



# Universiteit Leiden

## ICT in Business

### Web Tracking Detection System (TDS)

An effective strategy to reduce systematic monitoring  
and profiling of user habits across websites

Name: R.J.W. (Rob) van Eijk

Student-no: s0895393

Date: 26/08/2011

1st supervisor: Prof.dr. J.N. Kok

2nd supervisor: Dr. M.R.V. Chaudron

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)  
Leiden University  
Niels Bohrweg 1  
2333 CA Leiden  
The Netherlands

## Disclaimer

The content of this thesis represents the views of the author and does not necessarily reflect the views of his employer, or any other organization the author is active in or is associated with.

# Abstract

This thesis is about tracking online user behavior. Tracking is a granular problem, that goes beyond the phenomenon of behavioral advertising. It takes place in the context of rapid technological, social and political developments.

The main contributions of this thesis are:

- A working definition for web tracking.
- An opt-out mechanism for web tracking based on regular expressions repository in combination with persistent opt-out cookies.
- We show that the current do-not-track-me register, an opt-out mechanism for online behavioral advertising (OBA) based on opt-out cookies, of IAB/EASA falls short on it's promise. The opt-out cookie expiration dates are inconsistent, often less then five years. This is in contrast with the NAI opt-out tool, where all opt-out cookies have a minimum expiration data of five years. Added to that, opt-out behavior is collected by third parties.
- We show that the interconnectedness between nodes expressed in the number of links per node is an indicator (clustering coefficient) for web tracking.
- We show that the number of filtered nodes with TDS as a percentage of the total number of nodes is an indicator for web tracking.
- Five rules for constraint based graph mining of HTTP header information show that with use of a confirmed web tracking repository, it is possible to identify new third party tracking activity with connectivity color maps and pattern matching.

This thesis explores the real problems of tracking. First, there is no widely accepted definition of tracking. Supplier centric stakeholders proclaim a narrow scope definition whereas user centric stakeholders are clear about a wide scope definition. The working definition is: "The non-consensual collection or processing or storing of data for the purpose of systematic monitoring or profiling of user's habits across websites"

The collection or processing or storing of data required by law, necessary for the purpose of information security (availability, integrity, confidentiality) or fraud prevention is outside of the scope of the working definition. This working definition isn't killing the OBA business and leaves open whether we get to a solution for the Do Not Track debate through self regulatory agreements or with help of additional laws.

Second, users are often not aware that the traces they leave behind online and offline are valuable data. The ways these data can be used is unknown to them.

Third, the traces users leave behind contain enough information to draw a meaningful picture about a user's behavior. In many cases digital traces contain unique identifiers that can easily be connected to an individual. But also when anonymous digital traces are being collected, connecting data to an individual is still possible. When a large dataset contains enough entropy, re-identification to a unique individual is possible.

In order to make a contribution to the ongoing debate, the thesis explores a filter. With this filter, we see that tracking can be detected and suppressed in real time with a filter. The effectiveness of the filter is measured with an experiment using content from 819 newspaper websites in the European Union. With help of a base dataset of confirmed tracking domains we will derive two indicators for web tracking from the results.

Finally the results are visualized in connectivity color maps. This opens the door to future work in the field of graph mining. With connectivity color maps we look for patterns. By applying constraint based pattern mining we will see that five patterns can be derived from the experiments that indicate web tracking.

Keywords: Web Tracking, Customer lifetime value, Profiling, Privacy, Data Protection.

# Contents

1	Introduction	1
1.1	Motivation for the research . . . . .	1
1.2	Structure of the thesis . . . . .	2
2	Preliminary	3
3	Towards a working definition for web tracking	10
3.1	Research questions . . . . .	10
3.2	Identification of actors . . . . .	11
3.3	Identification of criteria . . . . .	12
3.4	Analysis of criteria for a working definition . . . . .	14
3.5	Working definition . . . . .	16
4	Web Tracking Detection System (TDS)	20
4.1	Assumptions for a filter to block web tracking . . . . .	20
4.2	Opt-out cookies . . . . .	21
4.3	Regular expressions . . . . .	27
5	Indicators for web tracking	31
5.1	Introduction . . . . .	31
5.2	Data acquisition and preparation . . . . .	31
5.3	Findings . . . . .	32
5.4	Conclusions . . . . .	35
6	Patterns in connectivity maps	37
6.1	Introduction . . . . .	37
6.2	Data acquisition and preparation . . . . .	38
6.3	Findings . . . . .	38
6.4	Conclusions . . . . .	42
7	Conclusion and future work	44
8	References	47
A	Results of the analysis of criteria for a working definition	51
B	Log files for creating nodes and links for EU newspaper websites	55
C	Tracking Detection System (TDS)	68

D	Source code: TDS cookies	70
E	Source code: TDS regular expressions	73
F	Source code: creating nodes and links	95
G	Source code: visualizing links and nodes with connectivity color maps	105

# 1 Introduction

## 1.1 Motivation for the research

Do Not Track (DNT) is a granular problem without a clear definition, that goes beyond the phenomenon of behavioral advertising. Web tracking takes place in the context of rapid technological and social developments. According to Koffijberg, Dekkers, Homburg, & van den Berg (2009, p. 1) such a context: *(...) leads to a society where people can be followed virtually anywhere and anytime, and where they leave numerous meaningful digital traces.* This is a free translation, the original Dutch wording is: *"(...) opvattingen over privacy in de context van de snelle technologisch-maatschappelijke ontwikkelingen die leiden naar een samenleving waarin mensen vrijwel overal en altijd in beeld zijn, gevolgd kunnen worden en veelzeggende digitale sporen achterlaten".*

This thesis is about tracking user behavior on the Internet. It is a snapshot of the ongoing discussion on tracking user behavior for marketing purposes. The reason for the research lies in the fact that the phenomenon of behavioral web tracking is a complex topic. The purpose of this thesis is to contribute to the ongoing debate on user privacy and the collection of online user habits. Most of the privacy and data protection principles on which legislation is based in both the European Union and the United States are being updated. As a result the debate has intensified.

We will start with exploring the criteria for a working definition which is not limited to online behavioral advertising. On 28/29 April 2011 the World Wide Web Consortium (W3C) Workshop on Web Tracking and User Privacy took place in Princeton (USA). I attended the workshop and participated in the panel discussion. On 22/23 June the Berkeley Center for Law and Technology (BCLT) organized together with University of Amsterdam's Institute for Information Law (IViR) the Online Tracking Protection & Browsers. This time, the discussion was hosted in Brussels. Many of the attendants of the debate in Princeton were present in Brussels. To come to a working definition, both workshops will serve in a case study.

Besides the working definition, this thesis intends to make a contribution to the ongoing debate by exploring an effective filter strategy. With this filter, we will see that tracking can be detected and suppressed in real time with a filter. The effectiveness of the filter is measured with an experiment using content from 819 newspaper websites in the European Union. Will will derive two indicators for web tracking from the results.

Finally we will visualize the results in connectivity color maps. This opens the door to future work in the field of graph mining. With connectivity color maps we look for patterns. By applying constraint based pattern mining we will see that patterns can be derived from the experiments that indicate web tracking.

## 1.2 Structure of the thesis

Next, a brief overview of the chapters in this thesis.

Chapter 2: We give an introduction to tracking user habits on the Internet. This chapter contains short history of the current debate to get grip on the main developments in tracking user behavior are brought forward. We will look at the flow of information from a user perspective. From there we will drill down to an example of targeted online behavior advertising (OBA).

Chapter 3: We look at the methodology to come to a working definition. The W3C workshop in Princeton and the Online Tracking Protection workshop in Brussels are used as a case study to analyze the different criteria for a working definition. Next we look at frequently used definitions in the debate, which are presented and critically discussed. The result of a stakeholder analysis is presented. The dominant criteria for a working definition for tracking are rooted in both workshops. Each round of discussion has been analyzed on dominant criteria and put into a table. This leads towards a working definition.

Chapter 4: In this chapter we will look at a web Tracking Detection System (TDS). With the working definition in place, we can explore if effective prevention of reduction of tracking user habits across websites is possible. First we will look at the assumptions for a filter with which web tracking can be blocked. Then we will explore blocking web tracking with opt-out cookies. In the last paragraph we will explore blocking web tracking with regular expressions. The thesis makes a contribution by exploring an opt-out mechanism. We will look at a filter that detects and suppresses tracking in real time. The design for the filter is based on managing privacy risk. The countermeasures for the reduction of the risk form the basis for the technical implementation.

Chapter 5: We will use the Web Tracking Detection System (TDS) in order to find indicators for web tracking. We will look at the effectiveness of the TDS filter. This is done with an experiment using content from 819 newspaper websites in the European Union. The methodology and the software to analyze HTTP headers are explained first. Then we look at the results.

Chapter 6: We will explore connectivity maps with colored nodes in order to look for patterns in clusters of links and nodes. From these patterns simple rules can be derived. These rules are the basis for graph mining and future work.

Chapter 7: This chapter concludes the research and looks at future work. It summarizes the research results and we will look critically at this thesis. Transparency and reproducibility and validity of the research will be addressed here as well.



## 2 Preliminary

First we start to mark the mostly stateless nature of the HTTP protocol and the role of cookies as a result. Then we will look at the beginning the Do-Not-Track debate followed by a short history leading to current developments. We will look at the flow of information from a user perspective. From there we will drill down to an example of targeted online behavior advertising (OBA).



Figure 1: Screen shot of a Dutch Donald Duck advertisement on the German Berliner Morgenpost (Source: morgenpost.de). The offer for a discounted subscription is the result of an automated decision which associates frequently visiting children's websites with buying intent. This is an example of targeted online behavior advertising (OBA).

The Hypertext Transfer Protocol (HTTP), which was initially standardized with RFC 2068 (Fielding, Gettys, Mogul, Frystyk, & Berners-Lee, 1997), is a mostly stateless protocol. More sophisticated web applications that need to maintain state use the cookie concept, defined in RFC 2109 (Kristol & Montulli, 1997). Cookies have found widespread usage in Web development and their current usage is being documented in Barth (2011).

Cookies have not only been used by web sites that the user explicitly wanted to connected to. Instead it has become common in Web deployment practice to “mash up” content from various other Web sites, including websites that provide advertising material. Over time the techniques for distributing information about user's web browsing behavior has become more sophisticated and researchers, such as Krishnamurthy & Wills (2009), have described the state-of-the-art.

Center for Democracy and Technology (2011) marks the beginning of the DNT discussion to be October 2007 when a coalition of public interest groups called on the Federal Trade Commission (FTC). With the publication of the preliminary FTC privacy report (Federal Trade Commission, 2010) in December 2010, which followed a series of round table discussions, concerns about the development in the area of user tracking on the Web has gotten the attention of the industry.

In discussions early 2011, the FCC reiterated its support for the Do Not Track (DNT) concept and articulated several success criteria for a solution towards tracking:

1. Implemented universally

2. Easy to use, find and understand
3. Persistent
4. Not only for use but also for collection
5. Effective and enforceable

In the meanwhile the European Commission has decided to tighten existing legislation by amending the e-Privacy Directive by the so-called “EU Cookie Directive” (European Commission, 2009). Implementation of the directive into national law by European member states is required by May 2011. The directive requires end user consent to the storing of cookies on a computer. This leads to active engagement of the political parties.

At the time of writing explicit consent for third party cookies for the purpose of tracking user behavior is part of the political debate. Because developments in the legislative process are currently very active we will leave these developments out of the scope of this thesis. The regulation of online tracking is addressed by (Tene & Polonesky, 2011). Their thesis discusses the range of policy considerations in Europe, United States and self regulation programs. Other leading document on regulation in the tracking and advertising context are Opinion 4/2010 (WP174) on the European code of conduct of FEDMA for the use of personal data in direct marketing (ARTICLE 29 Data Protection Working Party, 2010b), Opinion 2/2010 (WP171) on online behavioral advertising (ARTICLE 29 Data Protection Working Party, 2010a) and Opinion 3/2003 (WP77) on the European code of conduct of FEDMA for the use of personal data in direct marketing (ARTICLE 29 Data Protection Working Party, 2003).

Shortly after the publication of the preliminary FTC report industry players reacted by initiating standardization and implementation efforts. The IETF submission by Mozilla (Mayer, Narayanan, & Stamm, 2011) suggested standardization of an HTTP header conveying a preference of the user not to be tracked (the “Do Not Track (DNT) header”). Microsoft submitted a similar contribution (Zeigler, Bateman, & Graff, 2011) to the W3C, which additionally contains a black list mechanism. The final FTC report is expected late 2011.

These Microsoft and IETF contributions and the Mozilla DNT contribution in particular raise a number of interesting challenges for the standardization community. In addition to the typical technical questions there are also questions about the interaction between the technical and the regulatory community.

At the moment of writing, Mozilla has made available a DNT version in Firefox for mobile devices. Other browser manufacturers still have to follow.

The description of Online Behavioral Advertising (OBA) with a bow-tie structure in figure 2 is derived from work by Broder (2000). The figure displays an overview of limits to which third party data flows can be detected in the HTTP header. The data flows categories are derived from the 2010 Display Advertising Eco-System Europe (Improve Digital, 2011). The left side of the bow-tie graphic (symbolized with “IN”)

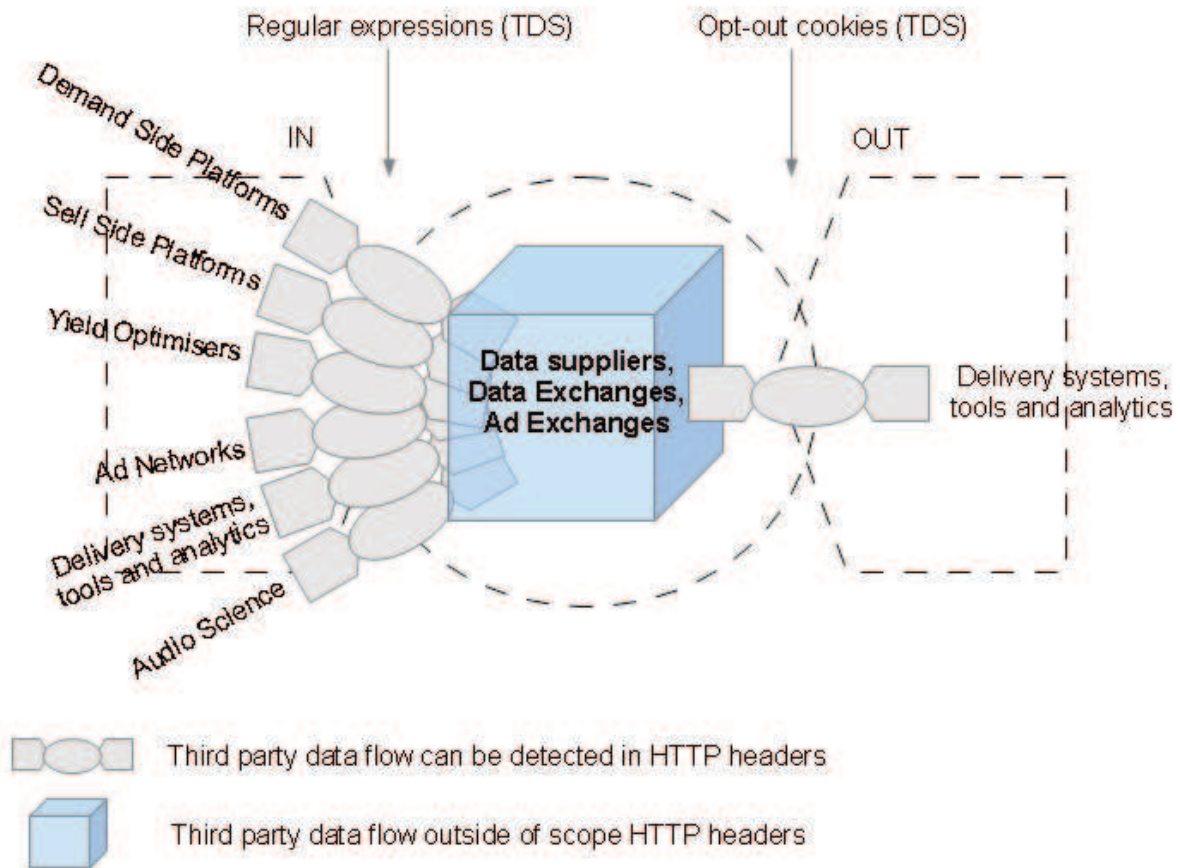


Figure 2: Bow-tie structure of Online Behavioral Advertising (OBA).

is the publisher's side, the right side (symbolized with "OUT") is the advertiser's side. An example of a publisher is a newspaper website (e.g. [telegraaf.nl](http://telegraaf.nl)) which rents space on his website where advertisers can display their content. Delivery systems, tools and analytics support the process of booking and managing online campaigns, forecasting, billing, verifying and measuring online advertising.

The regular expressions (TDS) and opt-out cookies (TDS) in figure 2 are compartments of a web filter. We will look at this filter in chapter 4.

Figure 3 is an attempt to illustrate the complexity involved in web tracking. The overview can be seen as a use case. A use case in systems engineering is a description of the interactions with a system. From a user perspective the complexity is hidden, the flows of information are often invisible. Figure 3 describes the interaction from a user with a website on a mobile smart phone.

The flow in figure 3 shows the data of an *Application Store*, in the case of this example, a shopping application. A shopping customer will visit the web shop using the *Mobile Application* on his smart phone. The user might see the web shop, without realizing that it is personalized based on his whereabouts. The Location Value Added Service takes care of this. Through *Data Analysis* of the items the user is interested in, *Advertising Services* can be notified through the *Mobile Network*. The user will see *Advertisements* as a result of his interests in products on the web shop. Payments are being handled

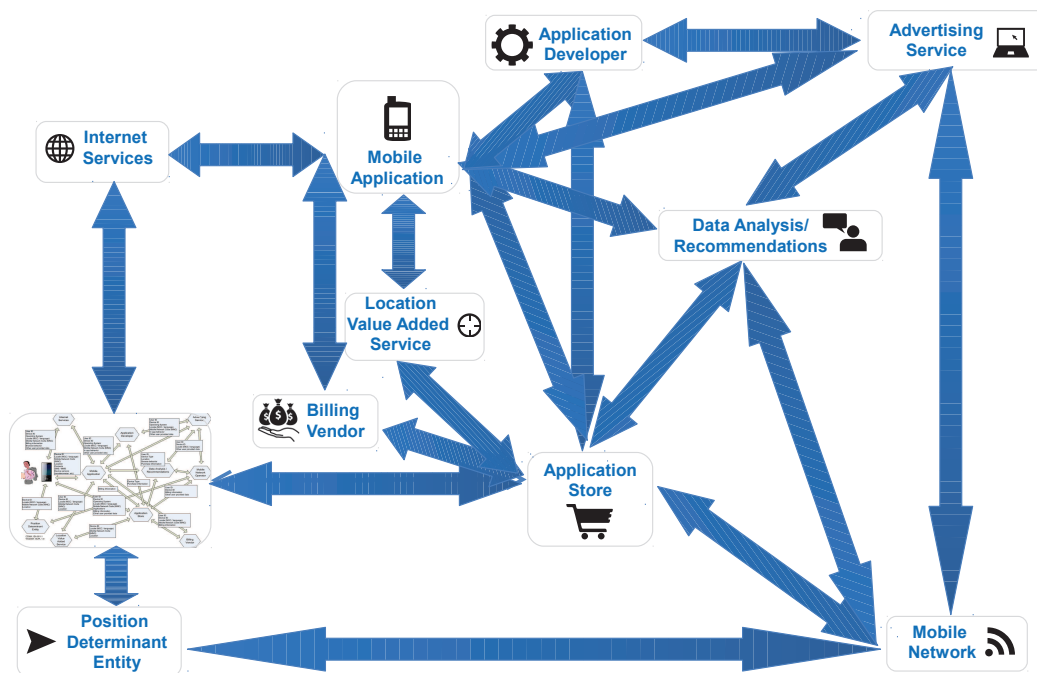


Figure 3: Technical overview. (Reproduced with permission of Jules Polonesky)

by a *Billing Vendor*.

