.google.com/fonts

[7] https://developers.google.com/speed/libraries/

5.6 Tracking and the media

Swedish media websites have been shown to have the highest amount of trackers per site among the datasets – both in general and for advertisement and analytics categories. Media and press are trying to be independent from and uninfluenced by their advertiser, despite being a source of income.

Advertisement choices are historically based on audience size and demographic, determined by readership questionnaires. Even publishers themselves couldn't know what pages and articles readers *actually* read, once the paper had left the press. Current tracker technology allow both publishers and advertisers to see *exactly* what users are reading online – down to time spent per paragraph if they wanted. It could also mean that this type of intimate knowledge of what news is popular, connected with the kind of click traffic advertisers are seeing, means that they have financial incentive to control exactly what the media *should* write – as opposed to should not write. Does bad news bring more *advertisement clicks* (or other measurable targets such as product purchases, as opposed to newspaper readers) than good? – spend more money advertising on articles about bad news [45]. This will eventually affect the publisher's advertisement income. Advertisers could also separately investigate "relatedness" [33] but use it to value advertisement and provide their own expanded article categorization with fine-grained details for further refinement.

5.7 Privacy tool reliability

Can a privacy tool using a *fixed blacklist* of domains to block be *trusted* – or can it only be trusted to be 10% effective (C.12)? Regular expression based blocking, such as EasyList used by AdBlock, might be more effective, as it can block resources by URL path separate from the URL domain name (7.5.1) – but it's no cure-all. It does seem as if the blacklist model needs to be improved – perhaps by using whitelisting instead of blacklisting. The question then becomes an issue of weighing a game of *cat and mouse* (5.2) – if the whitelist is shared by many users – against *convenience* – if each user maintains their own whitelist. At the moment it seems convenience and blacklists are winning, at the cost of playing cat and mouse with third parties who end up being blocked.

5.8 Open source contributions

During the development of code for this thesis, other projects have been utilized. In good open source manners, those projects should be improved when possible.

5.8.1 The HAR specification

After looking at further processing of the data, some improvements might be suggested.

One such suggestion might be to add an absolute/resolved version of `response.redirectURL`, as specification 1.2 seems to be unclear wether or not it should be kept as-is from the HTTP `Location` header or browser's `redirectURL` values – both of which possibly is relative. As subsequent HTTP requests are hard to refer to without relying either on exact request ordering (the executed redirect always coming exactly as the next entry) or at least having the URL resolved (preferably by the browser) before writing it to the HAR data. Current efforts in `netsniff.js` (B.2.2) to resolve relative URLs using a separate javascript library have proven inexact when it comes to matching against the browser's executed URL, differing for example in wether trailing slashes are kept for domain root requests or not. What would be even better, is a way to refer to the reason for the HTTP request, be it an HTML tag, a script call or an HTTP redirect – but that could to be highly implementation dependent per browser.

5.8.2 phantomjs

While `netsniff.js` (B.2.2) from the phantomjs example library has been improved in several ways, patches have not yet been submitted. Since it is only an example from their side, a more developed version might no longer serve the same purpose – educating new users on the possibilities of phantomjs. An attempt to break the code down and separate pure bug fixes from other improvements might help. The version written for this thesis is released under the same license as the original, so reuse should not be a problem for those interested.

5.8.3 jq

Using jq as the main program for data transformation and aggregation has given me a fair amount of knowledge of real world usage of the jq domain-specific language (DSL). Bugs and inconsistencies have been reported, and input regarding for example code sharing through a package management system and (semantic) versioning has been given. Some of the reusable jq code and helper scripts written for the thesis has been packaged for easy reuse, and more is on the way.

5.8.4 Disconnect

Disconnect relies heavily on their blocking list (A.3), as it is the base for both the service of blocking external resources and presenting statistics to the user. While preparing (B.2.3) and analyzing (B.2.3) the blocking list, a number of errors and inconsistencies were found. Unfortunately, the maintainers do not seem very active in the project, and even trivial data encoding errors were not patched over a month after submission. According to Disconnect's Eason Goodale in an email conversation 2014-08-13, the team has been concentrating on a second version of Disconnect as well as other projects. While patches can be submitted through Disconnect's Github project pages, Goodale's reply seems to indicate they will not be accepted in a timely fashion and perhaps irrelevant by the time the next generation is released to the public.

5.8.5 Public Suffix

A tool that parses the public suffix list from its original format to a JSON lookup object format has been written. Using that tool an inconsistency in the data was detected – the TLD .engineering being included twice instead of .engineer and .engineering separately. This had already been detected and reported by others, but it can be used to detect future inconsistencies in an automated manner.

5.9 Platform for Privacy Preferences (P3P) Project[8] HTTP header analysis

P3P is a way for websites to declare their policies and intentions for data collected from web users. It is declared in a machine-readable format, as an XML file and in a compact encoding

[8] http://www.w3.org/P3P/

as an HTTP header. W3C's work started in 1997 and P3P 1.0 became a W3C recommendation in 2002. It never gained enough momentum and the work with P3P 1.1 was suspended in 2006. P3P is still implemented by many websites, even though it may not follow the originally intended usage.

In conversations with Dwight Hunter, privacy policy researcher, he mentioned that P3P policies are seen as a good technical solution to policy problems in research he had read. Thesis data shows that this is not always true; there are policy-wise useless P3P headers being sent from some webpages, most probably to bypass Internet Explorer's (not all versions) strict cookie rules for third-party site without a P3P HTTP header. This has been highlighted by Microsoft in 2012, pointing at Google's P3P use.

> By default, IE blocks third-party cookies unless the site presents a P3P Compact Policy Statement indicating how the site will use the cookie and that the site's use does not include tracking the user. Google's P3P policy causes Internet Explorer to accept Google's cookies even though the policy does not state Google's intent.[9]

Looking at collected HAR data there are many examples of P3P headers. In the dataset "se.2014-07-10.random.100000-http" from 2014-09-01 with about 1,944,000 recorded requests, about 90,000 present a P3P policy. There are about 550 unique values, including example values shown in Table 5.5.

"Potato" comes from an example in a discussion regarding Internet Explorer and cookie blocking.[10] Other examples include `CP="This is not a P3P policy! It is used to bypass IEs problematic handling of cookies"`, `CP="This is not a P3P policy. Work on P3P has been suspended since 2006: http://www.w3.org/P3P/"`, `CP="This is not a P3P policy. P3P is outdated."`, `CP=\"Thanks IE8\` (which is a malformed value), `CP="No P3P policy because it has been deprecated"`.

This is but one example of where quantitive analysis of real-world web pages shows differences between technical, intended or perceived usage. While P3P may be an outdated example that has been researched [25], it shows how automated, generic tooling can help researchers a lot in their understanding of usage in the wild.

[9] http://blogs.msdn.com/b/ie/archive/2012/02/20/google-bypassing-user-privacy-settings.aspx

[10] *Cookie blocked/not saved in IFRAME in Internet Explorer* http://stackoverflow.com/a/16475093

Count	Value
4 231	CP="This is not a P3P policy! See `http://support.google.com/accounts/bin/answer.py?answer=151657&hl=en` for more info."
2 219	CP="This is not a P3P policy! See `http://www.google.com/support/accounts/bin/answer.py?hl=en&answer=151657` for more info."
855	CP="NO P3P POLICY"
701	CP='Olark does not have a P3P policy. Learn why here: `http://olark.com/p3p`'
138	CP:"BrowseAloud has no dedicated privacy protection policy"
135	CP="Potato"
128	CP="Facebook does not have a P3P policy. Learn why here: `http://fb.me/p3p`"
113	CP="This is not a P3P policy. See `http://acxiom.com/About-Acxiom/Privacy/` for more information."

Table 5.5: Top P3P values

Chapter 6

Related work

Privacy research, tracker research and internet measurement can be challenging, as has been shown by others. As .SE have in-house experts, their information was very valuable at an early stage – several pitfalls may have been avoided. This chapter discusses results in comparison to others' experience, methodology limitations and puts the work in a context.

6.1 *.SE Health Status*

While .SE themselves have written reports analyzing the technical state of services connected to .se domains, *.SE Health Status* [27, 28, 30, 31, 32, 29], the focus has not been on exploring the web services connected to these domains. The research is focused on statistics about usage and security in DNS, IP, web and e-mail; the target audience is IT strategists, executives and directors. Data for the reports is analyzed and summarized by Anne-Marie Eklund Löwinder, a world-renown DNS and security expert[1], while the technical aspects and tools are under the supervision of Patrik Wallström, a well known DNSSEC expert and free and open source software advocate[2].

The thesis subject has been selected to be in line with the .SE reports, but focusing on web issues; code may be reused and results may be included in future reports. The *.SE Health Status* reports do offer some groundwork in terms of selecting and grouping Swedish domains, HTTPS usage and Google Analytics coverage [30, 31, 32] which have been discussed in Section 5.1. The report is based on data collected from around 900 .se domain names deemed of importance to the Swedish society as a whole, as well as random selection of 1% of the registered .se domain names.

Results for the .se zone and curated lists have during meetings with Wallström and Eklund Löwinder been reported to be reasonable regarding comparable results, such as reachability, HTTPS adaptation and Google Analytics coverage.

Thesis input (domain lists) preparation was automated based on .SE:s internally used data formats. As the thesis is more detailed in analyzing web content than previous reports, there is not yet enough historic data to show a change over time.

[1] https://www.iis.se/bloggare/anne-marie/

[2] https://www.iis.se/bloggare/pawal/

6.2 *Characterizing Organizational use of Web-based Services*

Gill and Arlitt *et al.* [15] analyze several HTTP datasets collected in HTTP proxies at different times and in different organizations. Two datasets are from 2008; one enterprise and one university. They are contrasted with a dataset collected at a residential cable modem ISP in 1997.

The paper introduces methods to identify and categorize the collected traffic to *services providers* (organizations) and *service instances* (a service with a specific use, possibly accessed through different domains names) by looking at over 2,000,000 unique HTTP `Host` header values in total. First domain components are grouped in different ways, such as *brands* (first part after the public suffix), then top results are manually consolidated to service providers. Further consolidation is done by looking at domain name system (DNS) and organization identifiers from Regional Internet Registry (RIR) entries. Services are then grouped to *service classes*; some automated grouping is also possible by looking at HTML metadata. This thesis chose to use the public suffix list for automatic domain part identification, down to the brand (primary domain) level (A.2). In combination with Disconnect's blocking list, organizations and a simple categorization is obtained (A.3). It is a less generic way and possibly not fully effective on historical data, but accurate for the amount of work put in, as manual grouping and classification work is avoided – and possibly improved – by using crowdsourcing.

The single domain which is easiest to compare is doubleclick.net, listed as a separate brand in the paper, is shown to have 19% of transactions in the 2008 datasets. Thesis numbers (11-53%) are higher for most datasets (C.11.2), but the paper datasets also contain repeated and continuous use of services – such as Facebook and client side applications – which may lower relative numbers for other services.

While a comparison between paper and thesis numbers would be possible for HTTP methods, HTTP status codes as well as content types, they would require additional, slightly modified analysis of existing requests.

6.3 *Challenges in Measuring Online Advertising Systems*

The paper *Challenges in Measuring Online Advertising Systems* [17] shows that identifying ads and how data collected from trackers affect ads has several challenges. This thesis does not look at *which* ads are shown to a user, but rather where ads are served from. Potentially relevant to this thesis would be for example DNS load-balancing by ad networks, cookies differing between browser instances and local proxies affecting the HTTP request. This is how they were considered and dealt with:

- Final dataset downloads were performed by a single machine (A.4.1) so there should not be any proxy partially affecting the results (A.4.2).
- Load-balancing such as using multiple (round-robin) DNS records might lead to varying results, depending on the remote system setup. Browser instances themselves are short-lived and share the same system-level DNS cache, so requests made within the DNS records' time to live (TTL) should act, and thus vary, uniformly.
- Cookies might affect which ad network is used by an ad network aggregator; in this case each request is made by a new browser instance without any cookies, so results should be random – at least from the client-side point of view (A.4.3). As domains/websites are not

reloaded, but rather requested four separate times (A.4.4), results may differ but in a way unaffected by browser history and cache (A.4.3).

6.4 .SE Domain Check

In order to facilitate repeatable and improvable analysis for this thesis, tools have been developed to perform the collection and aggregation steps automatically. .SE already has a set of tools that run monthly; integration and interoperability will smooth the process and continuous usage. There is also a public .SE tool to allow website owners to test their own sites, *Domänkollen*[3], which might benefit from some of the code developed within the scope of this thesis.

6.5 Cookie syncing

A recent large-scale study by Acar *et al.* [1] included a cookie syncing privacy analysis. It was shown that unique user identifiers were shared between different third parties. IDs can be shared in different ways. If both third parties exist on the same page, they can be shared through scripts or by looking for any IDs in the location URL. They can also be shared by one third-party sending requests to a second third-party (known as a fourth-party), either by leaking the location URL as an HTTP referrer or by embedding it in the request URL. In crawls of Alexa's top 3,000 domains, one third-party script in particular sends requests with synced IDs to 25 domains; the IDs were eventually shared with 43 domains. They also showed that a user's browsing history reconstruction rate rose from 1.4% to 11% when backend/server-to-server overlaps were modeled.

- The study used a modified Firefox browser to look at values stored in primarily cookies. As all HTTP requests are recorded in this thesis, including HTTP cookie headers, a limited version of the same study could be performed.
- Acar *et al.* looks at in-browser scripting utilizing for example `localStorage`, `canvas` fingerprinting and ID storage in external plugins such as Flash. While that might be possible, the modifications that would need to be made to phantomjs are non-trivial, and my current scope does not allow for that. With their research as a base, cookie respawning and sharing could possibly be confirmed using this thesis' code as a external tool using a different browser platform.

6.6 HTTP Archive[4]

In an effort to measure web page speed on the internet, initially developed in October 2010, the HTTP Archive collects HAR data and runs analyses on them. Unfortunately, their official data dumps are in a custom format, not the original HAR files, but there are some direct comparisons to be made with their *Interesting stats*[5] aggregate data.

- Pages Using Google Libraries API
- HTTPS Requests
- Total Requests per Page

[3] "*Domain Check*." Domänkollen was not publicly released at the time of writing.

[4] http://httparchive.org/

[5] http://httparchive.org/interesting.php

It seems there are unofficial, not yet fully developed, exports of HAR data. Unfortunately they weren't made available until late in the thesis process, and could not be used for software validation and comparison.

Chapter 7

Conclusions and future work

In this chapter we first summarize our main conclusions, list unanswered questions, and then we discuss promising directions for future work. There are many potential improvements which could help other researchers, as well as refine the analysis of data already collected for this thesis.

7.1 Conclusions

The use of external resources and known trackers is very high. While it has been a trend to outsource resource hosting and to use third-party services, it was previously unknown to what extent. It has now been shown that most websites use external resources in some form – almost 80% of the 123,000 responding domains looking at the most common variation, HTTP-www C.5. This broad non-governmental tracking should be a concern for privacy minded individuals much as government controlled surveillance is.

This concern should be even higher on HTTPS enabled websites. Such sites have made an active choice to install encryption to avoid passive surveillance and stave off potential attacks – but 94% of HTTPS-www variation domains use external resources C.5.

It seems using a blacklist to stop trackers is the wrong way to go about it. Even with the crowdsourced list used by popular privacy tool Disconnect, the blocking list *only detects 10%* of external primary domains as trackers for top website datasets (4.3.4). Looking at the sheer number of external domains in use, it is easy to understand why blocking high-profile targets seems like a good option – but if 90% of external domains aren't listed even as known, *desirable* content, the blacklisting effort seems futile. Further research could use other blacklists and compare effectiveness (7.5.1).

7.2 Open questions

There are further questions in line with current results which could potentially be answered with additional research.

- Could any external resources actually be considered internal, despite being loaded from external domains?

- Could internal resources also be seen as external, if another organization manage the servers?

- If a resource whitelist is used, how large does it have to be to get a functional website – and how to determine functionality?
- How can more domains be connected to organizations, in terms of ownership, as in Disconnect's blocking list?
- How to determine if a resource
  - Crosses Sweden's borders in transit?
  - Is handled by an organization with base or ownership outside of Sweden?
- What user data could potentially be collected, and subsequently inferred?
- To what extent can the average Swedish internet user's browsing habits be correlated across the most commonly visited webpages?
- Can the same techniques be applied to data from countries other than Sweden?

7.3 Improving information sharing

7.3.1 Creating an information website

Despite thesis code being open source, much of the thesis data is hard to retrieve, analyze and process for individuals. A separate tool performing the work for anyone should be created. Apart from presenting data already collected as a part of the thesis, it could accept user input to analyze individual domains. With several domains as input, any overlap can be detected and presented to the user as an information sharing graph. One of the inspirations for this thesis was Collusion[1], which is a tool to dynamically display from which external domains a page retrieves resources right in the browser. A version of the same tool could be built, where instead of letting the user's browser retrieve sites the server would do it. This way a non-technical user does not have to "risk" anything by visiting web pages, and their relationship could be displayed anyways. Collecting data server-side also allows for cached lookups and a grander scope, where further relationships apart from the first hand ones could be suggested. "If you frequently visit these sites, you might also be visiting theses sites – click to display their relationships as well."

Over time and with user input, the dataset collected on the server would increase, and a historical graph relating to both results shown in this thesis and the relationships between sites can be created. This is similar to what both the HTTP Archive (6.6) is doing on a large scale but with slightly different focus, and what the *.SE Health Status* is doing but on a less continuous basis and with a shifting focus.

7.3.2 Package publicly available data

Datasets based on publicly available datasets can be packaged for other researchers to analyze. While fresh data would be better, a larger dataset can take time to download on a slow connection or computer and all software may not be available to the researcher. It lowers the step in for others who might be interested in the same kind of research, which might lead to the software used in this thesis being improved.

[1] http://collusion.toolness.org/

7.3.3 Code documentation

With some 75 scripts written and released as open source, the need for documentation has gradually increased. The reason for not writing proper documentation – not having direct collaborators writing code – is a hinderance for future collaborators or users to get started. While code documentation has not been an explicit part of the thesis plan, it can be seen as an important step for future usage. The code is not magic in any way, but if understanding the functionality of a file required reading over a hundred lines of code instead of two or three lines of comments, it means a rather steep learning curve for something that is supposed to be simple.

7.4 Improving domain retrieval

7.4.1 Automated testing

So far all testing of har-heedless and phantomjs has been done manually. It has proven to be a working setup, as thesis results are based on these tools, but the features are to be considered fragile as there are no regression tests. Automated tests of the different levels (shell scripts, `netsniff.js` (B.2.2), screenshots, error handling) might help achieve stability in case of for example future improvements of phantomjs. Tests might include setting up a web server with test pages serving different kinds of content, as well as different kinds of errors. During mass downloading of domains phantomjs has been observed outputting error messages, such as failed JPEG image decoding and unspecified crashes. The extent of these errors have so far not been examined, as they have ended up being clumped together with external errors such as network or remote server failures.

7.4.2 Investigating failed domains

There are many reasons domain retrieval could fail, but for top or curated domain lists the chances of the site being down are considerably lower than for randomly selected domains. Each website has been requested up to three times, in order to avoid intermittent problems (A.4.6). Despite this, certain sites do not respond to requests from the automated software. There are several ways for a remote system to detect requests from automation software, with the simplest one being looking at the HTTP `User-Agent` browser make/model identifier string.

Automated downloading of webpages, especially downloading several in short succession, can be seen by site and service owners as disruptive by using system resources and skewing statistical data. Traversing different pages on a single website can also be detected by looking at for example navigational patterns [42, 26]. By only downloading the domain root page and associated resources this tool might not fall into that category of detection.

As some sites respond to desktop browser requests, but not har-heedless' requests, it is believed they have implemented certain "protection" from this kind of software. In respect of their wish not to serve automated requests, har-heedless' browser has not been modified, for example by using a different `User-Agent` string, to try to avoid these measures.

7.4.3 Browser plugins

If possible, a set of common browser plugins could be installed into phantomjs. The first that comes into mind is Adobe Flash, which is sometimes used to display dynamic ads. Flash also has the ability to request resources from other domains, so it might affect results to not render them. An additional problem might be that Flash has its own cookie system, which used storage external to the browser. This brings a new set of potential problems, as Flash cookies are a big part of evercookies and cookie respawning [1]. This means that a headless browser without persistent storage might end up having identifier cookies set in Flash storage, thus being easily and uniquely identified on subsequent visits. While this might not affect this thesis much, as external plugins have not been installed, it might affect other kinds of research being conducted based on the same tools.

7.4.4 System language

Tests were run on English language systems, without making any customizations to phantomjs' settings or HTTP `Accept-Language` headers. While sites have been downloaded from around the world, localized domains might behave differently depending on user language. Google has a recommendation saying that they will prioritize TLDs specific to a region with a certain language (such as .se and Sweden) for users sending `Accept-Language` prioritizing Swedish[2]. This stems from them seeing that localized results have a higher usage rate.

7.4.5 System fonts

Some of the difference between site screenshots and manually browsing to a site is in the fonts displayed. Most of the domains have been downloaded on a headless server, where fonts have not mattered to the system owner. Installing additional fonts commonly available on average user systems might reduce perceived difference.

7.4.6 Do Not Track

While the HTTP header *Do Not Track* (`DNT`) has not been set, it would have been interesting to look at the difference in response from remote services. Detecting usage of the server-response header *Tracking Status Value* (`TSV`) would be a good start[3]. As cookie headers can be analyzed, the difference could have been detected both per origin domain and per connected service. See also the P3P analysis (5.9) for a related header.

7.4.7 Using more scalable software

While invoking phantomjs on a single machine is often enough (A.4.1, A.4.5), that level of computing power is not always enough to download large, or continuously monitor, domain lists in a timely manor. While downloaded HAR file output is easy enough to combine from different machines, it might be worth investigating already parallelized software, such as spookystuff[4]. Built to spider and extract data from websites using multiple – even thousands – of machines on Amazon's cloud computing platform, it could enable analysis of for example entire TLD zones.

7.4.8 Domain lists

There are other domain lists that might have been suitable in this thesis. One curated top list is the KIA index, a list of top sites in Sweden aggregating statistics from different curated sites' underlying analytics tools.[5] Other TLD zone files, both from other countries and more generic ones, could be used as well. For example the new generic TLDs could be compared to older ones.

[2] http://googlewebmastercentral.blogspot.se/2010/03/working-with-multi-regional-websites.html

[3] http://www.w3.org/2011/tracking-protection/drafts/tracking-dnt.html

[4] https://github.com/tribbloid/spookystuff

[5] http://www.kiaindex.net/

7.5 Improved data analysis and accuracy

Improvements to the data transformation and analysis steps.

7.5.1 Refined ad and privacy blocking lists

There are several lists of known ads and privacy invading trackers in use in blocking software than Disconnect (A.3). One of the most popular ones is EasyList[6], which exists in several varieties – privacy blocking, country specific, social media content blocking and others. They were considered, but in the end not incorporated because of the filter list format. It is a mixture of HTML element and URL blocking, and it lacks the connection between blocks and corresponding organization[7]. There is also Ghostery[8], which uses a proprietary list that also contains organizations, but it has not been used because of licensing issues[9].

In another approach, future research could use *whitelisting* to try to determine challenges in detecting and recording *desirable* and *functional* external resources. An example of whitelisting already exists in Disconnect's content category, which might be a good start (5.4).

On a technical level, some blocking rule formats have also posed a problem in terms of implementation into the current data processing framework that is har-dulcify. It relies heavily on jq (B.1.3), which does not have a public release that implements in regular expressions support, a major part of some blocking lists. The idea is that ads and related resources are filtered matching requests' complete URL against the blocking rules, which are a mixture of both more coarse and more fine-grained than Disconnect's domain based rules. One example is `/ads/`, matching a folder name that suggests that all URLs containing this particular path substring is serving advertisements. The thought of using a single path substring to block advertisements served from any domain is more coarse than pinpointing a single domain, but it is also more specific as it would not block legitimate content from another subfolder on the same domain. This way general ad serving systems can be blocked, while domains that serve both ads and content is still allowed serve the content without interfering with the blocking of ads.

> At the time of writing, jq is released as version 1.4. Support for regular expressions is planned for version 1.5.

7.5.2 Automated testing

Data transformations have been written in a semi-structured manner, with separate files for most separate tasks, often executed in serial stages. Each task accepts a certain kind of data as input for the transformation to work correctly – but as both input and output from separate stages looks very similar, it is hard to tell which kind of data it accepts and what the expected output is – and if a change in one stage will affect later stages. Writing automated tests for each stage would have helped during both adding functionality and refactoring the structure. At times, there have been rather time-consuming problems with unexpected or illegal input from real world sites – extracting that kind of input to create a test suite would have sped up fixes and raised confidence in that the input would be handled appropriately and output would still be correct. So far that has not been done, and much of the opportunity to gain from tests have been lost as work has progressed past each problem.

[6] https://easylist.adblockplus.org/

[7] https://easylist-downloads.adblockplus.org/easylist.txt

[8] https://www.ghostery.com/

[9] https://www.ghostery.com/eula

One solution to validating both input and output would have been to create JSON schemas[10] for each transformation. This kind of verification can easily be automated, and it will help any future changes.

7.5.3 Code reuse

Much of the code written in shell scripts, both Bash and jq code, is duplicated between files. While common functionality suitable for pipelining have been broken out, shared functions have not. Bash provides the `source` command for sharing functionality. Code sharing in jq through the use of modules and packages is still under development, but there is a way to load a single external command file. This file can be precomposed externally by concatenating files with function definitions first, and the actual usage of those functions second. The improvement was postponed due to the relative little reuse in early scripting and bright outlook on modules and packages support. As the number of scripts grew, code sharing/composition possibilities grew as well – and with them possible improvements in development speed, consistency and correctness. At this stage, software stability is more important for the final dataset download and analysis, and code refactoring can only be postponed. Foreseeing a greater reuse of JSON and jq tools, a separate open source project has been started – jq-hopkok [11] – where some scripts have been collected. Many functions and utilities local to har-dulcify are project-agnostic, and thus suitable objects to move to jq-hopkok for ease composition.

At the time of writing, jq is released as version 1.4. Support for modules is planned for a version after 1.5. Packages/package managers are external to the jq core, and do not follow the same planning.

7.5.4 Ignore domains without content

Many domains do not contain any actual content. Examples include web server placeholder pages ("Installation succeeded"), domain listings ("Index of /"), parked domains ("This domain was purchased from ...") and advertisement domains (such as Google Adsense for Domains[12], now retired, or similar Adsense usage). There is a Sweden-centric list of site titles for recognized non-content pages available internally at .SE, but it has not been incorporated.

7.5.5 Implement public suffix rules, use non-ICANN suffixes

As datasets have been analyzed, public suffix rules have proven to work in general, with co.uk and similar second level domains being properly grouped and primary domains extracted. There are still traces of the wildcard rules (A.2) in the data though, which means that while numbers are low, there are domains for which the public suffix rules have not been properly applied.

Other potential improvements would be implementing the non-ICANN, private suffixes. This would for example lower the aggregate numbers for cloudfront.net and amazonaws.com as primary domains in the aggregates, focusing on the fact that subdomains belongs to different organizations. Disconnect's dataset, which lists cloudfront.net as a single tracker entity would still present Amazon as a the single tracking organization behind the domain though, which might be a bit misleading. It is true that they can read traffic data in the same way other web hosting

[10] http://json-schema.org/

[11] https://github.com/joelpurra/jq-hopkok

[12] http://www.google.com/domainpark/index.html

and cloud services host can read traffic data to their customers, but customers' common domain name suffix has little to do with it.

7.5.6 Separating analysis steps

As some of the analysis relies on aggregate numbers, such as domain and requests counts, they expect the entire dataset to be available at the start of the analysis. Saving these numbers in an intermediate step would allow further dataset refinement without having to perform the same filtering multiple times, and thus ease second-level filtering. Custom data questions (A.5.6) are one example, as they need to carry the number of (non-failed) domains throughout the calculations, even though only a small subset of the data is interesting, in order to present them as part of the output.

Another example is the current analysis split into unfiltered and non-failed domains, which then is further split into unfiltered, internal and external resources (B.2.3). If the first step had been made as a separate filtering step instead of an integrated part, further analysis would have been clearer as well as easier to modularize.

Saving the intermediate filtering results would simplify selecting for example non-failed domains which only use secure resources (C.7), to look at their usage of internal/external/tracker usage compared to insecure domains (C.4). While redirects have been analyzed to some extent (C.8), another interesting idea would be to select domains with redirects (C.3) and perhaps consume the redirects – especially from secure domain variations looking at secure redirects – before performing further analysis.

7.5.7 Dynamic grouping, domain selection

The har-heedless software was built to download domains based on a simple list, put them in folders per list and list variation (A.4.4) and then har-dulcify use the most recent HAR data to perform the aggregate analysis for the entire dataset. The overlap between different domain lists have been downloaded multiple times, and each round of downloads have started from zero to be sure that the domain list results represent a specific point in time. It would be beneficial to download each unique domain once then select domains as belonging to a specific analysis group, currently represented by domain lists, after they have been downloaded. This would allow a more dynamic grouping, and possibly re-arranging of domains to enable more interesting second-level analysis.

- All curated or top lists grouped together.
- All random TLD zone selections grouped together.
- All sites from, or popular in, Sweden grouped together.
- Use classifications from one dataset variation as the basis of selection in another, for example looking at how many domains which respond to HTTPS requests also redirect their HTTP redirects to use HTTPS.
- Exclude domains which are deemed to be have no content, such as parked domains and other placeholders (7.5.4).
- Use of external tools to classify domains based on, for example, web server IP-addresses mapped to a geographical locations[13], domain registrar or HTTPS certificate properties such as validity.

[13] https://en.wikipedia.org/wiki/Geolocation_software

- Similarity selection – allowing a selection of domains to be the basis of finding similar domains. Similarities could be based on number of HTTP requests, specific external service usage or even individual unique IDs in query strings for services such as Google Analytics and Google AdWords.

7.5.8 Incremental adding and updating of domains

Another improvement would be to ensure that all data mapping steps are built in such a way that a single domain can be excluded or included from the results. This would enable single domains to be updated, perhaps as part of continuous analysis or from user input (7.3.1), without having to recalculate all steps for the entire dataset results. While most incremental updates are a matter of easy addition and subtraction, some late analysis steps introduce coverage calculations and other arithmetical divisions, which may cause some data/precision loss if reversed. If these data reductions can be deferred to a separate mapping step, computing time might be acceptable.

Bibliography

[1] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. The web never forgets: Persistent tracking mechanisms in the wild. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 674–689, New York, NY, USA, 2014. ACM.

[2] Marianne Ahlgren and Pamela Davidsson. Svenskarna och politiken på internet – delaktighet, påverkan och övervakning. Technical report, .SE The Internet Infrastructure Foundation, 2014.

[3] Esma Aïmeur and Manuel Lafond. The scourge of internet personal data collection. In *Proceedings of the International Conference on Availability, Reliability and Security*, ARES '13, pages 821–828, Washington, DC, USA, 2013. IEEE Computer Society.

[4] T. Berners-Lee, R. Fielding, and H. Frystyk. Hypertext Transfer Protocol – HTTP/1.0. RFC 1945 (Informational), May 1996.

[5] T. Berners-Lee, R. Fielding, and L. Masinter. Uniform Resource Identifier (URI): Generic Syntax. RFC 3986 (INTERNET STANDARD), January 2005. Updated by RFCs 6874, 7320.

[6] T. Bray. The JavaScript Object Notation (JSON) Data Interchange Format. RFC 7159 (Proposed Standard), March 2014.

[7] Markus Bylund. *Personlig integritet på nätet.* FORES, 1 edition, 2013.

[8] D. Cooper, S. Santesson, S. Farrell, S. Boeyen, R. Housley, and W. Polk. Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile. RFC 5280 (Proposed Standard), May 2008. Updated by RFC 6818.

[9] Zhicheng Dou, Ruihua Song, and Ji-Rong Wen. A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of International World Wide Web Conference*, WWW '07, pages 581–590, New York, NY, USA, 2007. ACM.

[10] Peter Eckersley. How unique is your web browser? Technical report, Electronic Frontier Foundation, 2009.

[11] Anja Feldmann, Nils Kammenhuber, Olaf Maennel, Bruce Maggs, Roberto De Prisco, and Ravi Sundaram. A methodology for estimating interdomain web traffic demand. In *Proceedings of ACM SIGCOMM Conference on Internet Measurement*, IMC '04, pages 322–335, New York, NY, USA, October 2004. ACM.

[12] R. Fielding and J. Reschke. Hypertext Transfer Protocol (HTTP/1.1): Conditional Requests. RFC 7232 (Proposed Standard), June 2014.

[13] R. Fielding and J. Reschke. Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content. RFC 7231 (Proposed Standard), June 2014.

[14] Olle Findahl and Pamela Davidsson. Svenskarna och internet 2014. Technical report, .SE The Internet Infrastructure Foundation, 2014.

[15] Phillipa Gill, Martin Arlitt, Niklas Carlsson, Anirban Mahanti, and Carey Williamson. Characterizing organizational use of web-based services: Methodology, challenges, observations, and insights. *ACM Trans. Web*, 5(4):19:1–19:23, October 2011.

[16] Phillipa Gill, Vijay Erramilli, Augustin Chaintreau, Balachander Krishnamurthy, Konstantina Papagiannaki, and Pablo Rodriguez. Best paper – follow the money: Understanding economics of online aggregation and advertising. In *Proceedings of ACM SIGCOMM Conference on Internet Measurement*, IMC '13, pages 141–148, New York, NY, USA, 2013. ACM.

[17] Saikat Guha, Bin Cheng, and Paul Francis. Challenges in measuring online advertising systems. In *Proceedings of ACM SIGCOMM Conference on Internet Measurement*, IMC '10, pages 81–87, New York, NY, USA, 2010. ACM.

[18] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. Measuring price discrimination and steering on e-commerce web sites. In *Proceedings of ACM SIGCOMM Conference on Internet Measurement*, IMC '14, pages 305–318, New York, NY, USA, 2014. ACM.

[19] Georgios Kontaxis, Michalis Polychronakis, Angelos D. Keromytis, and Evangelos P. Markatos. Privacy-preserving social plugins. In *Proceedings of USENIX Conference on Security Symposium*, Security'12, pages 30–30, Berkeley, CA, USA, 2012. USENIX Association.

[20] Balachander Krishnamurthy. Privacy and online social networks: Can colorless green ideas sleep furiously? *IEEE Security and Privacy*, 11(3):14–20, May 2013.

[21] Balachander Krishnamurthy and Craig E. Wills. Analyzing factors that influence end-to-end web performance. In *Proceedings of International World Wide Web Conference*, pages 17–32, Amsterdam, The Netherlands, The Netherlands, 2000. North-Holland Publishing Co.

[22] Balachander Krishnamurthy and Craig E. Wills. Cat and mouse: Content delivery tradeoffs in web access. In *Proceedings of International World Wide Web Conference*, WWW '06, pages 337–346, New York, NY, USA, 2006. ACM.

[23] Balachander Krishnamurthy and Craig E. Wills. Characterizing privacy in online social networks. In *Proceedings of Workshop on Online Social Networks*, WOSN '08, pages 37–42, New York, NY, USA, 2008. ACM.

[24] Saurabh Kumar and Mayank Kulkarni. Graph based techniques for user personalization of news streams. In *Proceedings of ACM India Computing Convention*, Compute '13, pages 12:1–12:7, New York, NY, USA, 2013. ACM.

[25] Pedro Giovanni Leon, Lorrie Faith Cranor, Aleecia M. McDonald, and Robert McGuire. Token attempt: The misrepresentation of website privacy policies through the misuse of p3p compact policy tokens. In *Proceedings of ACM Workshop on Privacy in the Electronic Society*, WPES '10, pages 93–104, New York, NY, USA, 2010. ACM.

[26] Anália G. Lourenço and Orlando O. Belo. Catching web crawlers in the act. In *Proceedings of International Conference on Web Engineering*, ICWE '06, pages 265–272, New York, NY, USA, 2006. ACM.

[27] Anne-Marie Eklund Löwinder. Health status 2008. Technical report, .SE The Internet Infrastructure Foundation, 2008.

[28] Anne-Marie Eklund Löwinder. Health status 2009. Technical report, .SE The Internet Infrastructure Foundation, 2009.

[29] Anne-Marie Eklund Löwinder. Health status 2013. Technical report, .SE The Internet Infrastructure Foundation, 2013.

[30] Anne-Marie Eklund Löwinder and Patrik Wallström. Health status 2010. Technical report, .SE The Internet Infrastructure Foundation, 2010.

[31] Anne-Marie Eklund Löwinder and Patrik Wallström. Health status 2011. Technical report, .SE The Internet Infrastructure Foundation, 2011.

[32] Anne-Marie Eklund Löwinder and Patrik Wallström. Health status 2012. Technical report, .SE The Internet Infrastructure Foundation, 2012.

[33] Yuanhua Lv, Taesup Moon, Pranam Kolari, Zhaohui Zheng, Xuanhui Wang, and Yi Chang. Learning to model relatedness for news recommendation. In *Proceedings of International World Wide Web Conference*, WWW '11, pages 57–66, New York, NY, USA, 2011. ACM.

[34] A. Soltani N. Good M. Ayenson, D. J. Wambach and C. J. Hoofnagle. Flash cookies and privacy ii: Now with html5 and etag respawning. In *Social Science Research Network Working Paper Series*, July 2011.

[35] Delfina Malandrino, Andrea Petta, Vittorio Scarano, Luigi Serra, Raffaele Spinelli, and Balachander Krishnamurthy. Privacy awareness about information leakage: Who knows what about me? In *Proceedings of ACM Workshop on Workshop on Privacy in the Electronic Society*, WPES '13, pages 279–284, New York, NY, USA, 2013. ACM.

[36] Jakub Mikians, László Gyarmati, Vijay Erramilli, and Nikolaos Laoutaris. Crowd-assisted search for price discrimination in e-commerce: First results. In *Proceedings of ACM Conference on Emerging Networking Experiments and Technologies*, CoNEXT '13, pages 1–6, New York, NY, USA, 2013. ACM.

[37] David Naylor, Alessandro Finamore, Ilias Leontiadis, Yan Grunenberger, Marco Mellia, Maurizio Munafò, Konstantina Papagiannaki, and Peter Steenkiste. The cost of the "s" in https. In *Proceedings of ACM International Conference on Emerging Networking Experiments and Technologies*, CoNEXT '14, pages 133–140, New York, NY, USA, 2014. ACM.

[38] Eli Pariser. *The filter bubble : what the Internet is hiding from you.* Penguin Press, New York, 2011.

[39] Arnold Roosendaal. Facebook tracks and traces everyone: Like this! Research Paper 03/2011, Tilburg Law School Legal Studies, November 30, 2010.