Looking more closely at the use case gives us figure 4. To make the application work and present personalized content at the same time, lots of data is being exchanged between different parties. It becomes apparent that the information is tied together with Unique ID's. A UDID can be anything, from a MAC address of a device, to a fingerprint of the web browser that a user is using to surf the Internet. The uniqueness of a MAC address is inherent to the working of the Internet. The possibility of singling out a browser has been explored by the EFF project Panopticlick. The project "tests your browser to see how unique it is based on the information it will share with sites it visits".

In many cases digital traces contain unique identifiers that can easily be connected to an individual. In other cases, when anonymous digital traces are being collected, connecting data to an individual is still possible. As explained by Ohm (2009), when a large dataset contains enough entropy, re-identification to a unique individual is possible. Ohm (2009) is right in addressing that a promise has been broken. According to ECP-ECN (2010) 70% of the Dutch have a problem with the data collection by companies and organizations.

Unique ID's are unique identification numbers that can often be linked to a natural person. Cortesi (2011) shows that mobile applications are very noisy. According to Smith (2010, p. 4) 68% of tested apps send UDID's upstream in the clear. Cortesi (2011) shows that 46% of applications that transmitted



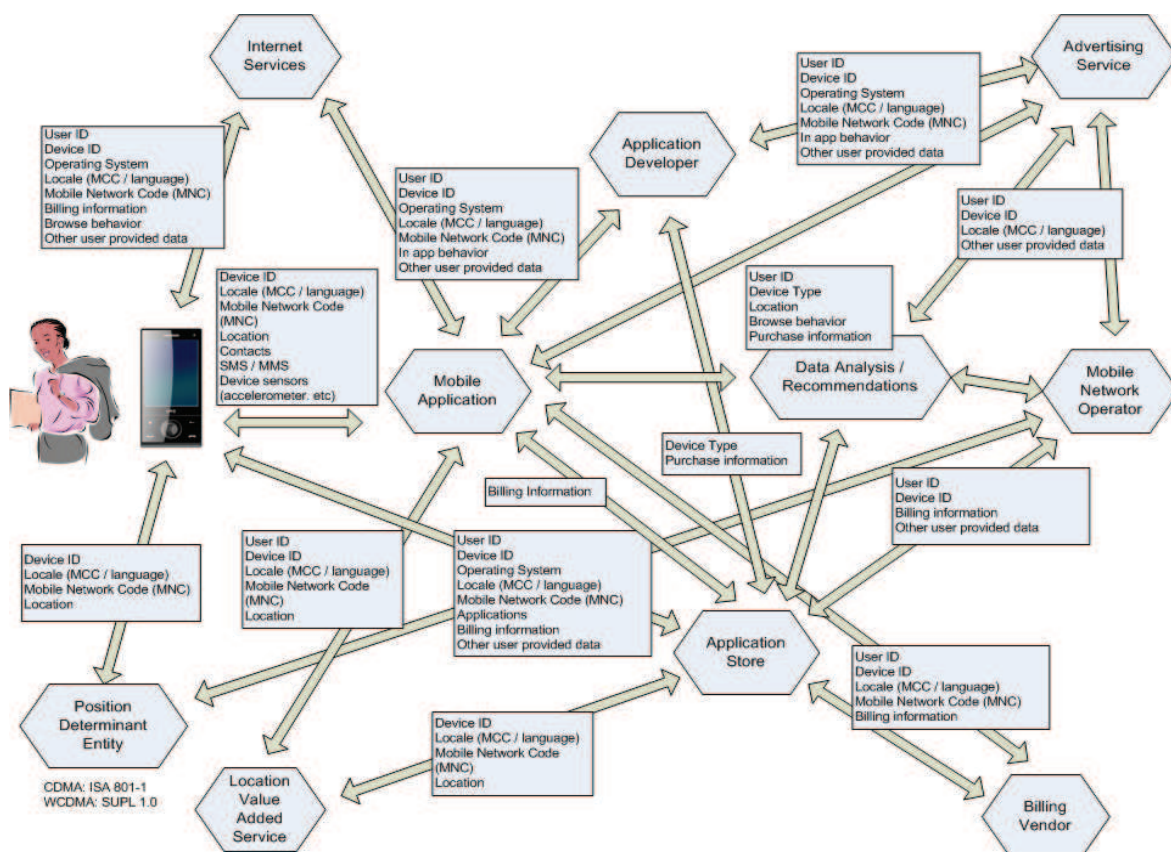


Figure 4: Detailed overview of information flow connected by User ID's and Device ID's. (Reproduced with permission of Jules Polonsky)

UDID's did so in clear text. 54% of applications transmitting UDID's used encryption for all UDID traffic. According to Cortesi the vast majority of applications sends UDID's to servers on the Internet. Also UDID-linked user information is aggregated in literally thousands of databases on the net.

In this context, UDID de-anonymization is a serious threat to user privacy. Also, it shows that the ongoing discussion in Europe about direct and indirect identifiability is relevant. UDID de-anonymization will often classify as personal data and therefore the data protection laws will apply. UDID de-anonymization can be used to create profiles.

With the amount of data increasing exponentially, the collection and use of information that contain identifiers is problematic according to many privacy experts. At the moment of writing this thesis, no simple solution has been agreed between stakeholders.

As a preparation, interviews have been held with a few privacy experts and transcribed a podcast on real-time customer intelligence (Network Security Podcast, Episode 241, 2011). The podcast gives a clear explanation of the marketing concept of customer value in both the on line and off line world. Furthermore, the podcast explores real-time customer intelligence, pulling multi-channel data into a single source, and explores into what this means for companies of all sizes. Another podcast on Unique

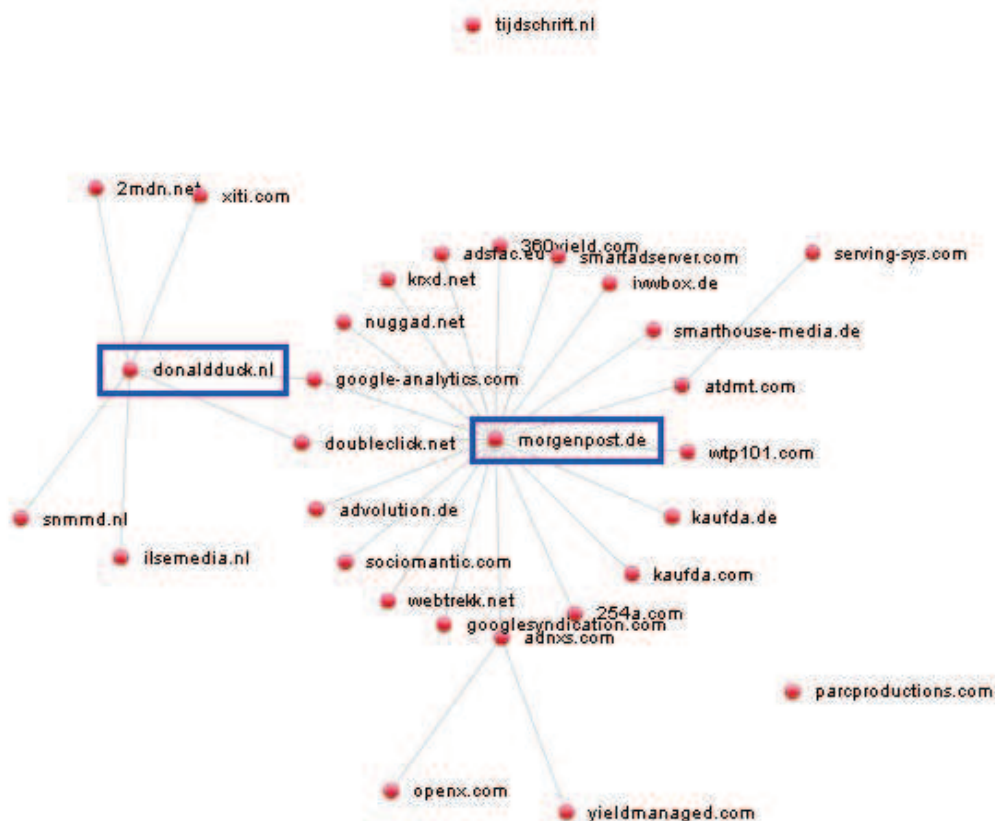


Figure 5: Example of information flow visiting the websites morgenpost.de and donaldduck.nl.

Identifiers (Beyond Web Analytics! Episode 45, 2011) has been transcribed. This podcast discusses state of the art de-anonymizing of Apple UDID's.

A good resource on profiling is Hildebrandt & Gutwirth (2008). It is not the technology itself, but the use of the technology that determines whether profiling practices are good or bad news. The visualization is an indicator of identifiable information being collected and used. Leenes uses the metaphor of a bubble chamber, an indicator for an unobservable phenomenon. In physics, a bubble chamber is a vessel filled with a very hot transparent liquid used to detect electrically charged particles moving through it. The comments of Leenes Hildebrandt & Gutwirth (2008, p. 296-298) apply to the illustration of figure 1 which is an indication of what really takes places in the background of your web browser is shown in figure 5.

Figure 1 shows an advertisement for a 13% discount on a Donald Duck subscription on the front page of the newspaper website Berliner Morgenpost. This is not a use case but a real world example. By analyzing web traffic entering and leaving the web browser, we create our own bubble chamber. The picture shows the origin and destination of content when visiting the websites morgenpost.de and donaldduck.nl. The front page of morgenpost.de is a rich mash-up with elements from various sources. The front page of donaldduck.nl pulls content from less different sources. However, the connections of

both websites link to doubleclick.net (Google) and google-analytics.com (Google). These links makes it possible to show targeted ads based on the fact that Google knows that both websites have been visited. Figures 5 and 1 are taken from experiments on analyzing HTTP traffic of visited websites. RAW traffic is processed with a handcrafted PERL script and visualized with Mike Bostock's D3 library. We will look into this in detail in chapter 4.

Before going to the next chapter it is important to note that figure 4 is not intended to limit a working definition for web tracking to data tied to unique identifiers. As described in ARTICLE 29 Data Protection Working Party (2010a), a list with websites a user has visited over time contains enough information to create a profile of a user's habits. However, because of the decreasing relevance of distinction between Personal Identifiable Information (PII) and Non-PII (Federal Trade Commission, 2010, p. 35), unique identifiers have been put more into the limelight of the debate. The proposed framework redefines Identifiable Information as "consumer data that can be reasonably linked to a specific consumer, computer or device".

### 3 Towards a working definition for web tracking

In this chapter we will look at the current debate on Do-Not-Track. On the basis of W3C Workshop on Web Tracking and User Privacy and the Online Tracking Protection & Browsers in Brussels we will formulate a working definition. In order to do that we first start with the research questions. Then we will identify stakeholders by grouping the organizations that are taking part to the debate. Next we will identify criteria that are being put forward in the workshops. We will come to the following working definition:

“The non-consensual collection or processing or storing of data for the purpose of systematic monitoring or profiling of user’s habits across websites”

The collection or processing or storing of data required by law, necessary for the purpose of information security (availability, integrity, confidentiality) or fraud prevention is outside of the scope of the working definition. This working definition isn’t killing the OBA business and leaves open whether we get to a solution for the Do Not Track debate through self regulatory agreements or with help of additional laws.

#### 3.1 Research questions

The reason for this research can *implicitly* be captured by the question:

*What is a good working definition of tracking?*

The purpose of this research is to provide insight into the different perspectives that stakeholders have about the collection and processing of the numerous meaningful digital traces left behind by users surfing the Internet. The main research question is:

*Which criteria can be identified that are important for one or more stakeholders?*

Tracking technology and Do Not Track are granular, therefor a limited scope is preferred. The scope is reflected in the breakdown of the main research question into the following sub questions:

*Which stakeholders can be identified?*

*What views do stakeholders have on tracking user behavior?*

Operationalizing these sub questions gives us the research questions and leads us to the operational model shown in figure 6. The operational research questions are:



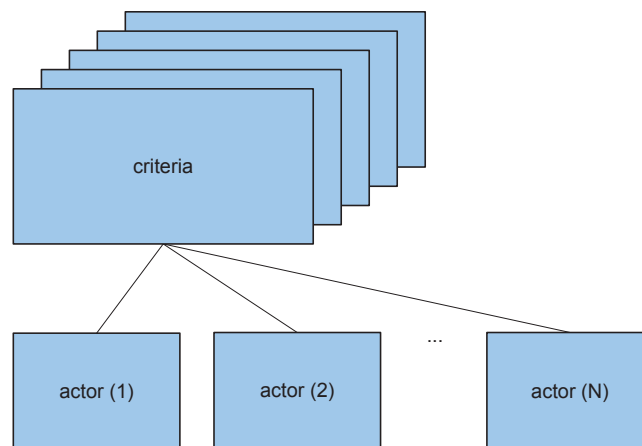


Figure 6: Operational model for a working definition of tracking.

*Which actors can be identified?*

*Which criteria can be identified?*

### 3.2 Identification of actors

In order to reduce the number of stakeholders, a grouping has been applied to the list of attendees. The resulting groups of stakeholders are the actors in this thesis. Grouping stakeholders is a commonly used technique in project management to describe a current situation.

The key aspect of the design of the research is to create the lens in such a way, that the researcher is able to look at different actors *in the same way*. Only by doing so it will be possible to compare the different groups and formulate a working definition. By viewing the stakeholders in the same way they become actors that can be compared to one another. The research validation consists of demonstrating that we are not comparing apples to oranges. Answering the research questions will lead us to a plausible scientific thesis.

Before addressing the research design, we will elaborate the concept of actors in relation to the participants to the workshop and stakeholders. We will do so by making use of existing project management literature to create an operational model.

A common way of grouping actors in a stakeholder analysis is described by Groote, Hugenholtz-Sasse, & Slikker (2000, pp. 108-118) and applied by for example van Aken (2002, pp. 73-80). The assumption here is that a group of actors have corresponding interests.

Another way of grouping is described by Morris (1988). Morris acknowledges that a strict grouping can not always be achieved. The boundaries between groups are not always sharp. Therefore we will take a practical approach in this research. The workshop attracted a broad collection of stakeholders. As

a result, a number of different perspectives were present, making the panel discussions interesting and valuable research data. The grouping of the list of attendees is based on the organization the attendees are representing. Another thing to keep in mind is that all attendees are users themselves as well.

First we divide the actors into three arbitrary groups. Group I represents the users. Group II represents the technology linking the user to group III. Group III represents the rest of the attendees. In van Aken (2002, p. 80) the difference between group II and III, from a project management point of view, is distinguished as group II having an internal project perspective, where as group II is having an external project perspective. Grouping the actors as shown in figure 7 corresponds with this difference in perspectives.

Then we try to get to a more granular grouping of organizations. Among the stakeholders were people from standardization bodies (ISOC, IETF, IAB, W3C) implementers from the mobile and desktop space, large and small content delivery providers, advertisement networks, search engines, policy and privacy experts, experts in consumer protection, and other parties with an interest in Web tracking technologies, including the developers and operators of Services on the Web that make use of tracking technologies for purposes other than to behavioral advertising.

The result of the grouping is shown in figure 7. A different grouping has been created can be found in Idate-TNO-IViR (2008, pp. 137-142). Instead of a stakeholder analysis, the economic concept of a value chain is used to distinguish the different stakeholders from one another. Other notable groupings are the preliminary staff report of the Federal Trade Commission (2010, appendix C1) which visualizes the personal data ecosystem, the Advertising Market Maps by LUMA Partners LLC (2010) and also Improve Digital (2011).

### 3.3 Identification of criteria

The methodology used is problem-finding research, as described by Verschuren & Doorewaard (2000, pp. 38-39). Data reduction takes place by comparing the dominant criteria and actors of the workshops that have taken place in Princeton with the dominant criteria and actors of the workshop that has taken place in Brussels. The audience was able to participate in the debate at both workshops. The research units represent the topic at hand.

The purpose of this type of case study, is to determine the criteria that underpin a problem. The purpose is not to solve the problem at hand. In the process of solving the problem, it first needs to be clear why something is a problem and why it is problematic. Solving the problem is outside the scope of this kind of research. The research only contributes to the process of solving of the problem.

A research question looking at the criteria and actors is relevant. It allows for looking at causal relations between criteria that are problematic and the reason(s) why that most probably is. This research methodology allows for looking at criteria that actors often are unaware of. It allows for deductive reasoning that something is a problem. In this way we are trying to frame the debate on web tracking. The purpose of this methodology is to uncover some of the true nature of the problem of web

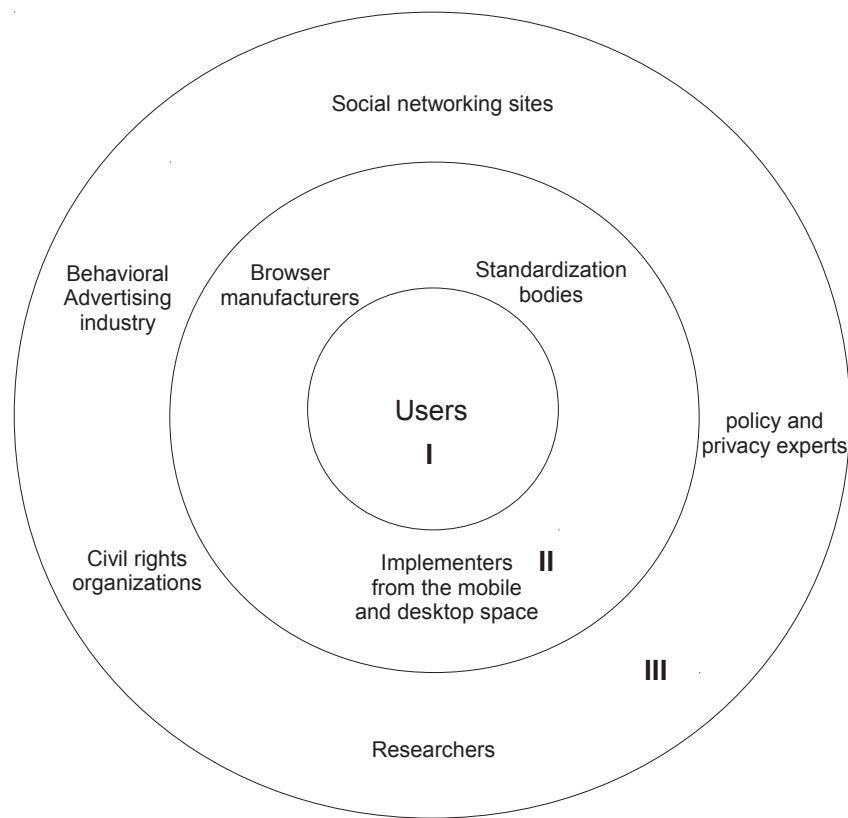


Figure 7: Visual representation of the different actors.

tracking.