[40] Diego Saez-Trumper, Yabing Liu, Ricardo Baeza-Yates, Balachander Krishnamurthy, and Alan Mislove. Beyond cpm and cpc: Determining the value of users on osns. In *Proceedings of ACM Conference on Online Social Networks*, COSN '14, pages 161–168, New York, NY, USA, 2014. ACM.

[41] H. Jeff Smith, Tamara Dinev, and Heng Xu. Information privacy research: An interdisciplinary review. In M. Lynne Markus and Paul Pavlou, editors, *MIS Quarterly*, volume 35, pages 989–1015, December 2011.

[42] Pang-Ning Tan and Vipin Kumar. Discovery of web robot sessions based on their navigational patterns. *Data Min. Knowl. Discov.*, 6(1):9–35, January 2002.

[43] O. Tange. Gnu parallel - the command-line power tool. *;login: The USENIX Magazine*, 36(1):42–47, Feb 2011.

[44] Anna Vapen, Niklas Carlsson, Anirban Mahanti, and Nahid Shahmehri. Third-party identity management usage on the web. In Michalis Faloutsos and Aleksandar Kuzmanovic, editors, *Passive and Active Measurement*, volume 8362 of *Lecture Notes in Computer Science*, pages 151–162. Springer International Publishing, 2014.

[45] Ingmar Weber, Venkata Rama Kiran Garimella, and Erik Borra. Mining web query logs to analyze political issues. In *Proceedings of ACM Web Science Conference*, WebSci '12, pages 330–334, New York, NY, USA, 2012. ACM.

[46] Nicholas Jing Yuan, Fuzheng Zhang, Defu Lian, Kai Zheng, Siyu Yu, and Xing Xie. We know how you live: Exploring the spectrum of urban lifestyles. In *Proceedings of ACM Conference on Online Social Networks*, COSN '13, pages 3–14, New York, NY, USA, 2013. ACM.

Appendix A

Methodology details

A.1 Domains

Table 3.1 has the details of the final domain lists in use, including full dataset size[1] [2] and selection method. Table 3.4 shows the top TLDs in the list of unique domains; while random TLD samples of course come from a single TLD, top lists are mixed. This list can be compared to the per-TLD (or technically public suffix) results in Table C.10, which shows the coverage of TLDs for external requests per dataset.

A.1.1 *.SE Health Status* domains

When .SE performs their annual *.SE Health Status* report measurements, they use an in-house curated list of domains of national interest. These domains are mostly from the .se zone and cover government, county, municipality, higher education, government-owned corporations, financial service, internet service provider (ISP), domain registrar, and media domains – see Table 3.5 for category domain counts and descriptions. Some domains overlap both within and between categories; domains have been deduplicated.

A.1.2 Random .se domains

The thesis was written in collaboration with .SE, which runs the .se TLD, and the work focusing on the state of Swedish domains. Early script development was done using a sample of 10,000 random domains, most often tested in groups of 100. A final sample of 100,000 domains was also provided. The .se TLD is to be considered Sweden-centric.

A.1.3 Random .dk domains

The Danish .dk TLD organization, DK Hostmaster A/S[3], helped out with a sample of 10,000 domains, chosen at random from the database of active domains in the zone. The .dk TLD is to be considered Denmark-centric.

[1] https://www.iis.se/domaner/statistik/tillvaxt/?chart=active

[2] https://stats.dk-hostmaster.dk/domains/total_domains/

[3] https://www.dk-hostmaster.dk/

A.1.4 Random .com, .net domains

The maintainers of the .com, .net and .name TLDs, Verisign, allow downloading of the complete zone file under an agreement. The .com zone is the largest one by far, and the .net zone is in the top 4.[4] This allows for a random selection of sites from around the world, even though usage is not geographically uniform – both in terms of registrations and actual usage.

A.1.5 Alexa Top 1,000,000 sites[5]

Alexa, owned by Amazon, is a well-known source of top sites in the world. It is used in many research papers, and can be seen as the standard dataset. Their daily 1-month average traffic rank top 1,000,000 list is freely available for download.[6] As Alexa distinguishes between a site and a domain, some domains with several popular sites are listed more than once. URL paths have been stripped and domains have been deduplicated before downloading.

A.1.6 Reach50 domains[7]

The top 50 sites in Sweden are presented by Webmie[8], who base their list on data from a user panel. The panelists have installed an extension into their browser, tracking their browsing habits by automated means. They also have results grouped by panelists categories: women, men, age 16-34, 35-54, 55+ but only the unfiltered top list is publicly available.

A.2 Public suffix list[9] [10]

In the domain name system, it is not always obvious what parts of a domain name are a public suffix and which are open for registration by Internet users. The main example is `example.co.uk`, where the *public suffix* `.co.uk` is different from the TLD `.uk`. Because HTTP cookies are based on domains names, it is important to browser vendors to be able to recognize which parts are public suffixes to be able to protect users against supercookies[11]; cookies which are scoped to a public suffix, and therefore readable across all websites under that public suffix. It should be noted that all subdomains do not have to point to servers owned by the same organization – it can be used as a way to allow tracking cookies for server-to-server tracking behind the scenes (external linkage) [20].

The same dataset is also useful for grouping domains without improperly counting `example.co.uk` as a *user-owned subdomain* of `.co.uk`, which would then render `.co.uk` as the most popular domain under the `.uk` TLD. Swedish examples include second level domains `.pp.se` for privately owned domains and `.tm.se` for trademarks[12]. These second level domains were more important before April 2003[13], when first level domain registration rules restricted registration to nation-wide companies, associations and authorities.

[4] http://www.keepalert.com/top-extension-ranking-july-2014-newgtlds

[5] http://www.alexa.com/topsites

[6] https://alexa.zendesk.com/hc/en-us/articles/200449834-Does-Alexa-have-a-list-of-its-top-ranked-websites-

[7] http://reach50.com/

[8] http://webmie.com/

[9] https://publicsuffix.org/

[10] https://en.wikipedia.org/wiki/Public_Suffix_List

[11] https://en.wikipedia.org/wiki/HTTP_cookie#Supercookie

[12] https://www.iis.se/data/barred_domains_list.txt

[13] https://en.wikipedia.org/wiki/.se#Pre_2003_system

The public suffix list (2014-07-24) contains 6,278 rules, against which domains are checked in one of the classification steps (B.2.3). It becomes the basis for the domain's division into public suffix and primary domain (first non-public suffix match), and subsequent grouping.

There is also an algorithm for wildcard rules which can have exceptions; this thesis has not implemented wildcards and exceptions in the classification step. There are 24 TLDs with wildcard public suffixes, and 8 non-TLD wildcards. Out of these 8 non-TLD wildcards, 1 is `*.sch.uk` and 7 are Japanese geographic areas. The 24 wildcards have 10 exception rules; 7 of them are Japanese cities grouped by the previously mentioned geographic areas and the remaining 3 seem to belong to ccTLD owner organizations.

Apart from ICANN domains, which have been implemented, there are also private domains considered public suffixes listed as rules. They are domains which have subdomains controlled by users/customers, for example joelpurra.github.io which is controlled by me but hosted by the code hosting service github.com. Other examples include cloud hosting/CDN services such as cloudfront.net, amazonaws.com, azurewebsites.net, fastly.net, herokuapp.com, blogs from several blogspot.TLD domains and dyndns.com's wide choice of dynamic domains. One example that looks like a technical choice in order to hinder accidental or malicious setting of cookies is googleapis.com, which is listed despite being (presumably) completely under Google's control.

A.3 Disconnect.me's blocking list

One of the most popular privacy tools is Disconnect.me, which blocks tracking sites by running as a browser plugin. Disconnect was started by ex-Google engineers, and still seems to have close ties to Google as the own domain disconnect.me is listed as a Google content domain in the blocking list.

The Disconnect software lets users block/unblock loading resources from specific third-party domains and domain categories. The dataset (2014-09-08) has a list of 2,149 domains used as the basis for the blocking. Each entry belongs to one of 980 organizations, which come with a link to their webpage. There is also a grouping into categories – see description and examples later in this chapter. Worth noting is that the content category is *not* blocked by default.

There are other open source alternatives to Disconnect's blocking list, but they use data formats that are not as easy to parse. The most popular ones also do not contain information about which organization each blocking rule belongs to. See Section 7.5.1.

A.3.1 Categories

Most domains and organizations by far are in the advertisement category. The reason the Disconnect category has so few organizations, is that it is treated as a special category (A.3.7) with only Google, Facebook and Twitter. See Table 3.2 for domain and organization count per category.

A.3.2 Domains per organization

The dataset shows 459 of the 980 organizations have more than one domain. One organization stands out, with 271 domains – Google. The biggest reason is that they own top level domains such as google.se and google.ch from over 200 TLDs. Yahoo comes in second with 71 domains, many of which are service-specific subdomains to yahoo.com, such as finance.yahoo.com and travel.yahoo.com. See Figure 3.1 for the distribution of organizations (y axis) with a certain number of associated domains (x axis), where Google is the datapoint to the far right on the x axis.

A.3.3 Organizations in more than one category

Some organizations are represented in more than one of the five Disconnect categories. Organizations represented in the content category may be blocked in part – but by serving content, they can achieve at least partial tracking. Yahoo has several ad services, several social services, several content services and a single analytics service, putting them in four categories. At least one organization, Google, is misrepresented in the categories; the special Disconnect category contains both their advertisement and analytics service domains (A.3.7). See Table 3.3 for organizations in more than one category, and which categories they are represented in.

A.3.4 Advertising

While this category has the most domains and organizations by far (see Table 3.2), many of the actors are unknown to the general public making it harder to know how information is collected and used. Several recognizable companies – such as AT&T, Deutsche Post DHL, eBay, Forbes, HP, IDG, Match.com, Monster, Opera, Salesforce.com, Telstra and Tinder – are listed with their primary domains. This suggests that they can follow their own customers across sites where their trackers are installed, without the use of more advanced techniques, such as cookie-sharing [1].

amazon-adsystem.com Amazon's ad delivery network. Several amazon.tld domains, such as `ca`, `co.uk` and `de` are also listed here – but amazon.com is not.

appnexus.com The AppNexus ad network.

imiclk.com Akamai's ad network Adroit.

overture.com Yahoo's ad network.

omniture.com Adobe's ad network.

tradedoubler.com The TradeDoubler ad network.

A.3.5 Analytics

Analytics services offer a simple way for website owners to gather data about their visitors. The service is often completely hosted on external servers, and the only connection is a javascript file loaded by the website. The script collects data, sends it back to the service and then presents aggregate numbers and graphs to the website owner.

alexa.com Amazon's web statistics service, considered an authority in web measurement. Alexa's statistics, in the form of their global top list, is also used as input for this thesis (A.1.5).

comscore.com Analytics service that also publishes statistics.

gaug.es GitHub's analytics service.

coremetrics.com Part of IBM's enterprise marketing services.

newrelic.com A suite of systems monitoring and analytics software, up to and including browsers.

nielsen.com Consumer studies.

statcounter.com Web statistics tool.

webtrends.com Digital marketing analytics and optimization across channels.

A.3.6 Content

Sites that deliver content. There is a wide variety of content, from images and videos to A/B testing, comment and help desk services. This category is not blocked by default.

apis.google.com One of Google's API domains.

brightcove.com Video hosting/monetization service.

disqus.com A third-party comment service.

flickr.com Flickr is a photo/video hosting site, owned by Yahoo.

googleapis.com One of Google's API domains, hosting third-party files/services such as Google Fonts and Google Hosted Libraries.

instagram.com Facebook's photo/video sharing site.

office.com Microsoft's Office suite online.

optimizely.com An A/B testing service.

truste.com Provides certification and tools for privacy policies in order to gain users' trust; "enabling businesses to safely collect and use customer data across web, mobile, cloud and advertising channels." This includes ways to selectively opt-out from cookies by feature level; required, functional or advertising.

tumblr.com A popular blogging platform.

uservoice.com A customer support service.

vimeo.com A video site.

www.google.com Google's main domain, which also hosts services such as search.

youtube.com One of Google's video sites.

A.3.7 Disconnect

A special category for non-content resources from Facebook, Google and Twitter. It seems to initially have been designed to block their respective like/+1/tweet buttons which seem to belong in the social category. As the category now contains many other known tracking domains from the same organizations, unblocking the social buttons also lets many other types of resources trough.

It is worth noting that this category includes google-analytics.com plus Google ad networks such as adwords.google.com, doubleclick.net and admob.com. It might have been more appropriate to have them in the analytics and advertisement categories respectively.

A.3.8 Social

Sites with an emphasis on social aspects. They often have buttons to vote for, recommend or share with others.

addthis.com A link sharing service aggregator.

digg.com News aggregator.

linkedin.com Professional social network.

reddit.com Social new and link sharing, and discussion.

A.4 Retrieving websites and resources

Websites based on lists of domains were downloaded using har-heedless (B.2.2).

A.4.1 Computer machines

Two computers were used to download web pages during development – one laptop machine and one server machine – see Table A.1 for specifications. The server is significantly more powerful than the laptop, and they downloaded a different number of web pages at a time. The final datasets were downloaded on the server.

A.4.2 Network connection

The laptop machine was connected by ethernet to the .SE office network, which is shared with employees' computers. The server machine was connected to server co-location network, which is shared with other servers. The .SE network technicians said load was kept very low, and only a few percent of the dedicated 100 Mbps per location was used. Both locations are in Stockholm city, and should therefore be well placed in regard to websites hosted in Sweden.

A.4.3 Software considerations

To expedite an automated and repeatable process, a custom set of scripts were written as the project har-heedless. The scripts are written using standard tools, available as open source and on multiple platforms.

Cookies

Cookies stored by a website may affect content upon requesting subsequent resources, and is one of the primary means of keeping track of a browser. Each browser instance has been started without any cookies, and while cookie usage has not been turned off, none have been stored after finalizing the web page request. A cookie analysis is still possible by looking at HTTP headers, but they have not been considered as an indicator of tracking as other techniques can serve the same purpose [1, 10].

Dynamic web pages

Previous efforts to download and analyze web pages by .SE used a static approach, analyzing the HTML by means of simple searches for `http://` and `https://` strings in HTML and CSS. It had proven hard to maintain, and the software project was abandoned before the thesis was started, but had not yet been replaced. In order to better handle the dynamic nature of modern web pages, the headless browser phantomjs (B.1.2) was chosen, as it would also download and execute javascript – a major component in both user interfaces as well as active trackers and ads.

Cached content

Many of the external resources will be overlapping between websites and domains, and downloading them multiple times can be avoided by caching the file the first time in a run. Keeping cached content would, depending on per-response cache settings and timeout, result in a different HTTP request and potentially different response. A file that has not changed on the server would generate an HTTP response status of 304 with no data, saving bandwidth and lowering transfer delays, where a changed file would generate a status 200 response with the latest version.

One of the techniques in determining if a locally cached file is the correct/latest version includes the HTTP `Etag` header [12], which is a string representation of a URL/file at a certain version. When content is transferred it may have an `Etag` attached; if the file is cacheable, the `Etag` is saved. Subsequent requests for the same, cached URL contain the `Etag` – and the server uses it to determine if a compact 304 response is enough or a full 200 response is necessary. It has been found that the `Etag` header can be used for cookieless cross-site tracking by using an arbitrarily chosen per-browser value instead of a file-dependent value [34]. This means that keeping a local file cache might affect how trackers respond; a file cache has not been implemented in har-heedless, making the browser amnesiac.

Flash files

Flash is a scriptable proprietary cross-platform vector based web technology owned by Adobe. Several kinds of content, including video players, games and ads, use Flash because it has historically been better suited than javascript for in-browser moving graphics and video. Flash usage has not been considered for this thesis as the technology isn not available on all popular web browsing platforms, notably Apple's iPad, and is being phased out by HTML 5 features such as <`canvas`> and <`video`> elements.

Combined javascript

A common technique for speeding up websites is to reduce the number of resources the browser needs to download, by combining or concatenating them in different ways depending on the file format. Javascript is a good example where there are potential benefits[14] since functionality

[14] https://developers.google.com/speed/docs/best-practices/rtt#CombineExternalJS

often is spread across several files, especially after the plugin style of frameworks such as jQuery[15] emerged. One concern is whether or not script concatenation on a web page would affect script analysis at a later stage, by reducing the number of third-party requests. While it is hard to analyze all scripts, based on their wide spread use, third-party scripts stay on their respective home servers as software as a service (SaaS) to enable faster update cycles and tracking of HTTP requests.

Google Tag Manager[16]

One of the concerns was Google Tag Manager (GTM), which a script aggregation service with asynchronous loading directed specifically to marketers. Google provides builtin support for their AdWords, Analytics (GA) and DoubleClick (DC) services. While simplifying management with only one `<script>` tag, each part should download separately and perform the same duties, including "calling home" to the usual addresses. In order to confirm this, a query was run on one of the datasets, se.2014-07-10.random.100000-http-www – see Table A.2.

The numbers point to that every domain that uses Google Tag Manager uses at least one of Google Analytics and DoubleClick, and will therefore not obscure information regarding which services are called when further analyzed.

`robots.txt` and `<meta name="robots" />`

Automated web spider/bot software can bring a significant load to a web server, as the spidering speed can exceed user browsing speed by far – a single web spider can potentially request thousands of pages in the span of minutes, effectively being a denial of service attack on an underpowered server. Some sites also have information considered sensitive to being available for spiders, or even for certain (kinds) of spiders such as image spiders. The choice to not serve spiders can stem from technical reasons (bandwidth, server load), to privacy (do not allow information to be indexed in search engines) and business (do not allow data to be collected and aggregated). In order to instruct web spiders not to get certain kind of material, there are two basic mechanisms – the special `robots.txt` file and the HTML header tag`<meta name="robots" />`. Both can contain instructions for certain bots not to index certain paths or pages, or not to follow further links stemming from a page.

Commercial software from search engines, information collectors and other software vendors take these explicit wishes from webmasters into consideration, but har-heedless has not. While it is automated software, it is not spidering software requesting many pages, following links to explore the site – it only accesses the front page of a domain, and resources explicitly requested by that one page. Information is not retrieved for indexing as such, as only HTTP request metadata is recorded. While some information requested to be not indexed might end up in the screenshots, they are kept in a format hard to re-parse for machines as non-public thesis data used for verification.

> Future versions of har-heedless might implement logic that checks `robots.txt` in a domain list preprocessing step, to determine wether or not the request should be made. The same list can be used to filter further resource requests made by phantomjs, perhaps considering also other domains' `robot.txt` files. The tag `<meta name="robots" />` can also be respected, perhaps in terms of not saving screenshots for `noindex` values.

[15] https://jquery.com/

[16] https://www.google.com/tagmanager/

Parallelizing downloads

A lot of the time spent downloading a web page is spent waiting for network resources, especially during timeouts during retrieval. To speed up downloading large amounts of web pages, parallelizing was employed using simple scripting techniques, starting multiple processes at once in batches, waiting for each batch to finish before starting the next. This was deemed inefficient, as the download and rendering speed of the slowest web page would be a bottle neck. In a later script versions, GNU `parallel` was used, with a setting to not start more jobs if the current CPU/system load was too high.

During initial parallelizing tests, the laptop machine was shown to be able to handle 100 domain downloads in parallel, but this was later scaled down to ensure system overload would not affect results, at the cost of significantly longer download times.

Screen size

When phantomjs is running, it emulates having a browser window by keeping an internal graphics buffer. Even if web page is not rendered and shown on screen, it still has a screen size, which affects layout. With a bit of javascript or responsive CSS, the screen size can affect downloaded resources. Javascript can be used to delay download of images and other resources that are *below the fold*, meaning outside of the initial view the user has without scrolling in any direction, as a page speed improvement. Responsive CSS adapts the page style to the screen size in order to increase usability for mobiles and other handheld devices, and might optimize the quality of downloaded images to match the screen size.

By default, phantomjs uses the physical computer's primary screen size. In order to reduce differences between the laptop and server machines, a fixed emulated window size of 1024x768 pixels was chosen. The basis for this is that 1024x768 pixel resolution screens have been the recommended screen size to design for[17] for a long time[18], and it still is a common screen size[19].

Scripted browser scrolling (vertically) through the page has not been performed, thus javascript scrolling events triggering for example downloading of images below the fold are not guaranteed.

Screenshots

In order to visually confirm that web pages have been downloaded correctly, a PNG screenshot can optionally be taken when the page has finished downloading. There is a processing cost associated with taking screenshots; it takes time for phantomjs to render the internal buffer as an image, convert it to base64 encoding for the augmented HAR data, jq to extract the image data, the system to convert the data back to binary and finally write it to disk. Extra processing is also needed to remove the screenshot from the augmented HAR file.

The emulated browser window size also sets a limit on the screenshot size when rendered, but the entire browser canvas is captured. This means that screenshots that are saved to disk in most cases extend beyond the viewport size, most often vertically; this corresponds to scrolling through the entire page.

The ratio of PNG to HAR data size on disk points to the compressed PNG files being up to 10 times the size of the uncompressed HAR files for certain types of pages, for example media domains in the *.SE Health Status* report dataset. See Table A.3.

[17] http://www.nngroup.com/articles/screen-resolution-and-page-layout/
[18] http://www.nngroup.com/articles/computer-screens-getting-bigger/
[19] http://gs.statcounter.com/#resolution-ww-monthly-201307-201407

A.4.4 Dataset variations

Each dataset has been used in four variations, effectively quadrupling the number of websites access attempts. Each of these variations have been downloaded and analyzed separately, then used for both intra- and inter-dataset comparisons.

- HTTP and HTTPS. Enables comparisons between secure and insecure origin website requests.
- Empty prefix versus www prefix/subdomain. The www subdomain has historically been used to denote a web server serving websites, and it still seems to have a higher usage than no prefix (C.2, C.8). The *.SE Health Status* reports specifically accesses websites on the www subdomain.

Each dataset in Table A.4 has the variation appended to the name in the detailed results, Chapter C.

A.4.5 System load during downloads

System load can affect the end results, if network timeouts occur during downloading and processing of domains' front pages. Apart from CPU and memory limitations, the other users of the .SE network should not be affected by tests.

System load on *NIX systems can be found using the `uptime` command.[20] Other processes were running at the time of these tests, so the numbers are not exclusive to the downloading. The complexity of the front pages of the currently processed domains affects the load, as well if screenshot generations is enabled. Final downloads were done with screenshots enabled. Table A.5 show loads for random samples in time.

> System load has been shown to vary greatly based on the type of request, mainly differing by HTTP and HTTPS response rates. High failure rates mean a lot of time is spent waiting until a set time limit/timeout has been reached. Increasing the upper bound on parallelism and instead dynamically adjusting the number of concurrent requests emphasizing system load should decrease time needed to download a set of domains with a large failure rate (C.2). Time limits can also be adjusted based on previous dataset results' actual reply timings found in HAR data, instead of setting a high and "safe" upper bound timeout, both for page timeouts and individual resources.

A.4.6 Failed domains

Some websites are not downloaded successfully, for different reasons. The DNS settings might not be correct, the server may be shut down, there might have been a temporary network timeout, there might have been a software error – or the server has been programmed to not respond to automated requests from phantomjs (B.1.2) and similar tools. Unfortunately, outside of local software errors which may result in parseable error messages, sources of the errors are hard to detect without an extensive external analysis of DNS settings and network connectivity – and even so, an automated analysis may include false negatives because of remote system automated request countermeasures.

[20] https://en.wikipedia.org/wiki/Load_(computing)

Machine	OS, version	CPU	Cores	Speed (MHz)	Memory (GB)
Laptop	Mac OS X 10.9.2 Mavericks	x86/64 bit	2	2 800	8
Server	Debian GNU/Linux 8 Jessie/Sid	x86/64 bit	4	2 500	16

Table A.1: Machine specifications

Name	Value	Percentage
Non-failed domains	77 261	100%
Domains with GTM	1 453	1.9%
GA/GTM domain coverage	1 229	85%
DC/GTM domain coverage	821	57%
GA+DC/GTM domain coverage	1 453	100%

Table A.2: Google Tag Manager versus Google Analytics and DoubleClick

Dataset	PNG files (MB)	HAR files (MB)	Ratio
2014-07-25 100k server	43 000	4 900	8.8:1
2014-08-04 health-status server	693	110	6.3:1

Table A.3: Output file size

Prefix/protocol	HTTP	HTTPS
(none)	http://	https://
www.	http://www.	https://www.

Table A.4: Dataset variations

Machine	Dataset	Concurrent downloads	Load	Load per core
Laptop	http	5	3.0	1.5
Laptop	http	10	7.0	3.5
Server	http	20	2.0	0.5
Server	https	75	0.5	0.1

Table A.5: System load

Each HTTP request has their HTTP status response recorded if it is available; absence or numbers outside the RFC7231 [13] range (100-599) indicates failure. Any error output the web page itself has produced, mostly because of javascript errors, have also been recorded in the HAR log or individual entry/request comment fields.

A distinction is made between *failed* and *unsuccessful* domains – unsuccessful domains rendered a complete response with an HTTP status that indicated that it was not successful. Domains that failed have been re-downloaded; it relieved some, but not all, failures.

The first round of retries rendered the greatest results, and subsequent retries are less successful. This seem to point to some intermittent failures being recoverable, and that some domains will not respond. Due to diminishing returns in the number of additional successful domains in each retry cycle, the number of retries was limited to two (B.2.1).

A.5 Analyzing resources

After downloading HAR files, they are processed using har-dulcify (B.2.3).

A.5.1 Screenshots

Screenshots were mainly used for verification during development, to see that the pages were loaded properly. While they have been retained, the manual inspection necessary makes it infeasible as a way to verify each and every domain's result.

A.5.2 Extracting HAR format parts

The HAR format specification includes fields that have to do with for example request/response timings and data sizes. Those, and other fields, are not analyzed in this thesis, so the first step is to extract the relevant information. These properties will be enough to see what kinds of resources are requested, if requests are successful and where the request is made to.

URL The request's URL. Recorded whether the request is successful or not. Although almost all requests are made to `http://` or `https://` a negligible amount of other and non-standard (sometimes misspelled) protocols have been registered. Urls starting with `data:` have been ignored, as they are page-internal.

Status The HTTP status code[21] found in the server's response. Defined as a 3-digit integer result, 100-599, grouped into classes by the first digit.

Mime-type The HTTP content-type header, which is the body internet media type (previously known as Multipurpose Internet Mail Extensions (MIME) type) combined with optional parameters, such as character encoding.

Referer The URL of the page that requested the resource. Can be used to build a tree of requests, but is limited by the fact that it requires HTML `<frame>` or `<iframe>` to differ from the origin page. The word referrer was misspelled `referer`[22] in the original proposal [4].

Redirect The URL of the new location of the requested resource, if defined by a 3xx (A.5.3) HTTP status.

[21] http://tools.ietf.org/html/rfc2616

[22] https://en.wiktionary.org/wiki/referer

A.5.3 Expanding parts

URL, referer, redirect

The URL format has several components [5], with interesting ones for standard web requests listed and/or split up further to get a more fine-grained analysis.

Scheme In this case, `http` and `https` protocols have been the most interesting to look at. Other interesting examples that can be found in the wild include `ftp` (for file downloads), `data` (for resources encoded into the URI) and `about` (mostly used for blank placeholder pages).

Domain For this thesis, the domain part has been of much interest, as it signifies the difference between internal and external resources.

Port While custom ports can be used, they usually implicitly default to 80 for HTTP and 443 for HTTPS.

Path The path specifies a folder or file on the server.

Querystring Most parameters sent back to servers are defined in an RFC compliant way, but there are other variants building on for example ‘/‘ as a pseudo-path separator.

Fragment The fragment is in the web context a client-only component, and is not to be sent back to the server as part of a request. The usage affects browsers' presentation, historically only by scrolling to a matching named element, but modern usage includes keeping browser state using javascript, for example following the web spider crawlable hash-bang syntax[23].

Status

The status value groups are defined by their first digit [13]. Groups outside of the defined range 100-599 are defined as null.

1xx	Informational
2xx	Successful
3xx	Redirection
4xx	Client error
5xx	Server error

Mime-type

The mime-type is grouped by their usage, which usually is the first group part. Table A.6 shows a selection of common types grouped together.

[23] https://developers.google.com/webmasters/ajax-crawling/docs/getting-started

A.5.4 Classification

Public Suffix List

The public suffix list is prepared for lookups per domain component. Each request's domain (including shorter domain components) is checked against it, and any matching public suffixes are kept in an array. A list of private suffixes, as in domain components not in the public suffix list, is also kept. The primary domain (first non-public suffix match, or the shortest private suffix) is extracted.

> In terms of grouping, it might be good to keep the primary (longest matching) public suffix separately. The public suffix list also contains special wildcard/exception rules and private suffixes (A.2). They have not been used in the thesis (7.5.5).

Basic

Simple properties in the request are checked, and their valued saved as a classification property. This property is used for grouping and for further analysis.

Successful request The status of the request is 200-299 or 304.

Unsuccessful request The status is 100-199 or 300-303 or 305-599.

Failed request The log file format is incomplete or the status is null, below 100 or above 599.

Same domain The request is to the same domain that was first visited.

Subdomain The request is to a subdomain of the domain that was first visited.

Superdomain The request is to a domain to which the origin domain is a subdomain. This basic superdomain classification is currently not checked against the public suffix list for invalid superdomains.

Same primary domain The request shared the same primary domain with the origin.

Internal domain The request is to the same domain, a subdomain or a superdomain of the domain that was first visited.

External domain The request is not to an internal domain.

Secure request The request is using HTTPS.

Insecure request The request is not using HTTPS.

Disconnect.me

The URLs in the extended data contains lists of domain components. As the disconnect list of blocked domains is prepared for lookups by domains, each of the matching domains (including shorter domain components) are extracted with organization, organization URL and domain category.

A.5.5 Analysis

An analysis, where request classifications are counted, summed and the coverage calculated, is performed as an automated step.

Origin

The origin domains are grouped separately from the requests that stemmed from them.

Requested URL counts

All requests are represented with their domain, and other classifications.

Distinct requested URLs

Requested URL counts can skew aggregate results if a single domains makes an excessive amount of requests to a certain URL/domain/tracker or in a certain classification. Reducing counts to boolean values, indicating at least one request matched the classification, gives the possibility to calculate coverage per domain later on.

Request/domain counts

All the numbers from all domains added together.

Request/domain coverage

The summed up counts divided by either the total request count (for requested URLs) or the number of domains in the current group (for distinct requested URLs). This gives a coverage percentage – either for the percentage of the number of requests, or domains that has the value.

Grouping

In order to not make too broad assumptions, some grouping was performed. The analysis was performed the same way on each of these groups (B.2.3). The origin domain's download status was checked, and grouped into both unfiltered and non-failed groups. The list of requested URLs was grouped into unfiltered, internal and external URLs.

A.5.6 Questions

Where the aggregate analysis is not enough, there are custom questions. These questions/queries can be executed against any previous intermediate step in the process, as they are saved to disk.

Google Tag Manager

One of the questions posed beforehand was if Google Tag Manager would have an impact upon results; it has been answered with the help of this data (A.4.3).

Origins with redirects

Looking at preliminary results, a large portion of domains yielded a redirect as the initial response. In order to look at these redirects specifically, and determine if they redirect to an internal or external domain, a specific question was written.

A.5.7 Multiset queries

After downloading several datasets, it is often interesting to compare them side by side. The multiset queries extract pieces of data from several datasets, and combine them into a single file.

Group	Example types
script	application/javascript, application/x-javascript, text/javascript
font	application/font-woff, application/x-woff, application/x-font-ttf, font/ttf, font/opentype
data	application/json, application/octet-stream, binary/octet-stream, application/xml
image	image/gif, image/jpeg, image/png, image/svg+xml
style	text/css
html	text/html, application/xhtml+xml
text	text/plain
document	application/pdf
object	application/x-shockwave-flash

Table A.6: Mime-type grouping

Appendix B

Software

Development was performed in the Mac OS X operating system while the server machine performing most of the downloading and analysis was running the Linux distribution Debian (A.4.1). The software is thought to be runnable on other Unix-like platforms with relative ease.

B.1 Third-party tools

In order to download and analyze tens of thousands of webpages in an automated fashion, a set of suitable tools were sought. Tools released as free and open source software have been preferred, found and selected; code written specifically for the thesis has also been released as such.

B.1.1 HTTP Archive (HAR) format[1]

In an effort to record and analyze network traffic as seen by individual browsers, the data/file format HTTP Archive (HAR) was developed. Browsers such as Google Chrome implement it as a complement to the network graph shown in the Developer Console, from where a HAR file can be exported. While constructed to analyze for example web performance, it also contains data suitable for this thesis: requested URLs and HTTP request/response headers such as referrer and content type. HAR files are based upon the JSON[2] standard [6], which is a Javascript object compatible data format commonly used to communicate dynamic data between client side scripts in browsers and web servers. The most recent specification at the time of writing was HAR 1.2.

B.1.2 phantomjs[3]

Accessing webpages is normally done by users in a graphical browser; the browser downloads then displays images, executes scripts, and plays videos. A browser is user friendly but not optimal for batch usage due to the overhead in constantly drawing results on screen and the lack of automation without external tools such as Selenium Webdriver[4]. While Webdriver can be used to control several kinds of browsers, such as Microsoft Internet Explorer, Mozilla Firefox, Google Chrome, they are not suitable for usage on a rack server that was not set up with "normal" browser usage in mind – that is with desktop software functionalities. A good alternative for

[1] http://www.softwareishard.com/blog/har-12-spec/

[2] http://json.org/

[3] http://phantomjs.org/

[4] http://docs.seleniumhq.org/projects/webdriver/

such servers is phantomjs, which is built as a command line tool without any graphical user interface. It acts like a browser internally, including rendering the webpage to a image buffer that is not displayed, and is controllable through the use of scripts. One such example script included in the default installation generates HAR files from a webpage visit. phantomjs has been implemented on top of the web page renderer Webkit library[5], also used in Apple Safari, Opera and previously Google Chrome.

> There are alternatives to phantomjs, but they have not been tested within the scope of the thesis. Future versions could try alternative automated browsers, such as SlimerJS[a] or both with CasperJS[b], to verify phantomjs' results.
>
> [a] http://slimerjs.org/
> [b] http://casperjs.org/

B.1.3 jq[6]

While there are command line tools to transform data in for example plain text, CSV and XML files, tools to work with JSON files are not as prevalent. One such tool gaining momentum is jq, which is implemented with a domain specific language (DSL) suitable for extracting or transforming data. The DSL is based around a set of filters, similar to pipes in the Unix world, transforming the input and passing it on to the next stage. jq performs well with large datasets, as it treats data as a stream where each top-level object is treated separately.

> At the time of writing, jq is released as version 1.4. Support for regular expressions is planned for version 1.5, which has been in the making for the duration of the thesis. As the thesis code is run on multiple machines/systems and expected to deliver the same results, using standardized packages has been a part of ensuring that.

B.1.4 GNU `parallel`[7]

To parallelize task execution, GNU `parallel` [43] has been used. It allows an input file to be distributed among several processes/CPU cores, and the results to be combined into a single file. It helps speed up downloading websites to create HAR files, and processing of the JSON data through jq, which is single threaded.

B.2 Code

In order to efficiently and repeatably download and analyze web pages, special tools have been written. Most of the code is written in bash[8] scripts utilizing external commands when possible, such as jq. The code for the jq commands has been embedded in the bash scripts.

The source code for the respective projects have been released to the public under GNU General Public License version 3.0 (GPL-3.0)[9], so other projects can make use of them as well.

[5] https://www.webkit.org/
[6] https://stedolan.github.io/jq/
[7] https://www.gnu.org/software/parallel/
[8] https://www.gnu.org/software/bash/
[9] https://www.gnu.org/licenses/gpl.html

B.2.1 har-portent[10]

A set of scripts that both downloads, retries failed downloads and analyzes websites in a single run – see har-heedless (B.2.2) and har-dulcify (B.2.3).

`domains/download-and-analyze-https-www-combos.sh` <`parallelism`> <`domainlists`>

Uses `domains/download-and-analyze.sh` to download four variations (A.4.4) of the same domains, so any differences between secure/insecure and www-prefixed domains can be observed. This is the true fire-and-forget script you want to run to download and analyze multiple large sets.

- `http://`
- `http://www.`
- `https://`
- `https://www.`

`domains/download-and-analyze.sh` <`prefix`> <`parallelism`> <`domainlists`>

Downloads a list of domains in parallel, with automatic per-prefix/input file folder and log file and creation. It also performs automatic retries for failed domains two times per dataset, with an increased parallelism as retries are mostly expected to yield another network timeout or error.

B.2.2 har-heedless[11]

Scriptable batch downloading of webpages to generate HTTP Archive (HAR) files, using phantomjs. With a simple text file with one domain name per line as input, har-heedless downloads all of their front pages. Downloads can be made either in serial or in parallel. The resulting HAR data is written in a folder structure per domain, with a timestamp in the file name. The script that extracts HAR data, `netsniff.js`, is based on example code shipped with phantomjs, but modified to be more stable.

`get/netsniff.js`

A modified version of `netsniff.js`[12] from the phantomjs project.

Some fixes include:

- Stable, logged output when resources failed to download.
- Adding error messages as HAR comments.
- Adding a base64 encoded page screenshot as an extended field.
- Waiting a period of time after downloading the web page before generating output, to let asynchronous downloading, processing and rendering finish.

[10] https://github.com/joelpurra/har-portent

[11] https://github.com/joelpurra/har-heedless

[12] https://github.com/ariya/phantomjs/blob/master/examples/netsniff.js

These patches have not yet been submitted to the phantomjs project. They should be split up into separate parts for the convenience of the project maintainers. A complete refactoring is also an alternative, but it might be less likely to be accepted.

`get/har.sh <url>`

Contains logic to run `netsniff.js` through phantomjs. If phantomjs crashed or otherwise encountered an error, a fallback HAR file is generated with a dummy response explaining that an error occurred.

`url/single.sh <domain> [--screenshot <true|false>]`

Downloads a URL of a single domain, and takes care of writing the HAR output to the correct folder and file. If a screenshot has been requested, it is extracted (and removed) from the extended HAR data and written to a separate file parallel to the resulting HAR file.

`url/parallel.sh [parallel-processes [--screenshot <true|false>]]`

Uses GNU `parallel` to download multiple webpages at a time. The number of separate processes running is adjusted per machine, depending on capacity, with `parallel-processes`.

`domain/parallel.sh <prefix> [parallel-processes [--screenshot <true|false>]]`

Download the front pages of a list of domains, in parallel, using a specific prefix, such as `https://www.`. See `url/parallel.sh`.

B.2.3 har-dulcify[13]

Extracts data from HTTP Archive (HAR) files for an aggregate analysis. HAR files by themselves contain too much data, so the relevant parts need to be extracted. The extracted parts are then broken down into smaller parts that are easier to group and analyze, and added to the data alongside the original. With the expanded data in place, requests are classified by basic measures and matched against external datasets. Scripts are written to perform only a limited task and instead be chained together by piping the data between them. As the scripts generally connect in only one way, the convenience scripts in the `one-shot/` folder are used the most. These convenience scripts also leave the files from partial executions, so they can be used for other kinds of analysis.

At the time of writing, there are 67 scripts in har-dulcify. Here is a selection with explanations.

`one-shot/all.sh [har-folder-path]`

Runs `preparations.sh`, `data.sh`, `aggregate.sh` and `questions.sh` based on data in the `har-folder-path` (defaulting to the current folder) outputting the results to the current folder.

`one-shot/preparations.sh`

Downloads, prepares and analyses third-party datasets, and puts them in the current folder for use by subsequent scripts.

[13] https://github.com/joelpurra/har-dulcify

`one-shot/data.sh [har-folder-path]`

Processes all HAR files in the `har-folder-path` (defaulting to the current folder), and puts the output in the same folder.

`one-shot/aggregate.sh`

Prepares data for aggregation by counting occurrences in each domain's data, then adds them together to a single file containing an aggregate analysis.

`domains/latest/all.sh`

Finds and lists the paths to the most recent HAR files, per domain.

`extract/request/parts.sh`

Extracts url, status, content-type (mime-type) and other individual pieces of data from the originating domain front page request and the subsequent requests.

`extract/request/expand-parts.sh`

Keeps the original data, but also expands the url and mime-type into their respective parts, and adds simple grouping to status and mime-type.

`classification/basic.sh`

Add basic classifications, such as if a request is internal or external to the originating domain, and if the request is secured with HTTPS.

`classification/disconnect/prepare-service-list.sh`

Prepares Disconnect's blocking list from the original format where blocked domains are stored deep in the structure, to one where domains are top level map keys, prepared for fast lookups.

`classification/disconnect/add.sh` <`prepared-disconnect-dataset-path`>

Matches each requested domain against Disconnect's list of domains to block, and adds the results to the output. Disconnect's original `service.json`[14] (or `disconnect-plaintext.json`[15]) needs to be prepared through `classification/disconnect/prepare-service-list.sh` before being used.

`classification/disconnect/analysis.sh`

Analyses Disconnect's blocking list and collects some aggregate numbers.

`classification/effective-tld/add.sh` <`prepared-disconnect-dataset-path`>

Matches each requested domain against Mozilla's Public Suffix list of effective top level domain names, and adds the results to the output. The original `effective_tld_names.dat`[16] needs to be prepared through `classification/effective-tld/prepare-list.sh` before being used.

[14] https://github.com/disconnectme/disconnect/raw/master/firefox/content/disconnect.safariextension/opera/chrome/data/servic

[15] https://services.disconnect.me/disconnect-plaintext.json

[16] https://publicsuffix.org/list/effective_tld_names.dat

`aggregate/prepare*.sh`

Running the aggregation before analysis is currently not possible in a single step, as jq requires all data for the reduce step to be in memory. The solution is to first map the data to a suitable format, and then reduce them in chunks repeatedly.

`aggregate/analysis.sh`

Takes counts and lists of values, and reduces them to easy to present values, percentages and top lists. Results are also grouped in order to draw separate conclusions regarding non-failed domains and internal/external resources. Below, a tree representation of the output after grouping is shown. The origin represents the original request of the subsequently requested URLs.

- Unfiltered origin domains
  - Origin
    - Counts
    - Coverage
  - Unfiltered URLs
    - Requested URLs
      - Counts
      - Coverage
    - Distinct requested URLs
      - Counts
      - Coverage
  - Internal URLs
    - Requested URLs
      - Counts
      - Coverage
    - Distinct requested URLs
      - Counts
      - Coverage
  - External URLs
    - Requested URLs
      - Counts
      - Coverage
    - Distinct requested URLs
      - Counts
      - Coverage
- Non-failed origin domains
  - Origin
    - Counts
    - Coverage

- Unfiltered URLs
    - Requested URLs
        - Counts
        - Coverage
    - Distinct requested URLs
        - Counts
        - Coverage
- Internal URLs
    - Requested URLs
        - Counts
        - Coverage
    - Distinct requested URLs
        - Counts
        - Coverage
- External URLs
    - Requested URLs
        - Counts
        - Coverage
    - Distinct requested URLs
        - Counts
        - Coverage

`questions/google-gtm-ga-dc.sh`

Analyze the impact of Google Tag Manager on coverage for other Google services, specifically Google Analytics and DoubleClick.

`questions/origin-redirects.sh`

Analyze requests to see if there are redirects from the origin page initially requested. One of the most interesting things to look at is wether or not domains redirect to or from secure https domains.

`questions/ratio-buckets.sh`

Collects occurrences of arbitrary things and ratios of other things per domain, puts them into 100+ counter buckets and calculates normalized and cumulative versions. Used to get the number of Disconnect organizations and ratios of internal and secure resources.

`multiset/*.sh`

A set of scripts that perform tasks on multiple datasets at once – an aggregate of aggregates usually. The scripts were developed late in the process, to extract small pieces of data per dataset for the report. The small pieces of data are combined to files with tab-separated values, which are the source of most data tables and figures in the report.

Appendix C

Detailed results

C.1 Differences between datasets

Domain lists chosen for this thesis come in three major categories – top lists, curated lists and random selection from zone files. While the top lists and curated lists are assumed to primarily contain sites with staff or enthusiasts to take care of them and make sure they are available and functioning, the domain lists randomly extracted from TLD zones might not. Results below seem to fall into groups of non-random and randomly selected domains – and result discussions often group them as such.

C.2 Failed versus non-failed

HAR data that does not have a parseable HTTP status outcome number (shown as `(null)` in C.3) is considered a failed request. In order to reduce temporary or intermittent problems, all domains that failed were retried up to two times (B.2.1).

Figure C.1 visualizes the percentage of requests in each HTTP status category and `null` for no response on the x axis. Non-random domains have a failure rate of below 15% for HTTP, and 70-90% for HTTPS, meaning that 30-10% implement HTTPS. Random zone domains have a failure rate of above 20% for HTTP and above 99% for HTTPS.

The very low HTTPS adoption rates among random sites is both surprising and not surprising – while larger sites might have felt the pressure to implement them, a non-professional site owner might see it as both an unnecessary technical challenge and an unnecessary additional cost. Most X.509 public key infrastructure (PKI) [8] certificates cost money to buy and install. With public IPv4 addresses running out and legacy browsers requiring one IP-address per HTTPS certificate, it can also lead to an additional fee for renting an exclusive IP-address. The random zone domains responding to HTTPS requests exhibit behavior more similar to top and curated sites – see how dataset variation lines follow each other in Figure C.3 and Figure C.8 – suggesting that a similar kind of effort has been put in developing them.

Dataset	Domains	Successful	Unsuccessful	Non-failed	Failed	Non-failure rate	Failure rate
alexa.2014-09-01.random.10000-http	9 777	4 186	4 030	8 216	1 561	0.840	0.160
alexa.2014-09-01.random.10000-http-www	9 779	5 922	2 571	8 493	1 286	0.868	0.132
alexa.2014-09-01.random.10000-https	9 952	418	717	1 135	8 817	0.114	0.886
alexa.2014-09-01.random.10000-https-www	9 951	754	470	1 224	8 727	0.123	0.877
alexa.2014-09-01.top.10000-http	9 750	2 468	6 077	8 545	1 205	0.876	0.124
alexa.2014-09-01.top.10000-http-www	9 759	5 980	2 702	8 682	1 077	0.890	0.110
alexa.2014-09-01.top.10000-https	9 971	651	1 856	2 507	7 464	0.251	0.749

Dataset	Domains	Successful	Unsuccessful	Non-failed	Failed	Non-failure rate	Failure rate
alexa.2014-09-01.top.10000-https-www	9 921	1 421	1 536	2 957	6 964	0.298	0.702
alexa.2014-09-01.top.dk.10000-http	2 584	807	1 456	2 263	321	0.876	0.124
alexa.2014-09-01.top.dk.10000-http-www	2 577	1 477	833	2 310	267	0.896	0.104
alexa.2014-09-01.top.dk.10000-https	2 637	114	225	339	2 298	0.129	0.871
alexa.2014-09-01.top.dk.10000-https-www	2 629	209	232	441	2 188	0.168	0.832
alexa.2014-09-01.top.se.10000-http	3 269	1 059	1 738	2 797	472	0.856	0.144
alexa.2014-09-01.top.se.10000-http-www	3 281	2 032	863	2 895	386	0.882	0.118
alexa.2014-09-01.top.se.10000-https	3 362	156	282	438	2 924	0.130	0.870
alexa.2014-09-01.top.se.10000-https-www	3 358	382	268	650	2 708	0.194	0.806
com.2014-08-29.random.10000-http	9 969	5 356	2 419	7 775	2 194	0.780	0.220
com.2014-08-29.random.10000-http-www	9 965	5 843	1 968	7 811	2 154	0.784	0.216
com.2014-08-29.random.10000-https	10 000	26	24	50	9 950	0.005	0.995
com.2014-08-29.random.10000-https-www	10 000	32	23	55	9 945	0.006	0.995
dk.2014-07-23.random.10000-http	9 973	4 952	2 228	7 180	2 793	0.720	0.280
dk.2014-07-23.random.10000-http-www	9 967	5 353	2 025	7 378	2 589	0.740	0.260
dk.2014-07-23.random.10000-https	10 000	13	10	23	9 977	0.002	0.998
dk.2014-07-23.random.10000-https-www	10 000	17	15	32	9 968	0.003	0.997
net.2014-08-29.random.10000-http	9 975	5 162	2 108	7 270	2 705	0.729	0.271
net.2014-08-29.random.10000-http-www	9 971	5 504	1 874	7 378	2 593	0.740	0.260
net.2014-08-29.random.10000-https	10 000	19	7	26	9 974	0.003	0.997
net.2014-08-29.random.10000-https-www	10 000	20	8	28	9 972	0.003	0.997
reach50.2014w35.se-http	47	6	37	43	4	0.915	0.085
reach50.2014w35.se-http-www	46	30	12	42	4	0.913	0.087
reach50.2014w35.se-https	48	5	13	18	30	0.375	0.625
reach50.2014w35.se-https-www	49	9	17	26	23	0.531	0.469
se.2014-07-10.random.100000-http	99 497	52 424	21 181	73 605	25 892	0.740	0.260
se.2014-07-10.random.100000-http-www	99 428	58 496	18 765	77 261	22 167	0.777	0.223
se.2014-07-10.random.100000-https	100 000	157	125	282	99 718	0.003	0.997
se.2014-07-10.random.100000-https-www	99 999	213	115	328	99 671	0.003	0.997
se.healthstatus.2013.counties-http	21	14	4	18	3	0.857	0.143
se.healthstatus.2013.counties-http-www	21	18	3	21	0	1.000	0.000
se.healthstatus.2013.counties-https	21	1	2	3	18	0.143	0.857
se.healthstatus.2013.counties-https-www	21	3	3	6	15	0.286	0.714
se.healthstatus.2013.domain-registrars-http	145	37	90	127	18	0.876	0.124
se.healthstatus.2013.domain-registrars-http-www	146	79	55	134	12	0.918	0.082
se.healthstatus.2013.domain-registrars-https	146	12	28	40	106	0.274	0.726
se.healthstatus.2013.domain-registrars-https-www	146	28	14	42	104	0.288	0.712
se.healthstatus.2013.financial-services-http	79	14	53	67	12	0.848	0.152
se.healthstatus.2013.financial-services-http-www	79	36	36	72	7	0.911	0.089
se.healthstatus.2013.financial-services-https	78	7	9	16	62	0.205	0.795
se.healthstatus.2013.financial-services-https-www	79	21	10	31	48	0.392	0.608

Dataset	Domains	Successful	Unsuccessful	Non-failed	Failed	Non-failure rate	Failure rate
se.healthstatus.2013.gocs-http	59	14	35	49	10	0.831	0.169
se.healthstatus.2013.gocs-http-www	60	43	14	57	3	0.950	0.050
se.healthstatus.2013.gocs-https	60	1	3	4	56	0.067	0.933
se.healthstatus.2013.gocs-https-www	60	7	2	9	51	0.150	0.850
se.healthstatus.2013.higher-education-http	48	14	26	40	8	0.833	0.167
se.healthstatus.2013.higher-education-http-www	48	37	10	47	1	0.979	0.021
se.healthstatus.2013.higher-education-https	49	2	7	9	40	0.184	0.816
se.healthstatus.2013.higher-education-https-www	49	16	8	24	25	0.490	0.510
se.healthstatus.2013.isps-http	20	4	14	18	2	0.900	0.100
se.healthstatus.2013.isps-http-www	20	12	7	19	1	0.950	0.050
se.healthstatus.2013.isps-https	20	2	4	6	14	0.300	0.700
se.healthstatus.2013.isps-https-www	20	8	2	10	10	0.500	0.500
se.healthstatus.2013.media-http	30	4	22	26	4	0.867	0.133
se.healthstatus.2013.media-http-www	32	18	10	28	4	0.875	0.125
se.healthstatus.2013.media-https	32	2	2	4	28	0.125	0.875
se.healthstatus.2013.media-https-www	32	2	3	5	27	0.156	0.844
se.healthstatus.2013.municipalities-http	286	176	73	249	37	0.871	0.129
se.healthstatus.2013.municipalities-http-www	288	245	26	271	17	0.941	0.059
se.healthstatus.2013.municipalities-https	290	21	23	44	246	0.152	0.848
se.healthstatus.2013.municipalities-https-www	290	34	20	54	236	0.186	0.814
se.healthstatus.2013.public-authorities-http	213	86	84	170	43	0.798	0.202
se.healthstatus.2013.public-authorities-http-www	214	149	54	203	11	0.949	0.051
se.healthstatus.2013.public-authorities-https	214	9	9	18	196	0.084	0.916
se.healthstatus.2013.public-authorities-https-www	214	25	12	37	177	0.173	0.827

Table C.1: Dataset HAR failure rates

During analysis har-dulcify splits results into unfiltered and filtered, non-failed origin domains. Unless otherwise mentioned, further results are presented based only on non-failed domains in each dataset, as failed origin requests add nothing to the further resource analysis.

There might be an overlap between failed HTTP/HTTPS and non-www/www dataset variations, but at the moment they are treated separately. If domains that failed to respond to HTTP requests are removed from the HTTPS set in a filter step, HTTPS failure statistics might be lower – and the same goes for www variations.

The difference between the number of domains in the dataset and the number of analyzed domains – for example 100,000 domains in “se.2014-07-10.random.100000-http-www” but 99,428 HAR files – is due to software crashes. A higher domain response rate means a higher risk of a crash; larger datasets have crashes in 0.25-2.5% of domains. Most crashes were observed to be caused by malformed image data triggering an uncaught software exception in the JPEG image decoder used by phantomjs. Looking at the resulting HAR files means that both local software errors and remote errors are considered failures. The analysis can be improved with additional testing and improvements to phantomjs; missing HAR files have been ignored in further analyses.

C.3 HTTP status codes

Choosing the term non-failed instead of successful when it comes to dividing and focusing result discussions has its basis in the HTTP standard, which defines a status code. Successful requests are generally shown with an HTTP status code of 200 (actually the entire 2xx group), or a 304 which means that a previously cached (presumably successful) result is still valid. Many sites respond with a 3xx status, which is not exactly successful as it does not contain actual content, but can not be considered a failure as it will most likely lead to another resource that is successful. While a status response of 4xx or 5xx shows there is a problem of some kind, for the purpose of this thesis a response that contains any HTTP status number is still considered a non-failure, as the remote system has responded with a proper HTTP response parseable by phantomjs and har-heedless.

The table below details the percentage of domains in each response code group, especially for 3xx (redirect) responses, as well as `null` for no response. Figure C.1 also visualize the results as percentages (x axis) per dataset.

The majority of domains domains are responding to HTTP requests during domain scanning, but many do not either by configuration or chance. While most non-failed domains produce 2xx responses, there is a high ratio of 3xx redirect class responses. With some 20% of random domains and around 50% of top sites redirecting their visitors, it begs further research. Many of them are 301, considered permanent, redirects; they indicate that the information supposedly found at the domain is actually found somewhere else. The difference between http and http-www datasets seems to suggest that redirects lead to the other (above 90%), or a secure variant; this thesis has looked at redirects within the same domain and to HTTPS variations (C.8).

Dataset	Domains	1xx	2xx	3xx	301	302	303	307	4xx	5xx	(null)
alexa.2014-09-01.random.10000-http	9 777	0.000	0.428	0.412	0.324	0.085	0.002	0.000	0.000	0.000	0.160
alexa.2014-09-01.random.10000-http-www	9 779	0.000	0.606	0.263	0.190	0.069	0.003	0.000	0.000	0.000	0.132
alexa.2014-09-01.random.10000-https	9 952	0.000	0.042	0.072	0.044	0.028	0.000	0.000	0.000	0.000	0.886
alexa.2014-09-01.random.10000-https-www	9 951	0.000	0.076	0.047	0.024	0.023	0.000	0.000	0.000	0.000	0.877
alexa.2014-09-01.top.10000-http	9 750	0.000	0.253	0.623	0.511	0.111	0.000	0.001	0.000	0.000	0.124
alexa.2014-09-01.top.10000-http-www	9 759	0.000	0.613	0.277	0.187	0.088	0.001	0.001	0.000	0.000	0.110
alexa.2014-09-01.top.10000-https	9 971	0.000	0.065	0.186	0.139	0.047	0.000	0.000	0.000	0.000	0.749
alexa.2014-09-01.top.10000-https-www	9 921	0.000	0.143	0.155	0.088	0.066	0.000	0.000	0.000	0.000	0.702
alexa.2014-09-01.top.dk.10000-http	2 584	0.000	0.312	0.563	0.461	0.100	0.002	0.001	0.000	0.000	0.124
alexa.2014-09-01.top.dk.10000-http-www	2 577	0.000	0.573	0.323	0.239	0.081	0.002	0.000	0.000	0.000	0.104
alexa.2014-09-01.top.dk.10000-https	2 637	0.000	0.043	0.085	0.060	0.025	0.000	0.000	0.000	0.000	0.871
alexa.2014-09-01.top.dk.10000-https-www	2 629	0.000	0.079	0.088	0.043	0.045	0.000	0.000	0.000	0.000	0.832
alexa.2014-09-01.top.se.10000-http	3 269	0.000	0.324	0.532	0.433	0.095	0.002	0.001	0.000	0.000	0.144
alexa.2014-09-01.top.se.10000-http-www	3 281	0.000	0.619	0.263	0.187	0.071	0.003	0.001	0.000	0.000	0.118
alexa.2014-09-01.top.se.10000-https	3 362	0.000	0.046	0.084	0.059	0.025	0.000	0.000	0.000	0.000	0.870
alexa.2014-09-01.top.se.10000-https-www	3 358	0.000	0.114	0.080	0.037	0.043	0.000	0.000	0.000	0.000	0.806
com.2014-08-29.random.10000-http	9 969	0.000	0.537	0.243	0.144	0.098	0.001	0.000	0.000	0.000	0.220
com.2014-08-29.random.10000-http-www	9 965	0.000	0.586	0.197	0.109	0.088	0.001	0.000	0.000	0.000	0.216
com.2014-08-29.random.10000-https	10 000	0.000	0.003	0.002	0.001	0.001	0.000	0.000	0.000	0.000	0.995
com.2014-08-29.random.10000-https-www	10 000	0.000	0.003	0.002	0.001	0.001	0.000	0.000	0.000	0.000	0.995
dk.2014-07-23.random.10000-http	9 973	0.000	0.497	0.223	0.160	0.062	0.002	0.000	0.000	0.000	0.280
dk.2014-07-23.random.10000-http-www	9 967	0.000	0.537	0.203	0.141	0.061	0.001	0.000	0.000	0.000	0.260
dk.2014-07-23.random.10000-https	10 000	0.000	0.001	0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.998
dk.2014-07-23.random.10000-https-www	10 000	0.000	0.002	0.002	0.000	0.001	0.000	0.000	0.000	0.000	0.997
net.2014-08-29.random.10000-http	9 975	0.000	0.517	0.211	0.125	0.085	0.001	0.000	0.000	0.000	0.271

Dataset	Domains	1xx	2xx	3xx	301	302	303	307	4xx	5xx	(null)
net.2014-08-29.random.10000-http-www	9 971	0.000	0.552	0.188	0.104	0.083	0.001	0.000	0.000	0.000	0.260
net.2014-08-29.random.10000-https	10 000	0.000	0.002	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.997
net.2014-08-29.random.10000-https-www	10 000	0.000	0.002	0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.997
reach50.2014w35.se-http	47	0.000	0.128	0.787	0.702	0.085	0.000	0.000	0.000	0.000	0.085
reach50.2014w35.se-http-www	46	0.000	0.652	0.261	0.196	0.065	0.000	0.000	0.000	0.000	0.087
reach50.2014w35.se-https	48	0.000	0.104	0.271	0.167	0.104	0.000	0.000	0.000	0.000	0.625
reach50.2014w35.se-https-www	49	0.000	0.184	0.347	0.143	0.204	0.000	0.000	0.000	0.000	0.469
se.2014-07-10.random.100000-http	99 497	0.000	0.527	0.213	0.152	0.060	0.001	0.000	0.000	0.000	0.260
se.2014-07-10.random.100000-http-www	99 428	0.000	0.588	0.189	0.130	0.058	0.001	0.000	0.000	0.000	0.223
se.2014-07-10.random.100000-https	100 000	0.000	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.997
se.2014-07-10.random.100000-https-www	99 999	0.000	0.002	0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.997
se.healthstatus.2013.counties-http	21	0.000	0.667	0.190	0.095	0.095	0.000	0.000	0.000	0.000	0.143
se.healthstatus.2013.counties-http-www	21	0.000	0.857	0.143	0.095	0.048	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.counties-https	21	0.000	0.048	0.095	0.000	0.095	0.000	0.000	0.000	0.000	0.857
se.healthstatus.2013.counties-https-www	21	0.000	0.143	0.143	0.000	0.143	0.000	0.000	0.000	0.000	0.714
se.healthstatus.2013.domain-registrars-http	145	0.000	0.255	0.621	0.462	0.159	0.000	0.000	0.000	0.000	0.124
se.healthstatus.2013.domain-registrars-http-www	146	0.000	0.541	0.377	0.253	0.123	0.000	0.000	0.000	0.000	0.082
se.healthstatus.2013.domain-registrars-https	146	0.000	0.082	0.192	0.144	0.048	0.000	0.000	0.000	0.000	0.726
se.healthstatus.2013.domain-registrars-https-www	146	0.000	0.192	0.096	0.055	0.041	0.000	0.000	0.000	0.000	0.712
se.healthstatus.2013.financial-services-http	79	0.000	0.177	0.671	0.532	0.139	0.000	0.000	0.000	0.000	0.152
se.healthstatus.2013.financial-services-http-www	79	0.000	0.456	0.456	0.278	0.152	0.025	0.000	0.000	0.000	0.089
se.healthstatus.2013.financial-services-https	78	0.000	0.090	0.115	0.064	0.051	0.000	0.000	0.000	0.000	0.795
se.healthstatus.2013.financial-services-https-www	79	0.000	0.266	0.127	0.051	0.076	0.000	0.000	0.000	0.000	0.608
se.healthstatus.2013.gocs-http	59	0.000	0.237	0.593	0.407	0.186	0.000	0.000	0.000	0.000	0.169
se.healthstatus.2013.gocs-http-www	60	0.000	0.717	0.233	0.133	0.100	0.000	0.000	0.000	0.000	0.050
se.healthstatus.2013.gocs-https	60	0.000	0.017	0.050	0.033	0.017	0.000	0.000	0.000	0.000	0.933
se.healthstatus.2013.gocs-https-www	60	0.000	0.117	0.033	0.017	0.017	0.000	0.000	0.000	0.000	0.850
se.healthstatus.2013.higher-education-http	48	0.000	0.292	0.542	0.438	0.104	0.000	0.000	0.000	0.000	0.167
se.healthstatus.2013.higher-education-http-www	48	0.000	0.771	0.208	0.146	0.063	0.000	0.000	0.000	0.000	0.021
se.healthstatus.2013.higher-education-https	49	0.000	0.041	0.143	0.102	0.041	0.000	0.000	0.000	0.000	0.816
se.healthstatus.2013.higher-education-https-www	49	0.000	0.327	0.163	0.041	0.122	0.000	0.000	0.000	0.000	0.510
se.healthstatus.2013.isps-http	20	0.000	0.200	0.700	0.500	0.200	0.000	0.000	0.000	0.000	0.100
se.healthstatus.2013.isps-http-www	20	0.000	0.600	0.350	0.150	0.200	0.000	0.000	0.000	0.000	0.050
se.healthstatus.2013.isps-https	20	0.000	0.100	0.200	0.150	0.050	0.000	0.000	0.000	0.000	0.700
se.healthstatus.2013.isps-https-www	20	0.000	0.400	0.100	0.000	0.100	0.000	0.000	0.000	0.000	0.500
se.healthstatus.2013.media-http	30	0.000	0.133	0.733	0.533	0.200	0.000	0.000	0.000	0.000	0.133
se.healthstatus.2013.media-http-www	32	0.000	0.563	0.313	0.125	0.156	0.000	0.031	0.000	0.000	0.125
se.healthstatus.2013.media-https	32	0.000	0.063	0.063	0.000	0.063	0.000	0.000	0.000	0.000	0.875
se.healthstatus.2013.media-https-www	32	0.000	0.063	0.094	0.031	0.063	0.000	0.000	0.000	0.000	0.844
se.healthstatus.2013.municipalities-http	286	0.000	0.615	0.255	0.178	0.077	0.000	0.000	0.000	0.000	0.129
se.healthstatus.2013.municipalities-http-www	288	0.000	0.851	0.090	0.035	0.056	0.000	0.000	0.000	0.000	0.059

Dataset	Domains	1xx	2xx	3xx	301	302	303	307	4xx	5xx	(null)
se.healthstatus.2013.municipalities-https	290	0.000	0.072	0.079	0.024	0.055	0.000	0.000	0.000	0.000	0.848
se.healthstatus.2013.municipalities-https-www	290	0.000	0.117	0.069	0.003	0.066	0.000	0.000	0.000	0.000	0.814
se.healthstatus.2013.public-authorities-http	213	0.000	0.404	0.394	0.258	0.136	0.000	0.000	0.000	0.000	0.202
se.healthstatus.2013.public-authorities-http-www	214	0.000	0.696	0.252	0.126	0.121	0.005	0.000	0.000	0.000	0.051
se.healthstatus.2013.public-authorities-https	214	0.000	0.042	0.042	0.005	0.037	0.000	0.000	0.000	0.000	0.916
se.healthstatus.2013.public-authorities-https-www	214	0.000	0.117	0.056	0.009	0.047	0.000	0.000	0.000	0.000	0.827

Table C.2: Dataset origin HTTP response code/group coverage

A further analysis might disregard 4xx and 5xx responses as well, but current numbers suggest the difference would not be significant. While this can be because of software problems, they do exist and therefore the software seems to work as intended.

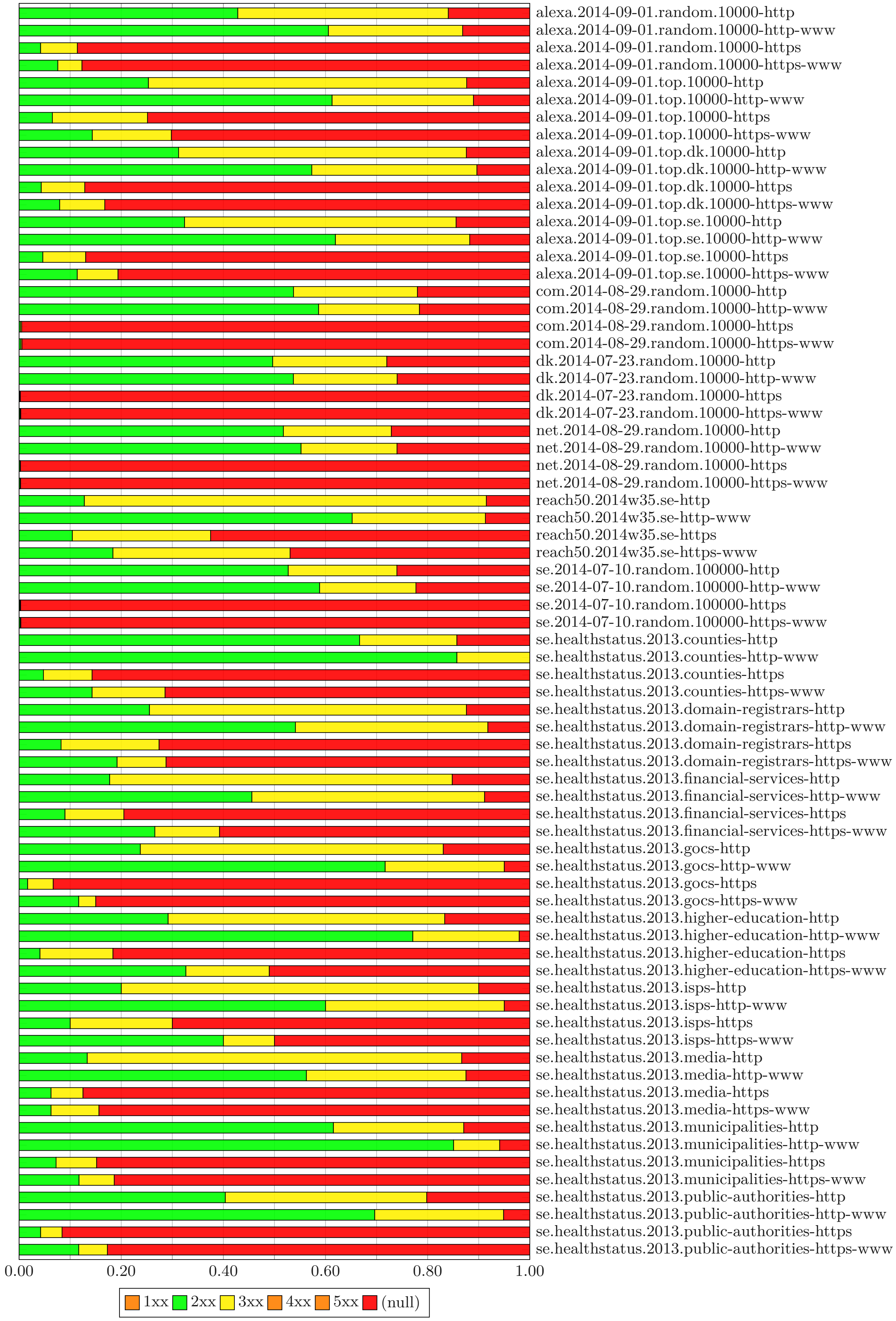

Figure C.1: Distribution of HTTP status codes

C.4 Internal versus external resources

The table shows non-failed origin domains having requests strictly to the same domain, subdomains, superdomains or same primary domain – jointly known as internal domains – or external (non-internal) domains. Origin domains that are not exclusively loading from either internal or external resources are loading from mixed domains. See also Figure C.2 visualizing the proportions (x axis) per dataset.

More detailed examples of distributions of resources are shown in Figure C.3, with ratio of internal resources per domain (as opposed to external resources) on the x axis, and what cumulative ratio of domains exhibit this property on the y axis. The leftmost marker in each dataset shows the ratio of domains (y axis) with 0% internal (strictly external) resources and the right hand marker 99% internal resources. To the right of the 99% marker are domains using strictly internal resources; the vertical difference (y axis) between the two markers shows the ratio of mixed resource usage.

Mixing resources from both internal and external domains is the most common way to compose a web page for datasets not randomly chosen from zones, although it is quite common for random domains as well. Random domains show relatively high tendencies to either extreme, with only either internal or external resources; in addition the usage of external resources is lower in HTTPS variations. As can be noted in both Figure C.2 and C.3, the top domains are very similar in terms of resource distribution between HTTP and HTTPS datasets. This means that active tracking on top sites is as prevalent when surfing a website over a secure, encrypted connection as on an insecure. Random .dk sites has the highest ratio of internal resources, but even so more than two thirds of domains load external resources. Strictly external plus mixed resource usage is above 80% in most and 90% in many datasets. The same is true for for HTTPS, confirming the thesis' hypotheses that tracking is installed through the use of external resources, deliberately or not, on HTTPS sites as well (2.1).

Dataset	Domains	Same domain	Subdomain	Superdomain	Same primary	Internal	Mixed	External
alexa.2014-09-01.random.10000-http	8 216	0.044	0.016	0.000	0.065	0.065	0.899	0.036
alexa.2014-09-01.random.10000-http-www	8 493	0.057	0.000	0.007	0.068	0.068	0.885	0.046
alexa.2014-09-01.random.10000-https	1 135	0.028	0.010	0.000	0.041	0.041	0.928	0.030
alexa.2014-09-01.random.10000-https-www	1 224	0.051	0.000	0.002	0.061	0.061	0.913	0.026
alexa.2014-09-01.top.10000-http	8 545	0.015	0.014	0.000	0.034	0.034	0.929	0.037
alexa.2014-09-01.top.10000-http-www	8 682	0.022	0.000	0.003	0.036	0.036	0.915	0.048
alexa.2014-09-01.top.10000-https	2 507	0.025	0.013	0.000	0.043	0.043	0.925	0.032
alexa.2014-09-01.top.10000-https-www	2 957	0.027	0.000	0.003	0.047	0.047	0.923	0.030
alexa.2014-09-01.top.dk.10000-http	2 263	0.027	0.019	0.000	0.050	0.050	0.920	0.030
alexa.2014-09-01.top.dk.10000-http-www	2 310	0.042	0.000	0.004	0.051	0.051	0.911	0.038
alexa.2014-09-01.top.dk.10000-https	339	0.033	0.012	0.000	0.054	0.054	0.919	0.027
alexa.2014-09-01.top.dk.10000-https-www	441	0.055	0.000	0.002	0.069	0.069	0.904	0.028
alexa.2014-09-01.top.se.10000-http	2 797	0.024	0.011	0.000	0.038	0.038	0.925	0.037
alexa.2014-09-01.top.se.10000-http-www	2 895	0.030	0.000	0.004	0.037	0.037	0.917	0.045
alexa.2014-09-01.top.se.10000-https	438	0.023	0.007	0.000	0.034	0.034	0.943	0.023
alexa.2014-09-01.top.se.10000-https-www	650	0.023	0.000	0.002	0.028	0.028	0.954	0.019
com.2014-08-29.random.10000-http	7 775	0.122	0.021	0.000	0.147	0.147	0.617	0.236
com.2014-08-29.random.10000-http-www	7 811	0.135	0.000	0.009	0.148	0.148	0.609	0.243
com.2014-08-29.random.10000-https	50	0.106	0.021	0.000	0.128	0.128	0.830	0.043
com.2014-08-29.random.10000-https-www	55	0.185	0.000	0.019	0.204	0.204	0.796	0.000
dk.2014-07-23.random.10000-http	7 180	0.285	0.025	0.000	0.316	0.316	0.371	0.313
dk.2014-07-23.random.10000-http-www	7 378	0.278	0.000	0.030	0.312	0.312	0.374	0.313
dk.2014-07-23.random.10000-https	23	0.261	0.043	0.000	0.304	0.304	0.652	0.043
dk.2014-07-23.random.10000-https-www	32	0.267	0.000	0.000	0.267	0.267	0.700	0.033

Dataset	Domains	Same domain	Subdomain	Superdomain	Same primary	Internal	Mixed	External
net.2014-08-29.random.10000-http	7 270	0.114	0.023	0.000	0.139	0.139	0.590	0.271
net.2014-08-29.random.10000-http-www	7 378	0.129	0.000	0.007	0.138	0.138	0.581	0.281
net.2014-08-29.random.10000-https	26	0.231	0.077	0.000	0.385	0.385	0.615	0.000
net.2014-08-29.random.10000-https-www	28	0.231	0.000	0.000	0.231	0.231	0.731	0.038
reach50.2014w35.se-http	43	0.000	0.000	0.000	0.000	0.000	0.953	0.047
reach50.2014w35.se-http-www	42	0.000	0.000	0.000	0.000	0.000	0.929	0.071
reach50.2014w35.se-https	18	0.000	0.056	0.000	0.056	0.056	0.833	0.111
reach50.2014w35.se-https-www	26	0.000	0.000	0.000	0.000	0.000	0.885	0.115
se.2014-07-10.random.100000-http	73 605	0.185	0.027	0.000	0.218	0.218	0.397	0.385
se.2014-07-10.random.100000-http-www	77 261	0.187	0.000	0.027	0.219	0.219	0.396	0.385
se.2014-07-10.random.100000-https	282	0.140	0.018	0.000	0.166	0.166	0.804	0.030
se.2014-07-10.random.100000-https-www	328	0.109	0.000	0.006	0.115	0.115	0.851	0.034
se.healthstatus.2013.counties-http	18	0.000	0.000	0.000	0.000	0.000	1.000	0.000
se.healthstatus.2013.counties-http-www	21	0.000	0.000	0.000	0.000	0.000	0.952	0.048
se.healthstatus.2013.counties-https	3	0.000	0.000	0.000	0.000	0.000	1.000	0.000
se.healthstatus.2013.counties-https-www	6	0.167	0.000	0.000	0.167	0.167	0.833	0.000
se.healthstatus.2013.domain-registrars-http	127	0.056	0.072	0.000	0.136	0.136	0.728	0.136
se.healthstatus.2013.domain-registrars-http-www	134	0.104	0.000	0.015	0.157	0.157	0.694	0.149
se.healthstatus.2013.domain-registrars-https	40	0.025	0.125	0.000	0.150	0.150	0.825	0.025
se.healthstatus.2013.domain-registrars-https-www	42	0.119	0.000	0.000	0.143	0.143	0.810	0.048
se.healthstatus.2013.financial-services-http	67	0.000	0.000	0.000	0.000	0.000	0.910	0.090
se.healthstatus.2013.financial-services-http-www	72	0.014	0.000	0.000	0.014	0.014	0.875	0.111
se.healthstatus.2013.financial-services-https	16	0.000	0.000	0.000	0.000	0.000	0.938	0.063
se.healthstatus.2013.financial-services-https-www	31	0.000	0.000	0.000	0.000	0.000	0.968	0.032
se.healthstatus.2013.gocs-http	49	0.000	0.020	0.000	0.020	0.020	0.878	0.102
se.healthstatus.2013.gocs-http-www	57	0.035	0.000	0.000	0.035	0.035	0.842	0.123
se.healthstatus.2013.gocs-https	4	0.000	0.000	0.000	0.000	0.000	1.000	0.000
se.healthstatus.2013.gocs-https-www	9	0.000	0.000	0.000	0.000	0.000	1.000	0.000
se.healthstatus.2013.higher-education-http	40	0.000	0.050	0.000	0.050	0.050	0.925	0.025
se.healthstatus.2013.higher-education-http-www	47	0.064	0.000	0.000	0.064	0.064	0.915	0.021
se.healthstatus.2013.higher-education-https	9	0.000	0.000	0.000	0.000	0.000	1.000	0.000
se.healthstatus.2013.higher-education-https-www	24	0.000	0.000	0.000	0.000	0.000	1.000	0.000
se.healthstatus.2013.isps-http	18	0.000	0.000	0.000	0.000	0.000	0.944	0.056
se.healthstatus.2013.isps-http-www	19	0.000	0.000	0.000	0.000	0.000	1.000	0.000
se.healthstatus.2013.isps-https	6	0.000	0.000	0.000	0.000	0.000	1.000	0.000
se.healthstatus.2013.isps-https-www	10	0.000	0.000	0.000	0.000	0.000	1.000	0.000
se.healthstatus.2013.media-http	26	0.038	0.000	0.000	0.038	0.038	0.885	0.077
se.healthstatus.2013.media-http-www	28	0.000	0.000	0.036	0.036	0.036	0.857	0.107
se.healthstatus.2013.media-https	4	0.000	0.000	0.000	0.000	0.000	1.000	0.000
se.healthstatus.2013.media-https-www	5	0.000	0.000	0.000	0.000	0.000	1.000	0.000
se.healthstatus.2013.municipalities-http	249	0.032	0.008	0.000	0.040	0.040	0.960	0.000

Dataset	Domains	Same domain	Subdomain	Superdomain	Same primary	Internal	Mixed	External
se.healthstatus.2013.municipalities-http-www	271	0.048	0.000	0.000	0.048	0.048	0.948	0.004
se.healthstatus.2013.municipalities-https	44	0.045	0.000	0.000	0.068	0.068	0.932	0.000
se.healthstatus.2013.municipalities-https-www	54	0.074	0.000	0.000	0.074	0.074	0.926	0.000
se.healthstatus.2013.public-authorities-http	170	0.035	0.012	0.000	0.047	0.047	0.853	0.100
se.healthstatus.2013.public-authorities-http-www	203	0.074	0.000	0.000	0.074	0.074	0.823	0.103
se.healthstatus.2013.public-authorities-https	18	0.000	0.111	0.000	0.167	0.167	0.833	0.000
se.healthstatus.2013.public-authorities-https-www	37	0.027	0.000	0.000	0.027	0.027	0.973	0.000

Table C.3: Internal versus external resources coverage

During analysis har-dulcify splits results into unfiltered, internal and external resources, each treated the same way in terms of classifications, allowing separate conclusions to be drawn. Further analysis showing internal/external ratios are generally calculated per domain which has at least one internal/external request. This makes a rather large difference for random zone domains, which have a rather high ratio of domains with no internal or no external resources (C.6).