According to Verschuren & Doorewaard (2000, pp. 169-176) a case study has six characteristics:

- *a small number of research units;*
- *a labor-intensive approach;*
- *more depth than breadth;*
- *a selective, or strategic sample;*
- *qualitative data and ditto research method;*
- *an open observation on location.*

This research design meets all six characteristics that Verschuren & Doorewaard (2000) describe for a case study. First, by taking the panel discussions as research units, the units stay small in number. We count nine rounds of panel discussions at the Princeton workshop, and four rounds at the workshop in Brussels. So we are looking at a total of 13 research units.

Second, preparing for the interviews with key actors and detailed note taking are labor intensive. Not only the collection of data is time consuming, so is the analysis of the data in order to be able to draw conclusions.

Third, as noted in the previous paragraph, a limited scope is preferred. By looking at criteria and actors only, the research will gain depth. Also, the position papers give an in depth perspective to the short presentations and often contain specific bibliographic references. The position papers have been reviewed by the program committee before acceptance.

Fourth, the strategic sampling of data has been done by the program committee. By organizing different rounds of discussions around strategic areas it is possible to cover the topic of the debate bit by bit.

Fifth, the workshops are taken to be the subject, the lens, through which the theoretical focus, the object, is viewed and explicated. The object can be viewed as the differences or similarities in views on web tracking between stakeholders.

And finally, because both workshop have been attended, this characteristic can be checked off.

Data reduction takes place by determining the criteria that different actors expressed in panel discussions. As noted in the previous paragraph 3.1, a key aspect of the design of this research is to create the lens in such a way, that the researcher is able to look at different actors *in the same way*. In order to achieve this research quality criteria, the approach of van Aken (2002, p. 79) is applied. The lens is shaped by criteria that form implicit and/or explicit points of view held by individual actors that can be derived from the research data.

Table 1 (p. 15) shows an example of a stakeholder analysis in two dimensions. On the horizontal axis, the different actors of figure 7 are listed. On the vertical axis, the different criteria "A" to "Z" are listed that will be derived from the panel discussions. In this example criteria "A" is shared by the actors "Browser manufacturers" and "Standardization bodies". Where as criteria "C" is only important for the "Implementers from the mobile and desktop space". Data reduction on the identified criteria will result in a working definition for tracking.

This research has been conducted in the period march - august 2011.

### 3.4 Analysis of criteria for a working definition

The results of the analysis of criteria for a working definition are shown in Appendix A (p. 51). We noticed that the implicit assumptions stakeholders have about each other are visible in the results. The criteria can be roughly divided into user centric and supplier centric. This is a common division in project management. Because the table in the appendix serves the purpose of data reduction, we will put the criteria into context first. The discussion about similarities and differences will follow in chapter 3.5.

The essence of the topic at hand is "not just bits on a wire, but also the broader meaning of do not

Table 1: Example of data reduction by using criteria and actors.

Criteria important for one or more actors	Browser manufacturers	Researchers	Civil rights organizations	Standardization bodies	Implementer	Policy and privacy experts	Behavioral advertising industry
A	X			X			
B				X			
C					X		
...							
Z			X			X	X

track.” (Roessler, 2011). “Personalization seems innocent, but may be close to unfair discrimination of consumers who no longer operate in a transparent market.” (Hustinx, 2011).

Vice-President of the European Commission responsible for the Digital Agenda Online privacy Ms. (Kroes, 2011) calls for three principles: “Transparency so that citizens know exactly what the deal is.”, “Fairness so that citizens are not forced into sharing their data.”, and “User control so that citizens can decide - in a simple and effective manner - what they allow others to know.”.

Commissioner of the Federal Trade Commission Ms. Brill calls for five principles: “it must be easy to use”, “it must be effective”, “it must be universal”, “it must be about the collection as well as use of information”, “it must be persistent”.

Different attempts to frame the topic into a working definition have been made. Most notably are:

- Hustinx (2011): “My short definition of OBA is that it is the presentation of targeted advertising on websites based on large scale tracking of consumer behavior online.”
- Soltani: “Tracking is about the collection and use of data tied to unique identifiers.”
- Brock (2011a), (Brock, 2011b): “Tracking is the non-consensual use or transfer of behavioral data collected across websites or applications as to an individual, computer or device.”
- Eckersley (2011): “Tracking is the retention of information that can be used to connect records of a person’s actions or reading habits across space, cyberspace, or time.”
- Center for Democracy and Technology (2011): “Tracking is the collection and correlation of data about the web-based activities of a particular user, computer, or device across non-commonly branded websites, for any purpose other than specifically excepted third-party ad reporting practices, narrowly scoped fraud prevention, or compliance with law enforcement requests.”

### 3.5 Working definition

The working definition proposed by the Center for Democracy and Technology (2011) is a starting point for many of the stakeholders. With the working definition as a boilerplate, CDT draws a sharp line as for which use of data falls within and which use falls beyond the scope. CDT has come with a proposal that includes and excludes criteria of a working definition of tracking.

CDT proposes to consider the following uses of data to be within the scope of the working definition of tracking:

- Include third-party online behavioral advertising;
- Include third-party behavioral data collection for first party uses;
- Include third-party behavioral data collection for other uses;
- Include behavioral data collected by first parties and transferred to third parties in identifiable form;
- Include demographic information appended to the user's device.

CDT proposes to consider the following uses of data to be outside of the scope of the working definition of tracking:

- Exclude third-party ad and content delivery;
- Exclude third-party analytics;
- Exclude third-party contextual advertising;
- Exclude first-party data collection and first-party use;
- Exclude federated identity transaction data;
- Exclude specially excepted third-party ad reporting;
- Exclude data collection required by law and for legitimate fraud prevention purposes.

(Barocas & Nissenbaum, 2011, pp. 3-4) point out that there are ethical issues at stake in online behavioral advertising (OBA). There is a difference between targeting and tracking. Targeting is about serving ads based upon interests inferred from online behavior. According to Nissenbaum, tracking is however "the relentless tracking and capture of online behavior".

An example of the difference between tracking and targeting can be explained by looking at the privacy statement of OpenX (Figure 8). When OpenX technology is used to serve an ad based upon interests inferred from online behavior on a website of one of its clients, this ad serving qualifies as

### 1. Information We Collect Through The OpenX Market.

(a) We may collect anonymous usage information that does not identify an individual User ("Non-Personal Information") when Publishers, advertisers, and ad networks use the OpenX Market. The fields of information that we collect may include, without limitation, the following:

Type	Field	Description
User Information	Unique Browser ID	A unique string that identifies this specific browser. This is used to provide more relevant advertising to the User.
	IP Address	A number which identifies a Users ISP. This is used for Geolocation information. <a href="#">Click here for more information.</a>
Page Information	Domain	The domain name of the website being viewed
	Page	The specific page of the website being viewed
	Referrer	The domain and page of the referring (previous) page being viewed
Ad Information	Size	The size of each of the ads on the page.
	Format	The ad formats acceptable on the web page
	Location	The location of each of the ads on the web page

This information is stored by OpenX and associated with Users by establishing an anonymous identification number for each user. When Users visit a Publisher's website, they may be identified by this anonymous identification number through the OpenX Cookie (further described below).

Figure 8: Collection and use based upon unique identifiers. (Source: openx.com)

targeting. When OpenX technology is used to collect data from multiple websites that a user has visited, this collection qualifies as tracking.

As with all technology, the essence of it is neutral (Arthur, 2009). It's the use of technology that has political, and cultural implications. Therefore, limiting the scope of the debate to cookies isn't correct. The example of OpenX illustrates that other technical means of collecting unique identifiers are possible. Using a unique string identifies a specific browser for the purpose of identifying a user has its implications. A user centric perspective means a wide scope beyond OBA and beyond the advertising



sector, a supplier centric perspective would imply a narrow scope, limited to OBA within the advertising sector.

The neutrality of technology itself is the main reason why a legal limitation of the use of technology should be based on open norms. It is important to notice that the use of tracking technology can be in line with the open norms in the new version of thesis 5(3) of the e-Privacy Directive, but only under specific conditions. Hustinx (2011) states that “activity is only allowed on condition that the user concerned has given his or her consent, having been provided with clear and comprehensive information in accordance with Directive 95/46/EX, *inter alia* about the purposes of the processing.”

Hustinx (2011) continues: “The new text of thesis 5(3) requires consent of the user concerned, which must be given before storing or accessing of information. The e-Privacy Directive also makes it clear that this consent should fulfill the requirements of thesis 2(h) of the Data Protection Directive, i.e. it should be a freely given, specific and informed indication of his wishes by which the user signifies his agreement to information being stored or access on his terminal.”. This statement is in line with the Opinion 15/2011 (WP187) on the definition of consent (ARTICLE 29 Data Protection Working Party, 2011).

The Netherlands is one of the first countries where the revised e-Privacy Directive will be implemented. The legislative proposal is to be discussed in the Senate. The legislative proposal has been amended by (van Bommel & van Dam, 2011) and now contains the specific open norm that systematic monitoring of the user’s behavior should be considered as processing of personal data under the Dutch Data Protection Directive.

Users are often not aware that the traces they leave behind online and offline are valuable data and the ways these online and offline data can be used is unknown to them. For example, ECP-ECN (2010, pp. 9-10) shows that only 28% of the users of a popular Dutch supermarket customer loyalty card are afraid of their data being used for something they do not want. This stands in sharp contrast with ECP-ECN (2010, pp. 15-16) which shows that 51% of the users of a popular Dutch social networking site are afraid of their data being used for something they do not want.

The traces users leave behind contain enough information to draw a meaningful picture about a user’s behavior. Therefore data protection laws and principles apply. The non-consensual collection of data should therefore be included in the working definition.

Looking at the results in appendix A, all stakeholders agree on the fact that the use of tracking technology needs limitation. Whether this is through new laws or self regulation is still under debate. Technological building blocks like the Do Not Track header could pave the way. But all stakeholders agree that the Do Not Track header is not necessarily the best mechanism. Whatever the technological building block will be to deal with tracking, all stakeholders agree that technological building blocks shouldn’t break any existing functionality.

Based upon the findings of the analysis of criteria for a working definition and taking all the similarities and differences into account we come to a working definition. The short definition of tracking is:



“The non-consensual collection or processing or storing of data for the purpose of systematic monitoring or profiling of user’s habits across websites”

The collection or processing or storing of data required by law, necessary for the purpose of information security (availability, integrity, confidentiality) or fraud prevention is outside of the scope of the working definition. This working definition isn’t killing the OBA business and leaves open whether we get to a solution for the Do Not Track debate through self regulatory agreements or with help of additional laws.

## 4 Web Tracking Detection System (TDS)

With the working definition in place, we can explore if effective prevention of reduction of tracking user habits across websites is possible. In this chapter we will look at a Web Tracking Detection System (TDS).

First we will look at the assumptions for a filter with which web tracking can be blocked. Then we will explore blocking web tracking with opt-out cookies. In the last paragraph we will explore blocking web tracking with regular expressions.

In the next chapter 5 we will use the Web Tracking Detection System (TDS) in order to find an indicator for web tracking. As figure 2 shows, much of it takes place in the 'black box' in the middle of the bow-tie.

### 4.1 Assumptions for a filter to block web tracking

As shown in Figure 2 blocking web tracking can be handled at both ends of the data flow. At the beginning of the bow-tie this is done with regular expressions, at the end with opt-out cookies. The basic assumptions of the filter strategy are:

1. appear to be a first time visitor.
2. limit the collection of data.
3. limit the use of data.

In order to accomplish this, a user needs to:

- Adjust the default settings in the browser
- Reduce data flying in and out of the browser from first and third parties by using regular expressions.
- Reduce the use and sharing of data which are not blocked by regular expressions by sending opt-out cookies.

We will explore these three tasks. First, a user needs to adjust the settings in the Firefox browser, check that the default browser settings (Tools|Options|Privacy tab) are set to custom setting 'Clear history when Firefox closes' set. Clearing all history when Firefox closes will result in appearing to be a first time visitor in most cases when visiting a website.

Second, to reduce data flying in and out of the browser we are going to see how using regular expressions work in paragraph 4.3. The purpose of TDS filter is to reduce the privacy risk of collection and use of data, often tied to unique identifiers, that can track you across different websites. A granular approach is needed in order to allow third party content without being tracked. A lot of websites consist of content that originates from dozens of different parties. Blocking ads is not part of the scope.

However an important side effect of limiting the collection of data with TDS, is less ads appearing on web pages you visit. Blocking the link between both donaldduck.nl and morgenpost.de with doubleclick.com in the example of figure 5 (p. 8) results in an ad-free morgenpost screen shot.

The Web Tracking Detection System (TDS) for Adblock contains an effective set of regular expressions. Note: do not mix with other rule sets because other rule sets can overrule the regular expressions with filter and allow rules resulting in a false sense of tracking protection.

Third, TDS uses the vehicle of cookies to communicate with first and third parties. The following paragraph 4.2 will explore this further.

The design for the filter is based on managing privacy risk. The risk management strategy is based on countermeasures for reduction. The purpose is to reduce tracking across websites. The aim is not to prevent such a thing from happening. Prevention implies the 100% detection and suppressing of tracking in real time. At the moment that is not a realistic goal. Chapter 4 looks at the quality of the filter and implicitly how far we get to the 100%.

Blocking content with regular expressions has not yet been identified in Cooper & Tschofenig (2011). Cooper & Tschofenig (2011, pp. 8-14) gives an overview of mechanisms to opt-out of web tracking. Paragraph 4.3 therefor is a contribution to the debate and will be submitted to the Network Working Group of Internet Engineering Task Force (IETF).

The flow of information when a user visits a website is visualized in figure 2 by using a bow-tie structure. Blocking content with regular expressions and blocking content with opt-out cookies differ from a conceptual perspective.

The Web Tracking Detection System consists of two elements, a public repository of regular expressions and a public repository of opt-out cookies. We will see that regular expressions are effective for blocking traffic on the 'IN' side of figure 2. We will also see that opt-out cookies are an effective risk control on the 'OUT' side.

Within the OSI stack, IP addresses are in layer 3 (network layer), and the HTTP headers are in layer 6 (presentation layer). We will explore the connectivity color maps of HTTP traffic. Both IP addresses and resolved host names are displayed in the connectivity color maps.

## 4.2 Opt-out cookies

We will explore blocking web tracking with opt-out cookies. As shown in Figure 2 the opt-out cookies is a way to handle web tracking at the end of the data flow. In this paragraph we will look at the opt-out register offered by industry. Then we will look at some of the reasons to use a public domain alternative to the industry opt-out register.

Before we look at the details of the proposed filter strategy, it is good to take note of current developments in the debate about browser settings. We will start with the FTC and look at recent

developments in the Netherlands.

In his Statement Commissioner William E. Kovacic ask for public comments (Federal Trade Commission, 2010, p. 112, appendix D-3). Most notably in the context of this thesis are the the questions:

- “How should policy makers go about identifying mainstream consumer expectations for purposes of setting default terms with respect to data collection and use?”
- “When should such default terms be based on considerations other than consumer expectations?”
- “Should the chosen default terms be immutable?”
- “If not, what steps should consumers be required to take to override the defaults?”

In the recent developments these questions have become an important topic in the debate. The first three questions are about the default settings of current browser technology. The last question is about what would technically be necessary for a user to do in order to give consent to the collection and use data.

Taking into account recent developments in the Dutch legislative process, the Explanatory Memorandum belonging to the Amendment of van Bommel & van Dam (2011) makes an important note on the default terms of current browser. In fact, the Minister of Economic Affairs Agriculture and Innovation has explained in the debate on June 18, 2011 that most of the current web browsers accept third party cookies by default. To quote the Explanatory Memorandum belonging to the Amendment of van Bommel & van Dam (2011): “ten aanzien van browsers geldt dat de huidige browsers, die vaak standaard zo zijn ingesteld dat zij alle cookies accepteren, niet geschikt zijn om toestemming te verlenen. Dit kan natuurlijk veranderen als nieuwe of vernieuwende browsers de mogelijkheid bieden om specifiek aan te geven van welke partij of welke website de gebruiker cookies wil accepteren, maar op dit moment is dat niet zo.”

Part of the discussion in the Netherlands has been about first and third party cookies in conjunction with the steps to override the defaults. van Bommel & van Dam (2011) notes in the quote that current browser technology has no ability to specifically indicate which party or which website the user wants to accept cookies from. In other words, the technical acceptance of cookies is equal to the consent of the user to collect and use his/her data. van Bommel & van Dam (2011) marks that if current browser technology isn't capable of doing so, future versions of web browsers might well do so. “This of course can change as new or innovative browsers will have the ability to specifically indicate which party or which website the user wants to accept cookies from, but today is not so.”

To summarize these developments in the current debate:

- Current default browser settings can not be considered as a means to express consent from a user for collecting and using data across websites for tracking or Online Behavioral Advertising (OBA);

- New or innovative browsers may have the capability for a user to be clear about whether consent has been given;
- Cookies currently are an important technical vehicle to communicate with the user on the collection and usage of data across websites for Online Behavioral Advertising (OBA).

IAB Europe takes cookies as a vehicle for communication seriously. IAB Europe has launched a website to give a user control over cookies. It lists affiliated companies who collect and use information to provide online behavioral advertising. The website states "Using this tool only applies to online behavioral advertising and will not affect other services that use the same technology, called cookies, such as email, shopping basket preferences and photo hosting. Web sites that you visit may also still collect information / use cookies for other purposes." The website is based upon the idea to give the user more control.

Let's take another example on user control and cookies as a vehicle for communication. Let's look at the privacy policy of the company called 247realmedia<sup>1</sup>, which states: "The use of cookies by our affiliates, service providers or tracking utility company is not covered by our Privacy Policy. We do not have access or control over these cookies." What we learn from this is that users have to act themselves. Taking into account that most websites pull content from numerous sources, this leads to a daunting task for a user, if he/she wants to take cookies as a vehicle for communication seriously.