Looking at internal versus external domains, with regards to the origin domain, is easy to do as it is only a matter of string comparison. The next step would be to use organization grouping, as seen in the Disconnect classifications (C.11). Private CDN domain detection requires more extensive reverse requests' domain usage mapping work plus manual classification work. While the work put in might only be effective for top organizations with many services, it adds to seeing legal entities as information receivers rather than the merely technical domain partitioning. It also adds more questions – could for example a private CDN domain hosted in another organization's datacenter be seen as both internal and external?

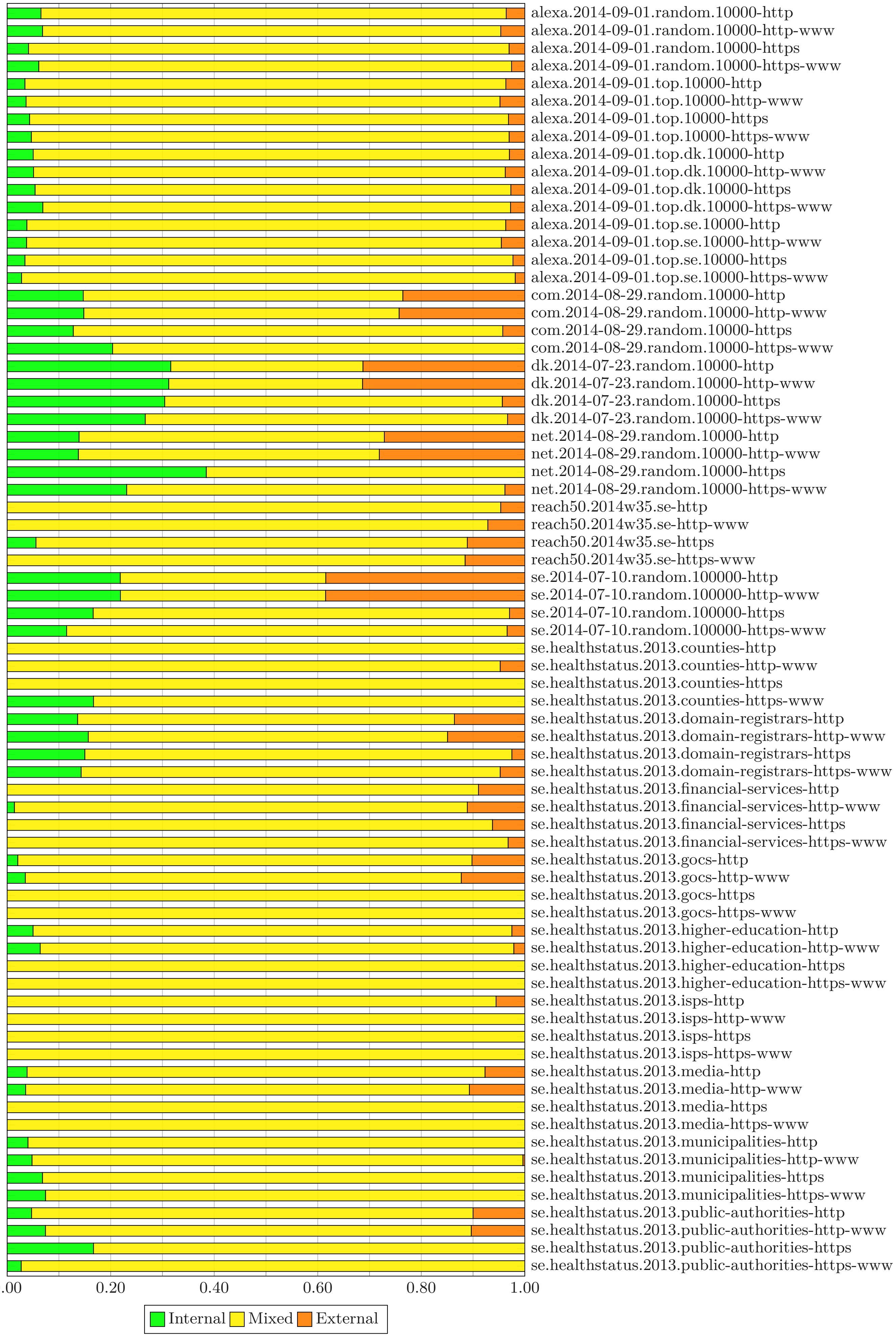

Figure C.2: Distribution of domains with strictly internal, mixed or strictly external resources

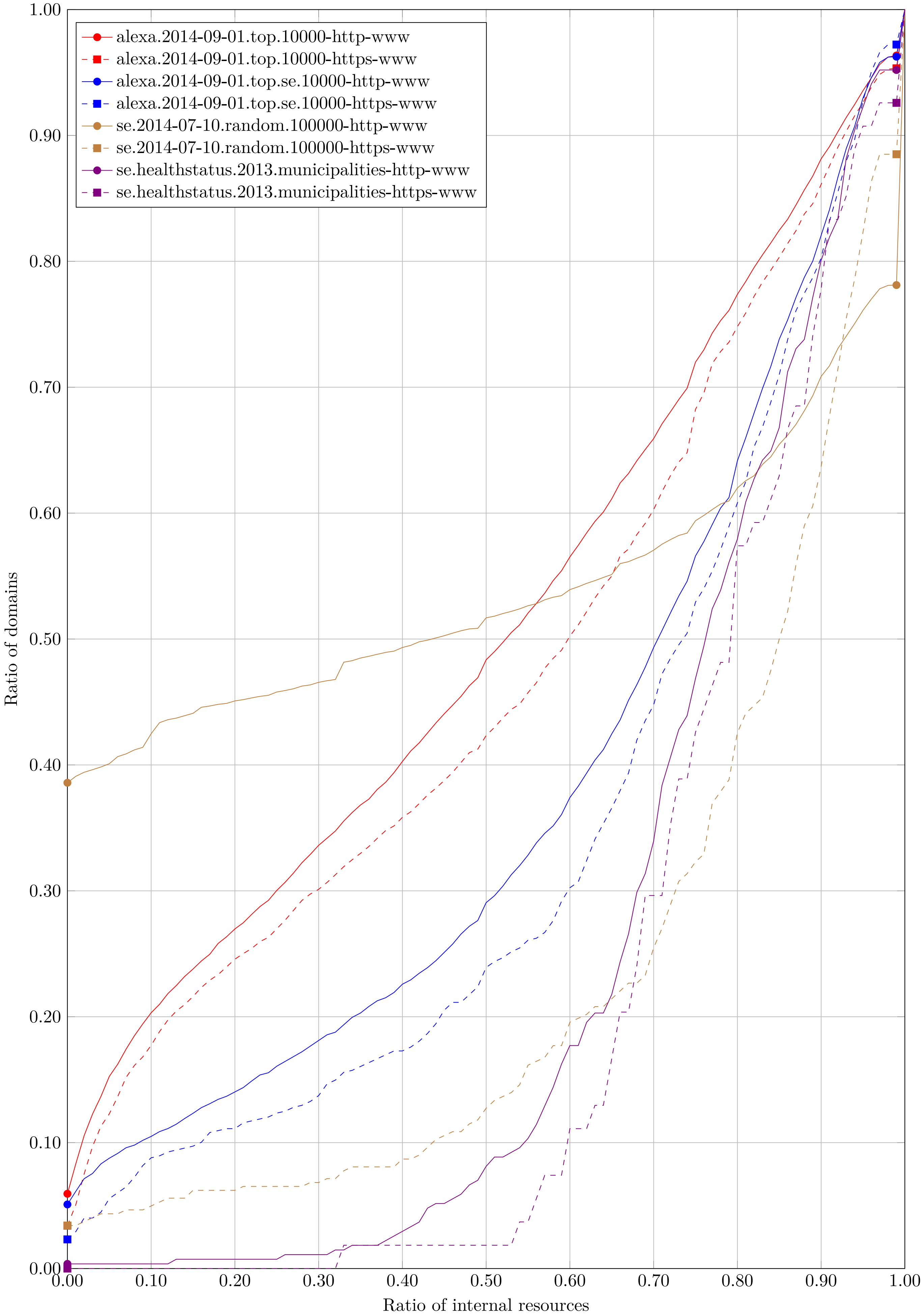

Figure C.3: Cumulative distribution of the ratio of internal resources per domain

C.5 Domain and request counts

The table shows counts of domains and domains which have at least one internal/external request, and to how many external domains, primary domains or Disconnect domains the requests was sent. Following that, counts of all requests, internal requests and external requests. External requests matching Disconnect's blocking lists are also shown. This table is mostly interesting to show the scale of the data collection, in terms of number of requests made and analyzed.

If we include all dataset variations 252,481 domains responded to the request; out of those 172,898 made at least one internal request and 199,358 external dittos. A total of 9,877,940 requests have been analyzed; 4,498,575 were internal, 5,379,365 external and 2,389,917 matched disconnect's blocking list.

See also (C.6) and (C.12).

Dataset	Domains	w/ int	w/ ext	Ext dom.	Ext prim.	Ext D dom.	All requests	Int	Ext	Disco.
alexa.2014-09-01.random.10000-http	8 216	7 829	7 591	14 257	7 312	704	610 150	343 214	266 936	166 702
alexa.2014-09-01.random.10000-http-www	8 493	8 009	7 825	14 478	7 501	704	627 185	355 413	271 772	169 685
alexa.2014-09-01.random.10000-https	1 135	1 084	1 072	3 071	1 454	370	86 124	49 816	36 308	23 599
alexa.2014-09-01.random.10000-https-www	1 224	1 182	1 139	2 406	1 233	368	87 423	60 773	26 650	16 764
alexa.2014-09-01.top.10000-http	8 545	8 156	8 176	22 212	8 335	755	899 404	408 553	490 851	274 782
alexa.2014-09-01.top.10000-http-www	8 682	8 190	8 289	22 661	8 544	760	912 709	415 958	496 751	276 636
alexa.2014-09-01.top.10000-https	2 507	2 398	2 369	7 217	2 909	542	207 090	93 986	113 104	67 788
alexa.2014-09-01.top.10000-https-www	2 957	2 849	2 801	8 017	3 120	569	243 323	117 947	125 376	73 239
alexa.2014-09-01.top.dk.10000-http	2 263	2 182	2 136	2 768	1 407	282	162 059	99 234	62 825	37 832
alexa.2014-09-01.top.dk.10000-http-www	2 310	2 212	2 182	2 850	1 483	284	165 023	100 543	64 480	38 373
alexa.2014-09-01.top.dk.10000-https	339	325	316	816	420	151	25 034	15 575	9 459	5 942
alexa.2014-09-01.top.dk.10000-https-www	441	424	406	997	516	176	29 901	18 596	11 305	6 901
alexa.2014-09-01.top.se.10000-http	2 797	2 687	2 684	4 681	2 199	342	209 012	117 914	91 098	52 345
alexa.2014-09-01.top.se.10000-http-www	2 895	2 756	2 779	4 751	2 207	351	212 700	121 479	91 221	52 398
alexa.2014-09-01.top.se.10000-https	438	427	422	990	524	167	32 271	19 899	12 372	7 104
alexa.2014-09-01.top.se.10000-https-www	650	636	630	1 237	651	199	45 199	28 286	16 913	9 510
com.2014-08-29.random.10000-http	7 775	5 575	6 222	6 329	3 713	404	226 636	76 167	150 469	55 666
com.2014-08-29.random.10000-http-www	7 811	5 546	6 241	6 339	3 717	405	230 039	78 086	151 953	55 955
com.2014-08-29.random.10000-https	50	45	41	127	84	47	2 251	1 654	597	446
com.2014-08-29.random.10000-https-www	55	54	43	163	99	49	2 650	1 930	720	477
dk.2014-07-23.random.10000-http	7 180	4 648	4 626	4 272	2 834	278	187 706	80 787	106 919	36 822
dk.2014-07-23.random.10000-http-www	7 378	4 763	4 773	4 378	2 894	275	190 186	82 052	108 134	35 960
dk.2014-07-23.random.10000-https	23	22	16	52	33	26	902	725	177	150
dk.2014-07-23.random.10000-https-www	32	29	22	81	54	32	1 337	921	416	257
net.2014-08-29.random.10000-http	7 270	4 871	5 757	6 206	3 806	412	192 646	56 364	136 282	48 379
net.2014-08-29.random.10000-http-www	7 378	4 867	5 839	6 311	3 889	411	196 301	58 205	138 096	49 471
net.2014-08-29.random.10000-https	26	26	16	49	26	21	1 299	1 071	228	203
net.2014-08-29.random.10000-https-www	28	25	20	62	34	27	1 568	1 210	358	291
reach50.2014w35.se-http	43	41	43	339	195	92	3 898	1 313	2 585	843
reach50.2014w35.se-http-www	42	39	42	342	194	92	3 645	1 135	2 510	801
reach50.2014w35.se-https	18	16	17	117	66	41	1 092	264	828	265
reach50.2014w35.se-https-www	26	23	26	139	83	40	1 436	455	981	303

Dataset	Domains	w/ int	w/ ext	Ext dom.	Ext prim.	Ext D dom.	All requests	Int	Ext	Disco.
se.2014-07-10.random.100000-http	73 605	43 216	54 882	24 289	15 746	496	1 931 501	782 998	1 148 503	395 347
se.2014-07-10.random.100000-http-www	77 261	45 312	57 547	25 366	16 546	502	2 006 337	807 160	1 199 177	406 990
se.2014-07-10.random.100000-https	282	263	226	393	235	94	14 140	10 726	3 414	1 962
se.2014-07-10.random.100000-https-www	328	311	285	546	340	124	17 686	13 057	4 629	2 451
se.healthstatus.2013.counties-http	18	18	18	34	23	10	921	726	195	105
se.healthstatus.2013.counties-http-www	21	20	21	39	27	11	1 066	809	257	133
se.healthstatus.2013.counties-https	3	3	3	6	5	2	156	137	19	7
se.healthstatus.2013.counties-https-www	6	6	5	15	11	4	240	191	49	20
se.healthstatus.2013.domain-registrars-http	127	108	108	216	148	66	6 418	4 459	1 959	886
se.healthstatus.2013.domain-registrars-http-www	134	114	113	214	144	62	6 627	4 565	2 062	872
se.healthstatus.2013.domain-registrars-https	40	39	34	124	86	46	2 342	1 620	722	430
se.healthstatus.2013.domain-registrars-https-www	42	40	36	116	79	40	2 439	1 833	606	327
se.healthstatus.2013.financial-services-http	67	61	67	137	97	49	3 260	2 319	941	378
se.healthstatus.2013.financial-services-http-www	72	64	71	144	97	50	3 518	2 491	1 027	415
se.healthstatus.2013.financial-services-https	16	15	16	47	37	24	881	696	185	95
se.healthstatus.2013.financial-services-https-www	31	30	31	71	50	32	1 504	1 148	356	228
se.healthstatus.2013.gocs-http	49	44	48	130	83	45	2 585	1 746	839	501
se.healthstatus.2013.gocs-http-www	57	50	55	150	95	47	2 925	1 894	1 031	577
se.healthstatus.2013.gocs-https	4	4	4	44	28	21	321	195	126	64
se.healthstatus.2013.gocs-https-www	9	9	9	65	44	27	567	377	190	91
se.healthstatus.2013.higher-education-http	40	39	38	73	53	24	2 064	1 685	379	270
se.healthstatus.2013.higher-education-http-www	47	46	44	74	52	26	2 305	1 886	419	308
se.healthstatus.2013.higher-education-https	9	9	9	38	25	16	571	442	129	104
se.healthstatus.2013.higher-education-https-www	24	24	24	63	45	22	1 291	1 038	253	182
se.healthstatus.2013.isps-http	18	17	18	111	76	47	1 150	757	393	271
se.healthstatus.2013.isps-http-www	19	19	19	135	92	55	1 209	735	474	317
se.healthstatus.2013.isps-https	6	6	6	84	63	41	523	323	200	152
se.healthstatus.2013.isps-https-www	10	10	10	89	66	43	669	448	221	163
se.healthstatus.2013.media-http	26	24	25	346	190	81	4 812	1 596	3 216	1 101
se.healthstatus.2013.media-http-www	28	25	27	378	207	79	5 507	1 676	3 831	1 234
se.healthstatus.2013.media-https	4	4	4	102	58	24	977	202	775	186
se.healthstatus.2013.media-https-www	5	5	5	95	59	28	868	316	552	204
se.healthstatus.2013.municipalities-http	249	249	239	207	113	39	14 028	10 162	3 866	2 367
se.healthstatus.2013.municipalities-http-www	271	270	258	203	113	39	14 749	10 827	3 922	2 447
se.healthstatus.2013.municipalities-https	44	44	41	67	41	18	2 603	2 047	556	305
se.healthstatus.2013.municipalities-https-www	54	54	50	73	42	18	3 001	2 347	654	394
se.healthstatus.2013.public-authorities-http	170	153	162	172	110	48	7 423	5 403	2 020	935
se.healthstatus.2013.public-authorities-http-www	203	182	188	170	111	48	8 297	6 123	2 174	945
se.healthstatus.2013.public-authorities-https	18	18	15	32	21	9	664	575	89	64
se.healthstatus.2013.public-authorities-https-www	37	37	36	63	41	23	1 596	1 315	281	200

Dataset	Domains	w/ int	w/ ext	Ext dom.	Ext prim.	Ext D dom.	All requests	Int	Ext	Disco.

Table C.4: Requests per domain

C.6 Requests per domain and ratios

Shown in the table below are ratios based on the request counts in Section C.5. First the ratio of domains with at least one internal/external request. Requests are then shown as average count per domain, internal per domain with internal requests and external per domain with external requests. External Disconnect requests are shown per domain with external requests. After that comes ratios of requests; internal/external/Disconnect out of all requests, external over internal and Disconnect request ratio out of the external requests.

It is interesting to look at the number of requests, as they differ between datasets. Random TLD domains have a low average of 26-29 requests per domain, random Alexa top sites an average 71-76 – but the very top of the Alexa top list comes in at 83-105. The smaller datasets Reach50 reaches 87, and the even smaller dataset *.SE Health Status*' media sets the record at 197 requests per domain!

Dataset	Domains	w/ int	w/ ext	A/d	I/di	E/de	D/de	I/A	E/A	D/A	E/I	D/E
alexa.2014-09-01.random.10000-http	8 216	0.953	0.924	74	44	35	22	0.563	0.437	0.273	0.778	0.625
alexa.2014-09-01.random.10000-http-www	8 493	0.943	0.921	74	44	35	22	0.567	0.433	0.271	0.765	0.624
alexa.2014-09-01.random.10000-https	1 135	0.955	0.944	76	46	34	22	0.578	0.422	0.274	0.729	0.650
alexa.2014-09-01.random.10000-https-www	1 224	0.966	0.931	71	51	23	15	0.695	0.305	0.192	0.439	0.629
alexa.2014-09-01.top.10000-http	8 545	0.954	0.957	105	50	60	34	0.454	0.546	0.306	1.201	0.560
alexa.2014-09-01.top.10000-http-www	8 682	0.943	0.955	105	51	60	33	0.456	0.544	0.303	1.194	0.557
alexa.2014-09-01.top.10000-https	2 507	0.957	0.945	83	39	48	29	0.454	0.546	0.327	1.203	0.599
alexa.2014-09-01.top.10000-https-www	2 957	0.963	0.947	82	41	45	26	0.485	0.515	0.301	1.063	0.584
alexa.2014-09-01.top.dk.10000-http	2 263	0.964	0.944	72	45	29	18	0.612	0.388	0.233	0.633	0.602
alexa.2014-09-01.top.dk.10000-http-www	2 310	0.958	0.945	71	45	30	18	0.609	0.391	0.233	0.641	0.595
alexa.2014-09-01.top.dk.10000-https	339	0.959	0.932	74	48	30	19	0.622	0.378	0.237	0.607	0.628
alexa.2014-09-01.top.dk.10000-https-www	441	0.961	0.921	68	44	28	17	0.622	0.378	0.231	0.608	0.610
alexa.2014-09-01.top.se.10000-http	2 797	0.961	0.960	75	44	34	20	0.564	0.436	0.250	0.773	0.575
alexa.2014-09-01.top.se.10000-http-www	2 895	0.952	0.960	73	44	33	19	0.571	0.429	0.246	0.751	0.574
alexa.2014-09-01.top.se.10000-https	438	0.975	0.963	74	47	29	17	0.617	0.383	0.220	0.622	0.574
alexa.2014-09-01.top.se.10000-https-www	650	0.978	0.969	70	44	27	15	0.626	0.374	0.210	0.598	0.562
com.2014-08-29.random.10000-http	7 775	0.717	0.800	29	14	24	9	0.336	0.664	0.246	1.976	0.370
com.2014-08-29.random.10000-http-www	7 811	0.710	0.799	29	14	24	9	0.339	0.661	0.243	1.946	0.368
com.2014-08-29.random.10000-https	50	0.900	0.820	45	37	15	11	0.735	0.265	0.198	0.361	0.747
com.2014-08-29.random.10000-https-www	55	0.982	0.782	48	36	17	11	0.728	0.272	0.180	0.373	0.663
dk.2014-07-23.random.10000-http	7 180	0.647	0.644	26	17	23	8	0.430	0.570	0.196	1.323	0.344
dk.2014-07-23.random.10000-http-www	7 378	0.646	0.647	26	17	23	8	0.431	0.569	0.189	1.318	0.333
dk.2014-07-23.random.10000-https	23	0.957	0.696	39	33	11	9	0.804	0.196	0.166	0.244	0.847
dk.2014-07-23.random.10000-https-www	32	0.906	0.688	42	32	19	12	0.689	0.311	0.192	0.452	0.618
net.2014-08-29.random.10000-http	7 270	0.670	0.792	26	12	24	8	0.293	0.707	0.251	2.418	0.355
net.2014-08-29.random.10000-http-www	7 378	0.660	0.791	27	12	24	8	0.297	0.703	0.252	2.373	0.358

Dataset	Domains	w/ int	w/ ext	A/d	I/di	E/de	D/de	I/A	E/A	D/A	E/I	D/E
net.2014-08-29.random.10000-https	26	1.000	0.615	50	41	14	13	0.824	0.176	0.156	0.213	0.890
net.2014-08-29.random.10000-https-www	28	0.893	0.714	56	48	18	15	0.772	0.228	0.186	0.296	0.813
reach50.2014w35.se-http	43	0.953	1.000	91	32	60	20	0.337	0.663	0.216	1.969	0.326
reach50.2014w35.se-http-www	42	0.929	1.000	87	29	60	19	0.311	0.689	0.220	2.211	0.319
reach50.2014w35.se-https	18	0.889	0.944	61	17	49	16	0.242	0.758	0.243	3.136	0.320
reach50.2014w35.se-https-www	26	0.885	1.000	55	20	38	12	0.317	0.683	0.211	2.156	0.309
se.2014-07-10.random.100000-http	73 605	0.587	0.746	26	18	21	7	0.405	0.595	0.205	1.467	0.344
se.2014-07-10.random.100000-http-www	77 261	0.586	0.745	26	18	21	7	0.402	0.598	0.203	1.486	0.339
se.2014-07-10.random.100000-https	282	0.933	0.801	50	41	15	9	0.759	0.241	0.139	0.318	0.575
se.2014-07-10.random.100000-https-www	328	0.948	0.869	54	42	16	9	0.738	0.262	0.139	0.355	0.529
se.healthstatus.2013.counties-http	18	1.000	1.000	51	40	11	6	0.788	0.212	0.114	0.269	0.538
se.healthstatus.2013.counties-http-www	21	0.952	1.000	51	40	12	6	0.759	0.241	0.125	0.318	0.518
se.healthstatus.2013.counties-https	3	1.000	1.000	52	46	6	2	0.878	0.122	0.045	0.139	0.368
se.healthstatus.2013.counties-https-www	6	1.000	0.833	40	32	10	4	0.796	0.204	0.083	0.257	0.408
se.healthstatus.2013.domain-registrars-http	127	0.850	0.850	51	41	18	8	0.695	0.305	0.138	0.439	0.452
se.healthstatus.2013.domain-registrars-http-www	134	0.851	0.843	49	40	18	8	0.689	0.311	0.132	0.452	0.423
se.healthstatus.2013.domain-registrars-https	40	0.975	0.850	59	42	21	13	0.692	0.308	0.184	0.446	0.596
se.healthstatus.2013.domain-registrars-https-www	42	0.952	0.857	58	46	17	9	0.752	0.248	0.134	0.331	0.540
se.healthstatus.2013.financial-services-http	67	0.910	1.000	49	38	14	6	0.711	0.289	0.116	0.406	0.402
se.healthstatus.2013.financial-services-http-www	72	0.889	0.986	49	39	14	6	0.708	0.292	0.118	0.412	0.404
se.healthstatus.2013.financial-services-https	16	0.938	1.000	55	46	12	6	0.790	0.210	0.108	0.266	0.514
se.healthstatus.2013.financial-services-https-www	31	0.968	1.000	49	38	11	7	0.763	0.237	0.152	0.310	0.640
se.healthstatus.2013.gocs-http	49	0.898	0.980	53	40	17	10	0.675	0.325	0.194	0.481	0.597
se.healthstatus.2013.gocs-http-www	57	0.877	0.965	51	38	19	10	0.648	0.352	0.197	0.544	0.560
se.healthstatus.2013.gocs-https	4	1.000	1.000	80	49	32	16	0.607	0.393	0.199	0.646	0.508
se.healthstatus.2013.gocs-https-www	9	1.000	1.000	63	42	21	10	0.665	0.335	0.160	0.504	0.479
se.healthstatus.2013.higher-education-http	40	0.975	0.950	52	43	10	7	0.816	0.184	0.131	0.225	0.712
se.healthstatus.2013.higher-education-http-www	47	0.979	0.936	49	41	10	7	0.818	0.182	0.134	0.222	0.735
se.healthstatus.2013.higher-education-https	9	1.000	1.000	63	49	14	12	0.774	0.226	0.182	0.292	0.806
se.healthstatus.2013.higher-education-https-www	24	1.000	1.000	54	43	11	8	0.804	0.196	0.141	0.244	0.719
se.healthstatus.2013.isps-http	18	0.944	1.000	64	45	22	15	0.658	0.342	0.236	0.519	0.690
se.healthstatus.2013.isps-http-www	19	1.000	1.000	64	39	25	17	0.608	0.392	0.262	0.645	0.669
se.healthstatus.2013.isps-https	6	1.000	1.000	87	54	33	25	0.618	0.382	0.291	0.619	0.760
se.healthstatus.2013.isps-https-www	10	1.000	1.000	67	45	22	16	0.670	0.330	0.244	0.493	0.738
se.healthstatus.2013.media-http	26	0.923	0.962	185	67	129	44	0.332	0.668	0.229	2.015	0.342
se.healthstatus.2013.media-http-www	28	0.893	0.964	197	67	142	46	0.304	0.696	0.224	2.286	0.322
se.healthstatus.2013.media-https	4	1.000	1.000	244	51	194	47	0.207	0.793	0.190	3.837	0.240
se.healthstatus.2013.media-https-www	5	1.000	1.000	174	63	110	41	0.364	0.636	0.235	1.747	0.370
se.healthstatus.2013.municipalities-http	249	1.000	0.960	56	41	16	10	0.724	0.276	0.169	0.380	0.612
se.healthstatus.2013.municipalities-http-www	271	0.996	0.952	54	40	15	9	0.734	0.266	0.166	0.362	0.624
se.healthstatus.2013.municipalities-https	44	1.000	0.932	59	47	14	7	0.786	0.214	0.117	0.272	0.549

Dataset	Domains	w/ int	w/ ext	A/d	I/di	E/de	D/de	I/A	E/A	D/A	E/I	D/E
se.healthstatus.2013.municipalities-https-www	54	1.000	0.926	56	43	13	8	0.782	0.218	0.131	0.279	0.602
se.healthstatus.2013.public-authorities-http	170	0.900	0.953	44	35	12	6	0.728	0.272	0.126	0.374	0.463
se.healthstatus.2013.public-authorities-http-www	203	0.897	0.926	41	34	12	5	0.738	0.262	0.114	0.355	0.435
se.healthstatus.2013.public-authorities-https	18	1.000	0.833	37	32	6	4	0.866	0.134	0.096	0.155	0.719
se.healthstatus.2013.public-authorities-https-www	37	1.000	0.973	43	36	8	6	0.824	0.176	0.125	0.214	0.712

Table C.5: Requests per domain and ratios

C.7 Insecure versus secure resources

Using HTTPS to secure the connection between site and site users is considered an effective way to avoid prying eyes on the otherwise technically quite open and insecure internet. Sites which handle sensitive information, such as e-commerce shops, online payment providers and of course banks often tout being secure to use – and they have strong financial incentives to provide a service that is (or at least comes across as) trustworthy. As browsers will warn users if a site secured with HTTPS loads resources over non-HTTPS connections, site developers will have to make sure each and every request is secure to avoid being labeled not trustworthy. This also applies to third-party services, which have to make sure to provide HTTPS in order to be able to continue providing services to sites making the switch to a fully secured experience.

One of the concerns with mixing in HTTP on an HTTPS site is that an attacker can use traffic sniffers to get a hold of sensitive information leaking out through HTTP, or man in the middle attacks on several kinds of resources to insert malicious code, even though the site is supposed to be protected.

The following table shows to what extent sites manage to take full advantage of HTTPS, and to which extent they fail in requesting either internal or external resources. Note that HTTP domains that redirect to HTTPS (C.8) right away and/or only load HTTPS resources are shown as fully secure, as the analysis excludes the origin request – although in general an initial request to HTTP can potentially nullify all subsequent security measures.

Figure C.4 shows an overview of ratio of domains with strictly secure requests, mixed security or strictly insecure requests (x axis) per dataset.

While the technology has been around a long time, it does not seem as if very many sites actually use HTTPS. Even origin sites that respond to HTTPS requests seem to either redirect to an HTTP site (C.8), or load at least some of its resources over non-HTTPS connections. Typing in an HTTPS address into the browser's address bar will actually only give full HTTPS security on 27-58% of the domains – a number where the random domains surprisingly beat the non-random ones.

The cumulative distribution of domains (y axis) with a certain ratio of secure resources (out of all requested resources) per domain (x axis) is shown on in Figure C.5. The first marker in each dataset shows the ratio of domains with no secure resources at all. The second marker shows 99% secure resources, which marks the start of fully secure domains. The vertical differences between the two markers for each dataset shows the range of sites with mixed security.

We can see that HTTPS datasets have much better security than their HTTP counterparts. Looking at se.2014-07-10.random.100000-http-www which has over 60% completely insecure domains, 30% mixed security and less than 10% domains with only secure resources. Comparing it with se.2014-07-10.random.100000-https-www, we see that it has less than 10% completely insecure domains, a bit more than 30% mixed security and over 55% domains with only secure resources. The ratios of completely insecure and completely secure domains have almost been reversed. We can also see that many domains in the HTTPS datasets have between 90% and 99% secure resources – around 25% of municipalities for example – which seems like a relatively small gap to close to get a completely secure site.

Why is adoption lower for top sites? As high-traffic sites they might have a high system load, and since HTTPS require some extra processing and data exchange, they might have deferred it until the security is *really* needed – such as when passwords of financial information is entered. Strict HTTPS performance concerns were dismissed by Google engineers in 2010[1] – and Google has since implemented HTTPS as an alternative for most and the default for some services[2]. HTTPS is also a positive "signal" in Google's PageRank algorithm, meaning the use of HTTPS will lead to a better position in Google's search results[3]. There are other effects on network services though, such as reduced ability for ISPs to cache results closer to network edges or companies to easily inspect filter traffic [37].

[1] https://www.imperialviolet.org/2010/06/25/overclocking-ssl.html

[2] http://gmailblog.blogspot.se/2010/01/default-https-access-for-gmail.html

[3] http://googleonlinesecurity.blogspot.in/2014/08/https-as-ranking-signal_6.html

Another concern is that curated domain lists seem to exhibit an even lower HTTPS adoption than both random and top domains – the domains have been selected as they are deemed important to the public in some way.

Dataset	Domains	Int insec	Mix int sec	Int sec	Ext insec	Mix ext sec	Ext sec	All insec	Mix sec	All sec
alexa.2014-09-01.random.10000-http	8 216	0.951	0.026	0.023	0.425	0.544	0.031	0.455	0.524	0.021
alexa.2014-09-01.random.10000-http-www	8 493	0.951	0.023	0.027	0.433	0.537	0.030	0.464	0.512	0.024
alexa.2014-09-01.random.10000-https	1 135	0.406	0.214	0.380	0.145	0.538	0.317	0.114	0.603	0.284
alexa.2014-09-01.random.10000-https-www	1 224	0.255	0.215	0.530	0.113	0.443	0.443	0.087	0.503	0.410
alexa.2014-09-01.top.10000-http	8 545	0.891	0.066	0.043	0.350	0.599	0.051	0.356	0.602	0.042
alexa.2014-09-01.top.10000-http-www	8 682	0.891	0.061	0.049	0.353	0.595	0.052	0.361	0.592	0.047
alexa.2014-09-01.top.10000-https	2 507	0.435	0.196	0.369	0.138	0.523	0.340	0.110	0.565	0.326
alexa.2014-09-01.top.10000-https-www	2 957	0.359	0.213	0.428	0.125	0.507	0.368	0.103	0.551	0.346
alexa.2014-09-01.top.dk.10000-http	2 263	0.943	0.024	0.033	0.391	0.569	0.040	0.413	0.556	0.031
alexa.2014-09-01.top.dk.10000-http-www	2 310	0.943	0.020	0.038	0.398	0.560	0.041	0.420	0.545	0.034
alexa.2014-09-01.top.dk.10000-https	339	0.351	0.200	0.449	0.114	0.528	0.358	0.102	0.572	0.326
alexa.2014-09-01.top.dk.10000-https-www	441	0.337	0.151	0.512	0.113	0.515	0.372	0.103	0.546	0.351
alexa.2014-09-01.top.se.10000-http	2 797	0.931	0.029	0.040	0.383	0.570	0.047	0.399	0.564	0.037
alexa.2014-09-01.top.se.10000-http-www	2 895	0.933	0.021	0.046	0.383	0.570	0.046	0.399	0.559	0.042
alexa.2014-09-01.top.se.10000-https	438	0.356	0.204	0.440	0.140	0.486	0.374	0.126	0.549	0.325
alexa.2014-09-01.top.se.10000-https-www	650	0.278	0.156	0.566	0.130	0.432	0.438	0.120	0.480	0.400
com.2014-08-29.random.10000-http	7 775	0.997	0.001	0.002	0.758	0.230	0.012	0.793	0.201	0.007
com.2014-08-29.random.10000-http-www	7 811	0.996	0.001	0.002	0.755	0.234	0.011	0.790	0.204	0.006
com.2014-08-29.random.10000-https	50	0.311	0.133	0.556	0.244	0.415	0.341	0.149	0.468	0.383
com.2014-08-29.random.10000-https-www	55	0.204	0.148	0.648	0.163	0.442	0.395	0.056	0.444	0.500
dk.2014-07-23.random.10000-http	7 180	0.997	0.000	0.002	0.640	0.331	0.029	0.753	0.235	0.012
dk.2014-07-23.random.10000-http-www	7 378	0.997	0.001	0.002	0.633	0.337	0.030	0.746	0.241	0.012
dk.2014-07-23.random.10000-https	23	0.227	0.182	0.591	0.063	0.250	0.688	0.087	0.348	0.565
dk.2014-07-23.random.10000-https-www	32	0.172	0.207	0.621	0.136	0.318	0.545	0.100	0.367	0.533
net.2014-08-29.random.10000-http	7 270	0.998	0.001	0.001	0.776	0.212	0.012	0.807	0.185	0.008
net.2014-08-29.random.10000-http-www	7 378	0.998	0.001	0.001	0.776	0.211	0.013	0.806	0.185	0.008
net.2014-08-29.random.10000-https	26	0.154	0.192	0.654	0.000	0.375	0.625	0.038	0.385	0.577
net.2014-08-29.random.10000-https-www	28	0.160	0.240	0.600	0.050	0.400	0.550	0.000	0.500	0.500
reach50.2014w35.se-http	43	0.829	0.098	0.073	0.442	0.442	0.116	0.442	0.512	0.047
reach50.2014w35.se-http-www	42	0.795	0.077	0.128	0.452	0.429	0.119	0.452	0.452	0.095
reach50.2014w35.se-https	18	0.250	0.063	0.688	0.118	0.471	0.412	0.111	0.444	0.444
reach50.2014w35.se-https-www	26	0.348	0.043	0.609	0.192	0.385	0.423	0.192	0.385	0.423
se.2014-07-10.random.100000-http	73 605	0.997	0.001	0.002	0.492	0.383	0.125	0.603	0.304	0.093
se.2014-07-10.random.100000-http-www	77 261	0.997	0.001	0.002	0.501	0.378	0.121	0.610	0.300	0.090
se.2014-07-10.random.100000-https	282	0.125	0.160	0.715	0.066	0.332	0.602	0.044	0.347	0.609
se.2014-07-10.random.100000-https-www	328	0.151	0.141	0.707	0.105	0.326	0.568	0.081	0.357	0.562
se.healthstatus.2013.counties-http	18	0.944	0.056	0.000	0.611	0.389	0.000	0.611	0.389	0.000
se.healthstatus.2013.counties-http-www	21	0.950	0.050	0.000	0.619	0.381	0.000	0.619	0.381	0.000

Dataset	Domains	Int insec	Mix int sec	Int sec	Ext insec	Mix ext sec	Ext sec	All insec	Mix sec	All sec
se.healthstatus.2013.counties-https	3	0.667	0.000	0.333	0.667	0.333	0.000	0.667	0.333	0.000
se.healthstatus.2013.counties-https-www	6	0.333	0.000	0.667	0.400	0.200	0.400	0.333	0.167	0.500
se.healthstatus.2013.domain-registrars-http	127	0.787	0.093	0.120	0.537	0.343	0.120	0.552	0.352	0.096
se.healthstatus.2013.domain-registrars-http-www	134	0.789	0.088	0.123	0.513	0.372	0.115	0.545	0.358	0.097
se.healthstatus.2013.domain-registrars-https	40	0.179	0.205	0.615	0.147	0.265	0.588	0.100	0.325	0.575
se.healthstatus.2013.domain-registrars-https-www	42	0.150	0.175	0.675	0.083	0.278	0.639	0.071	0.333	0.595
se.healthstatus.2013.financial-services-http	67	0.820	0.131	0.049	0.612	0.224	0.164	0.582	0.373	0.045
se.healthstatus.2013.financial-services-http-www	72	0.828	0.078	0.094	0.620	0.225	0.155	0.597	0.306	0.097
se.healthstatus.2013.financial-services-https	16	0.067	0.133	0.800	0.063	0.375	0.563	0.063	0.438	0.500
se.healthstatus.2013.financial-services-https-www	31	0.133	0.100	0.767	0.097	0.290	0.613	0.097	0.323	0.581
se.healthstatus.2013.gocs-http	49	0.977	0.000	0.023	0.500	0.479	0.021	0.510	0.469	0.020
se.healthstatus.2013.gocs-http-www	57	0.980	0.000	0.020	0.509	0.473	0.018	0.526	0.456	0.018
se.healthstatus.2013.gocs-https	4	0.250	0.000	0.750	0.000	0.250	0.750	0.000	0.250	0.750
se.healthstatus.2013.gocs-https-www	9	0.000	0.111	0.889	0.000	0.444	0.556	0.000	0.556	0.444
se.healthstatus.2013.higher-education-http	40	0.897	0.103	0.000	0.711	0.289	0.000	0.650	0.350	0.000
se.healthstatus.2013.higher-education-http-www	47	0.891	0.109	0.000	0.682	0.318	0.000	0.638	0.362	0.000
se.healthstatus.2013.higher-education-https	9	0.222	0.444	0.333	0.333	0.333	0.333	0.111	0.667	0.222
se.healthstatus.2013.higher-education-https-www	24	0.250	0.083	0.667	0.208	0.375	0.417	0.208	0.375	0.417
se.healthstatus.2013.isps-http	18	0.765	0.118	0.118	0.389	0.500	0.111	0.333	0.556	0.111
se.healthstatus.2013.isps-http-www	19	0.789	0.105	0.105	0.368	0.579	0.053	0.316	0.632	0.053
se.healthstatus.2013.isps-https	6	0.333	0.167	0.500	0.000	0.667	0.333	0.000	0.667	0.333
se.healthstatus.2013.isps-https-www	10	0.300	0.200	0.500	0.100	0.500	0.400	0.000	0.700	0.300
se.healthstatus.2013.media-http	26	0.958	0.042	0.000	0.160	0.840	0.000	0.192	0.808	0.000
se.healthstatus.2013.media-http-www	28	0.960	0.040	0.000	0.148	0.852	0.000	0.179	0.821	0.000
se.healthstatus.2013.media-https	4	0.500	0.250	0.250	0.000	1.000	0.000	0.000	1.000	0.000
se.healthstatus.2013.media-https-www	5	0.400	0.400	0.200	0.000	1.000	0.000	0.000	1.000	0.000
se.healthstatus.2013.municipalities-http	249	0.992	0.000	0.008	0.498	0.494	0.008	0.518	0.474	0.008
se.healthstatus.2013.municipalities-http-www	271	0.989	0.000	0.011	0.519	0.469	0.012	0.542	0.446	0.011
se.healthstatus.2013.municipalities-https	44	0.273	0.205	0.523	0.268	0.561	0.171	0.227	0.568	0.205
se.healthstatus.2013.municipalities-https-www	54	0.259	0.130	0.611	0.160	0.600	0.240	0.148	0.611	0.241
se.healthstatus.2013.public-authorities-http	170	0.967	0.013	0.020	0.636	0.321	0.043	0.653	0.324	0.024
se.healthstatus.2013.public-authorities-http-www	203	0.973	0.005	0.022	0.649	0.314	0.037	0.675	0.305	0.020
se.healthstatus.2013.public-authorities-https	18	0.389	0.000	0.611	0.333	0.400	0.267	0.333	0.333	0.333
se.healthstatus.2013.public-authorities-https-www	37	0.216	0.135	0.649	0.139	0.444	0.417	0.162	0.514	0.324

Table C.6: Secure versus insecure resources coverage

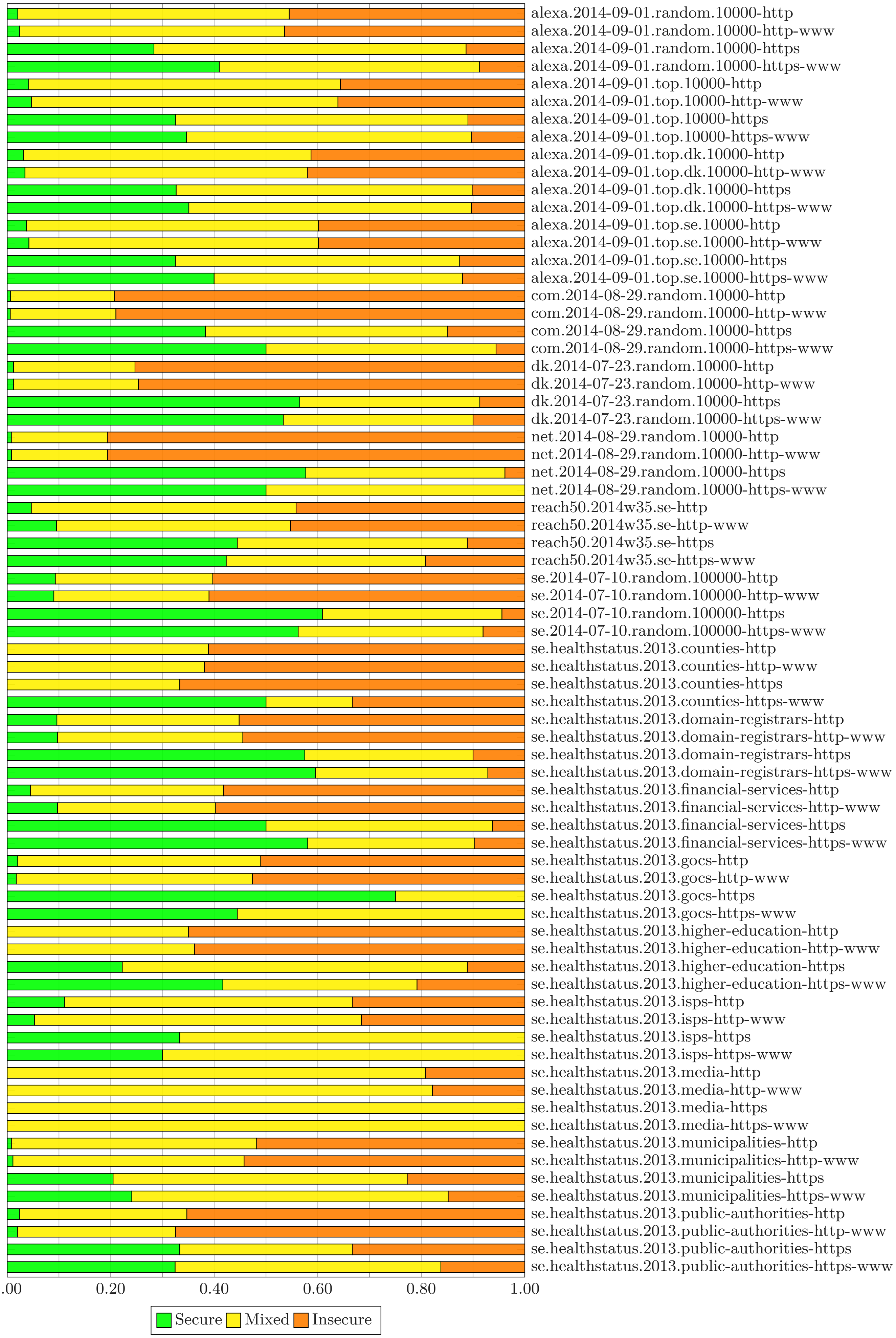

Figure C.4: Distribution of domains with strictly secure, mixed or strictly insecure resources

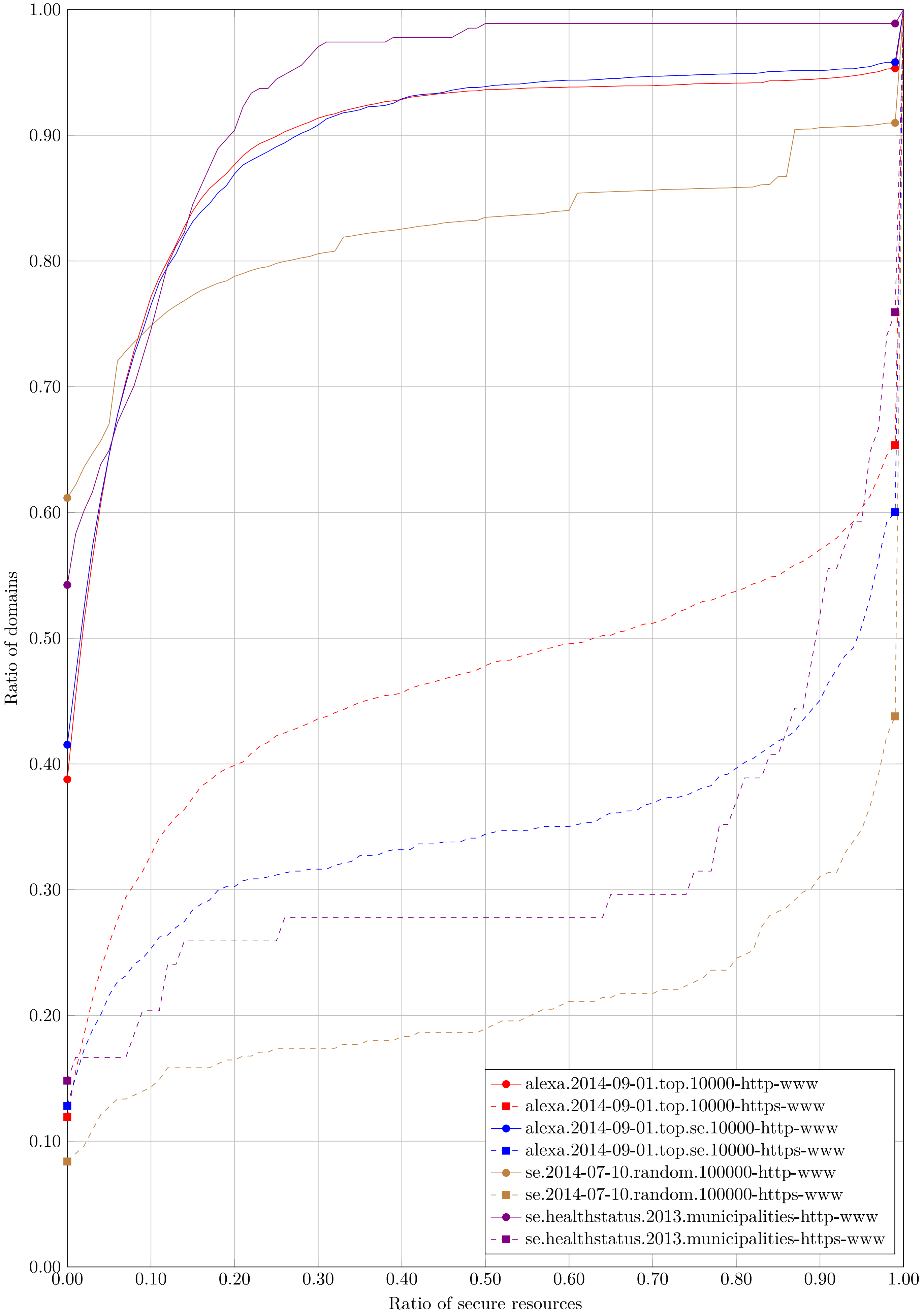

Figure C.5: Cumulative distribution of the ratio of secure resources per domain

C.8 HTTP, HTTPS and redirects

The table shows domains, domains with redirect responses to the origin request, the ratio of domains with redirects and the average length of the redirect chain per domain with redirects. Shown next is the ratio of redirected domains making strictly internal, domains mixing internal and external, and strictly external redirect URLs. The same goes for insecure, mixed security and strictly secure redirects – plus a column with the ratio of domains where the final redirect is to a secure URL. The last column shows the ratio of domains with mismatched redirect URLs without a subsequently requested URL.

Figure C.6 shows the distribution of strictly secure, mixed security and strictly insecure redirects (x axis) per dataset. The percentage of mismatched URLs is also shown. An additional mark shows the percentage of final secure redirects with an `x`, as not all mixed redirects lead to secure sites – sometimes they lead back to an insecure site.

Most domains that redirect make a single redirect, but every few sites make more than one; the average is around 1.23 redirects. With 49-71% of top sites in the HTTP variations redirecting, but only 30-36% of HTTP-www variations redirecting, it seems that the www subdomain still is in use, rather than not using any subdomain. For random domains numbers are a bit more even, with 24-31% of both HTTP and HTTP-www variations redirecting. More detailed data than in the below table clearly shows that domains generally pick either www or no subdomain, and redirect from one to the other; the www subdomain is most common as the final destination, especially for HTTPS variations.

A difference between top and random domains, is that top domains keep their redirects mostly internal, while random domains redirect elsewhere. Curated sites are in between, with a portion of HTTP-www variations redirecting externally. This seems to suggest that top sites has contents, while random domains to a larger extent do not – 39-60% of redirected domains end up being aliases for another domain than a user would have typed in.

When it comes to security, it is no surprise to see HTTP variations redirect mostly to other insecure URLs. The extent of domains implementing HTTPS but then redirecting to HTTP is more surprising – only in 23-35% of top sites will let you stay on a fully secure connection – and that number excludes mixed resource security in later stages of the browsing experience (C.7). An additional few percent of HTTPS domains even mix security usage during redirects and take a *detour* over an insecure URL in the process of redirecting the user to the final, secure destination – meaning that even if you type in a secure address *and* end up on a secure address, you may have passed through something completely insecure along the way. While this might be considered a corner case, it nullifies some of the security measures put in place by HTTPS, and could leak for example a carelessly set session identifier.

Here it is a surprise to see that financial institutions in the *.SE Health Status* domain lists do not take advantage of HTTPS, by redirecting users entering through HTTP – at less than 20% fully secured redirects and a surprising ratio of mixed redirects, they are at about the same level as general Swedish top sites. Even more surprising is that they even elect to redirect users away from HTTPS enabled pages to insecure variants in 60% of HTTPS-www domains. Counties, higher education and media are worse yet – all make no effort in redirecting users from HTTP to HTTPS.

Dataset	Domains	w/ R	DWR/D	R/DWR	I	Mix I+E	E	Insec	Mix sec	Sec	Final sec	Mism
alexa.2014-09-01.random.10000-http	8 216	4 030	0.491	1.137	0.912	0.009	0.057	0.918	0.015	0.045	0.058	0.023
alexa.2014-09-01.random.10000-http-www	8 493	2 571	0.303	1.134	0.858	0.011	0.096	0.872	0.011	0.083	0.090	0.036
alexa.2014-09-01.random.10000-https	1 135	717	0.632	1.262	0.803	0.004	0.162	0.632	0.075	0.265	0.294	0.031
alexa.2014-09-01.random.10000-https-www	1 224	470	0.384	1.232	0.923	0.009	0.036	0.615	0.047	0.304	0.321	0.034
alexa.2014-09-01.top.10000-http	8 545	6 077	0.711	1.203	0.935	0.011	0.036	0.905	0.019	0.058	0.075	0.018
alexa.2014-09-01.top.10000-http-www	8 682	2 702	0.311	1.243	0.865	0.011	0.087	0.789	0.024	0.150	0.171	0.038
alexa.2014-09-01.top.10000-https	2 507	1 856	0.740	1.318	0.952	0.009	0.023	0.602	0.088	0.292	0.314	0.018
alexa.2014-09-01.top.10000-https-www	2 957	1 536	0.519	1.243	0.923	0.011	0.035	0.688	0.056	0.227	0.245	0.031
alexa.2014-09-01.top.dk.10000-http	2 263	1 456	0.643	1.210	0.935	0.008	0.036	0.911	0.014	0.054	0.067	0.021
alexa.2014-09-01.top.dk.10000-http-www	2 310	833	0.361	1.194	0.872	0.011	0.073	0.833	0.017	0.106	0.121	0.044
alexa.2014-09-01.top.dk.10000-https	339	225	0.664	1.342	0.942	0.009	0.027	0.529	0.111	0.338	0.387	0.022
alexa.2014-09-01.top.dk.10000-https-www	441	232	0.526	1.276	0.892	0.013	0.039	0.651	0.043	0.250	0.276	0.056
alexa.2014-09-01.top.se.10000-http	2 797	1 738	0.621	1.162	0.948	0.004	0.032	0.906	0.012	0.065	0.077	0.017
alexa.2014-09-01.top.se.10000-http-www	2 895	863	0.298	1.146	0.898	0.002	0.079	0.819	0.010	0.149	0.160	0.021

Dataset	Domains	w/ R	DWR/D	R/DWR	I	Mix I+E	E	Insec	Mix sec	Sec	Final sec	Mism
alexa.2014-09-01.top.se.10000-https	438	282	0.644	1.273	0.968	0.000	0.014	0.564	0.071	0.348	0.379	0.018
alexa.2014-09-01.top.se.10000-https-www	650	268	0.412	1.216	0.937	0.000	0.022	0.653	0.037	0.269	0.276	0.041
com.2014-08-29.random.10000-http	7 775	2 419	0.311	1.205	0.571	0.015	0.387	0.931	0.027	0.014	0.039	0.028
com.2014-08-29.random.10000-http-www	7 811	1 968	0.252	1.227	0.467	0.007	0.493	0.918	0.033	0.017	0.046	0.034
com.2014-08-29.random.10000-https	50	24	0.480	1.333	0.917	0.000	0.000	0.583	0.083	0.250	0.250	0.083
com.2014-08-29.random.10000-https-www	55	23	0.418	1.217	0.913	0.000	0.043	0.522	0.087	0.348	0.348	0.043
dk.2014-07-23.random.10000-http	7 180	2 228	0.310	1.263	0.440	0.013	0.490	0.905	0.016	0.022	0.036	0.058
dk.2014-07-23.random.10000-http-www	7 378	2 025	0.274	1.298	0.365	0.010	0.558	0.893	0.018	0.023	0.039	0.068
dk.2014-07-23.random.10000-https	23	10	0.435	1.100	0.900	0.000	0.100	0.400	0.100	0.500	0.600	0.000
dk.2014-07-23.random.10000-https-www	32	15	0.469	1.267	0.733	0.000	0.067	0.333	0.067	0.400	0.467	0.200
net.2014-08-29.random.10000-http	7 270	2 108	0.290	1.237	0.467	0.015	0.496	0.932	0.035	0.012	0.046	0.021
net.2014-08-29.random.10000-http-www	7 378	1 874	0.254	1.259	0.377	0.003	0.594	0.918	0.041	0.015	0.055	0.026
net.2014-08-29.random.10000-https	26	7	0.269	1.000	0.857	0.000	0.000	0.429	0.000	0.429	0.429	0.143
net.2014-08-29.random.10000-https-www	28	8	0.286	1.000	0.750	0.000	0.125	0.625	0.000	0.250	0.250	0.125
reach50.2014w35.se-http	43	37	0.860	1.189	0.946	0.000	0.027	0.865	0.027	0.081	0.108	0.027
reach50.2014w35.se-http-www	42	12	0.286	1.250	0.833	0.000	0.083	0.583	0.000	0.333	0.333	0.083
reach50.2014w35.se-https	18	13	0.722	1.462	0.923	0.000	0.077	0.308	0.000	0.692	0.692	0.000
reach50.2014w35.se-https-www	26	17	0.654	1.059	0.882	0.000	0.059	0.471	0.000	0.471	0.471	0.059
se.2014-07-10.random.100000-http	73 605	21 181	0.288	1.240	0.502	0.012	0.436	0.915	0.015	0.020	0.034	0.051
se.2014-07-10.random.100000-http-www	77 261	18 765	0.243	1.263	0.410	0.004	0.534	0.909	0.016	0.023	0.039	0.053
se.2014-07-10.random.100000-https	282	125	0.443	1.320	0.872	0.008	0.056	0.256	0.056	0.624	0.640	0.064
se.2014-07-10.random.100000-https-www	328	115	0.351	1.270	0.878	0.017	0.070	0.417	0.035	0.513	0.522	0.035
se.healthstatus.2013.counties-http	18	4	0.222	1.250	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.counties-http-www	21	3	0.143	1.000	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.counties-https	3	2	0.667	1.000	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.counties-https-www	6	3	0.500	1.000	1.000	0.000	0.000	0.667	0.000	0.333	0.333	0.000
se.healthstatus.2013.domain-registrars-http	127	90	0.709	1.478	0.789	0.033	0.156	0.778	0.056	0.144	0.200	0.022
se.healthstatus.2013.domain-registrars-http-www	134	55	0.410	1.618	0.655	0.000	0.327	0.655	0.073	0.255	0.327	0.018
se.healthstatus.2013.domain-registrars-https	40	28	0.700	1.214	0.893	0.071	0.036	0.321	0.071	0.607	0.607	0.000
se.healthstatus.2013.domain-registrars-https-www	42	14	0.333	1.214	0.857	0.000	0.143	0.500	0.071	0.429	0.429	0.000
se.healthstatus.2013.financial-services-http	67	53	0.791	1.453	0.868	0.038	0.094	0.830	0.113	0.057	0.170	0.000
se.healthstatus.2013.financial-services-http-www	72	36	0.500	1.194	0.750	0.028	0.222	0.750	0.056	0.194	0.250	0.000
se.healthstatus.2013.financial-services-https	16	9	0.563	1.222	0.889	0.000	0.111	0.222	0.000	0.778	0.778	0.000
se.healthstatus.2013.financial-services-https-www	31	10	0.323	1.300	0.900	0.000	0.100	0.500	0.100	0.400	0.400	0.000
se.healthstatus.2013.gocs-http	49	35	0.714	1.286	0.857	0.000	0.114	0.943	0.000	0.029	0.029	0.029
se.healthstatus.2013.gocs-http-www	57	14	0.246	1.571	0.500	0.000	0.429	0.857	0.000	0.071	0.071	0.071
se.healthstatus.2013.gocs-https	4	3	0.750	1.000	1.000	0.000	0.000	0.333	0.000	0.667	0.667	0.000
se.healthstatus.2013.gocs-https-www	9	2	0.222	2.000	1.000	0.000	0.000	0.000	0.000	1.000	1.000	0.000
se.healthstatus.2013.higher-education-http	40	26	0.650	1.269	0.962	0.000	0.038	1.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.higher-education-http-www	47	10	0.213	1.200	0.900	0.000	0.100	1.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.higher-education-https	9	7	0.778	1.714	1.000	0.000	0.000	0.429	0.286	0.286	0.286	0.000

Dataset	Domains	w/ R	DWR/D	R/DWR	I	Mix I+E	E	Insec	Mix sec	Sec	Final sec	Mism
se.healthstatus.2013.higher-education-https-www	24	8	0.333	1.250	1.000	0.000	0.000	0.875	0.000	0.125	0.125	0.000
se.healthstatus.2013.isps-http	18	14	0.778	1.286	1.000	0.000	0.000	0.857	0.000	0.143	0.143	0.000
se.healthstatus.2013.isps-http-www	19	7	0.368	1.143	1.000	0.000	0.000	0.714	0.000	0.286	0.286	0.000
se.healthstatus.2013.isps-https	6	4	0.667	1.000	1.000	0.000	0.000	0.500	0.000	0.500	0.500	0.000
se.healthstatus.2013.isps-https-www	10	2	0.200	1.000	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.media-http	26	22	0.846	1.045	0.864	0.045	0.091	1.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.media-http-www	28	10	0.357	1.100	0.700	0.000	0.300	1.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.media-https	4	2	0.500	1.500	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.media-https-www	5	3	0.600	1.000	1.000	0.000	0.000	0.667	0.000	0.333	0.333	0.000
se.healthstatus.2013.municipalities-http	249	73	0.293	1.055	0.986	0.014	0.000	0.973	0.000	0.027	0.027	0.000
se.healthstatus.2013.municipalities-http-www	271	26	0.096	1.077	0.923	0.000	0.038	0.846	0.000	0.115	0.115	0.038
se.healthstatus.2013.municipalities-https	44	23	0.523	1.304	1.000	0.000	0.000	0.522	0.130	0.348	0.348	0.000
se.healthstatus.2013.municipalities-https-www	54	20	0.370	1.050	1.000	0.000	0.000	0.700	0.000	0.300	0.300	0.000
se.healthstatus.2013.public-authorities-http	170	84	0.494	1.143	0.810	0.012	0.179	0.940	0.024	0.036	0.060	0.000
se.healthstatus.2013.public-authorities-http-www	203	54	0.266	1.222	0.630	0.000	0.370	0.889	0.019	0.093	0.111	0.000
se.healthstatus.2013.public-authorities-https	18	9	0.500	1.111	1.000	0.000	0.000	0.778	0.000	0.222	0.222	0.000
se.healthstatus.2013.public-authorities-https-www	37	12	0.324	1.083	1.000	0.000	0.000	0.667	0.000	0.333	0.333	0.000

Table C.7: Origin domains with redirects

Analyzing domains redirecting to secure URLs would be a good candidate for a refined selection (7.5.6). Do the redirects help, or is the net result that insecure resources are loaded anyways?

The mismatched redirect and request URLs are in part due to the HAR standard not defining recorded redirect URLs as strictly absolute, and phantomjs returning unparsed/unresolved URLs when a redirect is initially detected in an HTTP response (5.8.1). Resolving redirect URLs outside of the browser means not all contexts and rules are considered, thus leading to errors. Both the thesis code, phantomjs software and HAR standard can be improved upon.

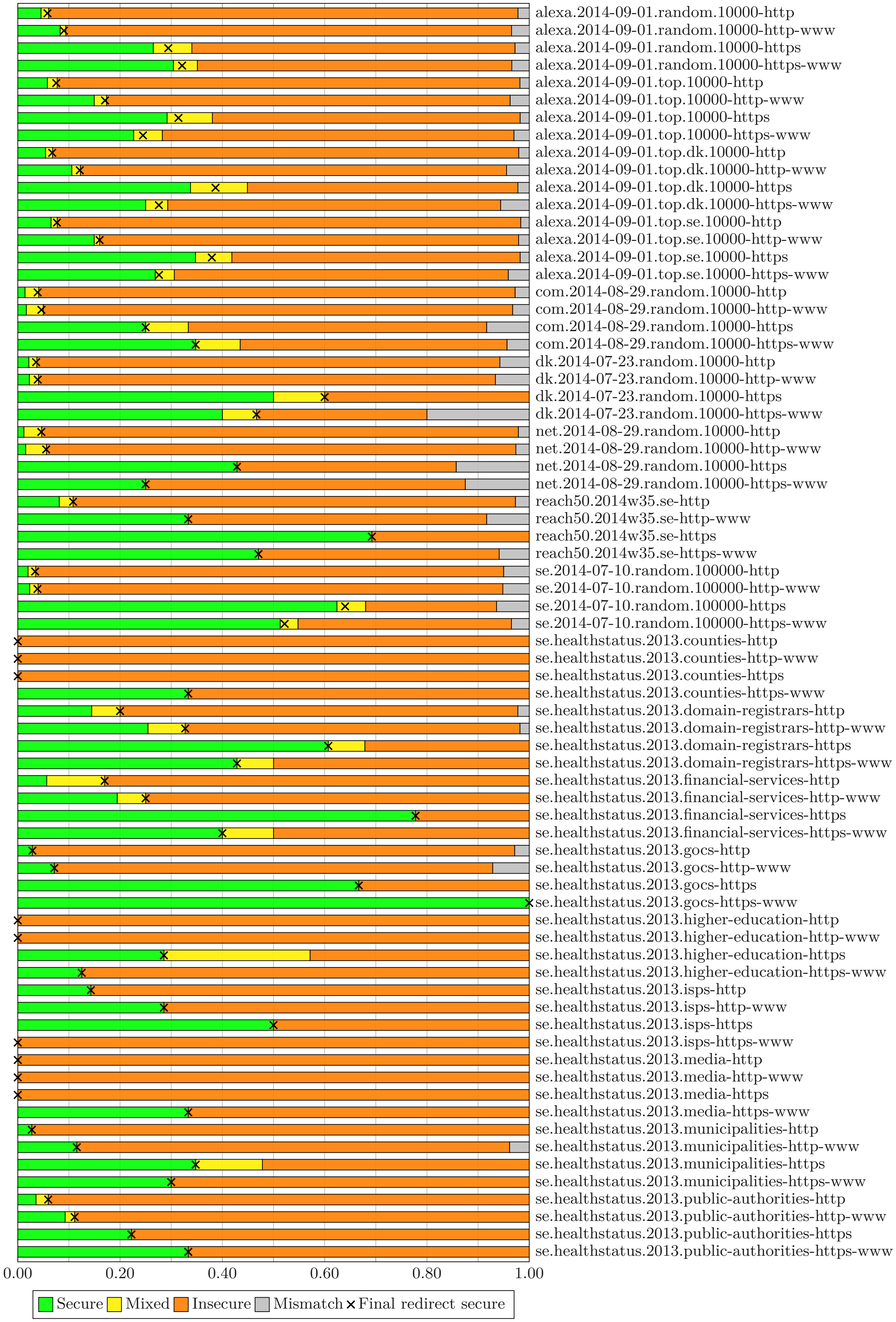

Figure C.6: Distribution of domains with strictly secure, mixed or strictly insecure redirects

C.9 Content type group coverage

Bytes sent from the server to the browser generally has an associated type, so the browser can parse and use them properly. The types can be grouped (A.5.3); incorrect or unknown types that did not match a group is shown as `(null)` below. The difference between what can be achieved between different types of resources makes the distribution interesting. Images [4] and text[5] loaded by a browser generally provide no additional way to load further resources, while html, scripts and styles do. While data resources can trigger downloading additional resources based on the logic that consumes the data, it still requires another type of resource present to do that.

Objects and external documents can also access additional resources, but the use of those types of resources has been very low in the extracted data. There might be several reasons, but the fact that the tests were run on a headless browser without additional plugins installed is probably the biggest in this case. An additional reason might be adoption of HTML5 and client side javascript instead of Flash for visual, dynamic material and animations. This evolution has been fueled by Apple's resistance towards supporting Flash on their handheld devices[6].

Note that web fonts have fairly low numbers here, but that they can also be served as styles which dynamically load additional fonts URLs. This is how Google Fonts do it, using the fonts.googleapis.com domain (5.4) to serve styles and gstatic.com to serve fonts dynamically selected to match to browser web font compatibility level, which could then be factored into these numbers (C.11.2).

C.9.1 Origin

Practically all successful origin requests result in a html response. The range is 84-100% html, with the difference being seemingly misconfigured responses, part of which are redirects without actual content (C.8).

C.9.2 Internal

The table below shows internal request ratios for domains with at least one internal request, excluding the origin request.

Dataset	Domains w/ int	html	script	style	image	data	text	font	object	document	(null)
alexa.2014-09-01.random.10000-http	7 591	0.610	0.812	0.872	0.927	0.070	0.078	0.024	0.003	0.000	0.378
alexa.2014-09-01.random.10000-http-www	7 825	0.475	0.825	0.885	0.939	0.070	0.077	0.024	0.004	0.000	0.382
alexa.2014-09-01.random.10000-https	1 072	0.669	0.806	0.784	0.946	0.105	0.084	0.030	0.002	0.000	0.409
alexa.2014-09-01.random.10000-https-www	1 139	0.572	0.887	0.901	0.943	0.116	0.110	0.038	0.003	0.000	0.428
alexa.2014-09-01.top.10000-http	8 176	0.818	0.816	0.770	0.872	0.186	0.152	0.041	0.004	0.000	0.375
alexa.2014-09-01.top.10000-http-www	8 289	0.627	0.825	0.780	0.883	0.189	0.152	0.042	0.004	0.000	0.380
alexa.2014-09-01.top.10000-https	2 369	0.832	0.795	0.710	0.851	0.209	0.157	0.048	0.002	0.000	0.382
alexa.2014-09-01.top.10000-https-www	2 801	0.720	0.821	0.741	0.878	0.211	0.179	0.051	0.002	0.000	0.418
alexa.2014-09-01.top.dk.10000-http	2 136	0.737	0.883	0.918	0.937	0.118	0.080	0.047	0.002	0.000	0.361
alexa.2014-09-01.top.dk.10000-http-www	2 182	0.541	0.891	0.926	0.944	0.120	0.083	0.048	0.002	0.000	0.366
alexa.2014-09-01.top.dk.10000-https	316	0.775	0.932	0.914	0.945	0.191	0.120	0.095	0.003	0.000	0.415
alexa.2014-09-01.top.dk.10000-https-www	406	0.672	0.927	0.906	0.932	0.158	0.099	0.078	0.002	0.000	0.455
alexa.2014-09-01.top.se.10000-http	2 684	0.716	0.873	0.909	0.926	0.144	0.114	0.041	0.001	0.000	0.380
alexa.2014-09-01.top.se.10000-http-www	2 779	0.489	0.884	0.916	0.932	0.144	0.112	0.041	0.002	0.000	0.374
alexa.2014-09-01.top.se.10000-https	422	0.752	0.923	0.911	0.956	0.232	0.103	0.056	0.000	0.000	0.454

[4] Scalable Vector Graphics (SVG) images can load resources.http://www.w3.org/TR/SVG/

[5] Unless improperly labeled during transfer and parsed as another format.