The following experiment demonstrates a tracking of user behavior across websites. Bol stores a unique Bol User Identifier (BUI=111.222.333.444.1311164917123123) in a cookie. The value of the cookie is the originating IP address (in this thesis anonymized to 111.222.333.444) of the user plus a unique session-identifier (in this thesis anonymized to 1311164917123123).

Bol.com also stores a unique identifier in its cookie (bn\_u1231231231231231231, in this thesis anonymized). This information is shared with third party company Baynote in a GET request. Baynote also receives the originating IP address of the user.

*http://bol-www.baynote.net/baynote/tags3/policy?customerldbol&codewww&subdomain  
&userId1231231231231231231&userPolicyRequested=true*

*GET /baynote/tags3/policy?customerldbol&codewww&subdomain&userId=1231231231231231231&  
userPolicyRequestedtrue HTTP/1.1*

*Host: bol-www.baynote.net*

*User-Agent: Mozilla/5.0 (Windows NT 5.1; rv:5.0) Gecko/20100101 Firefox/5.0*

*Accept: \*/\**

*Accept-Language: en-us,en;q=0.5*

---

<sup>1</sup> <http://www.247realmedia.com/EN-US/privacy-policy.html>

Accept-Encoding: gzip, deflate

Accept-Charset: ISO-8859-1,utf-8;q=0.7,\*;q=0.7

Connection: keep-alive

Referer: http://www.bol.com/nl/verkopenviabol/index.html

Finally, we look at an example that makes clear that relying on cookies alone as a means to communicate is not enough. As a company called Convertro<sup>2</sup> explains on their website: “The Convertro Visitor Tracking Code completely eliminates the issues of cookie dependency. This means that a visitor can clear his/her browser cookies and cache or toggle between browsers and still appear as a unique visitor. We also have at least a 50% success rate tracking the same visitor across machines if they do end up converting. The result is a very rich set of conversion data that most accurately represents the true sources that drove any particular conversion.”

The advertising industry was not inactive in light of the increasing concerns on cookies and offers a track-me-not register. The track-me-not register provides a mechanism to opt-out with cookies.

There are a number of reasons why Beef Taco is an essential component of the Web Tracking Detection System. First, due to the fact that cookies set through the IAB industry supplied website youronlinechoices.com is inconsistent on the expiration period. The results of an experiment on the expiration date of the opt-out cookies on the youronlinechoices.com website is shown in the table 2 on the next page.

Table 2 shows the value of all opt-out cookies in bold and *italic*. The value of the cookies when opt-in is the modus operandi are shown in normal font.

The column ‘Auto-expire’ shows the expiration date of the cookies. For instance the opt-out cookie for the host .bt.ilsemedia.nl with the name Wlopt-out has an expiration date of 360 days. This is in sharp contrast with for example the opt-out cookie for the host .criteo.com with the name opt-out which has an expiration date of 1827 days. Criteo follows the requirements set by the Network Advertising Initiative. The minimum expiration period for cookies on the NAI Opt-out web page is five years.

A second important reason for an alternative to the youronlinechoices.com website is that it tracks your opt-out behavior. The website contains multiple java-scripts that share information with confirmed tracking domains. These domains are confirmed by for example organizations like PrivacyChoice and Evidon.

Third, the website states “Using this tool only applies to online behavioral advertising and will not affect other services that use the same technology, called cookies, such as email, shopping basket preferences and photo hosting. Web sites that you visit may also still collect information / use cookies for other purposes.”<sup>3</sup> and lacks information or a privacy statement on the use of tracking technology on its own web page.

Fourth, the website places a number of persistent identifiers, even when opt-out cookies have been

---

<sup>2</sup><http://www.convertro.com/visitor-tracking.html>

<sup>3</sup><http://www.youronlinechoices.com/uk/your-ad-choices>

set. This results in the effect that when data is exchanged with the tracking domain, all cookies are being send, including unnecessary persistent unique identifiers.

The patched Beef Taco browser extension keeps your opt-outs permanent, sets a long expiration period and above all, has an community driven repository which is open source. Beef Taco gives you a clean collection of custom opt-out cookies with a consistent expiration date of well over five years, and without any unnecessary persistent unique identifiers. All the experiments in this thesis have been using the opt-out cookie set from version 3.6. Since the opt-out cookie repository is a community effort, the number of opt-out cookies is growing over time.

Next we will reflect on the objectivity of the tools applied for keeping the opt-out cookies. (draft cooper, tschofenig) give an overview of universal opt-out mechanisms for web tracking. They address the existence of other browser extensions for various browser platforms. For example Targeted Advertising Cookie Opt-Out (TACO) for the Firefox and Google Chrome, Keep My Opt-Outs (KMOO) for Chrome and Keep MORE Opt-Outs, developed by PrivacyChoice.

The main reason to use BeefTaco is that it is forked to perform just one simple task: keeping persistent opt-out cookies. The design philosophy is “one tool for one task. Other browser extensions might work in the same way. For our experiments, the BeefTaco extensions is fit for purpose.

For the sake of objectivity, the point to be stressed is that there is no preference for a particular browser add-on. What is essential is the ability of setting a custom expiration period and keeping the persistent cookies available, even after clearing the browser history and after restarting the web browser. This ability is regardless of the browser of choice. The main reason for choosing to do all experiments with Firefox is that the browser with its ability to be extended with our preferred add-ons is fit for purpose.

For the filtering experiment you need to install the Firefox Add-on Beef Taco<sup>4</sup> with IAB Europe opt-out cookies. A patch has been written to include the European IAB cookies. As of release 3.6 (July 23, 2011) this patch is included in the release. The source code of the supplied patch is included in appendix D (p. 70).

Double check that the browser settings are set to ‘Accept third-party cookies’. A technical issue overlooked in the public and political debate on cookies is that rejecting third party cookies results in opt-out cookies not being stored. Even if you had opt-out cookies in your browser, as is the case with add-ons like Beef Taco, setting the browser to ‘Reject third-party cookies’ will prevent the opt-out cookies from being send with the HTTP header!

---

<sup>4</sup><https://addons.mozilla.org/en-US/firefox/addon/beef-taco-targeted-advertising/>

Table 2. Results of analysis of opt-out cookies track-me-not register (Source: youronlinechoices.com)

id	last accessed	host	Auto- expire [days]	id	name	value	host
85	12-06-11 17:17	.adgenie.co.uk	365	85	_ngtid	ODMuMTYzLjE1LjY4-pohzFqApwzfwaxAEqBhawjEbkqsCwBgBH	.adgenie.co.uk
<b>94</b>	<b>12-06-11 17:20</b>	<b>.adgenie.co.uk</b>	<b>365</b>	<b>94</b>	<b>_ngtid</b>	<b>DEnviGmpuk</b>	<b>.adgenie.co.uk</b>
89	12-06-11 17:17	.adnxs.com	1	89	sess	1	.adnxs.com
90	12-06-11 17:17	.adnxs.com	90	90	uuid2	8900409985090056389	.adnxs.com
<b>108</b>	<b>12-06-11 17:20</b>	<b>.adnxs.com</b>	<b>1</b>	<b>108</b>	<b>sess</b>	<b>1</b>	<b>.adnxs.com</b>
<b>109</b>	<b>12-06-11 17:20</b>	<b>.adnxs.com</b>	<b>1826</b>	<b>109</b>	<b>uuid2</b>	<b>-1</b>	<b>.adnxs.com</b>
8	12-06-11 17:17	.advertising.aol.com	23	8	SESSff329d810a46b3a1bf645141daed34cf	de1729a91ebab011dd8717e81620c601	.advertising.aol.com
<b>8</b>	<b>12-06-11 17:20</b>	<b>.advertising.aol.com</b>	<b>23</b>	<b>8</b>	<b>SESSff329d810a46b3a1bf645141daed34cf</b>	<b>de1729a91ebab011dd8717e81620c601</b>	<b>.advertising.aol.com</b>
<b>102</b>	<b>12-06-11 17:20</b>	<b>.adviva.net</b>	<b>1825</b>	<b>102</b>	<b>ADVIVA</b>	<b>NOTRACK</b>	<b>.adviva.net</b>
11	12-06-11 17:17	.aol.com	730	11	s_vi	[CS]v1 26FA6C23851D2F68-6000012FA02800BA[CE]	.aol.com
12	12-06-11 17:17	.aol.com	730	12	s_pers	%20s_getnr%3D1307891782423-New%7C1370963782423%3B%20s_nrgvo%3DNew%7C1370963782424%3B	.aol.com
<b>11</b>	<b>12-06-11 17:20</b>	<b>.aol.com</b>	<b>730</b>	<b>11</b>	<b>s_vi</b>	<b>[CS]v1 26FA6C23851D2F68-6000012FA02800BA[CE]</b>	<b>.aol.com</b>
<b>12</b>	<b>12-06-11 17:20</b>	<b>.aol.com</b>	<b>730</b>	<b>12</b>	<b>s_pers</b>	<b>%20s_getnr%3D1307891782423-New%7C1370963782423%3B%20s_nrgvo%3DNew%7C1370963782424%3B</b>	<b>.aol.com</b>
<b>119</b>	<b>12-06-11 17:20</b>	<b>.atdmt.com</b>	<b>1827</b>	<b>119</b>	<b>TOptOut</b>	<b>1</b>	<b>.atdmt.com</b>
<b>118</b>	<b>12-06-11 17:20</b>	<b>.bing.com</b>	<b>1827</b>	<b>118</b>	<b>TOptOut</b>	<b>1</b>	<b>.bing.com</b>
7	12-06-11 17:17	.blinkx.com	365	7	bsid	a42c34b6c999dc498ed615bd8c0d5f09	.blinkx.com
86	12-06-11 17:17	.blinkx.com	1095	86	up	on	.blinkx.com
<b>7</b>	<b>12-06-11 17:20</b>	<b>.blinkx.com</b>	<b>365</b>	<b>7</b>	<b>bsid</b>	<b>a42c34b6c999dc498ed615bd8c0d5f09</b>	<b>.blinkx.com</b>
<b>112</b>	<b>12-06-11 17:20</b>	<b>.blinkx.com</b>	<b>1095</b>	<b>112</b>	<b>up</b>	<b>off</b>	<b>.blinkx.com</b>
60	12-06-11 17:17	.bt.ilsemedia.nl	0	60	iabtoken	e005179a3c9a0dc16d8bc663d41295e1	.bt.ilsemedia.nl
<b>99</b>	<b>12-06-11 17:20</b>	<b>.bt.ilsemedia.nl</b>	<b>0</b>	<b>99</b>	<b>iabtoken</b>	<b>936178cf7780373773f70946d6aa1ffe</b>	<b>.bt.ilsemedia.nl</b>
<b>100</b>	<b>12-06-11 17:20</b>	<b>.bt.ilsemedia.nl</b>	<b>360</b>	<b>100</b>	<b>WIOPtOut</b>	<b>1</b>	<b>.bt.ilsemedia.nl</b>
83	12-06-11 17:17	.criteo.com	1827	83	opt	*1OLcLD%2fflVt6BafC5UUEGckw%3d%3d	.criteo.com
<b>96</b>	<b>12-06-11 17:20</b>	<b>.criteo.com</b>	<b>1827</b>	<b>96</b>	<b>optout</b>	<b>1</b>	<b>.criteo.com</b>
91	12-06-11 17:17	.delivery.ctasnet.com	1000	91	RTC8	optedin	.delivery.ctasnet.com
<b>111</b>	<b>12-06-11 17:20</b>	<b>.delivery.ctasnet.com</b>	<b>1000</b>	<b>111</b>	<b>RTC8</b>	<b>a_</b>	<b>.delivery.ctasnet.com</b>
2	12-06-11 17:17	.doubleclick.net	0	2	pm_sess	ACi0TCi7JpLa6WaiemNmjrkerlmzpUWqohV7C4VIEuWQG8PR0sL8NSJ6C4x0L-X3V2NjT_KAYfx	.doubleclick.net
84	12-06-11 17:17	.doubleclick.net	730	84	id	221470d41f010029 t=1307891860 et=730 cs=usthkkj9	.doubleclick.net
<b>2</b>	<b>12-06-11 17:20</b>	<b>.doubleclick.net</b>	<b>0</b>	<b>2</b>	<b>pm_sess</b>	<b>ACi0TCi7JpLa6WaiemNmjrkerlmzpUWqohV7C4VIEuWQG8PR0sL8NSJ6C4x0L-X3V2NjT_KAYfx</b>	<b>.doubleclick.net</b>
<b>106</b>	<b>12-06-11 17:20</b>	<b>.doubleclick.net</b>	<b>14</b>	<b>106</b>	<b>OPT_OUT</b>	<b>1</b>	<b>.doubleclick.net</b>
<b>113</b>	<b>12-06-11 17:20</b>	<b>.doubleclick.net</b>	<b>7090</b>	<b>113</b>	<b>OPT_OUT</b>	<b>1</b>	<b>.doubleclick.net</b>
<b>110</b>	<b>12-06-11 17:20</b>	<b>.fastclick.net</b>	<b>3650</b>	<b>110</b>	<b>fastclick</b>	<b>optout</b>	<b>.fastclick.net</b>
14	12-06-11 17:17	.google.com	730	14	PREF	ID=58a4f081d54af6ce:TM=1307891823:LM=1307891823:S=IX0MwMK6mDIF3vXX	.google.com
<b>14</b>	<b>12-06-11 17:17</b>	<b>.google.com</b>	<b>730</b>	<b>14</b>	<b>PREF</b>	<b>ID=58a4f081d54af6ce:TM=1307891823:LM=1307891823:S=IX0MwMK6mDIF3vXX</b>	<b>.google.com</b>
<b>114</b>	<b>12-06-11 17:20</b>	<b>.live.com</b>	<b>1827</b>	<b>114</b>	<b>TOptOut</b>	<b>1</b>	<b>.live.com</b>
<b>120</b>	<b>12-06-11 17:20</b>	<b>.microsoft.com</b>	<b>1827</b>	<b>120</b>	<b>TOptOut</b>	<b>1</b>	<b>.microsoft.com</b>
<b>116</b>	<b>12-06-11 17:20</b>	<b>.msn.com</b>	<b>1827</b>	<b>116</b>	<b>TOptOut</b>	<b>1</b>	<b>.msn.com</b>
<b>103</b>	<b>12-06-11 17:20</b>	<b>.nuggad.net</b>	<b>3651</b>	<b>103</b>	<b>nuggstopp</b>	<b>true</b>	<b>.nuggad.net</b>
<b>107</b>	<b>12-06-11 17:20</b>	<b>.realmedia.com</b>	<b>3652</b>	<b>107</b>	<b>RMOPTOUT</b>	<b>3</b>	<b>.realmedia.com</b>
45	12-06-11 17:17	.revsci.net	11680	45	NETID01	TfTYgBMBFAoAAHk@AAQAAAAAT	.revsci.net
<b>45</b>	<b>12-06-11 17:20</b>	<b>.revsci.net</b>	<b>11680</b>	<b>45</b>	<b>NETID01</b>	<b>TfTYgBMBFAoAAHk@AAQAAAAAT</b>	<b>.revsci.net</b>
13	12-06-11 17:16	.scorecardresearch.com	730	13	UID	1fa4694c-82.94.229.25-1307891783	.scorecardresearch.com
<b>13</b>	<b>12-06-11 17:16</b>	<b>.scorecardresearch.com</b>	<b>730</b>	<b>13</b>	<b>UID</b>	<b>1fa4694c-82.94.229.25-1307891783</b>	<b>.scorecardresearch.com</b>
<b>117</b>	<b>12-06-11 17:20</b>	<b>.specificclick.net</b>	<b>1825</b>	<b>117</b>	<b>ADVIVA</b>	<b>NOTRACK</b>	<b>.specificclick.net</b>
<b>115</b>	<b>12-06-11 17:20</b>	<b>.specificmedia.com</b>	<b>1825</b>	<b>115</b>	<b>ADVIVA</b>	<b>NOTRACK</b>	<b>.specificmedia.com</b>
4	12-06-11 17:17	.unanimis.co.uk	3653	4	OPTOUT_SECURITY_TOKEN	7p1dxKpnmpQ8HrI0sKJUdjxXOCdSS85S3ERksUTm9cFRNjwYAS8QcXZerJvA8	.unanimis.co.uk
<b>4</b>	<b>12-06-11 17:20</b>	<b>.unanimis.co.uk</b>	<b>3653</b>	<b>4</b>	<b>OPTOUT_SECURITY_TOKEN</b>	<b>True</b>	<b>.unanimis.co.uk</b>
<b>101</b>	<b>12-06-11 17:20</b>	<b>.unanimis.co.uk</b>	<b>3653</b>	<b>101</b>	<b>OPTOUT</b>	<b>True</b>	<b>.unanimis.co.uk</b>
88	12-06-11 17:17	.yahoo.com	731	88	B	40o1mtp6v9m23&b=4&s=rq	.yahoo.com
<b>104</b>	<b>12-06-11 17:20</b>	<b>.yahoo.com</b>	<b>7305</b>	<b>104</b>	<b>AO</b>	<b>o=1</b>	<b>.yahoo.com</b>
<b>105</b>	<b>12-06-11 17:20</b>	<b>.yahoo.com</b>	<b>731</b>	<b>105</b>	<b>B</b>	<b>40o1mtp6v9m23&amp;b=4&amp;d=4auM3vprYH0wsQ--&amp;s=e1</b>	<b>.yahoo.com</b>
6	12-06-11 17:17	delivery.ctasnet.com	365	6	OAID	a8217e3080bd974009c6415867f1ee68	delivery.ctasnet.com
<b>6</b>	<b>12-06-11 17:20</b>	<b>delivery.ctasnet.com</b>	<b>365</b>	<b>6</b>	<b>OAID</b>	<b>a8217e3080bd974009c6415867f1ee68</b>	<b>delivery.ctasnet.com</b>
87	12-06-11 17:17	notrack.adviva.net	0	87	ansv4_uid	myvalue	notrack.adviva.net
<b>87</b>	<b>12-06-11 17:20</b>	<b>notrack.adviva.net</b>	<b>0</b>	<b>87</b>	<b>ansv4_uid</b>	<b>myvalue</b>	<b>notrack.adviva.net</b>
93	12-06-11 17:17	notrack.specificclick.net	0	93	smu	myvalue	notrack.specificclick.net
<b>93</b>	<b>12-06-11 17:20</b>	<b>notrack.specificclick.net</b>	<b>0</b>	<b>93</b>	<b>smu</b>	<b>myvalue</b>	<b>notrack.specificclick.net</b>
92	12-06-11 17:17	notrack.specificmedia.com	0	92	smu	myvalue	notrack.specificmedia.com
<b>92</b>	<b>12-06-11 17:20</b>	<b>notrack.specificmedia.com</b>	<b>0</b>	<b>92</b>	<b>smu</b>	<b>myvalue</b>	<b>notrack.specificmedia.com</b>
82	12-06-11 17:17	track.adform.net	60	82	C	1	track.adform.net
95	12-06-11 17:20	track.adform.net	730	95	C	3	track.adform.net



### 4.3 Regular expressions

We will explore blocking web tracking with regular expressions. First we why the regular expressions are an important contribution of this thesis. Next we will look at what web tracking HTTP content can be blocked and how this is done. Then we will look at the benefits of blocking with regular expressions. After that we will compare the approach with other similar approaches in a short reflection. The reason of the reflection on other approaches is to have an objective comparison. We will conclude this paragraph with notes on how to install the TDS rule set of regular expressions.