[6] https://www.apple.com/hotnews/thoughts-on-flash/

Dataset	Domains w/ int	html	script	style	image	data	text	font	object	document	(null)
alexa.2014-09-01.top.se.10000-https-www	630	0.568	0.923	0.917	0.950	0.215	0.101	0.064	0.000	0.000	0.445
com.2014-08-29.random.10000-http	6 222	0.579	0.381	0.465	0.806	0.029	0.022	0.004	0.003	0.000	0.194
com.2014-08-29.random.10000-http-www	6 241	0.517	0.392	0.478	0.814	0.030	0.023	0.004	0.003	0.000	0.191
com.2014-08-29.random.10000-https	41	0.578	0.867	0.956	0.956	0.067	0.089	0.067	0.000	0.000	0.444
com.2014-08-29.random.10000-https-www	43	0.519	0.778	0.944	0.926	0.074	0.037	0.056	0.000	0.000	0.389
dk.2014-07-23.random.10000-http	4 626	0.334	0.525	0.657	0.918	0.056	0.033	0.006	0.002	0.000	0.153
dk.2014-07-23.random.10000-http-www	4 773	0.279	0.527	0.653	0.912	0.056	0.030	0.006	0.002	0.000	0.152
dk.2014-07-23.random.10000-https	16	0.545	0.864	0.955	0.909	0.091	0.000	0.227	0.000	0.000	0.182
dk.2014-07-23.random.10000-https-www	22	0.655	0.828	0.931	0.931	0.103	0.034	0.172	0.000	0.000	0.207
net.2014-08-29.random.10000-http	5 757	0.605	0.317	0.407	0.815	0.019	0.016	0.003	0.002	0.000	0.172
net.2014-08-29.random.10000-http-www	5 839	0.556	0.329	0.421	0.825	0.019	0.015	0.002	0.001	0.000	0.176
net.2014-08-29.random.10000-https	16	0.423	0.692	0.885	0.885	0.077	0.038	0.038	0.000	0.000	0.385
net.2014-08-29.random.10000-https-www	20	0.480	0.800	0.920	1.000	0.120	0.000	0.040	0.000	0.000	0.480
reach50.2014w35.se-http	43	0.951	0.732	0.585	0.878	0.195	0.463	0.049	0.000	0.000	0.366
reach50.2014w35.se-http-www	42	0.641	0.744	0.538	0.872	0.179	0.436	0.051	0.000	0.000	0.487
reach50.2014w35.se-https	17	0.938	0.500	0.438	0.563	0.188	0.125	0.000	0.000	0.000	0.313
reach50.2014w35.se-https-www	26	0.870	0.609	0.478	0.739	0.087	0.174	0.087	0.000	0.000	0.348
se.2014-07-10.random.100000-http	54 882	0.399	0.580	0.716	0.870	0.049	0.046	0.006	0.001	0.000	0.175
se.2014-07-10.random.100000-http-www	57 547	0.336	0.585	0.710	0.867	0.048	0.045	0.006	0.001	0.000	0.173
se.2014-07-10.random.100000-https	226	0.612	0.871	0.905	0.954	0.129	0.065	0.038	0.000	0.000	0.365
se.2014-07-10.random.100000-https-www	285	0.534	0.871	0.913	0.955	0.154	0.074	0.039	0.000	0.000	0.373
se.healthstatus.2013.counties-http	18	0.278	1.000	1.000	1.000	0.111	0.000	0.056	0.000	0.000	0.389
se.healthstatus.2013.counties-http-www	21	0.250	1.000	1.000	1.000	0.100	0.000	0.050	0.000	0.000	0.350
se.healthstatus.2013.counties-https	3	0.667	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.333
se.healthstatus.2013.counties-https-www	5	0.500	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.333
se.healthstatus.2013.domain-registrars-http	108	0.769	0.870	0.926	0.954	0.093	0.111	0.009	0.000	0.000	0.269
se.healthstatus.2013.domain-registrars-http-www	113	0.474	0.877	0.921	0.974	0.079	0.105	0.009	0.000	0.000	0.298
se.healthstatus.2013.domain-registrars-https	34	0.872	0.872	0.897	0.923	0.103	0.128	0.026	0.000	0.000	0.308
se.healthstatus.2013.domain-registrars-https-www	36	0.600	0.900	1.000	0.975	0.150	0.075	0.025	0.000	0.000	0.400
se.healthstatus.2013.financial-services-http	67	0.820	0.951	0.951	0.951	0.197	0.049	0.082	0.000	0.000	0.328
se.healthstatus.2013.financial-services-http-www	71	0.516	0.953	0.969	0.953	0.203	0.047	0.063	0.000	0.000	0.313
se.healthstatus.2013.financial-services-https	16	0.667	1.000	1.000	1.000	0.200	0.133	0.067	0.000	0.000	0.533
se.healthstatus.2013.financial-services-https-www	31	0.433	1.000	1.000	1.000	0.300	0.100	0.100	0.000	0.000	0.467
se.healthstatus.2013.gocs-http	48	0.750	0.977	0.977	0.977	0.136	0.136	0.091	0.000	0.000	0.295
se.healthstatus.2013.gocs-http-www	55	0.300	0.960	0.960	0.980	0.120	0.120	0.100	0.000	0.000	0.320
se.healthstatus.2013.gocs-https	4	1.000	1.000	1.000	1.000	0.000	0.000	0.250	0.000	0.000	0.500
se.healthstatus.2013.gocs-https-www	9	0.444	0.889	0.889	1.000	0.222	0.111	0.111	0.000	0.000	0.667
se.healthstatus.2013.higher-education-http	38	0.718	0.949	0.974	0.974	0.205	0.128	0.077	0.000	0.000	0.256
se.healthstatus.2013.higher-education-http-www	44	0.326	0.957	0.978	1.000	0.174	0.109	0.065	0.000	0.000	0.283
se.healthstatus.2013.higher-education-https	9	0.889	1.000	1.000	1.000	0.333	0.222	0.000	0.000	0.000	0.222
se.healthstatus.2013.higher-education-https-www	24	0.500	0.958	1.000	1.000	0.208	0.125	0.083	0.000	0.000	0.292

Dataset	Domains w/ int	html	script	style	image	data	text	font	object	document	(null)
se.healthstatus.2013.isps-http	18	0.882	0.941	1.000	1.000	0.176	0.059	0.059	0.000	0.000	0.529
se.healthstatus.2013.isps-http-www	19	0.632	0.895	0.947	0.947	0.211	0.053	0.053	0.000	0.000	0.368
se.healthstatus.2013.isps-https	6	0.833	1.000	1.000	1.000	0.167	0.000	0.167	0.000	0.000	0.500
se.healthstatus.2013.isps-https-www	10	0.500	1.000	1.000	1.000	0.200	0.100	0.100	0.000	0.000	0.500
se.healthstatus.2013.media-http	25	0.958	0.958	0.917	0.958	0.208	0.583	0.167	0.000	0.000	0.500
se.healthstatus.2013.media-http-www	27	0.680	1.000	0.920	1.000	0.320	0.640	0.200	0.000	0.000	0.560
se.healthstatus.2013.media-https	4	0.750	1.000	0.750	1.000	0.000	0.500	0.000	0.000	0.000	1.000
se.healthstatus.2013.media-https-www	5	1.000	1.000	1.000	1.000	0.200	0.600	0.200	0.000	0.000	0.800
se.healthstatus.2013.municipalities-http	239	0.402	0.968	0.992	0.996	0.040	0.020	0.012	0.000	0.000	0.281
se.healthstatus.2013.municipalities-http-www	258	0.215	0.959	0.993	1.000	0.033	0.019	0.011	0.000	0.000	0.315
se.healthstatus.2013.municipalities-https	41	0.614	0.955	1.000	1.000	0.000	0.023	0.045	0.000	0.000	0.318
se.healthstatus.2013.municipalities-https-www	50	0.500	0.944	1.000	1.000	0.019	0.037	0.056	0.000	0.000	0.278
se.healthstatus.2013.public-authorities-http	162	0.529	0.954	0.980	1.000	0.085	0.052	0.020	0.000	0.000	0.242
se.healthstatus.2013.public-authorities-http-www	188	0.313	0.940	0.978	1.000	0.071	0.049	0.022	0.000	0.000	0.225
se.healthstatus.2013.public-authorities-https	15	0.556	0.944	1.000	0.944	0.111	0.056	0.056	0.000	0.000	0.278
se.healthstatus.2013.public-authorities-https-www	36	0.432	0.973	1.000	1.000	0.108	0.081	0.027	0.000	0.000	0.324

Table C.8: Content type group coverage (internal)

C.9.3 External

The table below shows external resources from each group enjoy almost the same coverage as their internal counterparts. Among non-zone datasets scripts often reach above 90% coverage, showing that active and popular web pages contain a lot of external dynamic material. Images, while not dynamic, as well as styles and html are also popular to load externally.

Dataset	Domains w/ ext	html	script	style	image	data	text	font	object	document	(null)
alexa.2014-09-01.random.10000-http	7 591	0.662	0.948	0.592	0.927	0.125	0.083	0.107	0.002	0.000	0.556
alexa.2014-09-01.random.10000-http-www	7 825	0.657	0.948	0.586	0.927	0.120	0.087	0.108	0.002	0.000	0.547
alexa.2014-09-01.random.10000-https	1 072	0.664	0.967	0.628	0.935	0.152	0.115	0.136	0.002	0.000	0.611
alexa.2014-09-01.random.10000-https-www	1 139	0.578	0.961	0.516	0.924	0.130	0.095	0.067	0.002	0.000	0.572
alexa.2014-09-01.top.10000-http	8 176	0.761	0.976	0.578	0.965	0.232	0.214	0.119	0.004	0.000	0.674
alexa.2014-09-01.top.10000-http-www	8 289	0.757	0.974	0.573	0.964	0.228	0.217	0.118	0.004	0.000	0.670
alexa.2014-09-01.top.10000-https	2 369	0.688	0.975	0.555	0.952	0.218	0.180	0.125	0.005	0.000	0.697
alexa.2014-09-01.top.10000-https-www	2 801	0.677	0.968	0.527	0.950	0.206	0.199	0.099	0.004	0.000	0.699
alexa.2014-09-01.top.dk.10000-http	2 136	0.636	0.968	0.574	0.932	0.128	0.107	0.079	0.001	0.000	0.524
alexa.2014-09-01.top.dk.10000-http-www	2 182	0.637	0.965	0.577	0.935	0.132	0.113	0.077	0.000	0.000	0.528
alexa.2014-09-01.top.dk.10000-https	316	0.611	0.968	0.525	0.940	0.136	0.139	0.085	0.000	0.000	0.589
alexa.2014-09-01.top.dk.10000-https-www	406	0.631	0.978	0.512	0.919	0.155	0.167	0.059	0.000	0.000	0.586
alexa.2014-09-01.top.se.10000-http	2 684	0.640	0.963	0.592	0.936	0.153	0.112	0.075	0.001	0.000	0.553
alexa.2014-09-01.top.se.10000-http-www	2 779	0.645	0.963	0.584	0.937	0.158	0.098	0.075	0.001	0.000	0.557
alexa.2014-09-01.top.se.10000-https	422	0.585	0.972	0.519	0.938	0.111	0.118	0.071	0.000	0.000	0.654

Dataset	Domains w/ ext	html	script	style	image	data	text	font	object	document	(null)
alexa.2014-09-01.top.se.10000-https-www	630	0.565	0.978	0.497	0.929	0.138	0.110	0.065	0.002	0.000	0.627
com.2014-08-29.random.10000-http	6 222	0.732	0.881	0.505	0.902	0.066	0.048	0.034	0.002	0.000	0.279
com.2014-08-29.random.10000-http-www	6 241	0.728	0.883	0.509	0.904	0.067	0.050	0.036	0.001	0.000	0.277
com.2014-08-29.random.10000-https	41	0.463	0.829	0.512	0.756	0.122	0.000	0.073	0.000	0.000	0.512
com.2014-08-29.random.10000-https-www	43	0.488	0.884	0.558	0.837	0.140	0.047	0.070	0.000	0.000	0.535
dk.2014-07-23.random.10000-http	4 626	0.569	0.797	0.654	0.805	0.073	0.041	0.038	0.001	0.000	0.309
dk.2014-07-23.random.10000-http-www	4 773	0.565	0.790	0.646	0.810	0.072	0.043	0.039	0.001	0.000	0.314
dk.2014-07-23.random.10000-https	16	0.500	0.938	0.438	0.875	0.000	0.063	0.063	0.000	0.000	0.438
dk.2014-07-23.random.10000-https-www	22	0.500	1.000	0.455	0.955	0.091	0.045	0.000	0.000	0.000	0.455
net.2014-08-29.random.10000-http	5 757	0.787	0.872	0.455	0.913	0.053	0.056	0.029	0.001	0.000	0.251
net.2014-08-29.random.10000-http-www	5 839	0.785	0.872	0.460	0.917	0.053	0.057	0.029	0.001	0.000	0.257
net.2014-08-29.random.10000-https	16	0.375	0.875	0.563	0.813	0.000	0.000	0.063	0.000	0.000	0.375
net.2014-08-29.random.10000-https-www	20	0.500	0.850	0.550	0.850	0.000	0.050	0.050	0.000	0.000	0.500
reach50.2014w35.se-http	43	0.744	0.953	0.651	0.953	0.209	0.140	0.093	0.000	0.000	0.628
reach50.2014w35.se-http-www	42	0.714	0.929	0.619	0.952	0.238	0.167	0.095	0.000	0.000	0.619
reach50.2014w35.se-https	17	0.588	0.941	0.647	0.882	0.235	0.000	0.118	0.000	0.000	0.706
reach50.2014w35.se-https-www	26	0.462	1.000	0.615	0.962	0.154	0.038	0.077	0.000	0.000	0.654
se.2014-07-10.random.100000-http	54 882	0.568	0.849	0.740	0.849	0.068	0.039	0.036	0.001	0.000	0.327
se.2014-07-10.random.100000-http-www	57 547	0.559	0.838	0.728	0.853	0.069	0.039	0.037	0.001	0.000	0.329
se.2014-07-10.random.100000-https	226	0.420	0.925	0.438	0.836	0.075	0.027	0.040	0.000	0.000	0.482
se.2014-07-10.random.100000-https-www	285	0.400	0.930	0.435	0.849	0.081	0.049	0.032	0.000	0.000	0.439
se.healthstatus.2013.counties-http	18	0.278	1.000	0.556	0.944	0.000	0.000	0.000	0.000	0.000	0.500
se.healthstatus.2013.counties-http-www	21	0.333	1.000	0.571	0.952	0.000	0.000	0.000	0.000	0.000	0.524
se.healthstatus.2013.counties-https	3	0.000	1.000	0.333	1.000	0.000	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.counties-https-www	5	0.200	1.000	0.600	1.000	0.000	0.000	0.000	0.000	0.000	0.400
se.healthstatus.2013.domain-registrars-http	108	0.472	0.917	0.519	0.917	0.093	0.056	0.065	0.000	0.000	0.296
se.healthstatus.2013.domain-registrars-http-www	113	0.478	0.903	0.531	0.894	0.088	0.062	0.053	0.000	0.000	0.292
se.healthstatus.2013.domain-registrars-https	34	0.529	0.941	0.500	0.971	0.147	0.147	0.088	0.000	0.000	0.500
se.healthstatus.2013.domain-registrars-https-www	36	0.361	0.944	0.472	0.972	0.139	0.167	0.056	0.000	0.000	0.333
se.healthstatus.2013.financial-services-http	67	0.433	0.896	0.343	0.940	0.164	0.119	0.015	0.000	0.000	0.343
se.healthstatus.2013.financial-services-http-www	71	0.423	0.873	0.366	0.944	0.127	0.127	0.014	0.000	0.000	0.366
se.healthstatus.2013.financial-services-https	16	0.125	0.875	0.313	0.938	0.125	0.125	0.000	0.000	0.000	0.500
se.healthstatus.2013.financial-services-https-www	31	0.323	0.935	0.290	0.935	0.226	0.161	0.032	0.000	0.000	0.484
se.healthstatus.2013.gocs-http	48	0.521	0.979	0.417	0.938	0.083	0.021	0.042	0.000	0.000	0.292
se.healthstatus.2013.gocs-http-www	55	0.509	0.982	0.436	0.945	0.127	0.036	0.055	0.000	0.000	0.291
se.healthstatus.2013.gocs-https	4	0.500	0.750	0.250	0.750	0.250	0.250	0.500	0.000	0.000	0.750
se.healthstatus.2013.gocs-https-www	9	0.556	1.000	0.556	0.889	0.111	0.111	0.333	0.000	0.000	0.667
se.healthstatus.2013.higher-education-http	38	0.447	0.974	0.474	0.947	0.026	0.000	0.026	0.000	0.000	0.368
se.healthstatus.2013.higher-education-http-www	44	0.432	0.955	0.477	0.909	0.023	0.000	0.023	0.000	0.000	0.386
se.healthstatus.2013.higher-education-https	9	0.556	1.000	0.444	1.000	0.000	0.000	0.000	0.000	0.000	0.556
se.healthstatus.2013.higher-education-https-www	24	0.542	1.000	0.333	0.958	0.000	0.000	0.042	0.000	0.000	0.500

Dataset	Domains w/ ext	html	script	style	image	data	text	font	object	document	(null)
se.healthstatus.2013.isps-http	18	0.556	1.000	0.389	0.944	0.278	0.111	0.000	0.000	0.000	0.389
se.healthstatus.2013.isps-http-www	19	0.632	1.000	0.368	1.000	0.211	0.158	0.000	0.000	0.000	0.526
se.healthstatus.2013.isps-https	6	1.000	1.000	0.333	1.000	0.333	0.167	0.000	0.000	0.000	0.667
se.healthstatus.2013.isps-https-www	10	0.500	1.000	0.300	0.900	0.400	0.200	0.000	0.000	0.000	0.500
se.healthstatus.2013.media-http	25	0.960	1.000	0.760	1.000	0.440	0.600	0.000	0.000	0.000	0.880
se.healthstatus.2013.media-http-www	27	0.963	1.000	0.815	1.000	0.444	0.556	0.000	0.000	0.000	0.889
se.healthstatus.2013.media-https	4	1.000	1.000	1.000	1.000	1.000	0.250	0.000	0.000	0.000	1.000
se.healthstatus.2013.media-https-www	5	1.000	1.000	0.800	1.000	0.400	0.200	0.000	0.000	0.000	1.000
se.healthstatus.2013.municipalities-http	239	0.481	0.979	0.661	0.967	0.063	0.013	0.021	0.000	0.000	0.393
se.healthstatus.2013.municipalities-http-www	258	0.469	0.977	0.655	0.950	0.066	0.012	0.019	0.000	0.000	0.364
se.healthstatus.2013.municipalities-https	41	0.390	1.000	0.610	0.951	0.024	0.000	0.000	0.000	0.000	0.463
se.healthstatus.2013.municipalities-https-www	50	0.360	1.000	0.640	0.940	0.040	0.000	0.000	0.000	0.000	0.460
se.healthstatus.2013.public-authorities-http	162	0.346	0.975	0.432	0.932	0.043	0.019	0.012	0.000	0.000	0.321
se.healthstatus.2013.public-authorities-http-www	188	0.346	0.989	0.415	0.920	0.043	0.016	0.011	0.000	0.000	0.314
se.healthstatus.2013.public-authorities-https	15	0.200	1.000	0.200	0.867	0.067	0.000	0.000	0.000	0.000	0.400
se.healthstatus.2013.public-authorities-https-www	36	0.306	0.972	0.306	0.889	0.056	0.028	0.000	0.000	0.000	0.472

Table C.9: Content type group coverage (external)

As some external file types can trigger further HTTP requests, it might be possible to build a hierarchy of requests. The easiest way is to look at resources loaded in HTML <`iframe`> (or the now less popular <`frame`>) as they will have an HTTP `referer` header set to the URL of the frame. Scripts and styles requesting other resources directly, without the use of frames, cannot be detected as a strict hierarchy. With the large number of requests made from different sites, URLs can be cleaned up (for example removing unique identifiers or looking at domain parts only) and connected in a graph and analyzed for similarities.

C.10 Public suffix coverage

Resources served from external URLs may well come from other public suffixes; here they have been grouped by TLD. The connections between datasets and TLDs is interesting; .se datasets load more from .se domains than others, and the equivalent is valid for .dk datasets. We can also see that despite Alexa's top 10,000 being an international list, nearly 19% of them use resources are loaded from .se domains. This points towards those sites being aware of the country of origin for the request, leading to localized content being served. It is also evident that the .com TLD is the most widespread for external resources – it beats same-TLD coverage.

Dataset	Domains w/ ext	se	dk	com	net	org	nu	uk	de	ru	jp	cn	br	fr
alexa.2014-09-01.random.10000-http	7 591	0.093	0.001	0.974	0.566	0.052	0.000	0.009	0.038	0.073	0.027	0.016	0.009	0.007
alexa.2014-09-01.random.10000-http-www	7 825	0.095	0.001	0.974	0.557	0.051	0.000	0.010	0.039	0.070	0.027	0.018	0.009	0.007
alexa.2014-09-01.random.10000-https	1 072	0.235	0.001	0.987	0.649	0.039	0.000	0.006	0.049	0.030	0.024	0.000	0.006	0.005
alexa.2014-09-01.random.10000-https-www	1 139	0.183	0.001	0.978	0.621	0.032	0.000	0.006	0.058	0.017	0.030	0.000	0.007	0.007
alexa.2014-09-01.top.10000-http	8 176	0.187	0.001	0.978	0.727	0.057	0.001	0.008	0.070	0.070	0.047	0.041	0.010	0.016
alexa.2014-09-01.top.10000-http-www	8 289	0.197	0.001	0.976	0.727	0.060	0.001	0.009	0.069	0.071	0.051	0.042	0.009	0.014
alexa.2014-09-01.top.10000-https	2 369	0.231	0.002	0.976	0.762	0.051	0.001	0.007	0.065	0.074	0.034	0.002	0.008	0.007

Dataset	Domains w/ ext	se	dk	com	net	org	nu	uk	de	ru	jp	cn	br	fr
alexa.2014-09-01.top.10000-https-www	2 801	0.239	0.002	0.970	0.752	0.046	0.000	0.011	0.072	0.057	0.032	0.003	0.007	0.010
alexa.2014-09-01.top.dk.10000-http	2 136	0.183	0.280	0.974	0.635	0.025	0.000	0.002	0.046	0.002	0.002	0.000	0.000	0.001
alexa.2014-09-01.top.dk.10000-http-www	2 182	0.183	0.291	0.973	0.633	0.025	0.000	0.001	0.044	0.002	0.001	0.000	0.000	0.001
alexa.2014-09-01.top.dk.10000-https	316	0.269	0.263	0.965	0.680	0.025	0.000	0.003	0.101	0.000	0.000	0.000	0.000	0.003
alexa.2014-09-01.top.dk.10000-https-www	406	0.286	0.249	0.956	0.690	0.025	0.000	0.002	0.079	0.000	0.000	0.000	0.000	0.002
alexa.2014-09-01.top.se.10000-http	2 684	0.390	0.033	0.980	0.612	0.028	0.021	0.003	0.069	0.002	0.003	0.000	0.000	0.002
alexa.2014-09-01.top.se.10000-http-www	2 779	0.388	0.032	0.979	0.610	0.026	0.021	0.003	0.055	0.003	0.002	0.000	0.000	0.002
alexa.2014-09-01.top.se.10000-https	422	0.370	0.040	0.986	0.711	0.024	0.017	0.000	0.085	0.002	0.000	0.000	0.002	0.002
alexa.2014-09-01.top.se.10000-https-www	630	0.389	0.030	0.983	0.687	0.030	0.013	0.000	0.076	0.002	0.000	0.000	0.002	0.003
com.2014-08-29.random.10000-http	6 222	0.021	0.000	0.954	0.572	0.019	0.000	0.018	0.016	0.004	0.009	0.011	0.003	0.005
com.2014-08-29.random.10000-http-www	6 241	0.022	0.000	0.952	0.563	0.018	0.000	0.019	0.017	0.004	0.008	0.012	0.003	0.006
com.2014-08-29.random.10000-https	41	0.098	0.000	1.000	0.366	0.000	0.000	0.000	0.000	0.024	0.024	0.000	0.000	0.000
com.2014-08-29.random.10000-https-www	43	0.070	0.000	0.977	0.372	0.023	0.000	0.000	0.000	0.023	0.023	0.000	0.000	0.000
dk.2014-07-23.random.10000-http	4 626	0.115	0.347	0.825	0.326	0.012	0.003	0.009	0.019	0.000	0.001	0.000	0.000	0.001
dk.2014-07-23.random.10000-http-www	4 773	0.111	0.347	0.826	0.320	0.012	0.003	0.009	0.017	0.000	0.001	0.000	0.000	0.001
dk.2014-07-23.random.10000-https	16	0.000	0.125	1.000	0.500	0.000	0.000	0.000	0.125	0.000	0.000	0.000	0.000	0.000
dk.2014-07-23.random.10000-https-www	22	0.045	0.182	0.955	0.500	0.000	0.000	0.000	0.045	0.000	0.000	0.000	0.000	0.000
net.2014-08-29.random.10000-http	5 757	0.020	0.000	0.934	0.603	0.025	0.000	0.017	0.032	0.006	0.020	0.009	0.002	0.006
net.2014-08-29.random.10000-http-www	5 839	0.020	0.000	0.933	0.596	0.025	0.000	0.017	0.035	0.005	0.021	0.010	0.003	0.006
net.2014-08-29.random.10000-https	16	0.000	0.000	1.000	0.500	0.063	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
net.2014-08-29.random.10000-https-www	20	0.050	0.000	0.950	0.550	0.050	0.000	0.000	0.050	0.000	0.000	0.000	0.000	0.000
reach50.2014w35.se-http	43	0.581	0.023	0.907	0.721	0.047	0.070	0.000	0.070	0.000	0.000	0.000	0.000	0.000
reach50.2014w35.se-http-www	42	0.548	0.000	0.905	0.667	0.048	0.071	0.000	0.024	0.000	0.000	0.000	0.000	0.000
reach50.2014w35.se-https	17	0.412	0.059	0.882	0.588	0.059	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
reach50.2014w35.se-https-www	26	0.385	0.038	0.923	0.500	0.038	0.038	0.000	0.000	0.000	0.000	0.000	0.000	0.000
se.2014-07-10.random.100000-http	54 882	0.498	0.009	0.810	0.273	0.011	0.015	0.008	0.009	0.001	0.000	0.000	0.000	0.000
se.2014-07-10.random.100000-http-www	57 547	0.495	0.009	0.814	0.268	0.011	0.015	0.007	0.009	0.001	0.000	0.000	0.000	0.000
se.2014-07-10.random.100000-https	226	0.217	0.004	0.982	0.504	0.013	0.022	0.009	0.027	0.000	0.000	0.000	0.004	0.000
se.2014-07-10.random.100000-https-www	285	0.253	0.011	0.975	0.505	0.021	0.021	0.004	0.039	0.000	0.000	0.000	0.004	0.000
se.healthstatus.2013.counties-http	18	0.333	0.000	1.000	0.111	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.counties-http-www	21	0.429	0.000	1.000	0.095	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.counties-https	3	0.333	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.counties-https-www	5	0.400	0.000	1.000	0.200	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.domain-registrars-http	108	0.315	0.028	0.926	0.417	0.009	0.000	0.000	0.019	0.000	0.000	0.000	0.000	0.009
se.healthstatus.2013.domain-registrars-http-www	113	0.319	0.027	0.929	0.416	0.009	0.000	0.000	0.009	0.000	0.000	0.000	0.000	0.018
se.healthstatus.2013.domain-registrars-https	34	0.382	0.029	0.912	0.588	0.000	0.000	0.000	0.029	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.domain-registrars-https-www	36	0.278	0.028	0.917	0.639	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.028
se.healthstatus.2013.financial-services-http	67	0.478	0.015	0.761	0.433	0.000	0.000	0.000	0.015	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.financial-services-http-www	71	0.465	0.014	0.761	0.408	0.000	0.000	0.000	0.014	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.financial-services-https	16	0.250	0.000	0.875	0.813	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.financial-services-https-www	31	0.323	0.000	0.935	0.645	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Dataset	Domains w/ ext	se	dk	com	net	org	nu	uk	de	ru	jp	cn	br	fr
se.healthstatus.2013.gocs-http	48	0.208	0.000	1.000	0.417	0.000	0.000	0.000	0.063	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.gocs-http-www	55	0.236	0.000	0.982	0.400	0.000	0.000	0.000	0.073	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.gocs-https	4	0.000	0.000	1.000	1.000	0.000	0.000	0.000	0.250	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.gocs-https-www	9	0.222	0.000	1.000	0.778	0.000	0.000	0.000	0.222	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.higher-education-http	38	0.132	0.026	1.000	0.316	0.026	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.higher-education-http-www	44	0.114	0.023	1.000	0.341	0.023	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.higher-education-https	9	0.000	0.000	1.000	0.444	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.higher-education-https-www	24	0.167	0.042	1.000	0.458	0.042	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.isps-http	18	0.333	0.056	1.000	0.667	0.000	0.000	0.000	0.222	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.isps-http-www	19	0.421	0.053	1.000	0.684	0.053	0.000	0.000	0.211	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.isps-https	6	0.667	0.000	1.000	1.000	0.000	0.000	0.000	0.333	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.isps-https-www	10	0.500	0.000	1.000	0.800	0.000	0.000	0.000	0.100	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.media-http	25	1.000	0.000	1.000	0.960	0.000	0.080	0.000	0.200	0.000	0.040	0.000	0.000	0.000
se.healthstatus.2013.media-http-www	27	1.000	0.074	1.000	0.963	0.000	0.074	0.000	0.259	0.000	0.037	0.000	0.000	0.000
se.healthstatus.2013.media-https	4	0.750	0.000	1.000	1.000	0.000	0.250	0.000	0.000	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.media-https-www	5	1.000	0.000	1.000	1.000	0.000	0.200	0.000	0.200	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.municipalities-http	239	0.456	0.050	0.992	0.343	0.004	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.municipalities-http-www	258	0.415	0.054	0.992	0.322	0.004	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.municipalities-https	41	0.512	0.049	1.000	0.317	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.municipalities-https-www	50	0.460	0.040	1.000	0.320	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.public-authorities-http	162	0.340	0.037	0.957	0.167	0.025	0.000	0.000	0.006	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.public-authorities-http-www	188	0.335	0.032	0.973	0.144	0.021	0.000	0.000	0.005	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.public-authorities-https	15	0.267	0.067	0.933	0.200	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
se.healthstatus.2013.public-authorities-https-www	36	0.278	0.028	0.972	0.250	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table C.10: Public suffixes in external resources

Future downloads could be performed from other countries to measure location-dependent service coverage.

C.11 Disconnect's blocking list matches

The table below shows coverage for requested URLs' domains matching Disconnect's blocking list (B.2.3), for internal and external resources separately as well as all resources together. Coverage per domain for top domains is shown later in this section C.11.2. As the blocking list contains details about which category and organization each domain belongs to (A.3), they have been used grouped to display aggregates per category (C.11.3) and organization (C.11.4) as well.

Figure C.7 shows the CDF of the percentage of domains (y axis) which have a certain ratio of all requests matching Disconnect's blocking list (x axis). The leftmost marker for each dataset shows 0% Disconnect matches, and the rightmost marker shows 99% matches.

About 20% of Alexa's top 10,000 sites make more than 50% of their requests to known tracker domains; similarly 50% of domains from the same datasets load more than 20% of their resources form known tracker domains. Other datasets rely less on this kind of external resource, at around or less than 10% using 50% or more tracker resources – but again the number of requests is less interesting than the number of organizations potentially collecting the leaked user traffic data.

Dataset	Domains	Dom w/ int	Int non-D	Some int D	Dom w/ ext	Ext non-D	Some ext D	All non-D	Some D
alexa.rnd.10k-h	8 216	7 829	0.994	0.006	7 591	0.051	0.949	0.113	0.887
alexa.rnd.10k-hw	8 493	8 009	0.994	0.006	7 825	0.053	0.947	0.118	0.882
alexa.rnd.10k-s	1 135	1 084	0.994	0.006	1 072	0.025	0.975	0.065	0.935
alexa.rnd.10k-sw	1 224	1 182	0.997	0.003	1 139	0.037	0.963	0.096	0.904
alexa.top.10k-h	8 545	8 156	0.961	0.039	8 176	0.047	0.953	0.078	0.922
alexa.top.10k-hw	8 682	8 190	0.962	0.038	8 289	0.049	0.951	0.082	0.918
alexa.top.10k-s	2 507	2 398	0.917	0.083	2 369	0.025	0.975	0.064	0.936
alexa.top.10k-sw	2 957	2 849	0.920	0.080	2 801	0.030	0.970	0.072	0.928
alexa.top.dk.10k-h	2 263	2 182	1.000	0.000	2 136	0.024	0.976	0.073	0.927
alexa.top.dk.10k-hw	2 310	2 212	1.000	0.000	2 182	0.027	0.973	0.076	0.924
alexa.top.dk.10k-s	339	325	0.997	0.003	316	0.032	0.968	0.084	0.916
alexa.top.dk.10k-sw	441	424	0.998	0.002	406	0.037	0.963	0.103	0.897
alexa.top.se.10k-h	2 797	2 687	1.000	0.000	2 684	0.032	0.968	0.069	0.931
alexa.top.se.10k-hw	2 895	2 756	1.000	0.000	2 779	0.032	0.968	0.068	0.932
alexa.top.se.10k-s	438	427	0.998	0.002	422	0.012	0.988	0.046	0.954
alexa.top.se.10k-sw	650	636	0.998	0.002	630	0.016	0.984	0.043	0.957
com.rnd.10k-h	7 775	5 575	1.000	0.000	6 222	0.157	0.843	0.281	0.719
com.rnd.10k-hw	7 811	5 546	1.000	0.000	6 241	0.161	0.839	0.285	0.715
com.rnd.10k-s	50	45	1.000	0.000	41	0.098	0.902	0.213	0.787
com.rnd.10k-sw	55	54	1.000	0.000	43	0.070	0.930	0.259	0.741
dk.rnd.10k-h	7 180	4 648	1.000	0.000	4 626	0.220	0.780	0.467	0.533
dk.rnd.10k-hw	7 378	4 763	1.000	0.000	4 773	0.229	0.771	0.470	0.530
dk.rnd.10k-s	23	22	1.000	0.000	16	0.000	1.000	0.304	0.696
dk.rnd.10k-sw	32	29	1.000	0.000	22	0.045	0.955	0.300	0.700
net.rnd.10k-h	7 270	4 871	1.000	0.000	5 757	0.170	0.830	0.285	0.715
net.rnd.10k-hw	7 378	4 867	1.000	0.000	5 839	0.171	0.829	0.285	0.715
net.rnd.10k-s	26	26	1.000	0.000	16	0.063	0.938	0.423	0.577
net.rnd.10k-sw	28	25	1.000	0.000	20	0.100	0.900	0.308	0.692
reach50.se-h	43	41	0.780	0.220	43	0.047	0.953	0.023	0.977
reach50.se-hw	42	39	0.744	0.256	42	0.048	0.952	0.024	0.976
reach50.se-s	18	16	0.688	0.313	17	0.059	0.941	0.111	0.889
reach50.se-sw	26	23	0.652	0.348	26	0.115	0.885	0.038	0.962
se.rnd.100k-h	73 605	43 216	1.000	0.000	54 882	0.233	0.767	0.400	0.600
se.rnd.100k-hw	77 261	45 312	1.000	0.000	57 547	0.241	0.759	0.407	0.593
se.rnd.100k-s	282	263	1.000	0.000	226	0.040	0.960	0.199	0.801
se.rnd.100k-sw	328	311	1.000	0.000	285	0.049	0.951	0.158	0.842
se.hs.counties-h	18	18	1.000	0.000	18	0.056	0.944	0.056	0.944
se.hs.counties-hw	21	20	1.000	0.000	21	0.048	0.952	0.048	0.952
se.hs.counties-s	3	3	1.000	0.000	3	0.000	1.000	0.000	1.000
se.hs.counties-sw	6	6	1.000	0.000	5	0.000	1.000	0.167	0.833
se.hs.registrars-h	127	108	1.000	0.000	108	0.083	0.917	0.208	0.792

Dataset	Domains	Dom w/ int	Int non-D	Some int D	Dom w/ ext	Ext non-D	Some ext D	All non-D	Some D
se.hs.registrars-hw	134	114	1.000	0.000	113	0.080	0.920	0.224	0.776
se.hs.registrars-s	40	39	1.000	0.000	34	0.088	0.912	0.225	0.775
se.hs.registrars-sw	42	40	1.000	0.000	36	0.083	0.917	0.214	0.786
se.hs.financial-h	67	61	1.000	0.000	67	0.194	0.806	0.194	0.806
se.hs.financial-hw	72	64	1.000	0.000	71	0.197	0.803	0.208	0.792
se.hs.financial-s	16	15	1.000	0.000	16	0.000	1.000	0.000	1.000
se.hs.financial-sw	31	30	1.000	0.000	31	0.000	1.000	0.000	1.000
se.hs.gocs-h	49	44	1.000	0.000	48	0.000	1.000	0.020	0.980
se.hs.gocs-hw	57	50	1.000	0.000	55	0.018	0.982	0.053	0.947
se.hs.gocs-s	4	4	1.000	0.000	4	0.000	1.000	0.000	1.000
se.hs.gocs-sw	9	9	1.000	0.000	9	0.000	1.000	0.000	1.000
se.hs.education-h	40	39	1.000	0.000	38	0.000	1.000	0.050	0.950
se.hs.education-hw	47	46	1.000	0.000	44	0.000	1.000	0.064	0.936
se.hs.education-s	9	9	1.000	0.000	9	0.000	1.000	0.000	1.000
se.hs.education-sw	24	24	1.000	0.000	24	0.000	1.000	0.000	1.000
se.hs.isps-h	18	17	1.000	0.000	18	0.000	1.000	0.000	1.000
se.hs.isps-hw	19	19	1.000	0.000	19	0.000	1.000	0.000	1.000
se.hs.isps-s	6	6	1.000	0.000	6	0.000	1.000	0.000	1.000
se.hs.isps-sw	10	10	1.000	0.000	10	0.000	1.000	0.000	1.000
se.hs.media-h	26	24	1.000	0.000	25	0.000	1.000	0.038	0.962
se.hs.media-hw	28	25	1.000	0.000	27	0.000	1.000	0.036	0.964
se.hs.media-s	4	4	1.000	0.000	4	0.000	1.000	0.000	1.000
se.hs.media-sw	5	5	1.000	0.000	5	0.000	1.000	0.000	1.000
se.hs.municipalities-h	249	249	1.000	0.000	239	0.038	0.962	0.076	0.924
se.hs.municipalities-hw	271	270	1.000	0.000	258	0.039	0.961	0.085	0.915
se.hs.municipalities-s	44	44	1.000	0.000	41	0.049	0.951	0.114	0.886
se.hs.municipalities-sw	54	54	1.000	0.000	50	0.040	0.960	0.111	0.889
se.hs.pubauth-h	170	153	1.000	0.000	162	0.062	0.938	0.106	0.894
se.hs.pubauth-hw	203	182	1.000	0.000	188	0.053	0.947	0.123	0.877
se.hs.pubauth-s	18	18	1.000	0.000	15	0.067	0.933	0.222	0.778
se.hs.pubauth-sw	37	37	1.000	0.000	36	0.028	0.972	0.054	0.946

Table C.11: Disconnect coverage for internal, external and all requests

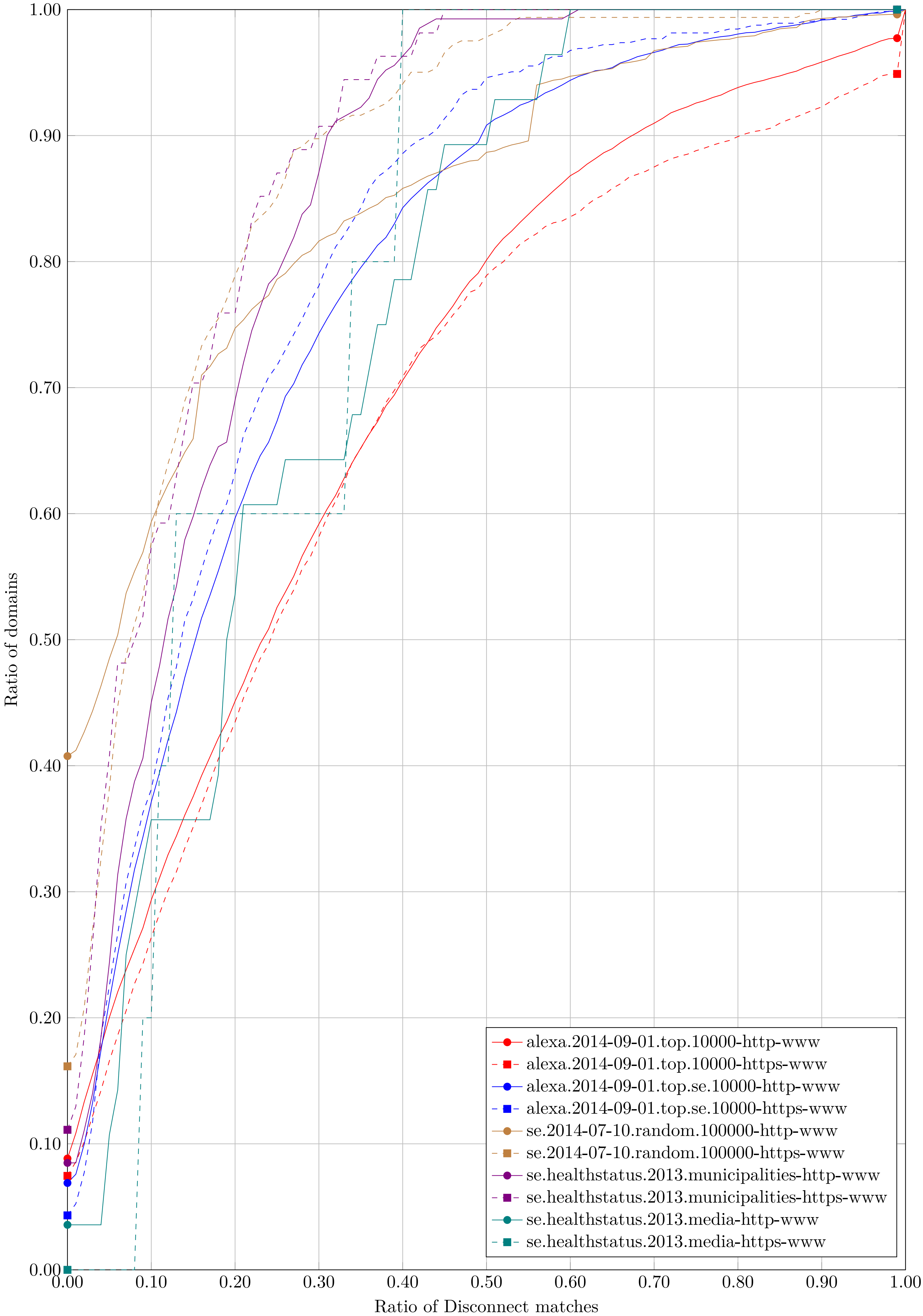

Figure C.7: Cumulative distribution of the ratio of Disconnect's blocking list matches for all resources per domain

C.11.1 Domain and organization counts

The below table shows the number of domains in the dataset, the number of requests classified as trackers in the Disconnect dataset (see also C.5 for other types of requests) followed by aggregate counts for domains, organizations and categories for these requests. Next we see Disconnect requests per domain in the dataset, and the same number per Disconnect organization. The last section contains the ratios out of the total 2,149 domains and 980 organizations (see A.3).

Figure C.8 shows the cumulative distribution of domains (y axis) with requests to between 0 and 99 organizations (x axis). Domains with 100 or more organizations are shown in the rightmost segment of the graph.

The *.SE Health Status'* media category has over 40 tracker requests on average – the highest number of tracker requests per domain, much higher than for example public authorities and random zone domains at 5-7, with top domains having 17-32 requests (C.5). What is more interesting than request counts is the number of tracker organizations per domain – while more information may be leaked as the number of requests to an organization increase, the same amount could potentially leak through with just one or two carefully composed requests to each organization.

While it is hard to compare Disconnect's organizations' coverage across datasets as the number of domains increases the chance of additional organizations being represented, it seems global top sites use a broad range of trackers – more than 500 of the 980 organizations. Looking at Figure C.8 we see that top sites have requests to more organizations than random domains do. Over 40% of random .se HTTP-www domains have no known trackers, about as many have one tracker and the top 1% have six or more. This is compared to 32% of Alexa's top site sharing information with more than five organizations, 10% have 13 or more, and 1% share information with at least 48 organizations. There are even a couple of domains among the Alexa sites which have more than 75 recognized tracker organizations just on the front page – a clear example of how it is impossible to tell where your browsing habits can end up, and even more so how it is used in a second stage. It is also clear that the non-zone domains have as much tracking on when using a secure connection as on an insecure connection, while the difference for zone domains can be explained by the very low HTTPS usage (C.2).

Dataset	Domains	D Requests	D Domains	D Orgs	D Cats	DR/d	(DR/d)/DO	DD/T	DO/T
alexa.2014-09-01.random.10000-http	7 591	166 702	704	481	5	21.960	0.046	0.328	0.491
alexa.2014-09-01.random.10000-http-www	7 825	169 685	704	476	5	21.685	0.046	0.328	0.486
alexa.2014-09-01.random.10000-https	1 072	23 599	370	272	5	22.014	0.081	0.172	0.278
alexa.2014-09-01.random.10000-https-www	1 139	16 764	368	276	5	14.718	0.053	0.171	0.282
alexa.2014-09-01.top.10000-http	8 176	274 782	755	505	5	33.608	0.067	0.351	0.515
alexa.2014-09-01.top.10000-http-www	8 289	276 636	760	515	5	33.374	0.065	0.354	0.526
alexa.2014-09-01.top.10000-https	2 369	67 788	542	388	5	28.615	0.074	0.252	0.396
alexa.2014-09-01.top.10000-https-www	2 801	73 239	569	413	5	26.147	0.063	0.265	0.421
alexa.2014-09-01.top.dk.10000-http	2 136	37 832	282	205	5	17.712	0.086	0.131	0.209
alexa.2014-09-01.top.dk.10000-http-www	2 182	38 373	284	206	5	17.586	0.085	0.132	0.210
alexa.2014-09-01.top.dk.10000-https	316	5 942	151	109	5	18.804	0.173	0.070	0.111
alexa.2014-09-01.top.dk.10000-https-www	406	6 901	176	127	5	16.998	0.134	0.082	0.130
alexa.2014-09-01.top.se.10000-http	2 684	52 345	342	245	5	19.503	0.080	0.159	0.250
alexa.2014-09-01.top.se.10000-http-www	2 779	52 398	351	255	5	18.855	0.074	0.163	0.260
alexa.2014-09-01.top.se.10000-https	422	7 104	167	114	5	16.834	0.148	0.078	0.116
alexa.2014-09-01.top.se.10000-https-www	630	9 510	199	146	5	15.095	0.103	0.093	0.149
com.2014-08-29.random.10000-http	6 222	55 666	404	273	5	8.947	0.033	0.188	0.279
com.2014-08-29.random.10000-http-www	6 241	55 955	405	277	5	8.966	0.032	0.188	0.283
com.2014-08-29.random.10000-https	41	446	47	26	5	10.878	0.418	0.022	0.027
com.2014-08-29.random.10000-https-www	43	477	49	28	5	11.093	0.396	0.023	0.029
dk.2014-07-23.random.10000-http	4 626	36 822	278	187	5	7.960	0.043	0.129	0.191
dk.2014-07-23.random.10000-http-www	4 773	35 960	275	187	5	7.534	0.040	0.128	0.191
dk.2014-07-23.random.10000-https	16	150	26	15	5	9.375	0.625	0.012	0.015

Dataset	Domains	D Requests	D Domains	D Orgs	D Cats	DR/d	(DR/d)/DO	DD/T	DO/T
dk.2014-07-23.random.10000-https-www	22	257	32	23	5	11.682	0.508	0.015	0.023
net.2014-08-29.random.10000-http	5 757	48 379	412	291	5	8.404	0.029	0.192	0.297
net.2014-08-29.random.10000-http-www	5 839	49 471	411	293	5	8.473	0.029	0.191	0.299
net.2014-08-29.random.10000-https	16	203	21	8	4	12.688	1.586	0.010	0.008
net.2014-08-29.random.10000-https-www	20	291	27	12	5	14.550	1.213	0.013	0.012
reach50.2014w35.se-http	43	843	92	66	5	19.605	0.297	0.043	0.067
reach50.2014w35.se-http-www	42	801	92	61	5	19.071	0.313	0.043	0.062
reach50.2014w35.se-https	17	265	41	24	5	15.588	0.650	0.019	0.024
reach50.2014w35.se-https-www	26	303	40	25	5	11.654	0.466	0.019	0.026
se.2014-07-10.random.100000-http	54 882	395 347	496	336	5	7.204	0.021	0.231	0.343
se.2014-07-10.random.100000-http-www	57 547	406 990	502	335	5	7.072	0.021	0.234	0.342
se.2014-07-10.random.100000-https	226	1 962	94	66	5	8.681	0.132	0.044	0.067
se.2014-07-10.random.100000-https-www	285	2 451	124	94	5	8.600	0.091	0.058	0.096
se.healthstatus.2013.counties-http	18	105	10	6	4	5.833	0.972	0.005	0.006
se.healthstatus.2013.counties-http-www	21	133	11	6	4	6.333	1.056	0.005	0.006
se.healthstatus.2013.counties-https	3	7	2	1	2	2.333	2.333	0.001	0.001
se.healthstatus.2013.counties-https-www	5	20	4	1	2	4.000	4.000	0.002	0.001
se.healthstatus.2013.domain-registrars-http	108	886	66	49	5	8.204	0.167	0.031	0.050
se.healthstatus.2013.domain-registrars-http-www	113	872	62	45	5	7.717	0.171	0.029	0.046
se.healthstatus.2013.domain-registrars-https	34	430	46	32	4	12.647	0.395	0.021	0.033
se.healthstatus.2013.domain-registrars-https-www	36	327	40	25	4	9.083	0.363	0.019	0.026
se.healthstatus.2013.financial-services-http	67	378	49	36	5	5.642	0.157	0.023	0.037
se.healthstatus.2013.financial-services-http-www	71	415	50	35	5	5.845	0.167	0.023	0.036
se.healthstatus.2013.financial-services-https	16	95	24	15	5	5.938	0.396	0.011	0.015
se.healthstatus.2013.financial-services-https-www	31	228	32	19	5	7.355	0.387	0.015	0.019
se.healthstatus.2013.gocs-http	48	501	45	28	5	10.438	0.373	0.021	0.029
se.healthstatus.2013.gocs-http-www	55	577	47	30	5	10.491	0.350	0.022	0.031
se.healthstatus.2013.gocs-https	4	64	21	11	4	16.000	1.455	0.010	0.011
se.healthstatus.2013.gocs-https-www	9	91	27	16	5	10.111	0.632	0.013	0.016
se.healthstatus.2013.higher-education-http	38	270	24	12	4	7.105	0.592	0.011	0.012
se.healthstatus.2013.higher-education-http-www	44	308	26	12	4	7.000	0.583	0.012	0.012
se.healthstatus.2013.higher-education-https	9	104	16	7	4	11.556	1.651	0.007	0.007
se.healthstatus.2013.higher-education-https-www	24	182	22	11	4	7.583	0.689	0.010	0.011
se.healthstatus.2013.isps-http	18	271	47	37	5	15.056	0.407	0.022	0.038
se.healthstatus.2013.isps-http-www	19	317	55	45	5	16.684	0.371	0.026	0.046
se.healthstatus.2013.isps-https	6	152	41	35	5	25.333	0.724	0.019	0.036
se.healthstatus.2013.isps-https-www	10	163	43	34	5	16.300	0.479	0.020	0.035
se.healthstatus.2013.media-http	25	1 101	81	57	5	44.040	0.773	0.038	0.058
se.healthstatus.2013.media-http-www	27	1 234	79	57	5	45.704	0.802	0.037	0.058
se.healthstatus.2013.media-https	4	186	24	15	4	46.500	3.100	0.011	0.015
se.healthstatus.2013.media-https-www	5	204	28	17	4	40.800	2.400	0.013	0.017

Dataset	Domains	D Requests	D Domains	D Orgs	D Cats	DR/d	(DR/d)/DO	DD/T	DO/T
se.healthstatus.2013.municipalities-http	239	2 367	39	19	5	9.904	0.521	0.018	0.019
se.healthstatus.2013.municipalities-http-www	258	2 447	39	19	5	9.484	0.499	0.018	0.019
se.healthstatus.2013.municipalities-https	41	305	18	9	4	7.439	0.827	0.008	0.009
se.healthstatus.2013.municipalities-https-www	50	394	18	9	4	7.880	0.876	0.008	0.009
se.healthstatus.2013.public-authorities-http	162	935	48	31	5	5.772	0.186	0.022	0.032
se.healthstatus.2013.public-authorities-http-www	188	945	48	31	5	5.027	0.162	0.022	0.032
se.healthstatus.2013.public-authorities-https	15	64	9	6	5	4.267	0.711	0.004	0.006
se.healthstatus.2013.public-authorities-https-www	36	200	23	14	5	5.556	0.397	0.011	0.014

Table C.12: Disconnect requests, organizations and categories counts and ratios

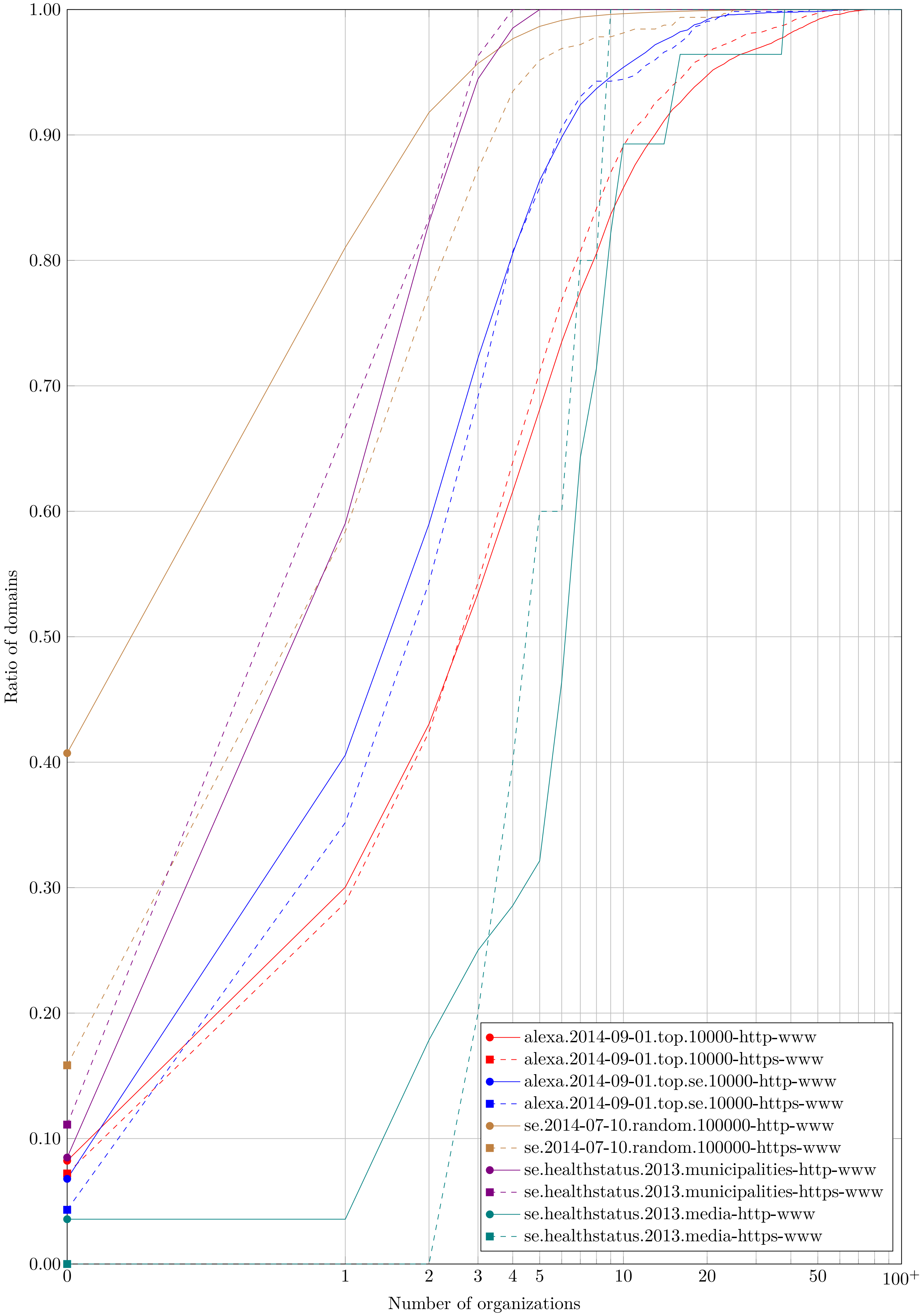

Figure C.8: Cumulative distribution of the number of organizations per domain

C.11.2 Top domains

A selection of domains, and their coverage across different datasets. Worth noting is that many recognizable domains belong to organization which have multiple domains (A.3.3), so their total organization coverage is higher. This first table excludes top Google domains, which dominate the top list for all datasets.

General

Facebook and Twitter have their like and tweet buttons which often are talked about in terms of social sharing, but it seems that their coverage is comparatively low except for the *.SE Health Status*' media category. AddThis' service includes some of Facebook's and Twitter's buttons' functionality, among other social sharing sites but they do not have quite the same coverage. Unfortunately, Facebook and Twitter are both in Disconnect's special Disconnect category (A.3.7), and are not reflected in the social category aggregates (C.11.3).

The domain cloudfront.net belonging to Amazon, is hosting services, file and data on behalf of other organizations on subdomains (A.2). While it is clear that Amazon can analyze and use traffic information from hosted services, the data itself can be assumed to belong to the hosted organizations. This is a strong reason why the domain is listed as content in Disconnect's blocking list – because of the variety of services being hosted on subdomains, it cannot be blocked as for example advertisement even if it is being hosted there. It can be seen as a flaw in the Disconnect way of blocking, even though listing individual subdomains in other categories might be a way to override the content bypass – but as these non-branded domains can be seen as throw-away domains, it can become a game of cat and mouse (5.2).

Dataset	Domains w/ ext	facebook.com	twitter.com	cloudfront.net	addthis.com	newrelic.com	optimizely.com	scorecardresearch.com
alexa.rnd.10k-h	7 591	0.260	0.168	0.077	0.076	0.033	0.010	0.121
alexa.rnd.10k-hw	7 825	0.260	0.169	0.076	0.075	0.031	0.010	0.121
alexa.rnd.10k-s	1 072	0.249	0.215	0.130	0.076	0.059	0.026	0.111
alexa.rnd.10k-sw	1 139	0.175	0.168	0.140	0.068	0.058	0.038	0.091
alexa.top.10k-h	8 176	0.362	0.228	0.188	0.081	0.079	0.058	0.230
alexa.top.10k-hw	8 289	0.360	0.229	0.190	0.083	0.076	0.057	0.228
alexa.top.10k-s	2 369	0.327	0.219	0.239	0.057	0.122	0.095	0.173
alexa.top.10k-sw	2 801	0.268	0.195	0.232	0.057	0.105	0.096	0.168
alexa.top.dk.10k-h	2 136	0.294	0.085	0.113	0.067	0.037	0.019	0.082
alexa.top.dk.10k-hw	2 182	0.287	0.081	0.110	0.068	0.039	0.018	0.083
alexa.top.dk.10k-s	316	0.228	0.082	0.177	0.066	0.066	0.047	0.089
alexa.top.dk.10k-sw	406	0.204	0.069	0.165	0.064	0.074	0.049	0.084
alexa.top.se.10k-h	2 684	0.310	0.123	0.108	0.082	0.069	0.023	0.107
alexa.top.se.10k-hw	2 779	0.309	0.121	0.109	0.080	0.067	0.023	0.107
alexa.top.se.10k-s	422	0.263	0.135	0.171	0.073	0.085	0.038	0.088
alexa.top.se.10k-sw	630	0.184	0.090	0.176	0.063	0.089	0.040	0.090
com.rnd.10k-h	6 222	0.076	0.049	0.035	0.028	0.025	0.008	0.041
com.rnd.10k-hw	6 241	0.073	0.048	0.033	0.028	0.023	0.008	0.040
com.rnd.10k-s	41	0.220	0.073	0.024	0.049	0.049	0.024	0.073
com.rnd.10k-sw	43	0.256	0.140	0.093	0.070	0.070	0.023	0.116
dk.rnd.10k-h	4 626	0.121	0.030	0.055	0.036	0.015	0.002	0.044
dk.rnd.10k-hw	4 773	0.118	0.030	0.055	0.036	0.016	0.002	0.044
dk.rnd.10k-s	16	0.125	0.000	0.000	0.063	0.063	0.000	0.063
dk.rnd.10k-sw	22	0.136	0.045	0.136	0.045	0.136	0.045	0.045
net.rnd.10k-h	5 757	0.067	0.043	0.021	0.023	0.016	0.005	0.031

Dataset	Domains w/ ext	facebook.com	twitter.com	cloudfront.net	addthis.com	newrelic.com	optimizely.com	scorecardresearch.com
net.rnd.10k-hw	5 839	0.068	0.045	0.021	0.025	0.016	0.005	0.032
net.rnd.10k-s	16	0.313	0.125	0.000	0.000	0.000	0.000	0.063
net.rnd.10k-sw	20	0.350	0.200	0.050	0.000	0.000	0.000	0.050
reach50.se-h	43	0.209	0.093	0.116	0.023	0.070	0.116	0.186
reach50.se-hw	42	0.214	0.095	0.095	0.024	0.024	0.119	0.214
reach50.se-s	17	0.059	0.059	0.176	0.059	0.059	0.118	0.353
reach50.se-sw	26	0.000	0.038	0.077	0.000	0.000	0.154	0.154
se.rnd.100k-h	54 882	0.117	0.039	0.048	0.025	0.014	0.004	0.037
se.rnd.100k-hw	57 547	0.114	0.037	0.046	0.025	0.014	0.004	0.037
se.rnd.100k-s	226	0.186	0.049	0.124	0.040	0.035	0.009	0.044
se.rnd.100k-sw	285	0.161	0.032	0.144	0.039	0.060	0.018	0.049
se.hs.counties-h	18	0.056	0.056	0.000	0.056	0.000	0.000	0.111
se.hs.counties-hw	21	0.095	0.048	0.000	0.048	0.000	0.000	0.095
se.hs.counties-s	3	0.000	0.000	0.000	0.000	0.000	0.000	0.000
se.hs.counties-sw	5	0.000	0.000	0.000	0.000	0.000	0.000	0.000
se.hs.registrars-h	108	0.157	0.093	0.056	0.019	0.009	0.009	0.019
se.hs.registrars-hw	113	0.159	0.088	0.053	0.009	0.009	0.009	0.009
se.hs.registrars-s	34	0.265	0.147	0.088	0.000	0.029	0.000	0.000
se.hs.registrars-sw	36	0.194	0.111	0.083	0.000	0.028	0.000	0.000
se.hs.financial-h	67	0.075	0.030	0.045	0.030	0.045	0.075	0.045
se.hs.financial-hw	71	0.070	0.028	0.042	0.028	0.042	0.070	0.042
se.hs.financial-s	16	0.000	0.063	0.063	0.063	0.063	0.125	0.125
se.hs.financial-sw	31	0.032	0.032	0.097	0.065	0.032	0.194	0.097
se.hs.gocs-h	48	0.063	0.042	0.125	0.104	0.042	0.021	0.104
se.hs.gocs-hw	55	0.073	0.073	0.127	0.091	0.036	0.018	0.109
se.hs.gocs-s	4	0.250	0.500	0.250	0.000	0.000	0.000	0.000
se.hs.gocs-sw	9	0.111	0.222	0.222	0.111	0.111	0.000	0.111
se.hs.education-h	38	0.158	0.132	0.026	0.132	0.000	0.000	0.132
se.hs.education-hw	44	0.159	0.114	0.023	0.114	0.000	0.000	0.159
se.hs.education-s	9	0.111	0.222	0.000	0.222	0.000	0.000	0.222
se.hs.education-sw	24	0.083	0.125	0.042	0.167	0.000	0.000	0.250
se.hs.isps-h	18	0.167	0.000	0.278	0.111	0.056	0.056	0.111
se.hs.isps-hw	19	0.211	0.000	0.263	0.105	0.053	0.053	0.105
se.hs.isps-s	6	0.167	0.000	0.167	0.167	0.167	0.167	0.167
se.hs.isps-sw	10	0.100	0.000	0.100	0.100	0.100	0.100	0.100
se.hs.media-h	25	0.600	0.280	0.480	0.160	0.200	0.120	0.160
se.hs.media-hw	27	0.630	0.259	0.481	0.148	0.185	0.111	0.333
se.hs.media-s	4	0.750	0.250	0.750	0.000	0.000	0.500	0.000
se.hs.media-sw	5	0.400	0.200	0.400	0.000	0.000	0.200	0.000
se.hs.municipalities-h	239	0.126	0.029	0.188	0.134	0.029	0.004	0.138
se.hs.municipalities-hw	258	0.089	0.019	0.190	0.136	0.039	0.004	0.140

Dataset	Domains w/ ext	facebook.com	twitter.com	cloudfront.net	addthis.com	newrelic.com	optimizely.com	scorecardresearch.com
se.hs.municipalities-s	41	0.122	0.024	0.146	0.098	0.000	0.000	0.122
se.hs.municipalities-sw	50	0.120	0.020	0.100	0.120	0.000	0.000	0.140
se.hs.pubauth-h	162	0.080	0.062	0.019	0.056	0.031	0.000	0.080
se.hs.pubauth-hw	188	0.064	0.043	0.021	0.048	0.032	0.000	0.074
se.hs.pubauth-s	15	0.067	0.000	0.000	0.067	0.000	0.000	0.133
se.hs.pubauth-sw	36	0.056	0.028	0.083	0.111	0.028	0.000	0.167

Table C.13: Top Disconnect domain match coverage

One thing to look closer at is cloudfront.net's subdomains, to analyze their content and which sites use them. Are they used as CDNs local to a single domain, or do third-party services host content there? If so, what kind of content? Are there similar patterns for Amazon AWS and other cloud services? See also public suffixes owned by private companies (A.2).

Google

Google has many domains in Disconnect's blocking list (A.3.2), and many in the top results. One of the reasons is that they have several popular services with content (5.4), but the top Disconnect-recognized domain in most datasets is google-analytics.com, with googleapis.com and www.google.com as the top result in the others.

Here we see DoubleClick's, one of Google's ad services, coverage. Unfortunately it has not been included in Disconnect's advertisement category, and it skews the numbers for the advertisement category (C.11.3), nor has Google Analytics been put in the analytics category. Google Analytics is served from its own domain, but Google is making a push to move site owners to the DoubleClick domain, where they have replicated the Google Analytics engine[7]. The reason given to site owners using Google Analytics is that the DoubleClick tracker offers implementors additional rich audience information such as age, gender and interests on top of their current technical analysis from Google Analytics. The underlying reason is possibly that Google wants DoubleClick, which brings a lot of income, to have greater coverage across websites – greater coverage meaning more and higher quality visitor information. While Google owns both services, site owner/visitor/usage policies[8] might prevent Google from crossmatching information between them without site owner consent. Cookies set to the doubleclick.net domain should be more valuable to Google than those set to google-analytics.com, as they translate to more directed ads [16].

Dataset	Domains w/ ext	www.google.com	doubleclick.net	google-analytics.com	googleapis.com	maps.google.com	youtube.com	google.se
alexa.rnd.10k-h	7 591	0.315	0.328	0.684	0.489	0.017	0.077	0.077
alexa.rnd.10k-hw	7 825	0.311	0.321	0.683	0.484	0.018	0.077	0.077
alexa.rnd.10k-s	1 072	0.348	0.382	0.737	0.502	0.014	0.101	0.147
alexa.rnd.10k-sw	1 139	0.299	0.375	0.766	0.434	0.016	0.075	0.176
alexa.top.10k-h	8 176	0.380	0.523	0.682	0.398	0.005	0.057	0.178
alexa.top.10k-hw	8 289	0.378	0.521	0.682	0.389	0.005	0.054	0.179
alexa.top.10k-s	2 369	0.366	0.529	0.688	0.403	0.005	0.061	0.223
alexa.top.10k-sw	2 801	0.367	0.530	0.667	0.378	0.006	0.052	0.229
alexa.top.dk.10k-h	2 136	0.308	0.389	0.736	0.559	0.021	0.057	0.166
alexa.top.dk.10k-hw	2 182	0.305	0.394	0.738	0.556	0.020	0.055	0.162

[7] *Update your Analytics tracking code to support Display Advertising* https://support.google.com/analytics/answer/2444872

[8] *Policy requirements for Google Analytics Advertising Features* https://support.google.com/analytics/answer/2700409

Dataset	Domains w/ ext	www.google.com	doubleclick.net	google-analytics.com	googleapis.com	maps.google.com	youtube.com	google.se
alexa.top.dk.10k-s	316	0.323	0.402	0.715	0.519	0.019	0.044	0.244
alexa.top.dk.10k-sw	406	0.333	0.441	0.729	0.468	0.017	0.044	0.264
alexa.top.se.10k-h	2 684	0.295	0.346	0.779	0.517	0.024	0.072	0.139
alexa.top.se.10k-hw	2 779	0.294	0.341	0.788	0.525	0.025	0.069	0.135
alexa.top.se.10k-s	422	0.308	0.370	0.822	0.526	0.033	0.057	0.190
alexa.top.se.10k-sw	630	0.325	0.403	0.798	0.470	0.030	0.056	0.229
com.rnd.10k-h	6 222	0.435	0.423	0.343	0.316	0.016	0.031	0.018
com.rnd.10k-hw	6 241	0.432	0.420	0.342	0.316	0.017	0.030	0.019
com.rnd.10k-s	41	0.244	0.146	0.585	0.537	0.024	0.098	0.098
com.rnd.10k-sw	43	0.209	0.163	0.581	0.465	0.023	0.093	0.070
dk.rnd.10k-h	4 626	0.197	0.118	0.428	0.470	0.018	0.036	0.101
dk.rnd.10k-hw	4 773	0.192	0.114	0.427	0.462	0.018	0.034	0.096
dk.rnd.10k-s	16	0.313	0.250	0.688	0.438	0.000	0.188	0.000
dk.rnd.10k-sw	22	0.273	0.227	0.727	0.364	0.000	0.182	0.045
net.rnd.10k-h	5 757	0.483	0.467	0.307	0.260	0.009	0.025	0.017
net.rnd.10k-hw	5 839	0.477	0.459	0.307	0.257	0.009	0.025	0.017
net.rnd.10k-s	16	0.125	0.125	0.688	0.750	0.063	0.000	0.000
net.rnd.10k-sw	20	0.150	0.150	0.600	0.650	0.050	0.000	0.050
reach50.se-h	43	0.186	0.326	0.512	0.279	0.000	0.023	0.163
reach50.se-hw	42	0.143	0.262	0.548	0.262	0.000	0.024	0.119
reach50.se-s	17	0.118	0.235	0.588	0.294	0.000	0.000	0.118
reach50.se-sw	26	0.077	0.192	0.577	0.308	0.000	0.038	0.077
se.rnd.100k-h	54 882	0.184	0.110	0.429	0.453	0.018	0.032	0.078
se.rnd.100k-hw	57 547	0.180	0.109	0.429	0.445	0.017	0.031	0.075
se.rnd.100k-s	226	0.164	0.142	0.801	0.487	0.022	0.031	0.088
se.rnd.100k-sw	285	0.161	0.193	0.772	0.439	0.021	0.021	0.116
se.hs.counties-h	18	0.222	0.000	0.833	0.444	0.000	0.000	0.000
se.hs.counties-hw	21	0.190	0.000	0.810	0.524	0.000	0.000	0.000
se.hs.counties-s	3	0.000	0.000	1.000	0.333	0.000	0.000	0.000
se.hs.counties-sw	5	0.200	0.000	1.000	0.600	0.000	0.000	0.000
se.hs.registrars-h	108	0.213	0.213	0.815	0.500	0.028	0.046	0.194
se.hs.registrars-hw	113	0.221	0.221	0.788	0.513	0.035	0.053	0.195
se.hs.registrars-s	34	0.324	0.382	0.824	0.441	0.059	0.118	0.324
se.hs.registrars-sw	36	0.250	0.333	0.806	0.444	0.028	0.083	0.222
se.hs.financial-h	67	0.149	0.179	0.597	0.209	0.000	0.030	0.104
se.hs.financial-hw	71	0.141	0.169	0.592	0.225	0.014	0.028	0.099
se.hs.financial-s	16	0.188	0.125	0.688	0.313	0.000	0.125	0.063
se.hs.financial-sw	31	0.290	0.226	0.742	0.258	0.000	0.065	0.194
se.hs.gocs-h	48	0.208	0.167	0.979	0.438	0.125	0.063	0.021
se.hs.gocs-hw	55	0.182	0.145	0.945	0.418	0.109	0.055	0.018
se.hs.gocs-s	4	0.000	0.250	0.750	0.250	0.250	0.000	0.000

Dataset	Domains w/ ext	www.google.com	doubleclick.net	google-analytics.com	googleapis.com	maps.google.com	youtube.com	google.se
se.hs.gocs-sw	9	0.000	0.222	0.667	0.222	0.111	0.000	0.000
se.hs.education-h	38	0.158	0.132	0.921	0.289	0.026	0.079	0.053
se.hs.education-hw	44	0.136	0.091	0.932	0.318	0.045	0.068	0.045
se.hs.education-s	9	0.222	0.111	1.000	0.222	0.000	0.111	0.000
se.hs.education-sw	24	0.167	0.208	0.958	0.292	0.000	0.125	0.083
se.hs.isps-h	18	0.222	0.333	0.722	0.444	0.056	0.000	0.222
se.hs.isps-hw	19	0.263	0.368	0.789	0.474	0.053	0.000	0.263
se.hs.isps-s	6	0.500	0.667	0.833	0.500	0.000	0.000	0.500
se.hs.isps-sw	10	0.400	0.400	0.700	0.500	0.100	0.000	0.400
se.hs.media-h	25	0.160	0.440	0.720	0.560	0.000	0.040	0.080
se.hs.media-hw	27	0.111	0.407	0.741	0.630	0.000	0.000	0.074
se.hs.media-s	4	0.000	0.250	0.750	0.750	0.000	0.000	0.000
se.hs.media-sw	5	0.000	0.400	1.000	0.800	0.000	0.000	0.000
se.hs.municipalities-h	239	0.368	0.013	0.828	0.552	0.042	0.025	0.013
se.hs.municipalities-hw	258	0.391	0.016	0.814	0.554	0.035	0.019	0.012
se.hs.municipalities-s	41	0.268	0.024	0.878	0.415	0.049	0.024	0.000
se.hs.municipalities-sw	50	0.260	0.020	0.880	0.480	0.040	0.000	0.000
se.hs.pubauth-h	162	0.105	0.031	0.840	0.401	0.025	0.037	0.000
se.hs.pubauth-hw	188	0.096	0.021	0.846	0.420	0.021	0.027	0.000
se.hs.pubauth-s	15	0.067	0.000	0.867	0.267	0.000	0.000	0.000
se.hs.pubauth-sw	36	0.028	0.028	0.806	0.278	0.000	0.000	0.000

Table C.14: Top Disconnect Google domain match coverage

C.11.3 Tracker categories

Disconnect's categories and their coverage across different datasets are shown in the table below, as well as the coverage of domains where any (at least one) external resource matches Disconnect's blocking list. As mentioned earlier, the special Disconnect category contains major Facebook, Google and Twitter domains – domains which could also have been listed as advertising, analytics or social domains (A.3.7). The content category, which bypasses Disconnect's blocking by default, can for this reason be seen as the most accurate in terms of coverage, as domains have presumably been added as content in a manual process of whitelisting (A.3.6).

Figure C.9 shows each category's coverage (x axis) per dataset. The grey bar in the background shows ratio of domains with any (at least one) external request matching Disconnect's blocking list for each dataset; it effectively shows the union of the coverage of all categories per domain. In some cases a single category has the same coverage as the union.

The highest coverage being connected with the Disconnect category explains the low coverage of advertising, analytics and social. If, for example, the two domains facebook.com and twitter.com would be included in the social category, coverage would be 35-56% percentage points higher for top domains and 9-11% percentage points higher for random domains (C.11.2) – and more accurate. The same goes for advertising and doubleclick.net (30-50%, 7-36%) and analytics and google-analytics.com (63-76%, 24-32%) (C.11.2).

What is surprising is the high coverage of content from known trackers. While the Disconnect category has the highest coverage overall, the content category is the second largest – significantly larger than the advertising, analytics and social categories in most datasets. While a large portion of this is due to extensive usage of Google's hosted services (C.11.2), all organizations with only content domains as well as those with "mixed" domains (Table 3.3) are let through to 67-78% of top domains and 38-56% or random domains. Mixing advertisement, or in this case tracking in general, with content has previously been discussed as a way for organizations to avoid in-browser blocking

(5.2) – and it seems prevalent.