As shown in figure 2 (p. 5) regular expressions are a way to handle web tracking at the beginning of the data flow. In this paragraph we will look at the mechanism of blocking HTTP content based on regular expressions. Together with the mechanism of blocking content based on opt-out cookies, these two elements lead to an opt-out mechanism for web tracking. An opt-out mechanism for web tracking based on regular expressions repository in combination with persistent opt-out cookies is one of the main contributions of this thesis.

The experiments in chapter 5 lead to an indicator for web tracking. These experiments rely on a filter with a reference set of web tracking domains. The data acquisition for the experiments is done with the Web Tracking Detection System (TDS). Besides that the experiments in chapter 5 lead to an indicator for web tracking, the also lead to an indication on the effectiveness of the filter. The quality of the filter is expressed in rate of decline of the indicator for web tracking.

The design philosophy of the Web Tracking Detection (TDS) rule set of regular expressions is: keep things simple but granular. The granularity of regular expressions allows to go deeper than the domain level of URLs. This depth is possible due to the mostly stateless nature of the current Internet. This depth is needed to detect scripting and other modus operandi which we will explore now.

Regular expressions are like a Swiss army knife when it comes to matching patterns. With regular expressions a variety of content can be detected. For example blocking images of advertisements, filter ads from specific domains, filter ads by URL, pop-up killing, web bug (1x1 sized images) blocking, shockwave elements, Java scripts, HTML5 pings, access to geo-location information and detecting the use of the originating ip address in a HTTP GET or POST string.

Next we will reflect on the objectivity of the tools applied for regular expressions for the purpose of a Web Detection System (TDS). The main reason to use Adblock is that it is designed to capture and block content before the content gets executed in the browser. Other browser extensions might work in the same way. For our experiments, the Adblock Plus browser extensions is fit for purpose.

For the sake of objectivity, the point to be stressed is that there is no preference for a particular browser add-on. What is essential is the ability of capturing and blocking based on regular expressions BEFORE the content gets executed in the browser. This ability is regardless of the browser of choice. The main reason for choosing to do all experiments with Firefox is that the browser with its ability to be extended with our preferred add-ons is fit for purpose.

Piggy backing on the add-ons is what Web Tracking Detection System (TDS) is about. Piggy backing

with a surgical knife so to speak. As we saw in figure 2, the main contribution lies in the question where you cut and the effectiveness of cutting with the surgical knife. Loading the browser extensions with a data designed for the purpose of detecting and blocking web tracking is the essence. Regular expressions cut before execution of the content in the browser. Opt-out cookies cut in the end of the content flow by preventing the “tailoring of advertising see on websites to your likely interests or preferences on the web browser you are currently using<sup>5</sup>.

Other browser add-ons are Ghostery by Evidon. Ghostery runs on Safari, Firefox, Chrome Opera and Internet Explorer. Trackerblock runs on Firefox and Internet Explorer. This extension also operates on regular expressions. The difference however, is that it’s repository has focus on the US based players.

Other tools are Trackerblock developed by PrivacyChoice and Trusted Protection Lists (TPL) for Internet Explorer. These approaches use regular expressions, but only on domain level by black/white-listing domains. This approach lacks the granularity which can be attained with regular expressions.

A different approach is not with use of add-ons, but with help of a browser proxy. Detecting and blocking HTTP data takes in that case place outside of the browser. An example of a browser proxy that is capable of handling regular expressions is Privoxy. Privoxy operates from a design point of view (figure 1) also at the IN side of the bow-tie. Privoxy can do extra things that Adblock can not do, e.g. blocking cookies, block the referrer in the HTTP header, handle referrer forging in the HTTP header, GIF de-animation, blocking fast HTTP redirects, handle Image tag reordering.

Disabling the referrer can also be done in most browsers without the help of browser extensions. In Firefox you follow the following three steps:

1. Type about:config
2. Find Network.http.sendRefererHeader
3. Set the entry to 0 to disable referrer

Other tools that we need to address are PRLifetime privacy dashboard, Prividor (PRLifetime Violation DetectOR) and MozBlock. The PRLifetime privacy dashboard has a different design philosophy and are primarily aimed at reporting and detecting. Blocking is possible, but on a website-per-website basis. This tool also makes browser functionality available for the casual user, instead of having to dive deep under the hood for example by using the about:config feature described in the instruction for disabling the referrer. Another example of features that otherwise is difficult to reach is the disabling of access to the DOM storage.

Prividor is a tool aimed at doing a privacy audit on a site-by-site basis. The tool is developed by the German Fraunhofer-Institut SIT in Munchen on behalf of the Federal Data Protection Commissioner.

MozBlock is based on the Adblock browser extension. Its scope is limited to block third party content (e.g. webbugs) embedded within web pages to reduce cross-side click stream tracking by

---

<sup>5</sup><http://www.youronlinechoices.com/opt-out-interface?lang=uk>

advertisers. Unfortunately this project is inactive. MozBlock is part of the MozPETs project: Mozilla Privacy Enhancement Technologies. The project evolved from a research project at the Darmstadt University of Technology.

From a technical viewpoint regular expressions are cryptic. For example:

```
/*.(gif|jpe?g|png|bmp|ico)($|?).+/  
/tracking.quisma.com/c.cfs?.+/  
/rc.bt.ilsemedia.nl/(Tag|Get)/ilsemedia/JS/
```

The first regular expression typically blocks web bugs (1x1 pixel images) that are stored on third party hosts. The second expression blocks Java script from a specific tracking host within a third party domain. The third example of blocking web tracking with regular expressions is aimed at Java scripts (JS) from a behavioral tracking (.bt.) host within a third party domain.

Blocking web tracking content with regular expressions is not aimed at blocking advertisements. In the example of figure 2, ad blocking takes place at the “OUT” side of the bow-tie diagram. However, blocking data flows at the “IN” side results in less ads being served at the “OUT” side.

Figure 9 demonstrates the effects of blocking web tracking content. Example (A) in the figure shows the way the website looks when no filtering is taking place. The website displays various targeted advertisements. Example (B) shows the way the website looks when only TDS opt-out cookies are enabled. The result is that ads are still being shown, although the ads are different because they are not based on online behavioral targeting information. Example (C) shows the website when both TDS opt-out cookies and TDS regular expressions are enabled. It shows the visual effect of the elements of TDS on the content being displayed. Although the content stays the same, the advertisements have disappeared. Regular expressions (with or without opt-out cookies) result in ads not being served on this specific website.

From using the Web Tracking Detection (TDS) approach, a few benefits can be derived:

- blocking content results in less content being transferred over the Internet;
- less content results in faster loading times of a web page and a quicker user experience;
- reduction of the leakage of users' behavior across websites;
- TDS regular expressions are not designed to block ads, less ads presented is a side effect caused by stopping the web tracking flow before it sets delivery systems, tools and analytics in motion;
- set and forget is part of the design philosophy. The rule set with regular expressions is automatically updated with a frequency of five days;

- the rule set is open source and can be maintained and crowd sourced as a community effort.

For the filtering experiment you need to install the Firefox Add-on Adblock Plus<sup>6</sup>. Do not install any rule sets at this point. The rule set of regular expressions in appendix E is a handcrafted set, designed to block web tracking HTTP headers. Using the TDS blocking rule set of regular expressions in combination with other rule sets like EasyPrivacy Tracking Protection List or Fanboy Tracking List is not recommended. Combining different lists results in overruling of regular expressions, i.e. white listing rules that allow domain traffic are dominant to blocking rules and as a result get executed first, or allowing domain traffic by hiding rules which hide the content in the web browser but doesn't block the web tracking HTTP traffic. Hiding rules are also dominant to blocking rules and as a result get executed first.

We have seen that misconfiguration of the rule sets with regular expressions could lead to white-list tracking domains and leaking behavioral data. The same can be concluded for misconfiguration of cookie settings. This could also lead to leaking behavioral data. In order to quantify the effectiveness of the configuration of the Web Tracking Detection System (TDS) we will look at the results of an experiment in the following chapter.

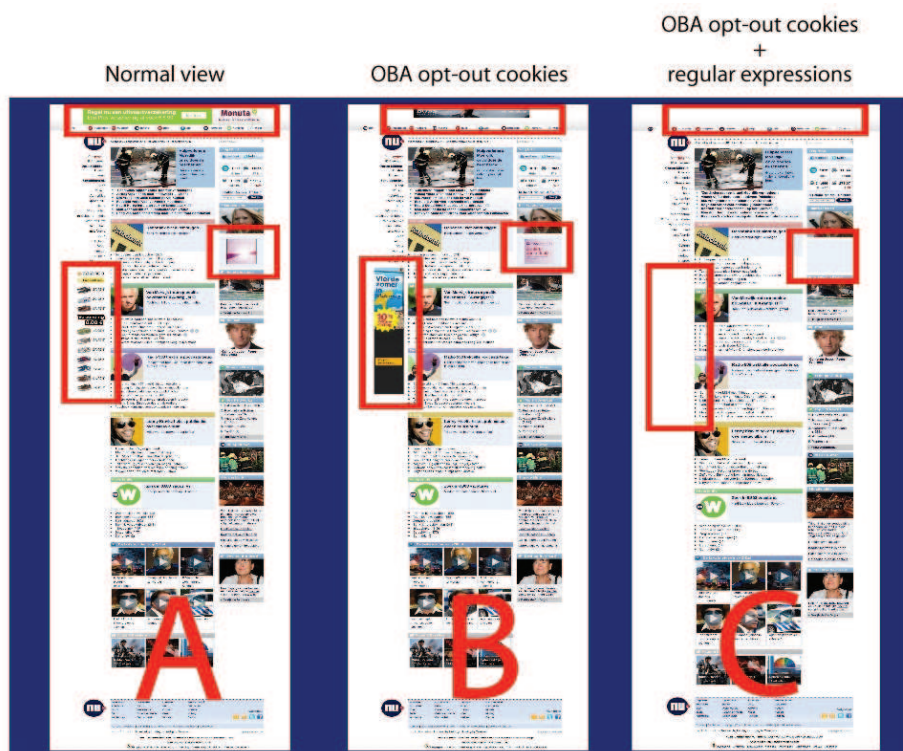


Figure 9: Effects of application of Web Tracking Detection System (TDS) elements to a web page (Source: www.nu.nl)

<sup>6</sup><https://addons.mozilla.org/nl/firefox/addon/adblock-plus/>

## 5 Indicators for web tracking

We will explore the quality of the Web Tracking Detection System (TDS) with an experiment. We will see that the experiments in this chapter lead to an indicator for web tracking. These experiments rely on a filter with a reference set of web tracking domains. The data acquisition for the experiments is done with a web browser by visiting a set of 819 newspaper websites across different countries in the European Union and European Economic Area. The experiments lead to an indication of the effectiveness of the filtering. This is done with the Web Tracking Detection System (TDS). The experiments also lead to an indicator for web tracking.

### 5.1 Introduction

In order to measure the effectiveness of the Web Tracking Detection System (TDS) an experiment has been set up. The tools used to conduct the experiment are:

- Web browser (Firefox version 5).
- Bookmarks pointing to 819 newspaper websites across different countries in the European Union and European Economic Area (Source: [www.eufeds.eu](http://www.eufeds.eu)).
- A browser extension to capture the HTTP headers of the generated traffic (LiveHTTPheaders version 0.17).
- Web Tracking Detection System (TDS) which consists of two browser extensions (AdBlock Plus version 1.3.9, BeefTaco version 1.3.6) and a rule set with regular expressions and a public repository with permanent opt-out cookies (version Aug 3, 2011).
- Base dataset as a reference for confirmed trackers (Source: PrivacyChoice, File: `trackers.json`, Retrieved at July 1, 2011).
- PERL script to process the captured HTTP headers and convert the results to JSON format to make visualization in connectivity color maps possible. (Appendix F).

The experiment has been conducted at Aug 3, 2011 and Aug 4, 2011.

### 5.2 Data acquisition and preparation

The logging of HTTP headers is done with the firefox add-on LiveHTTPheaders. In order to have a large dataset the Eufeds dataset is used. This dataset spans across different EU or EEA countries. Eufeds is a tool provided by the European Journalism Centre.

According to its website "The European Journalism Centre (EJC) is an independent, international, non-profit institute dedicated to the highest standards in journalism, primarily through the further training of journalists and media professionals".

The hyper links per country on the website [www.eufeds.eu](http://www.eufeds.eu) are used as base data set. In order to be able to reproduce the results of the experiments, the URLs's of the base data set are bookmarked within the web browser. Opening all bookmarks at the same time will start the data capturing. At the moment that all web pages in the tabs are loaded, all tabs are reloaded through a right mouse click and selecting the menu item "reload all tabs". Figure 17 (p. 55) shows how this is done for all the national and regional newspapers of Estonia.

For the transparency and reproducibility of the experiment, the processing log files for creating nodes and links for the EU newspaper websites are added to appendix B. Besides the number of nodes and links the log files show the name of the file to which the captured HTTP data was saved to, the number of processed HTTP headers, the number of unique nodes, the item number of unique links connecting the nodes, the number of nodes found from the base set of confirmed trackers (PrivacyChoice) and the time it took to process the data.

To quantify the number of links per node with and without the Web Tracking Detection System (TDS), the PERL program in appendix F does the counting during the processing of the HTTP headers:

```
$dbh->disconnect();
my $proc = time - $start;
if ($proc) {
    $log = $log . "$cnt_headers headers";
    $log = $log . "$cnt_names nodes";
    $log = $log . "$cnt_domains links";
    $log = $log . "$cnt_tracker confirmed trackers";
    $log = $log . "job completed succesfully in $proc seconds";
}
```

The from the PERL program resulting numbers are entered in a spreadsheet. Appendix B contains all the log files and the numbers obtained by the processing of the captured HTTP headers while visiting the bookmarked EU newspaper websites.

### 5.3 Findings

The results of the processing have been analyzed in a spreadsheet. Figure 10 shows four columns. The first column [A] is the number of bookmarks of [eufeds.eu](http://eufeds.eu) newspaper websites opened per county. The domains are called nodes. Column [B] is the number of nodes from which third-party content is loaded when visiting the bookmarked home-pages of the newspaper websites without the Web Tracking



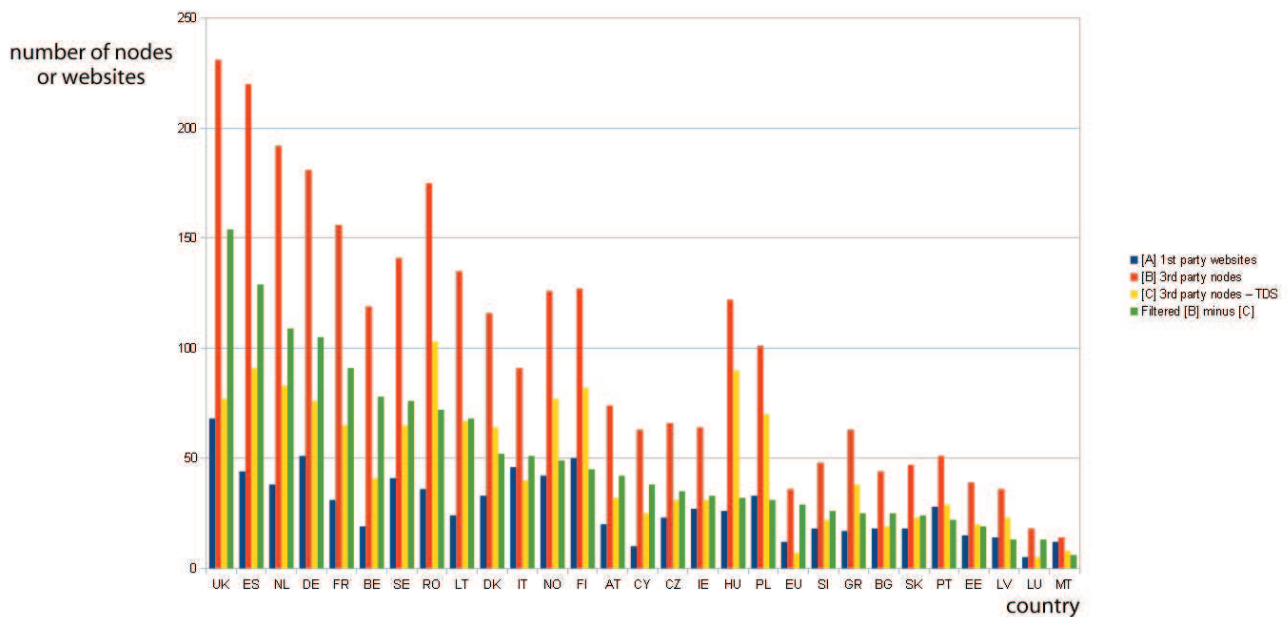


Figure 10: Traffic per country in absolute numbers.

Detection System (TDS). This can be anything from an advertisement to a Java script or a web bug (1x1 pixel). The third column shows the number of nodes from which third-party content is loaded but now the websites have been filtered with the Web Tracking Detection System (TDS). The last column [D] shows the amount of third-party nodes that have been blocked by the Web Tracking Detection System (TDS). This number is calculated by subtracting column [C] from [B]. This number is an indication for the amount of tracking that is taking place because the

The graph is sorted on column [C]. The graph shows the countries in the order of countries where the Web Tracking Detection System (TDS) is most effective. It also shows which countries have a lot of content being served to a user originating from third-party nodes. We can see that the UK, Spain, The Netherlands, Romania, Germany and France have relatively high totals of number of third-party nodes in comparison to the other countries.

In order to compare the effectiveness of the filter, the percentage of filtered nodes has been calculated. This is done by dividing the total number of nodes (column [A] + [B]) and the total number of filtered nodes (column [D]). The percentage filtered nodes compared to the total number of nodes is shown in figure 11. For example on the bookmarks of newspaper sites in Bulgaria, the Web Tracking Detection System (TDS) filters 40% of the traffic. The total range of effectiveness of the Web Tracking Detection System (TDS) is between 22% and 60% as a percentage of the number of filtered nodes by total number of nodes. From this we conclude that the Web Tracking Detection System (TDS) is working and efficient.