Dataset	Domains	Any	Disconnect	Content	Advertising	Analytics	Social
alexa.2014-09-01.random.10000-http	7 591	0.949	0.813	0.731	0.304	0.273	0.160
alexa.2014-09-01.random.10000-http-www	7 825	0.947	0.809	0.724	0.296	0.270	0.157
alexa.2014-09-01.random.10000-https	1 072	0.975	0.901	0.806	0.289	0.289	0.136
alexa.2014-09-01.random.10000-https-www	1 139	0.963	0.896	0.731	0.269	0.240	0.117
alexa.2014-09-01.top.10000-http	8 176	0.953	0.849	0.737	0.539	0.457	0.126
alexa.2014-09-01.top.10000-http-www	8 289	0.951	0.849	0.729	0.537	0.452	0.129
alexa.2014-09-01.top.10000-https	2 369	0.975	0.857	0.798	0.526	0.423	0.103
alexa.2014-09-01.top.10000-https-www	2 801	0.970	0.846	0.765	0.535	0.417	0.099
alexa.2014-09-01.top.dk.10000-http	2 136	0.976	0.896	0.780	0.290	0.204	0.093
alexa.2014-09-01.top.dk.10000-http-www	2 182	0.973	0.895	0.774	0.293	0.212	0.092
alexa.2014-09-01.top.dk.10000-https	316	0.968	0.864	0.744	0.424	0.253	0.079
alexa.2014-09-01.top.dk.10000-https-www	406	0.963	0.874	0.729	0.426	0.268	0.074
alexa.2014-09-01.top.se.10000-http	2 684	0.968	0.901	0.751	0.298	0.250	0.108
alexa.2014-09-01.top.se.10000-http-www	2 779	0.968	0.902	0.754	0.297	0.249	0.108
alexa.2014-09-01.top.se.10000-https	422	0.988	0.950	0.806	0.320	0.249	0.090
alexa.2014-09-01.top.se.10000-https-www	630	0.984	0.940	0.762	0.321	0.254	0.086
com.2014-08-29.random.10000-http	6 222	0.843	0.692	0.702	0.100	0.314	0.045
com.2014-08-29.random.10000-http-www	6 241	0.839	0.687	0.699	0.099	0.309	0.045
com.2014-08-29.random.10000-https	41	0.902	0.707	0.780	0.171	0.171	0.073
com.2014-08-29.random.10000-https-www	43	0.930	0.721	0.744	0.209	0.279	0.070
dk.2014-07-23.random.10000-http	4 626	0.780	0.520	0.600	0.067	0.096	0.056
dk.2014-07-23.random.10000-http-www	4 773	0.771	0.513	0.590	0.066	0.096	0.056
dk.2014-07-23.random.10000-https	16	1.000	0.875	0.750	0.188	0.188	0.063
dk.2014-07-23.random.10000-https-www	22	0.955	0.773	0.727	0.273	0.227	0.045
net.2014-08-29.random.10000-http	5 757	0.830	0.697	0.702	0.097	0.327	0.038
net.2014-08-29.random.10000-http-www	5 839	0.829	0.692	0.701	0.095	0.324	0.040
net.2014-08-29.random.10000-https	16	0.938	0.750	0.813	0.000	0.063	0.063
net.2014-08-29.random.10000-https-www	20	0.900	0.750	0.800	0.050	0.050	0.050
reach50.2014w35.se-http	43	0.953	0.744	0.605	0.581	0.512	0.047
reach50.2014w35.se-http-www	42	0.952	0.714	0.571	0.571	0.500	0.048
reach50.2014w35.se-https	17	0.941	0.706	0.529	0.471	0.471	0.118
reach50.2014w35.se-https-www	26	0.885	0.615	0.538	0.423	0.423	0.038
se.2014-07-10.random.100000-http	54 882	0.767	0.533	0.585	0.061	0.093	0.040
se.2014-07-10.random.100000-http-www	57 547	0.759	0.531	0.575	0.061	0.092	0.040
se.2014-07-10.random.100000-https	226	0.960	0.867	0.708	0.097	0.111	0.066
se.2014-07-10.random.100000-https-www	285	0.951	0.853	0.649	0.161	0.151	0.060
se.healthstatus.2013.counties-http	18	0.944	0.833	0.500	0.000	0.111	0.111
se.healthstatus.2013.counties-http-www	21	0.952	0.857	0.571	0.000	0.095	0.095
se.healthstatus.2013.counties-https	3	1.000	1.000	0.333	0.000	0.000	0.000

Dataset	Domains	Any	Disconnect	Content	Advertising	Analytics	Social
se.healthstatus.2013.counties-https-www	5	1.000	1.000	0.600	0.000	0.000	0.000
se.healthstatus.2013.domain-registrars-http	108	0.917	0.843	0.657	0.130	0.065	0.019
se.healthstatus.2013.domain-registrars-http-www	113	0.920	0.841	0.655	0.115	0.062	0.009
se.healthstatus.2013.domain-registrars-https	34	0.912	0.853	0.588	0.176	0.029	0.000
se.healthstatus.2013.domain-registrars-https-www	36	0.917	0.833	0.611	0.194	0.056	0.000
se.healthstatus.2013.financial-services-http	67	0.806	0.672	0.433	0.239	0.194	0.030
se.healthstatus.2013.financial-services-http-www	71	0.803	0.662	0.437	0.254	0.183	0.028
se.healthstatus.2013.financial-services-https	16	1.000	0.750	0.688	0.375	0.313	0.063
se.healthstatus.2013.financial-services-https-www	31	1.000	0.806	0.677	0.387	0.323	0.065
se.healthstatus.2013.gocs-http	48	1.000	1.000	0.646	0.229	0.125	0.104
se.healthstatus.2013.gocs-http-www	55	0.982	0.982	0.618	0.236	0.127	0.109
se.healthstatus.2013.gocs-https	4	1.000	1.000	0.750	0.750	0.250	0.000
se.healthstatus.2013.gocs-https-www	9	1.000	1.000	0.667	0.667	0.222	0.111
se.healthstatus.2013.higher-education-http	38	1.000	0.947	0.605	0.000	0.132	0.158
se.healthstatus.2013.higher-education-http-www	44	1.000	0.977	0.591	0.000	0.159	0.182
se.healthstatus.2013.higher-education-https	9	1.000	1.000	0.556	0.000	0.222	0.333
se.healthstatus.2013.higher-education-https-www	24	1.000	1.000	0.708	0.000	0.250	0.292
se.healthstatus.2013.isps-http	18	1.000	0.778	0.778	0.444	0.222	0.111
se.healthstatus.2013.isps-http-www	19	1.000	0.842	0.789	0.526	0.263	0.105
se.healthstatus.2013.isps-https	6	1.000	0.833	0.833	0.833	0.333	0.167
se.healthstatus.2013.isps-https-www	10	1.000	0.700	0.800	0.700	0.300	0.200
se.healthstatus.2013.media-http	25	1.000	0.920	0.800	0.960	0.600	0.160
se.healthstatus.2013.media-http-www	27	1.000	0.926	0.852	0.926	0.556	0.148
se.healthstatus.2013.media-https	4	1.000	1.000	1.000	0.750	0.250	0.000
se.healthstatus.2013.media-https-www	5	1.000	1.000	0.800	1.000	0.600	0.000
se.healthstatus.2013.municipalities-http	239	0.962	0.858	0.695	0.013	0.205	0.142
se.healthstatus.2013.municipalities-http-www	258	0.961	0.845	0.702	0.012	0.209	0.140
se.healthstatus.2013.municipalities-https	41	0.951	0.902	0.561	0.000	0.146	0.122
se.healthstatus.2013.municipalities-https-www	50	0.960	0.920	0.600	0.000	0.160	0.140
se.healthstatus.2013.public-authorities-http	162	0.938	0.846	0.500	0.037	0.130	0.086
se.healthstatus.2013.public-authorities-http-www	188	0.947	0.851	0.495	0.037	0.128	0.080
se.healthstatus.2013.public-authorities-https	15	0.933	0.867	0.333	0.067	0.133	0.133
se.healthstatus.2013.public-authorities-https-www	36	0.972	0.833	0.444	0.056	0.250	0.167

Table C.15: Disconnect category match coverage

The current analysis performed for this thesis is built in such a way that the Disconnect blocking list used for matching can easily be replaced with an updated version. This also opens up the possibility of using a locally modified blocking list, re-categorizing each of the Disconnect category's domains as either advertising, analytics or social. The per-organization aggregate analysis would still produce the same numbers (C.11.4).

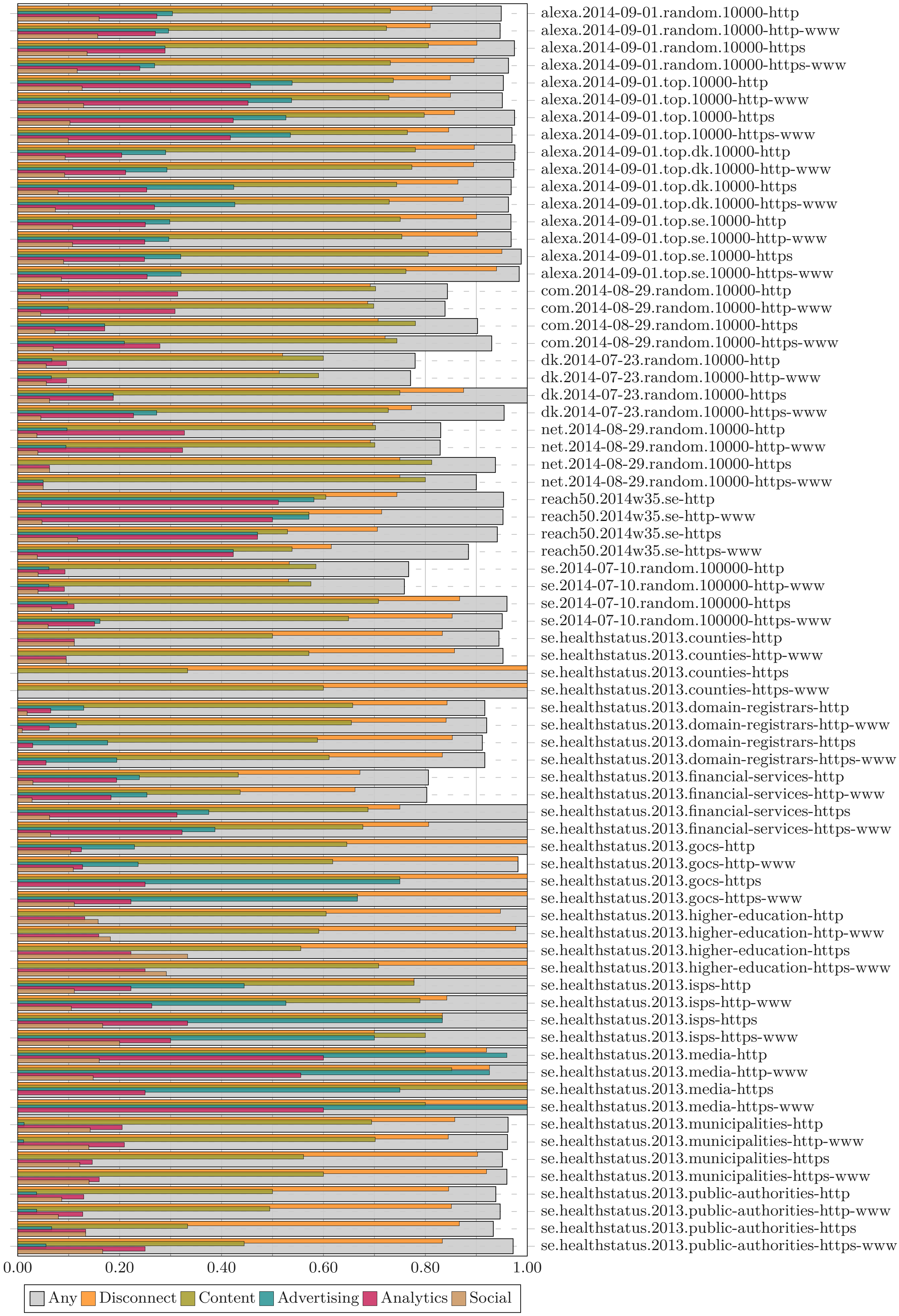

Figure C.9: Ratio of domains with requests to Disconnect's categories

C.11.4 Top organizations

A selection of organizations, and their coverage across different datasets. Facebook and Twitter are often touted as the big social network sites, but in terms of domain coverage they are far behind Google.

Figure C.10 shows the coverage of these three organizations (x axis) per dataset. As these organizations, and most of their domains, are included in the Disconnect category of Disconnect's blocking list (A.3.7), an `x` has been added to show their collective Disconnect category coverage (C.11.3).

Google is very popular with all Alexa and most Swedish curated datasets have a coverage above 80% – and many closer to 90%. Random domains have a lower reliance on Google at 47-62% – still about half of all domains. Apart from the .SE Health Status list of Swedish media domains, Facebook doesn't reach 40% in top or curated domains. Facebook coverage on random zone domains is 6-10%, which is also much lower than Google's numbers. Twitter generally has even lower coverage, at about half of that of Facebook on average. As can be seen, Google alone oftentimes has a coverage higher than the domains in the Disconnect category – it shows that Google's content domains are in use (A.3.3). In fact, at around 90% of the total tracker coverage, Google's coverage approaches that of the union of *all other* known trackers.

Dataset	Domains w/ ext	Google	Facebook	Twitter	Microsoft	Amazon	Adobe	Yahoo!	AddThis	AppNexus	comScore	Quantcast
alexa.rnd.10k-h	7 591	0.879	0.281	0.174	0.005	0.097	0.032	0.029	0.076	0.070	0.122	0.041
alexa.rnd.10k-hw	7 825	0.876	0.279	0.175	0.005	0.093	0.032	0.030	0.075	0.066	0.122	0.041
alexa.rnd.10k-s	1 072	0.932	0.351	0.219	0.006	0.148	0.033	0.022	0.076	0.070	0.116	0.034
alexa.rnd.10k-sw	1 139	0.924	0.305	0.173	0.006	0.144	0.042	0.022	0.068	0.079	0.095	0.030
alexa.top.10k-h	8 176	0.888	0.384	0.234	0.015	0.203	0.104	0.044	0.081	0.177	0.231	0.106
alexa.top.10k-hw	8 289	0.883	0.383	0.235	0.016	0.205	0.106	0.043	0.083	0.172	0.229	0.106
alexa.top.10k-s	2 369	0.921	0.412	0.228	0.014	0.252	0.088	0.039	0.057	0.167	0.175	0.108
alexa.top.10k-sw	2 801	0.903	0.377	0.202	0.015	0.245	0.109	0.044	0.058	0.164	0.171	0.092
alexa.top.dk.10k-h	2 136	0.944	0.316	0.088	0.010	0.114	0.037	0.012	0.067	0.099	0.089	0.004
alexa.top.dk.10k-hw	2 182	0.941	0.311	0.084	0.004	0.111	0.036	0.012	0.068	0.091	0.091	0.004
alexa.top.dk.10k-s	316	0.902	0.348	0.082	0.025	0.177	0.063	0.006	0.066	0.149	0.089	0.003
alexa.top.dk.10k-sw	406	0.901	0.328	0.069	0.015	0.165	0.086	0.010	0.064	0.140	0.089	0.005
alexa.top.se.10k-h	2 684	0.933	0.333	0.131	0.026	0.108	0.044	0.016	0.082	0.122	0.109	0.010
alexa.top.se.10k-hw	2 779	0.932	0.335	0.129	0.016	0.109	0.044	0.017	0.080	0.114	0.108	0.008
alexa.top.se.10k-s	422	0.957	0.405	0.140	0.031	0.171	0.064	0.009	0.073	0.130	0.088	0.002
alexa.top.se.10k-sw	630	0.948	0.354	0.097	0.038	0.176	0.057	0.013	0.063	0.121	0.094	0.005
com.rnd.10k-h	6 222	0.774	0.083	0.053	0.004	0.039	0.014	0.018	0.028	0.016	0.041	0.018
com.rnd.10k-hw	6 241	0.771	0.080	0.051	0.004	0.038	0.014	0.019	0.028	0.015	0.041	0.017
com.rnd.10k-s	41	0.854	0.293	0.098	0.000	0.024	0.000	0.024	0.049	0.049	0.073	0.024
com.rnd.10k-sw	43	0.860	0.302	0.163	0.000	0.093	0.023	0.023	0.070	0.093	0.116	0.023
dk.rnd.10k-h	4 626	0.736	0.131	0.032	0.003	0.055	0.018	0.011	0.036	0.016	0.047	0.006
dk.rnd.10k-hw	4 773	0.726	0.127	0.031	0.003	0.055	0.017	0.011	0.036	0.016	0.047	0.006
dk.rnd.10k-s	16	0.938	0.125	0.000	0.000	0.000	0.063	0.000	0.063	0.063	0.063	0.000
dk.rnd.10k-sw	22	0.909	0.136	0.045	0.000	0.136	0.045	0.000	0.045	0.045	0.045	0.000
net.rnd.10k-h	5 757	0.773	0.073	0.046	0.003	0.026	0.016	0.016	0.023	0.018	0.031	0.011
net.rnd.10k-hw	5 839	0.769	0.074	0.047	0.003	0.026	0.019	0.015	0.025	0.018	0.032	0.011
net.rnd.10k-s	16	0.938	0.313	0.125	0.000	0.000	0.063	0.000	0.000	0.000	0.063	0.000
net.rnd.10k-sw	20	0.850	0.350	0.200	0.000	0.050	0.050	0.000	0.000	0.050	0.050	0.000
reach50.se-h	43	0.791	0.209	0.116	0.023	0.116	0.116	0.023	0.023	0.186	0.209	0.047
reach50.se-hw	42	0.762	0.238	0.119	0.048	0.119	0.119	0.024	0.024	0.143	0.238	0.048

Dataset	Domains w/ ext	Google	Facebook	Twitter	Microsoft	Amazon	Adobe	Yahoo!	AddThis	AppNexus	comScore	Quantcast
reach50.se-s	17	0.706	0.118	0.118	0.000	0.176	0.059	0.059	0.059	0.059	0.353	0.059
reach50.se-sw	26	0.769	0.038	0.077	0.000	0.077	0.038	0.038	0.000	0.038	0.192	0.038
se.rnd.100k-h	54 882	0.712	0.128	0.041	0.003	0.048	0.011	0.006	0.025	0.011	0.037	0.006
se.rnd.100k-hw	57 547	0.705	0.125	0.039	0.003	0.046	0.011	0.006	0.025	0.011	0.037	0.006
se.rnd.100k-s	226	0.912	0.288	0.049	0.013	0.124	0.027	0.009	0.040	0.022	0.044	0.000
se.rnd.100k-sw	285	0.909	0.235	0.032	0.011	0.144	0.039	0.004	0.039	0.042	0.049	0.000
se.hs.counties-h	18	0.833	0.056	0.056	0.000	0.000	0.000	0.000	0.056	0.000	0.111	0.000
se.hs.counties-hw	21	0.857	0.095	0.048	0.000	0.000	0.000	0.000	0.048	0.000	0.095	0.000
se.hs.counties-s	3	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
se.hs.counties-sw	5	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
se.hs.registrars-h	108	0.898	0.185	0.093	0.019	0.056	0.065	0.000	0.019	0.065	0.019	0.000
se.hs.registrars-hw	113	0.903	0.186	0.088	0.009	0.053	0.053	0.000	0.009	0.062	0.009	0.000
se.hs.registrars-s	34	0.912	0.324	0.147	0.029	0.088	0.029	0.000	0.000	0.147	0.000	0.000
se.hs.registrars-sw	36	0.889	0.278	0.111	0.000	0.083	0.083	0.000	0.000	0.111	0.000	0.000
se.hs.financial-h	67	0.701	0.075	0.030	0.000	0.045	0.119	0.015	0.030	0.060	0.045	0.000
se.hs.financial-hw	71	0.704	0.070	0.028	0.000	0.042	0.127	0.014	0.028	0.070	0.042	0.000
se.hs.financial-s	16	0.813	0.125	0.063	0.000	0.063	0.188	0.000	0.063	0.063	0.125	0.000
se.hs.financial-sw	31	0.871	0.097	0.032	0.000	0.097	0.194	0.000	0.065	0.065	0.097	0.000
se.hs.gocs-h	48	1.000	0.083	0.042	0.000	0.125	0.000	0.000	0.104	0.063	0.104	0.000
se.hs.gocs-hw	55	0.964	0.091	0.073	0.000	0.127	0.000	0.000	0.091	0.055	0.109	0.000
se.hs.gocs-s	4	1.000	0.250	0.500	0.000	0.250	0.000	0.000	0.000	0.000	0.000	0.000
se.hs.gocs-sw	9	1.000	0.222	0.222	0.000	0.222	0.000	0.000	0.111	0.111	0.111	0.000
se.hs.education-h	38	1.000	0.158	0.132	0.000	0.026	0.000	0.026	0.132	0.000	0.132	0.000
se.hs.education-hw	44	0.977	0.205	0.114	0.000	0.023	0.000	0.023	0.114	0.000	0.159	0.000
se.hs.education-s	9	1.000	0.222	0.222	0.000	0.000	0.000	0.000	0.222	0.000	0.222	0.000
se.hs.education-sw	24	1.000	0.250	0.125	0.000	0.042	0.000	0.042	0.167	0.000	0.250	0.000
se.hs.isps-h	18	0.889	0.167	0.000	0.056	0.278	0.111	0.000	0.111	0.167	0.111	0.000
se.hs.isps-hw	19	0.895	0.211	0.000	0.053	0.263	0.105	0.000	0.105	0.263	0.105	0.000
se.hs.isps-s	6	0.833	0.167	0.000	0.167	0.167	0.167	0.000	0.167	0.500	0.167	0.000
se.hs.isps-sw	10	0.800	0.100	0.000	0.000	0.100	0.200	0.000	0.100	0.300	0.100	0.000
se.hs.media-h	25	0.880	0.600	0.280	0.000	0.480	0.080	0.000	0.160	0.280	0.280	0.000
se.hs.media-hw	27	0.926	0.630	0.259	0.037	0.481	0.074	0.000	0.148	0.222	0.370	0.000
se.hs.media-s	4	0.750	0.750	0.250	0.000	0.750	0.000	0.000	0.000	0.250	0.000	0.000
se.hs.media-sw	5	1.000	0.600	0.200	0.000	0.400	0.000	0.000	0.000	0.200	0.200	0.000
se.hs.municipalities-h	239	0.929	0.134	0.029	0.004	0.188	0.008	0.004	0.134	0.000	0.138	0.000
se.hs.municipalities-hw	258	0.934	0.097	0.019	0.004	0.190	0.012	0.008	0.136	0.000	0.140	0.000
se.hs.municipalities-s	41	0.927	0.146	0.024	0.000	0.146	0.000	0.000	0.098	0.000	0.122	0.000
se.hs.municipalities-sw	50	0.940	0.160	0.020	0.000	0.100	0.000	0.000	0.120	0.000	0.140	0.000
se.hs.pubauth-h	162	0.914	0.080	0.080	0.000	0.019	0.025	0.012	0.056	0.006	0.080	0.000
se.hs.pubauth-hw	188	0.926	0.064	0.064	0.000	0.021	0.027	0.011	0.048	0.005	0.074	0.000
se.hs.pubauth-s	15	0.933	0.067	0.000	0.000	0.000	0.000	0.000	0.067	0.000	0.133	0.000

Dataset	Domains w/ ext	Google	Facebook	Twitter	Microsoft	Amazon	Adobe	Yahoo!	AddThis	AppNexus	comScore	Quantcast
se.hs.pubauth-sw	36	0.917	0.056	0.056	0.000	0.083	0.056	0.000	0.111	0.000	0.167	0.000

Table C.16: Top Disconnect organization match coverage

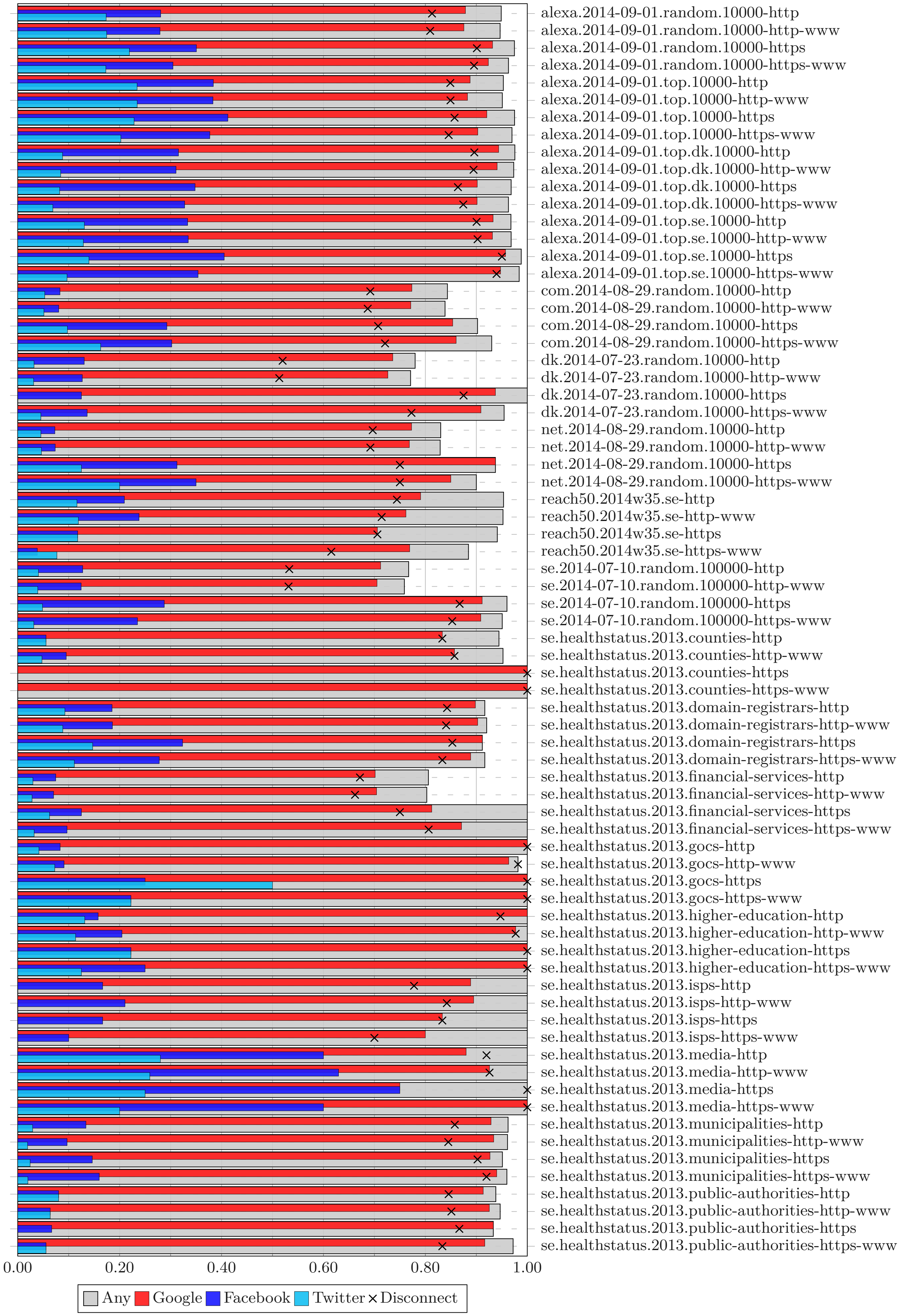

Figure C.10: Ratio of domains with requests to Google, Facebook and Twitter

C.12 Undetected external domains

While all external resources are considered trackers, parts of this thesis has concentrated on using Disconnect.me's blocking list for tracker verification. But how effective is that list of 2,149 *known* and *recognized* tracker domains across the datasets? Below is a comparison with the unique external primary domain count in each dataset; while the count is lower than the total number of external domains as it excludes subdomains (A.2), it matches the blocking style of Disconnect better.

The table below shows the number of unique external domains requested in each dataset, unique external primary domains and unique domains marked as trackers in Disconnect's blocking list. The difference between the number of Disconnect's detected tracker domains and external and primary domains respectively is shown next. The next column group shows the ratio of detected Disconnect domains over all external domains. Lastly, the ratio of domains detected as well as domains undetected by Disconnect, over the number of primary domains.

Figure C.11 shows the ratio of detected and undetected primary domains (x axis) per dataset.

While some of the domains which have not been matched by Disconnect are private/internal CDNs, the fact that less than 10% of external domains are blocked in top website HTTP datasets is notable. The blocking results are also around 10% or lower for random domain HTTP datasets, but it seems it might be connected to the number of domains in the dataset. Only 3% of the 15,746 external primary domains in .se 100k random domain HTTP dataset were detected. Smaller datasets, including HTTPS datasets with few reachable websites, have a higher detection rate at 30% and more.

Can a privacy tool using a *fixed blacklist* of domains to block be *trusted* – or can it only be trusted to be 10% effective? Regular expression based blocking, such as EasyList used by AdBlock, might be more effective, as it can block resources by URL path separate from the URL domain name (7.5.1) – but it's no cure-all. It does seem as if the blacklist model needs to be improved – perhaps by using whitelisting instead of blacklisting. The question then becomes an issue of either *cat and mouse* (5.2) – if the whitelist is shared by many users – or *convenience* – if each user maintains their own whitelist. At the moment it seems convenience and blacklists are winning, at the cost of playing cat and mouse with third parties who end up being blocked.

Dataset	Domains	Ext dom.	Prim.	D dom.	D diff ext.	D diff prim.	D/ext.	Prim. det.	Undet.
alexa.2014-09-01.random.10000-http	8 216	14 257	7 312	704	13 553	6 608	0.049	0.096	0.904
alexa.2014-09-01.random.10000-http-www	8 493	14 478	7 501	704	13 774	6 797	0.049	0.094	0.906
alexa.2014-09-01.random.10000-https	1 135	3 071	1 454	370	2 701	1 084	0.120	0.254	0.746
alexa.2014-09-01.random.10000-https-www	1 224	2 406	1 233	368	2 038	865	0.153	0.298	0.702
alexa.2014-09-01.top.10000-http	8 545	22 212	8 335	755	21 457	7 580	0.034	0.091	0.909
alexa.2014-09-01.top.10000-http-www	8 682	22 661	8 544	760	21 901	7 784	0.034	0.089	0.911
alexa.2014-09-01.top.10000-https	2 507	7 217	2 909	542	6 675	2 367	0.075	0.186	0.814
alexa.2014-09-01.top.10000-https-www	2 957	8 017	3 120	569	7 448	2 551	0.071	0.182	0.818
alexa.2014-09-01.top.dk.10000-http	2 263	2 768	1 407	282	2 486	1 125	0.102	0.200	0.800
alexa.2014-09-01.top.dk.10000-http-www	2 310	2 850	1 483	284	2 566	1 199	0.100	0.192	0.808
alexa.2014-09-01.top.dk.10000-https	339	816	420	151	665	269	0.185	0.360	0.640
alexa.2014-09-01.top.dk.10000-https-www	441	997	516	176	821	340	0.177	0.341	0.659
alexa.2014-09-01.top.se.10000-http	2 797	4 681	2 199	342	4 339	1 857	0.073	0.156	0.844
alexa.2014-09-01.top.se.10000-http-www	2 895	4 751	2 207	351	4 400	1 856	0.074	0.159	0.841
alexa.2014-09-01.top.se.10000-https	438	990	524	167	823	357	0.169	0.319	0.681
alexa.2014-09-01.top.se.10000-https-www	650	1 237	651	199	1 038	452	0.161	0.306	0.694
com.2014-08-29.random.10000-http	7 775	6 329	3 713	404	5 925	3 309	0.064	0.109	0.891
com.2014-08-29.random.10000-http-www	7 811	6 339	3 717	405	5 934	3 312	0.064	0.109	0.891
com.2014-08-29.random.10000-https	50	127	84	47	80	37	0.370	0.560	0.440
com.2014-08-29.random.10000-https-www	55	163	99	49	114	50	0.301	0.495	0.505
dk.2014-07-23.random.10000-http	7 180	4 272	2 834	278	3 994	2 556	0.065	0.098	0.902

Dataset	Domains	Ext dom.	Prim.	D dom.	D diff ext.	D diff prim.	D/ext.	Prim. det.	Undet.
dk.2014-07-23.random.10000-http-www	7 378	4 378	2 894	275	4 103	2 619	0.063	0.095	0.905
dk.2014-07-23.random.10000-https	23	52	33	26	26	7	0.500	0.788	0.212
dk.2014-07-23.random.10000-https-www	32	81	54	32	49	22	0.395	0.593	0.407
net.2014-08-29.random.10000-http	7 270	6 206	3 806	412	5 794	3 394	0.066	0.108	0.892
net.2014-08-29.random.10000-http-www	7 378	6 311	3 889	411	5 900	3 478	0.065	0.106	0.894
net.2014-08-29.random.10000-https	26	49	26	21	28	5	0.429	0.808	0.192
net.2014-08-29.random.10000-https-www	28	62	34	27	35	7	0.435	0.794	0.206
reach50.2014w35.se-http	43	339	195	92	247	103	0.271	0.472	0.528
reach50.2014w35.se-http-www	42	342	194	92	250	102	0.269	0.474	0.526
reach50.2014w35.se-https	18	117	66	41	76	25	0.350	0.621	0.379
reach50.2014w35.se-https-www	26	139	83	40	99	43	0.288	0.482	0.518
se.2014-07-10.random.100000-http	73 605	24 289	15 746	496	23 793	15 250	0.020	0.032	0.968
se.2014-07-10.random.100000-http-www	77 261	25 366	16 546	502	24 864	16 044	0.020	0.030	0.970
se.2014-07-10.random.100000-https	282	393	235	94	299	141	0.239	0.400	0.600
se.2014-07-10.random.100000-https-www	328	546	340	124	422	216	0.227	0.365	0.635
se.healthstatus.2013.counties-http	18	34	23	10	24	13	0.294	0.435	0.565
se.healthstatus.2013.counties-http-www	21	39	27	11	28	16	0.282	0.407	0.593
se.healthstatus.2013.counties-https	3	6	5	2	4	3	0.333	0.400	0.600
se.healthstatus.2013.counties-https-www	6	15	11	4	11	7	0.267	0.364	0.636
se.healthstatus.2013.domain-registrars-http	127	216	148	66	150	82	0.306	0.446	0.554
se.healthstatus.2013.domain-registrars-http-www	134	214	144	62	152	82	0.290	0.431	0.569
se.healthstatus.2013.domain-registrars-https	40	124	86	46	78	40	0.371	0.535	0.465
se.healthstatus.2013.domain-registrars-https-www	42	116	79	40	76	39	0.345	0.506	0.494
se.healthstatus.2013.financial-services-http	67	137	97	49	88	48	0.358	0.505	0.495
se.healthstatus.2013.financial-services-http-www	72	144	97	50	94	47	0.347	0.515	0.485
se.healthstatus.2013.financial-services-https	16	47	37	24	23	13	0.511	0.649	0.351
se.healthstatus.2013.financial-services-https-www	31	71	50	32	39	18	0.451	0.640	0.360
se.healthstatus.2013.gocs-http	49	130	83	45	85	38	0.346	0.542	0.458
se.healthstatus.2013.gocs-http-www	57	150	95	47	103	48	0.313	0.495	0.505
se.healthstatus.2013.gocs-https	4	44	28	21	23	7	0.477	0.750	0.250
se.healthstatus.2013.gocs-https-www	9	65	44	27	38	17	0.415	0.614	0.386
se.healthstatus.2013.higher-education-http	40	73	53	24	49	29	0.329	0.453	0.547
se.healthstatus.2013.higher-education-http-www	47	74	52	26	48	26	0.351	0.500	0.500
se.healthstatus.2013.higher-education-https	9	38	25	16	22	9	0.421	0.640	0.360
se.healthstatus.2013.higher-education-https-www	24	63	45	22	41	23	0.349	0.489	0.511
se.healthstatus.2013.isps-http	18	111	76	47	64	29	0.423	0.618	0.382
se.healthstatus.2013.isps-http-www	19	135	92	55	80	37	0.407	0.598	0.402
se.healthstatus.2013.isps-https	6	84	63	41	43	22	0.488	0.651	0.349
se.healthstatus.2013.isps-https-www	10	89	66	43	46	23	0.483	0.652	0.348
se.healthstatus.2013.media-http	26	346	190	81	265	109	0.234	0.426	0.574
se.healthstatus.2013.media-http-www	28	378	207	79	299	128	0.209	0.382	0.618

Dataset	Domains	Ext dom.	Prim.	D dom.	D diff ext.	D diff prim.	D/ext.	Prim. det.	Undet.
se.healthstatus.2013.media-https	4	102	58	24	78	34	0.235	0.414	0.586
se.healthstatus.2013.media-https-www	5	95	59	28	67	31	0.295	0.475	0.525
se.healthstatus.2013.municipalities-http	249	207	113	39	168	74	0.188	0.345	0.655
se.healthstatus.2013.municipalities-http-www	271	203	113	39	164	74	0.192	0.345	0.655
se.healthstatus.2013.municipalities-https	44	67	41	18	49	23	0.269	0.439	0.561
se.healthstatus.2013.municipalities-https-www	54	73	42	18	55	24	0.247	0.429	0.571
se.healthstatus.2013.public-authorities-http	170	172	110	48	124	62	0.279	0.436	0.564
se.healthstatus.2013.public-authorities-http-www	203	170	111	48	122	63	0.282	0.432	0.568
se.healthstatus.2013.public-authorities-https	18	32	21	9	23	12	0.281	0.429	0.571
se.healthstatus.2013.public-authorities-https-www	37	63	41	23	40	18	0.365	0.561	0.439

Table C.17: Requests per domain and ratios

While only aggregate numbers per dataset have been presented here, it would be interesting to use the full list of undetected primary domains to improve Disconnect's blocking list. While it is an endless endeavor, sorting by number of occurrences would at least give a hint as to which domains might be useful to block. The same list could be used to classify some of the domains as private/internal CDNs.

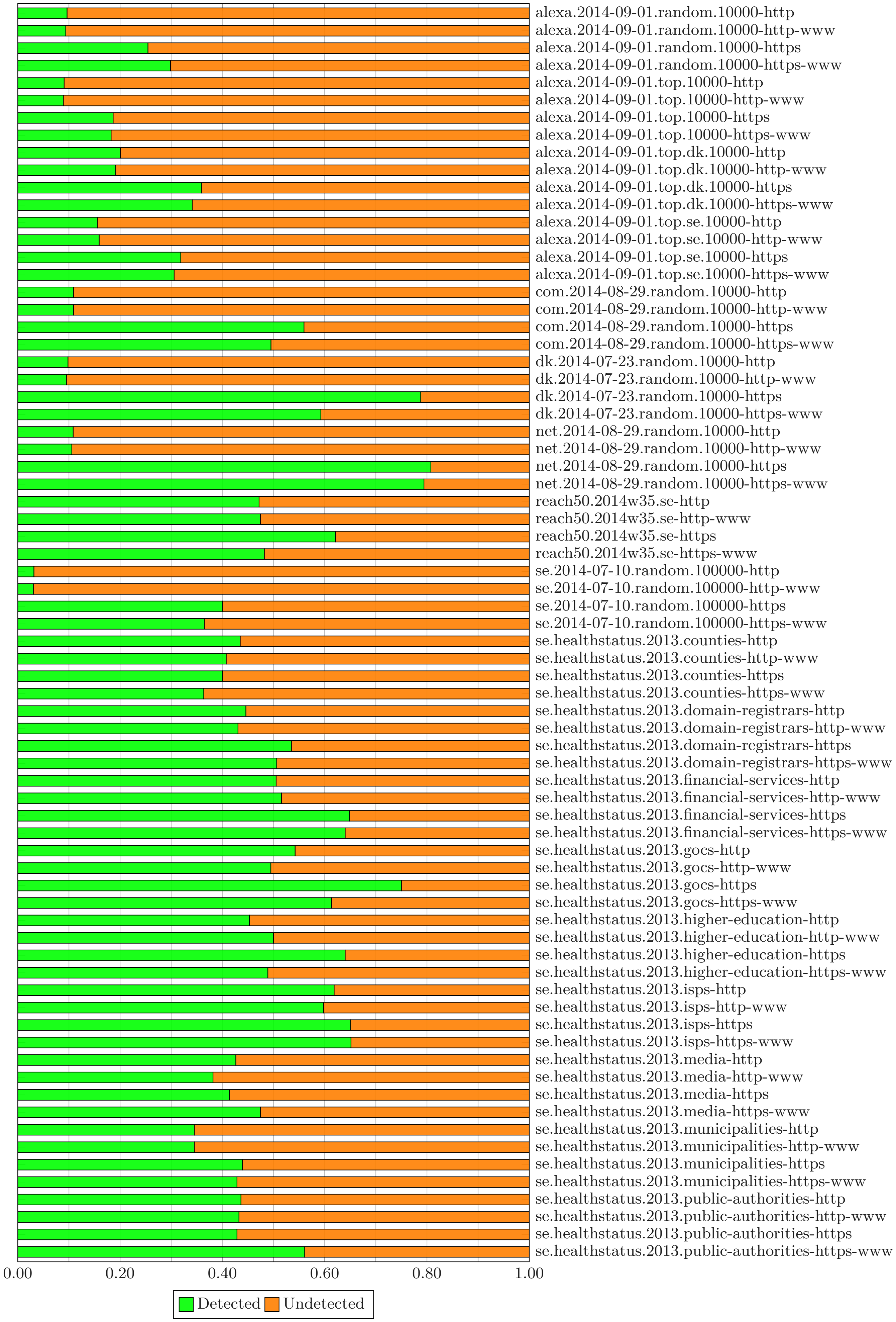

Figure C.11: Distribution of external primary domains detected/undetected by Disconnect

The end

You can find updated document versions, as well as source code and datasets online[9]. Thank you for reading this far – feedback would be very much appreciated!

[9] http://joelpurra.com/projects/masters-thesis/

This page intentionally left blank.

Almost.