The next step is to look at the connectivity of the nodes. First have a look at the network diagrams in

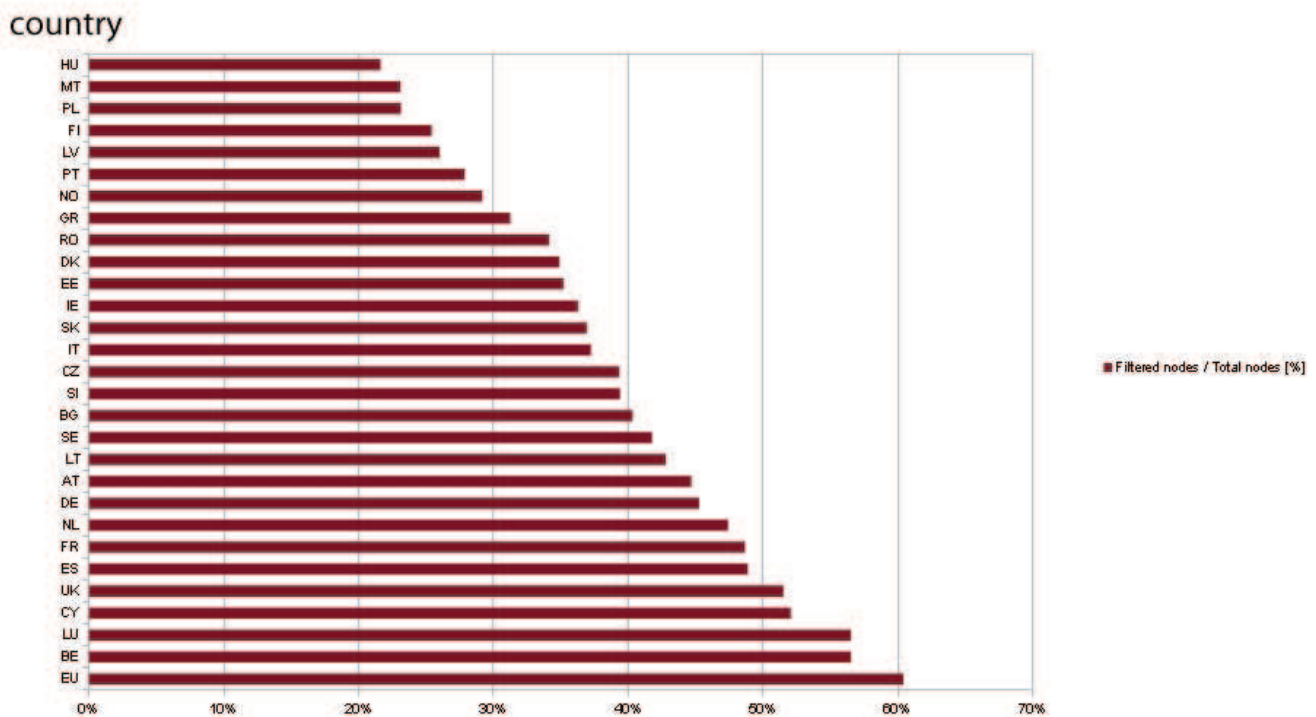


Figure 11: Effectiveness of the Web Tracking Detection System (TDS) as a percentage of the number of filtered nodes by total number of nodes.

appendix B that visualize the results of the processing for the experiments. Luxembourg is a country with a few newspapers whereas the UK is a country with many. The network maps in the appendix visualize the effect of the Web Tracking Detection System (TDS), which very much resembles the image of “cleaning the floor by sweeping with a broom”.

We will now try to quantify this effect. Figure 12 shows the number of links per node with and without the Web Tracking Detection System (TDS). The numbers have been calculated by dividing the total number of nodes and the total number of links connecting the nodes.

Figures 10 and 11 only give us information on the number of nodes. Figure 12 gives us information on the interconnectedness of the nodes. We will call the interconnectedness links. The higher the number of links per node, the more interconnected the nodes are.

Countries with high interconnectedness (many links) between nodes are Finland, Sweden, UK, Spain and Germany. Without the Web Tracking Detection System (TDS) the range of links per node varies between 1.3 and 3.8. Applying the Web Tracking Detection System (TDS) the range of links per node varies between 1.2 and 2.7. This is a significant reduction of interconnectedness.

Figure 12 can be linked back to figure 11 because the Web Tracking Detection System (TDS) only filters the nodes from the base reference set of confirmed tracking domains. The base dataset is a reference for confirmed trackers. Countries with a high number of links per node have also a high percentage of effectiveness for the application of the Web Tracking Detection System (TDS).



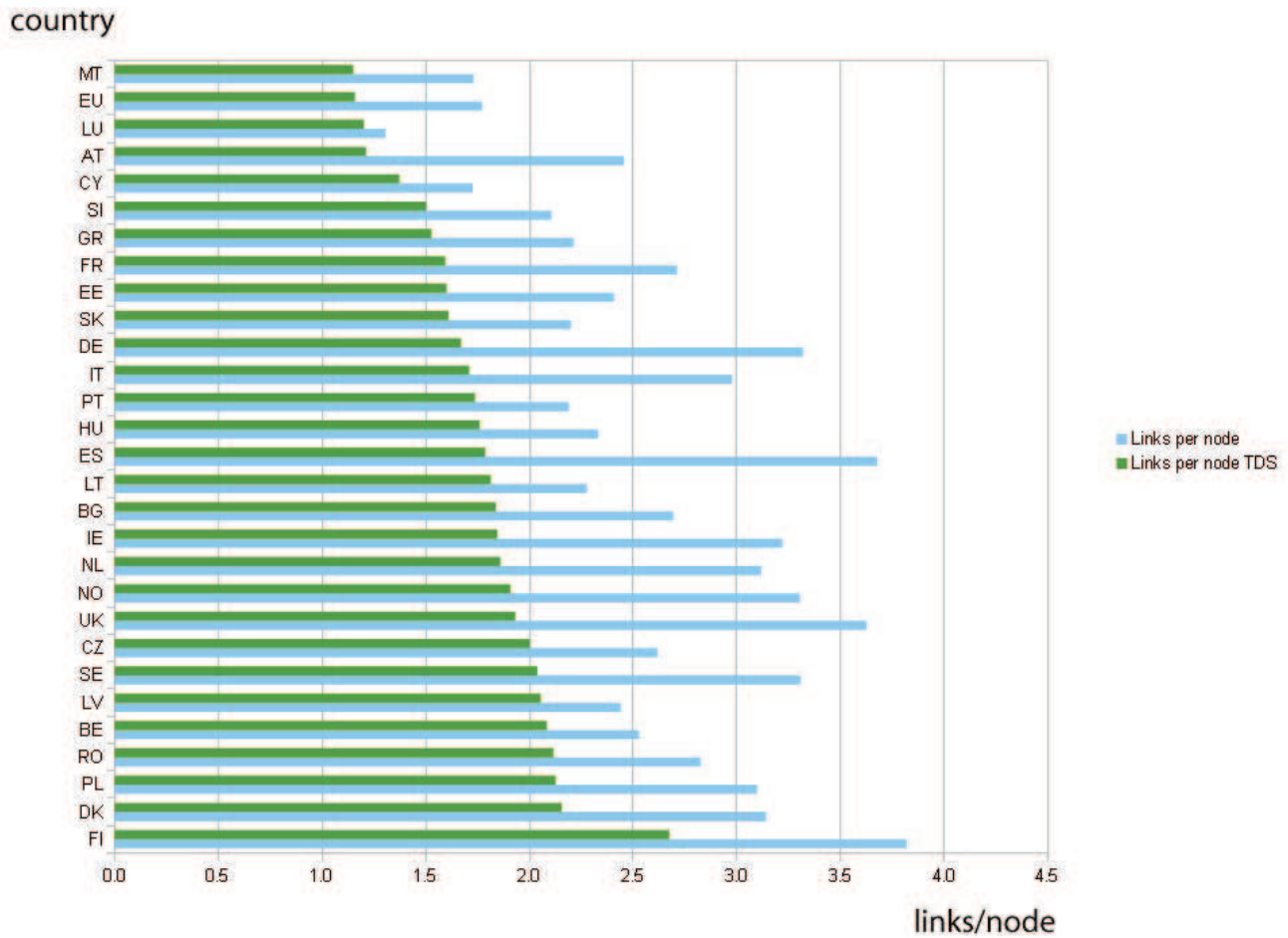


Figure 12: Links per node per country.

Sweden has 3.3 links per node without TDS, 2.0 links per node with TDS and a 42% of effectiveness on blocking third party tracking. UK has 3.6 links per node without TDS, 1.9 links per node with TDS and a 52% of effectiveness on blocking third party tracking. Spain has 3.7 links per node without TDS, 1.8 links per node with TDS and a 49% of effectiveness on blocking third party tracking. And Germany has 3.3 links per node without TDS, 1.7 links per node with TDS and a 45% of effectiveness on blocking third party tracking.

## 5.4 Conclusions

We have explored the leaking of behavioral data while surfing across rich context websites by using the Web Tracking Detection System (TDS). We have surveyed 819 EU newspaper sites. Furthermore, we have used a filter based on regular expressions and persistent opt-out cookies. The experiment provides data about the use of tracking via third-party domains.

We have looked for an indication of web tracking activity. We can conclude that web tracking activity

can be expressed by two indicators:

1. the interconnectedness between nodes expressed in the number of links per node;
2. the number of filtered nodes with TDS as a percentage of the total number of nodes.

The findings of our experiment can be summarized with the following itemized list:

- We have seen that the total range of effectiveness of the Web Tracking Detection System (TDS) is between 22% and 60% as a percentage of the number of filtered nodes by total number of nodes (figure 11).
- We have also seen that without the Web Tracking Detection System (TDS) the range of links per node varies between 1.3 and 3.8 (figure 12).
- Applying the Web Tracking Detection System (TDS) the range of links per node varies between 1.2 and 2.7 (figure 12).

From this we can conclude with deductive reasoning that the Web Tracking Detection System (TDS) is efficient.

The only out layer is Finland. Finland has 3.8 links per node without TDS, 2.7 links per node with TDS but only a 25% of effectiveness on blocking third party tracking.

The hypothesis is that in Finland tracking is being deployed on a large scale which is not detected by the rule set of regular expressions and the set of opt-out cookies used in the Web Tracking Detection System (TDS) in the experiment. The assumption is that third-parties involved in tracking are not within the base set of confirmed tracking we have used for our experiment as a referential dataset. Figure 16 (p. 42) shows a circle of green dots, which represent the bookmarked newspaper websites. If the hypothesis stands, then the blue dots within the green circle are active in web tracking activities. In chapter 6 we will look for an explanation for this phenomenon.

## 6 Patterns in connectivity maps

Having processed the data of the 819 newspapers, we will revisit analyzing the data by visualizing the nodes and links that connect the nodes. We will look at different visualization software. The methodology used in this chapter is constrained based pattern matching. We will look for patterns that repeat itself in the connectivity maps used to visualize the data. Having a set of rich data we will see that colors add an extra dimension to graphs. This leads to five patterns of interconnected nodes that indicate web tracking by third-party domains. Verification of the hypothesis that in Finland tracking is being deployed on a large scale concludes this chapter.

### 6.1 Introduction

We will start by looking at different software that could help us visualizing the data we have collected and processed in the previous chapter. The following software solutions have been reviewed:

- D3<sup>7</sup>, a Java Script library that allows for efficient manipulation of documents based on data.
- VOSviewer<sup>8</sup>, created by the Center for Science and Technology Studies at Leiden University. The software can be used to construct, view and explore maps based on network data.
- Gephi<sup>9</sup> which is open source graph visualization and manipulation software. Gephi is an interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs.
- HistCite<sup>10</sup> is a flexible software solution to aid researchers in visualizing the results of literature searches in the Web of Science.
- Orange<sup>11</sup>, open source data visualization, analysis and data mining through visual programming or Python scripting.
- SCI<sup>2</sup> Tool<sup>12</sup>, which supports network analysis.

Of all the software that has been reviewed, the D3 Java Script library has proven to get the job done. The choice for D3 is arbitrary. The main reasons for preferring this tool is the well documented Application Programming Interface (API) and the rich collection of examples.

---

<sup>7</sup><http://mbostock.github.com/d3/>

<sup>8</sup><http://www.vosviewer.com/>

<sup>9</sup><http://gephi.org/>

<sup>10</sup>[http://thomsonreuters.com/products\\_services/science/science\\_products/a-z/histcite/](http://thomsonreuters.com/products_services/science/science_products/a-z/histcite/)

<sup>11</sup><http://orange.biolab.si/>

<sup>12</sup><https://sci2.cns.iu.edu/user/index.php>

With D3, data can be displayed in a force-diagram. A force-diagram is a physical simulation of charged particles and springs. A force-diagram places nodes that are related in closer proximity, while unrelated nodes are farther apart.

By preparing the data into three groups and adding a force-value, the force-diagram will spread out into a pattern because of the grouping of related and unrelated nodes. In the experiments in this chapter, the following grouping has been applied to the data:

- Group zero: visited web pages
- Group one: base reference data set of confirmed trackers (PrivacyChoice dataset)
- Group two: other nodes

These groups have been assigned colors to make the nodes stand out in a connectivity map. The colors in the figures in this chapter map to these groups: Group 0 is displayed in green, group 1 in purple and group two in blue.

## 6.2 Data acquisition and preparation

We reuse the dataset that has been acquired in the previous chapter. The bookmarks per country on the website [www.eufedds.eu](http://www.eufedds.eu) are used as the data set. Opening all bookmarks at the same time will start the data capturing. At the moment that all web pages in the tabs are loaded, all tabs are reloaded through a right mouse click and selecting the menu item “reload all tabs”.

We reuse the data that has been processed in the previous chapter. With a PERL script (source code: appendix F we have converted the captured HTTP header data into a JSON. A JSON (JavaScript Object Notation) is a lightweight data-interchange format. Appendix F (p. 95) illustrates an example of the JSON format used to feed the D3 library. The JSON file is one of the output files of the PERL script.

The JSON is then loaded with a web page containing HTML code and Java Scripting. The HTML code uses the D3 library. Appendix G (105) contains the source code that has been written to visualize the JSON format within a web browser. This way the hidden relationship between HTTP headers can be visualized.

## 6.3 Findings

By applying color to connectivity maps, a new dimension is added. This enables constraint-based graph pattern mining. Borgwardt & Yan (2009, p. 40) has performed research on constraint-based graph pattern mining. They postulates that:

- Highly connected subgraphs in a large graph usually are not artifacts (group, functionality);

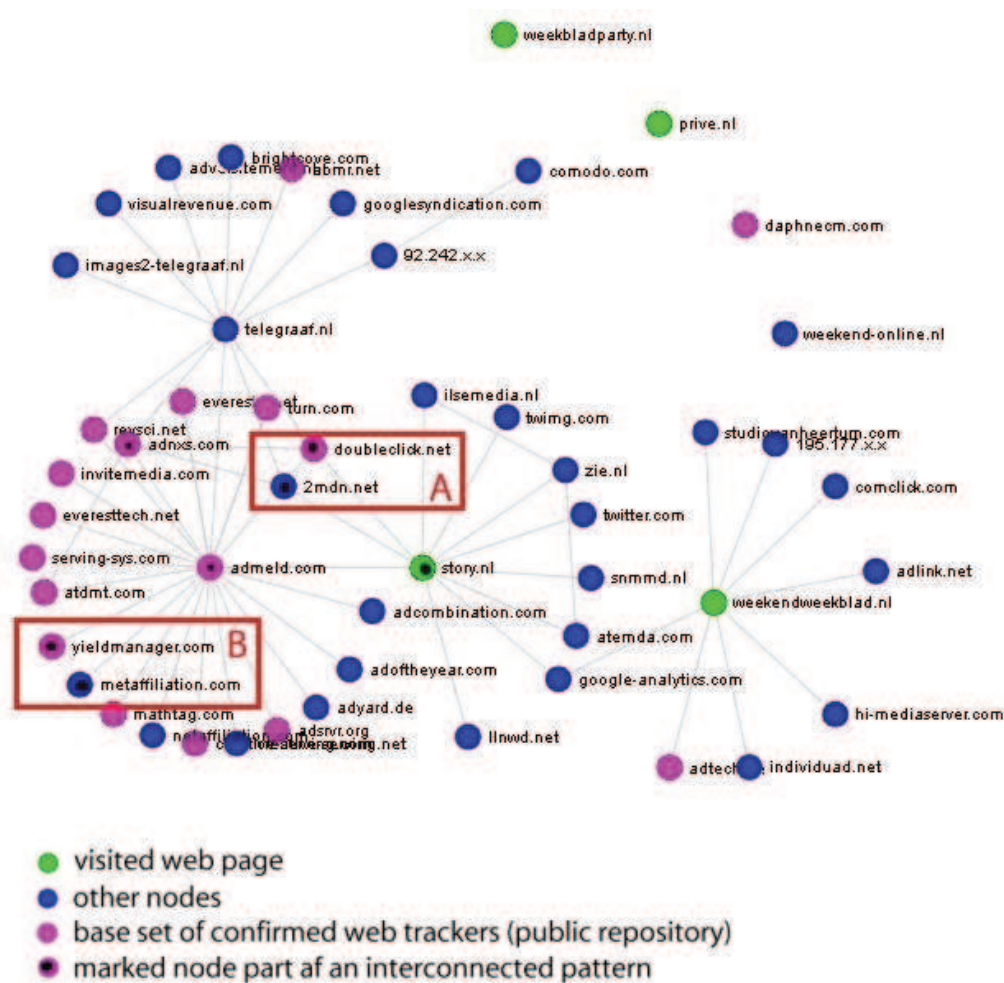


Figure 13: Visualized HTTP traffic for four websites (story.nl, weekbladparty.nl, privé.nl and weekendblad.nl) loaded into four tabs at the same time in the web browser without TDS filtering. The patterns A and B are identified for constraint-based graph pattern mining.

- Recurrent patterns discovered in multiple graphs are more robust than the pattern mined from a single graph.

The results of our experiment are best explained by looking at the graphics. Figure 13 shows two interesting patterns. Pattern A connects the green node story.nl with the purple node adnxs.com via either the purple node doubleclick.net or the blue node 2mdn.net. This pattern has been seen in other connectivity maps in our dataset. It is a common pattern in the EU newspaper dataset.

Pattern B shows an ad network, in this case admeld.com. Different third-party nodes communicate with admeld.com. The pattern shows a green node story.nl that is connected to a purple node yieldmanager.com and a blue node metaaffiliation.com via the purple node admeld.com. This is also a common pattern seen in the EU newspaper dataset.

Figure 13 shows three more patterns of interest. The first is pattern C which cross-links the green nodes lameuse.be and lessor.be with the purple node adtech.de and the blue node groupolitan.be.

This is also a common pattern in our dataset.

Pattern D is less commonly observed. It links a blue node to two purple nodes. this blue node is not connected to any other nodes. The assumption is that if both purple nodes are confirmed trackers, AND the blue node which is connected to both trackers must be a tracker too.

Pattern E links a green node nyan.ax with a purple node addthis.com in line with a blue node. The principle of inheritance of the attribute of the purple node applies in this pattern.

Figure 15 shows five new rules that can be derived from the patterns A, B from figure 13 and the patterns C, D and E from figure 14. These simple rules are an important contribution of this thesis.

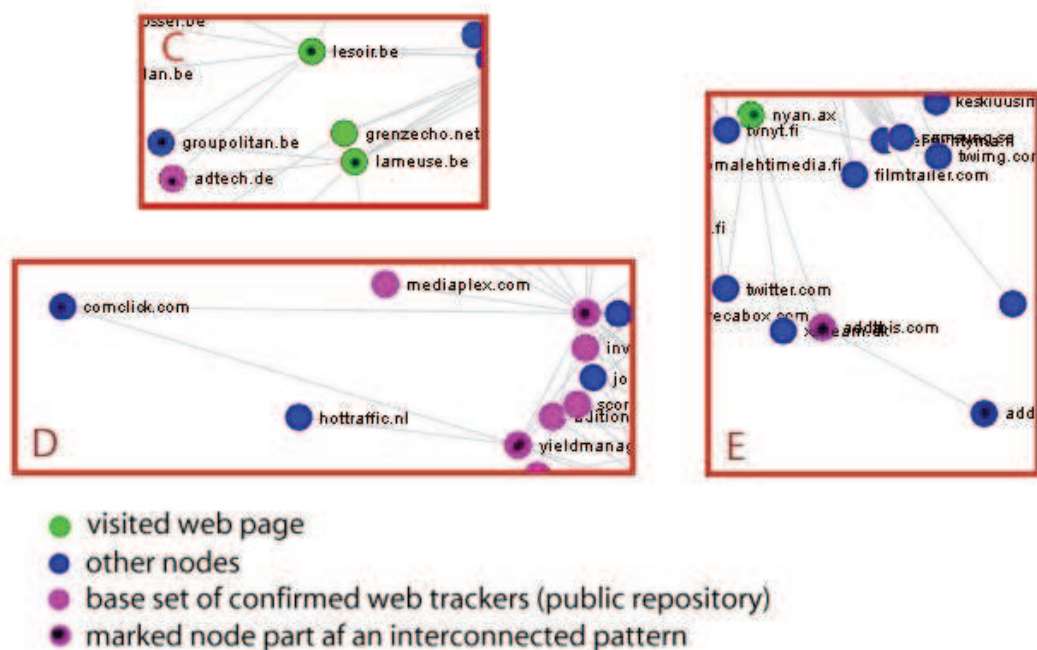


Figure 14: Constraint-based graph pattern mining: patterns C, D and E.

Let us try to explain the patterns in plain English. This effort is not to be considered to be a boolean description nor a mathematical exercise of logic, but rather an informal way to explain the meaning of the patterns in a generic way.

- Rule A means “when a green node is connected to a purple node by two parallel links, one via a purple node AND one via a blue node; then the blue node can be considered to be a purple one”.
- Rule B means “when a green node is connected to a purple node AND a blue node, via a purple node; then the blue node can be considered to be a purple one”.
- Rule C means “when two green nodes are connected to the same purple node AND connected to the same blue node; then the blue node can be considered to be a purple one”.

- Rule D means “when two purple nodes are connected to just one blue node; then the blue node can be considered to be a purple one”.
- Rule E means “when a green node is connected to a blue node via a purple node; then the blue node can be considered to be a purple one”.

With “then the blue node can be considered to be a purple one” we mean that these nodes can track a user’s web behavior across web sites. If you wish, this knew knowledge could be used as a staring point for a feedback loop. The philosophy of a feedback loop is common in measurement and control systems theory. The blue node can be added to the repository of confirmed tracker domains.

A feedback expressed in regular expressions and/or an opt-out cookie requires some more work. For turning a blue note into an effective regular expression that can be used in Web Tracking Detection System (TDS), the details of the HTTP headers related to the blue node need to be investigated. For turning a blue note in to a persistent opt-out cookie the web page with the privacy statement of the blue domain will have to be visited.

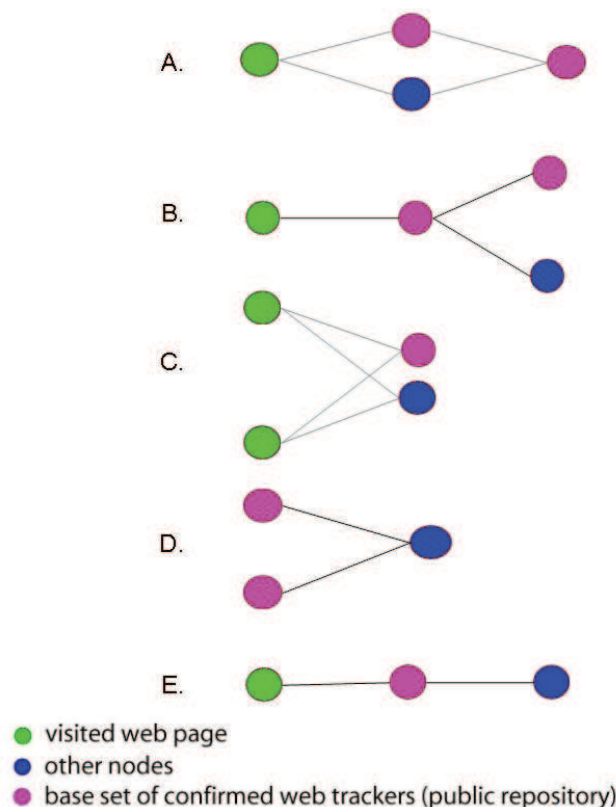


Figure 15: Constraint-based graph pattern mining: five new rules for the blue colored nodes. In each of the cases the blue node can be added to the public repository of confirmed web tracking domains. These rules allow for further investigation of the captured HTTP traffic of the blue colored node, in order to create a granular regular expression.



## 6.4 Conclusions

With this new knowledge, an explanation is found for Finland being the only statistical out layer in figure 12 (p. 35). Finland has 3.8 links per node without TDS, 2.7 links per node with TDS but only a 25% of effectiveness on blocking third party tracking.

The hypothesis is that in Finland tracking is being deployed on a large scale. However the third-parties involved in tracking are not within the base set of confirmed tracking we have used for our experiment. Figure 16 (p. 42) shows a circle of green dots<sup>13</sup>, which represent the bookmarked newspaper websites. If the hypothesis stands, then the blue nodes within the green circle should be involved in web tracking activities.

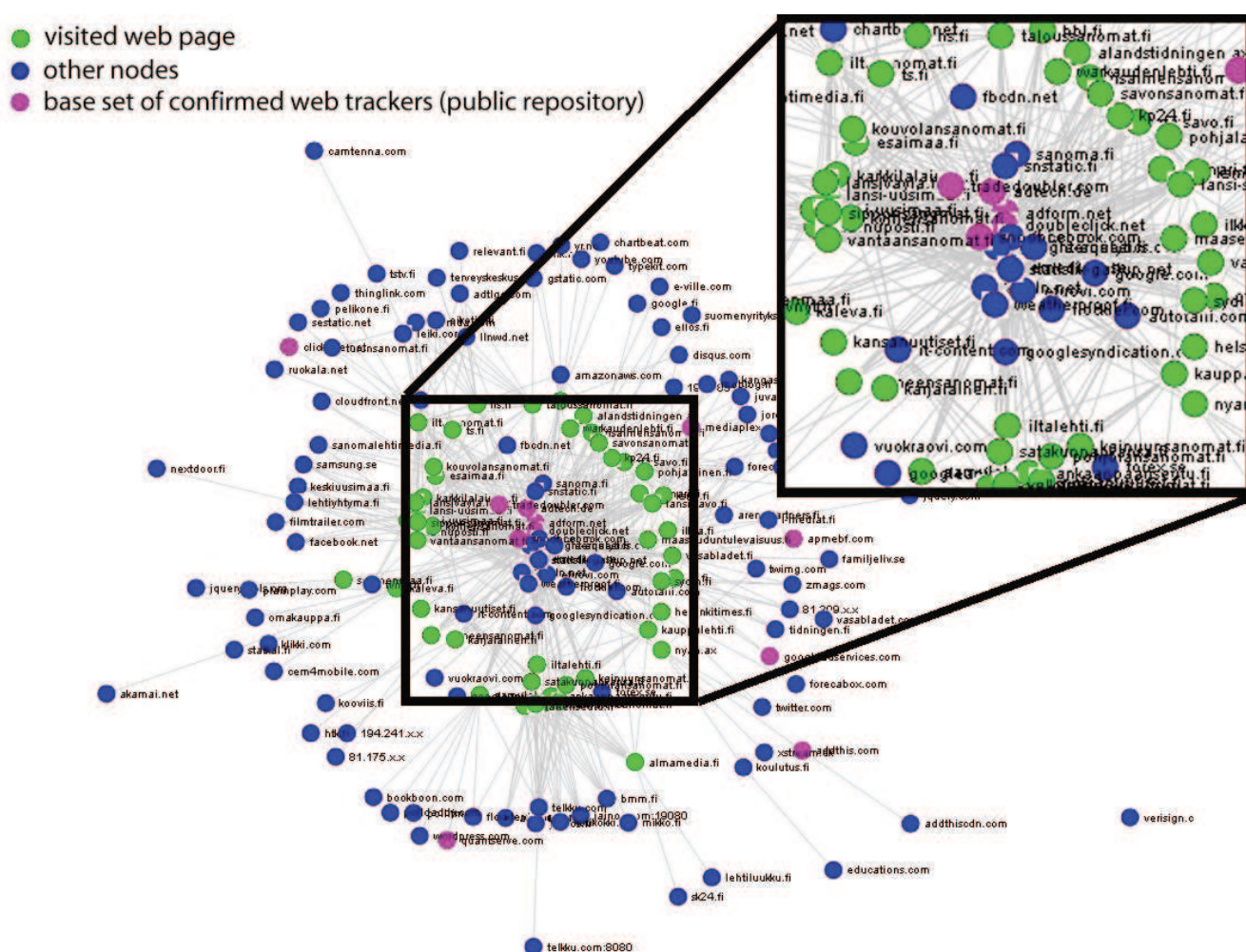


Figure 16: connectivity color map for Finland (FI) (not filtered by TDS). The blue nodes in the green circle (nicknamed 'The Eye') are of type 'C' which means that these nodes can track a user's web behavior across web sites.

<sup>13</sup>The force graph can be viewed interactively on <http://www.blaeu.com/uploads/tracking/eu.finland-colored.html> by dragging a node with a mouseclick.

Inspection of the nodes learns that all blue nodes are of type C. When we apply the rules that we have derived in figure 15 we can conclude that the blue dots within the green circle in figure 16 are to be considered to be involved in web tracking activities.

This also sets open a door to future research. Constraint-based graph pattern mining enables the applications of existing mathematical graph theory to HTTP traffic in the context of web tracking.

## 7 Conclusion and future work

This thesis is about tracking online user behavior. We have explored the criteria for a working definition which is not limited to online behavioral advertising. The thesis attempts to provide insight into the different perspectives that stakeholders have about the importance of data of user's surfing behavior. The implicit research question is therefor: What is a good working definition of tracking? What we can notice is that stakeholders have implicit assumptions about data like "data is free to flow because click-stream data isn't personal data". We have seen that criteria can be grouped together even more than shown in figure 7: user centric stakeholders and supplier centric stakeholders.

We have seen that that the real problem of tracking lies in three aspects:

- At the moment there is no widely accepted definition of tracking. Supplier centric stakeholders proclaim a narrow scope definition whereas user centric stakeholders are clear about a wide scope definition.
- Users are often not aware that the traces they leave behind online and offline are valuable data and the ways these online and offline data can be used is unknown to them.
- The traces users leave behind contain enough information to draw a meaningful picture about a user's behavior. Therefor data protection laws and principles apply.

The working definition we have come to is: "The non-consensual collection or processing or storing of data for the purpose of systematic monitoring or profiling of user's habits across websites".

The collection or processing or storing of data required by law, necessary for the purpose of information security (availability, integrity, confidentiality) or fraud prevention is outside of the scope of the working definition. This working definition isn't killing the OBA business. It also leaves open whether a solution for the Do Not Track debate is reached through self regulatory agreements or with help of additional laws.

This thesis also makes a contribution to the ongoing debate, by showing that tracking can be detected and suppressed effectively in real time with a filter. The effectiveness of the filter is measured with an experiment using content from 819 newspaper websites in the European Union.

The experiments show that there are methods to counter profiling. The hypotheses is that if a user is able to block web content just before execution in the browser, tracking will be reduced. The purpose of the Web Tracking Detection System is to give user control over blocking or allowing tracking. This opt-out mechanism based on regular expressions and opt-out cookies will be submitted to the IETF (Cooper & Tschofenig, 2011).

The main contributions of this thesis are:

- A working definition for web tracking.
- An opt-out mechanism for web tracking based on regular expressions repository in combination with persistent opt-out cookies.
- We show that the current do-not-track-me register, an opt-out mechanism for online behavioral advertising (OBA) based on opt-out cookies, of IAB/EASA falls short on it's promise. The opt-out cookie expiration dates are inconsistent, often less then five years. This is in contrast with the NAI opt-out tool, where all opt-out cookies have a minimum expiration data of five years. Added to that, opt-out behavior is collected by third parties.
- We show that the interconnectedness between nodes expressed in the number of links per node is an indicator (clustering coefficient) for web tracking.
- We show that the number of filtered nodes with TDS as a percentage of the total number of nodes is an indicator for web tracking.
- Five rules for constraint based graph mining of HTTP header information show that with use of a confirmed web tracking repository, it is possible to identify new third party tracking activity with connectivity color maps and pattern matching.

Transparency and reproducibility are important factors in a scientific approach. Most of the research data of the W3C workshop in Princeton are available on the companion website of the workshop (URL [www.w3.org/2011/track-privacy/](http://www.w3.org/2011/track-privacy/)). Position papers, agenda and accepted presentations can be downloaded. A report has been released based on the minutes and IRC log from April 28th and April 29th. The companion website of the Online Tracking Protection & Browsers workshop is less informative. However Roessler (2011) prepared a good overview that can serve as minutes.

For the experiments in chapter 5 and 6 the tools for data acquisition have been discussed. The tools are either in the open source domain, or in the appendix of this thesis. Appendix F contains the source code for the processing of the HTTPheader data. The logs of the processing that lead to conclusions can be verified in appendix B.

Usability is another important factor. Looking back at the six months of research it has become clear to me that having a broadly shared working definition is an essential step in the process of contributing to the solving the complex social problem of web tracking. Having said that, we have seen that a working definition shouldn't be too broad or too narrow.

The significant results for the two indicators and the five rules for constraint based graph mining have been tested multiple times during the month of August 2011, on three different locations with different hardware. The reliability of the experiments in this thesis is to be tested in future experiments. The results look future proof. Up to now the results have proven to be significant and repeatable.

Future work is suggested to rerun the experiment described in chapter 5. Instead of using both the regular expression and the persistent opt-out cookies, it is suggested to explore the effect on the number of links per node by just using the opt-out cookie as the filter. The experiment leads to quantitative data of the effect of opt-out cookies on web tracking across countries. This is relevant in the current debate on the do-not-track register that has been proposed by the advertising industry.

Future work is suggested to explore the possibilities of graph-mining. Kreyszig (1993, pp. 1133 - 1175) shows flow augmenting paths and other algorithms that need to be explored on usability for graph mining the colored connectivity data. The software to process the HTTP headers (Appendix F) already stores the information in which data flows by keeping track of source and destination of a HTTP header.

Future work is needed to create an algorithm to automatically discover new tracking domains. With the help of such an algorithms the quality of a Web Tracking Detection System (TDS) can be improved. Future work is needed to find out whether prevention (100% blocking) of Web Tracking is possible.

The reason for the Donald Duck offer is because of visiting different categories of interest based web content in the week before conducting the newspaper experiment. So far the following categories have been explored: sports, magazines for children, motorcycle & car, women's magazines, men's magazines, wellness, TV guides, girl's magazines, gossip, glossies, food & drink, magazines concerning raising children, living, animals, computer & Internet & audio & video, art & culture and opinion magazine. The results of these data have been analyzed casually and look promising for discovery of more constrained based patterns. Future work is needed to look into these and other data sets.

## Acknowledgments

I thank Joost Kok and Michel Chaudron, for inspiring me to write one extra technical chapter. The colored network maps were a great source of renewed enthusiasm. I thank Jules Polonesky, director of The Future of Privacy Forum (FPF) for your permission to use your diagrams (figure 3 and 4), Hannes Tschofenig for giving me the opportunity to present at the W3C workshop, Roger Coopmans for allowing me to find a balance between work and study and Adriaan in 't Groen for your hospitality to let me write in a monastic setting at Campus The Hague. Finally I thank you, Delia for your unconditional support throughout this project.

## 8 References

- Arthur, W. B. (2009). *The nature of technology: what it is and how it evolves*. Simon and Schuster.
- ARTICLE 29 Data Protection Working Party (2003). Opinion 3/2003 (WP77) on the European code of conduct of FEDMA for the use of personal data in direct marketing. URL [http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2003/wp77\\_en.pdf](http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2003/wp77_en.pdf).
- ARTICLE 29 Data Protection Working Party (2010a). Opinion 2/2010 (WP171) on online behavioural advertising. URL [http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2010/wp171\\_en.pdf](http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2010/wp171_en.pdf).
- ARTICLE 29 Data Protection Working Party (2010b). Opinion 4/2010 (WP174) on the European code of conduct of FEDMA for the use of personal data in direct marketing. URL [http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2010/wp174\\_en.pdf](http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2010/wp174_en.pdf).
- ARTICLE 29 Data Protection Working Party (2011). Opinion 15/2011 (WP187) on the definition of consent. URL [http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2011/wp187\\_en.pdf](http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2011/wp187_en.pdf).
- Barocas, S. & Nissenbaum, H. (2011). On Notice: The Trouble with Notice and Consent. URL [www.nyu.edu/projects/nissenbaum/papers/ED\\_SII\\_On\\_Notice.pdf](http://www.nyu.edu/projects/nissenbaum/papers/ED_SII_On_Notice.pdf).
- Barth, A. (2011). HTTP State Management Mechanism. URL [tools.ietf.org/id/draft-ietf-httpstate-cookie-23.txt](http://tools.ietf.org/id/draft-ietf-httpstate-cookie-23.txt). IETF draft (work in progress).
- Beyond Web Analytics! Episode 45 (2011). Beyond web analytics! episode 45: Real-time customer intelligence. URL [www.beyondwebanalytics.com/2011/05/07/episode-45/](http://www.beyondwebanalytics.com/2011/05/07/episode-45/).
- Borgwardt, K. & Yan, X. (2009). Biological network analysis: Graph mining. URL <http://agbs.kyb.tuebingen.mpg.de/wikis/bg/BNA-4.pdf>.
- Brock, J. (2011a). Mobile tracking privacy: Three thoughts | PrivacyChoice blog. URL <http://blog.privacychoice.org/2011/03/06/mobile-tracking-privacy-three-thoughts/>.
- Brock, J. (2011b). A working definition of Do not track | PrivacyChoice blog. URL <http://blog.privacychoice.org/2011/03/22/a-working-definition-of-do-not-track/>.
- Broder, A. (2000). Graph structure in the web. *Computer networks*, 33(1), 309.



- Center for Democracy and Technology (2011). What does 'Do not track' mean? a scoping proposal ver. 2.0 | center for democracy & technology. URL <http://www.cdt.org/DNT-2>.
- Cooper, A. & Tschofenig, H. (2011). Overview of Universal Opt-Out Mechanisms for Web Tracking. URL [tools.ietf.org/html/draft-cooper-web-tracking-opt-outs-00.txt](http://tools.ietf.org/html/draft-cooper-web-tracking-opt-outs-00.txt). IETF draft (work in progress).
- Cortesi, A. (2011). How UDIDs are used: a survey. URL [corte.si/posts/security/apple-udid-survey/index.html](http://corte.si/posts/security/apple-udid-survey/index.html).
- Eckersley, P. (2011). What does the "Track" in "Do not track" mean? | electronic frontier foundation. URL <http://www.eff.org/deeplinks/2011/02/what-does-track-do-not-track-mean>.
- ECP-ECN (2010). Onderzoek vijf privacy-intrusive toepassingen. URL [http://www.ecp.nl/sites/default/files/Rapport\\_enquete\\_toepassingen\\_ICT\\_en\\_privacy\\_-\\_januari\\_2010.pdf](http://www.ecp.nl/sites/default/files/Rapport_enquete_toepassingen_ICT_en_privacy_-_januari_2010.pdf).
- European Commission (2009). Directive 2009/136/EC of the European Parliament and of the Council of 25 November 2009 amending Directive 2002/22/EC on universal service and users rights relating to electronic communications networks and services, Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector and Regulation (EC) No 2006/2004 on cooperation between national authorities responsible for the enforcement of consumer protection laws. Official Journal L 337/11, 18/12/2009.
- Federal Trade Commission (2010). Protecting Consumer Privacy in an Era of Rapid Change; A Proposed Framework for Businesses and Policymakers. URL [www.ftc.gov/opa/2010/12/privacyreport.shtm](http://www.ftc.gov/opa/2010/12/privacyreport.shtm).
- Fielding, R., Gettys, J., Mogul, J., Frystyk, H., & Berners-Lee, T. (1997). Hypertext Transfer Protocol -- HTTP/1.1. RFC 2068, Request For Comments.
- Groote, G., Hugenholtz-Sasse, C., & Slikker, P. (2000). *Projecten leiden : methoden en technieken voor projectmatig werken*. Utrecht: Het Spectrum, 10th completely revised ed.
- Hildebrandt, M. & Gutwirth, S. (2008). *Profiling the European citizen*. [New York]: Springer.
- Hustinx, P. (2011). 11-07-07\_Speech\_Edinburgh\_EN.pdf. URL [http://www.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/EDPS/Publications/Speeches/2011/11-07-07\\_Speech\\_Edinburgh\\_EN.pdf](http://www.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/EDPS/Publications/Speeches/2011/11-07-07_Speech_Edinburgh_EN.pdf).
- Idate-TNO-IViR (2008). User-created-content: Supporting a participative information society. URL [http://ec.europa.eu/information\\_society/eeurope/i2010/docs/studies/ucc-final\\_report.pdf](http://ec.europa.eu/information_society/eeurope/i2010/docs/studies/ucc-final_report.pdf).



- Improve Digital (2011). Improve digital advertising market map europe 2010. URL <http://www.improvedigital.com/market-map-2010>.
- Koffijberg, J., Dekkers, S., Homburg, G., & van den Berg, B. (2009). *Niets te verbergen en toch bang*. The Hague: College bescherming persoonsgegevens (Dutch DPA). Research conducted by Regioplan Beleidsonderzoek.
- Kreyszig, E. (1993). *Advanced engineering mathematics*. New York [etc.]: Wiley, seventh ed. ed.
- Krishnamurthy, B. & Wills, C. (2009). Privacy diffusion on the web: a longitudinal perspective. In *Proceedings of the 18th international conference on World wide web*. WWW '09, 541550. ACM ID: 1526782.
- Kristol, D. & Montulli, L. (1997). HTTP State Management Mechanism. RFC 2109, Request For Comments.
- Kroes, N. (2011). EUROPA - press releases - neelie kroes Vice-President of the european commission responsible for the digital agenda online privacy reinforcing trust and confidence online tracking protection & browsers workshop brussels, 22 june 2011. URL <http://europa.eu/rapid/pressReleasesAction.do?reference=SPEECH/11/461>.
- LUMA Partnets LLC (2010). Luma display ad tech landscape for adexchanger. URL <http://www.adexchanger.com/wp-content/uploads/2010/09/LUMA-Display-Ad-Tech-Landscape-for-AdExchanger.jpg>.
- Mayer, J., Narayanan, A., & Stamm, S. (2011). Do not track: A universal third-party web tracking opt out. URL [www.ietf.org/id/draft-mayer-do-not-track-00.txt](http://www.ietf.org/id/draft-mayer-do-not-track-00.txt). IETF draft (work in progress).
- Morris, P. (1988). Managing interfaces. In *Project management handbook*. New York: Van Nostrand Reinhold.
- Network Security Podcast, Episode 241 (2011). Network security podcast, episode 241. URL [netsecpodcast.com/?p=772](http://netsecpodcast.com/?p=772).
- Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. In *UCLA Law Review*. Los Angeles, CA (USA), vol. 57, 1710. URL [ssrn.com/abstract=1450006](http://ssrn.com/abstract=1450006).
- Roessler, T. (2011). Do Not Track: The Regulators' Challenge. URL [www.w3.org/QA/2011/06/do\\_not\\_track\\_the\\_regulators\\_ch.html](http://www.w3.org/QA/2011/06/do_not_track_the_regulators_ch.html).
- Smith, E. (2010). iphone applications & privacy issues: An analysis of application transmission of iphone unique device identifiers (udids). URL [www.psk1.us/wp/wp-content/uploads/2010/09/iPhone-Applications-Privacy-Issues.pdf](http://www.psk1.us/wp/wp-content/uploads/2010/09/iPhone-Applications-Privacy-Issues.pdf).

- Tene, O. & Polonesky, J. (2011). To Track or Do Not Track: Advancing Transparency and Individual Control in Online Behavioral Advertising. Draft.
- van Aken, T. (2002). *De weg naar projectsucces : eerder via werkstijl dan via instrumenten*. Utrecht: De Tijdstroom, 3th revised ed.
- van Bommel & van Dam (2011). Wijziging van de telecommunicatiewet ter implementatie van de herziene telecommunicatierichtlijnen; amendement; gewijzigd amendement van bommel c.s. ter vervanging van nr. 14 over toestemming voor het plaatsen van cookies. URL <https://zoek.officielebekendmakingen.nl/dossier/32549/kst-32549-34?resultIndex=17&sorttype=1&sortorder=4>.
- Verschuren, P. & Doorewaard, H. (2000). *Het ontwerpen van een onderzoek (Designing a research project)*. Utrecht: LEMMA, 3th ed.
- Zeigler, A., Bateman, A., & Graff, E. (2011). Web Tracking Protection. Microsoft W3C Member Submission. URL [www.w3.org/Submission/web-tracking-protection/](http://www.w3.org/Submission/web-tracking-protection/).

## A Results of the analysis of criteria for a working definition

The results of the analysis of the debate on Do Not Track has been added to this appendix. These criteria lead to the working definition for web tracking. The working definition is:

“The non-consensual collection or processing or storing of data for the purpose of systematic monitoring or profiling of user’s habits across websites”

Criteria	Browser manufacturers	Researchers	Civil rights organizations	Standardization bodies	Implementers	Policy and privacy experts	Behavioral advertising industry
Not limited to just cookies (Kroes)				X		X	
Not limited to just the advertising sector (Kroes, Brock)			X			X	
Industry led self regulation solution (Madelin)						X	
Deadline self regulatory solution by june 2012 (Kroes)						X	
Deadline self regulatory solution by end of 2011 (Brill)						X	
Must be easy to use (Brill)						X	
Must be effective (Brill)						X	
Must be universal (Brill)						X	
Must be about the collection as well as use of information (Brill, Kroes)						X	
Must be persistent (Brill)						X	
Clear process towards self-regulation (Madelin)						X	
Self regulation needs to include responsibility and accountability of controllers (Madelin, Hustinx)						X	
A technology framework needs policies	X	X	X	X	X	X	X
Best Practice Recommendation and Framework on behavioural advertising (IAB EASA)					X	X	X
Transparency (Kroes, Hustinx, Madelin)						X	
User Control (Kroes, Hustinx)						X	
Fairness (Kroes, Hustinx)						X	
Article 5(3) of the revised e-Privacy directive implies consent of the user having been provided with clear and comprehensive information in accordance with Directive 95/46/EX, inter alia about the purposes of the processing. (Hustinx)						X	
Consent should be freely given, specific and an informed indication of his wishes (Hustinx, EFF clear and non-confusing opt-back-in)						X	

Criteria	Browser manufacturers	Researchers	Civil rights organizations	Standardization bodies	Implementers	Policy and privacy experts	Behavioral advertising industry
Browser settings currently do not express user's consent but might do so in the (near) future (van Bommel)				X		X	
Privacy by default browser settings (Hustinx)	X			X		X	
Individual, computer or device (Brill, Brock)			X			X	
Collection and use of data through unique identifiers (Soltani, Brock)		X	X				
Include third-party online behavioral advertising (CDT)			X				
Include thirdparty behavioral data collection for first party uses (CDT)			X				
Include third-party behavioral data collection for other uses (CDT)			X				
Include behavioral data collected by first parties and transferred to third parties in identifiable form (CDT)			X				
Include demographic information appended to the user's device (CDT)			X				
Exclude third-party ad and content delivery(CDT)			X				
Exclude third-party analytics (CDT)			X				
Exclude third-party contextual advertising (CDT)			X				
Exclude first-party data collection and first-party use (CDT,EFF)			X				
Exclude federated identity transaction data (CDT)			X				
Exclude specially excepted third-party ad reporting (CDT)			X				
Exclude data collection required by law and for legitimate fraud prevention purposes (CDT,EFF)			X				
The DNT header is not necessarily the best mechanism (EFF)	X	X	X	X	X	X	X
Technical solutions should not break existing functionality	X			X	X		X

Criteria	Browser manufacturers	Researchers	Civil rights organizations	Standardization bodies	Implementers	Policy and privacy experts	Behavioral advertising industry
User expectation research is important to take into account (McDonald, Nissenbaum)		X					
Include economical and innovation aspects to collect data (IAB)		X				X	X
Standardized technical building blocks should enable a simple solution	X			X	X		

**Tabel 2 Criteria for a working definition on tracking**

## B Log files for creating nodes and links for EU newspaper websites

The processing log files for creating nodes and links for the EU newspaper websites are added to the appendix because of transparency and reproducibility of the research. Besides the number of nodes and links the log files shows the name of the liveHTTPHeader file, the number of processed HTTP headers, the number of unique nodes, the item number of unique links connecting the nodes, the number of nodes found from the base set of confirmed trackers (PrivacyChoice) and the time it took to process the data.

The network diagrams on the following page visualize the results of the processing for the experiments in chapter 5. Luxembourg is a country with a few newspapers whereas the UK is a country with many. The network maps visualize the effect of the Web Tracking Detection System (TDS), which very much resembles sweeping the floor with a broom.

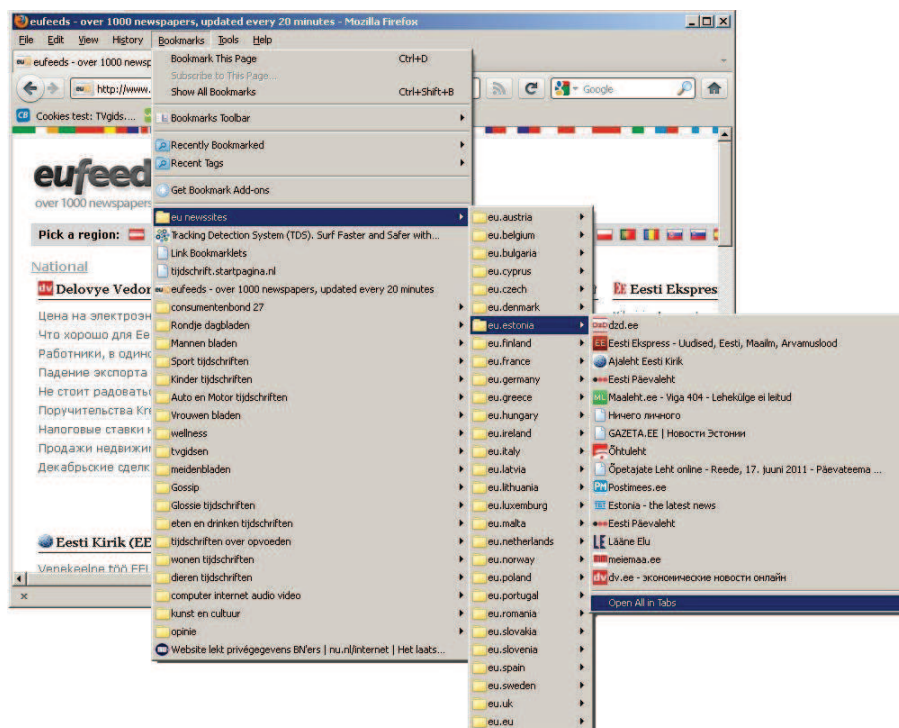
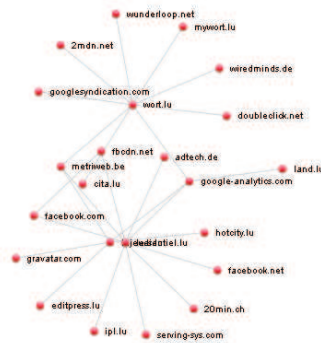


Figure 17: The hyper links per country on the website [www.eufeds.eu](http://www.eufeds.eu) are used as base data set. In order to be able to reproduce the results of the experiments, the URL's of the base data set are bookmarked within the web browser. Opening all bookmarks at the same time will start the data capturing. Data capturing is done with the browser extension liveHTTPheaders. At the moment that all web pages in the tabs are loaded, all tabs are reloaded through a right mouseclick and selecting the menu item "reload all tabs".



# Luxemburg



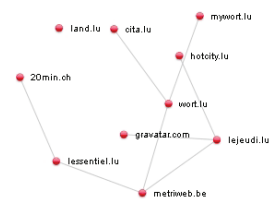
start processing ./data/eu.luxemburg

1181 headers  
23 nodes  
30 links  
4 confirmed trackers

job completed succesfully in 64 seconds

Aug 3, 2011

# Luxemburg (TDS)



start processing ./data/eu.luxemburg-TDS

1051 headers  
10 nodes  
12 links  
0 confirmed trackers

job completed succesfully in 57 seconds

Aug 4, 2011

# UK



start processing ./data/eu.uk

16993 headers

299 nodes

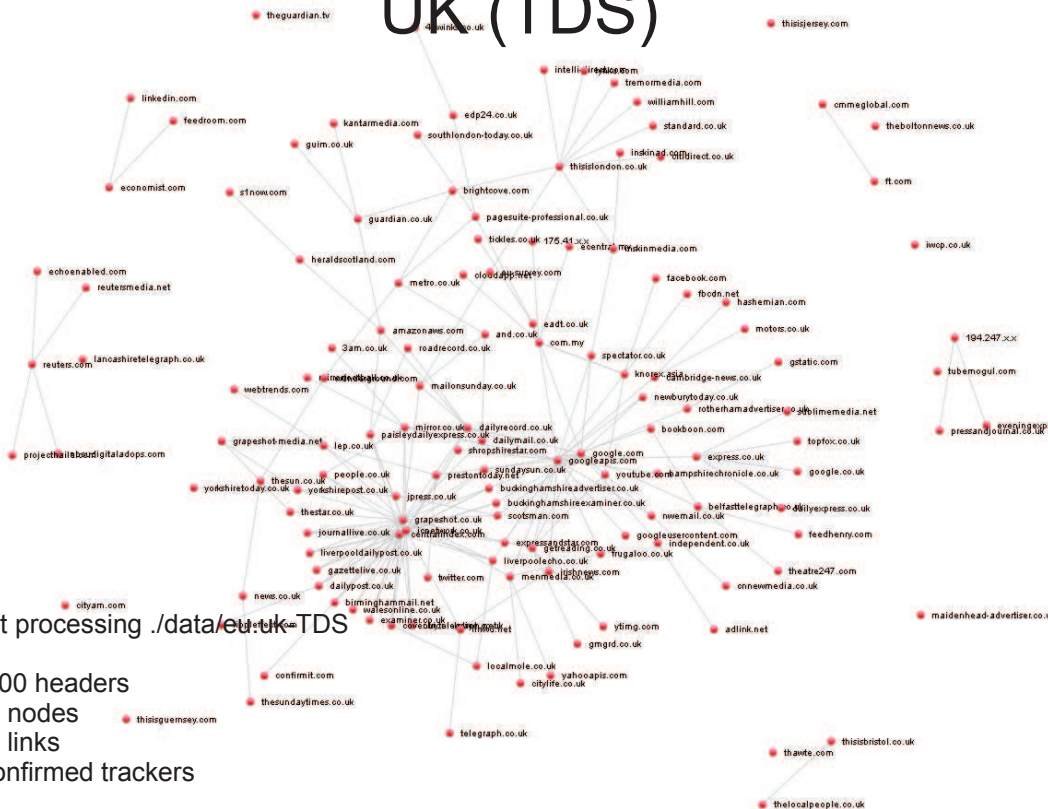
1084 links

90 confirmed trackers

job completed succesfully in 381 seconds

Aug 3, 2011

# UK (TDS)



start processing ./data/eu.uk TDS

11000 headers

145 nodes

280 links

3 confirmed trackers

job completed succesfully in 188 seconds

Aug 4, 2011

```
start processing ./data/eu.austria-TDS

5025 headers
52 nodes
83 links
1 confirmed trackers

job completed succesfully in 107 seconds

start processing ./data/eu.austria

6614 headers
94 nodes
231 links
21 confirmed trackers

job completed succesfully in 163 seconds

start processing ./data/eu.belgium-TDS

6116 headers
60 nodes
125 links
0 confirmed trackers

job completed succesfully in 122 seconds

start processing ./data/eu.belgium

7487 headers
138 nodes
349 links
50 confirmed trackers

job completed succesfully in 176 seconds

start processing ./data/eu.bulgaria-TDS

2683 headers
37 nodes
68 links
1 confirmed trackers

job completed succesfully in 76 seconds

start processing ./data/eu.bulgaria

3724 headers
62 nodes
167 links
11 confirmed trackers

job completed succesfully in 103 seconds

start processing ./data/eu.cyprus-TDS

1805 headers
```

```
35 nodes
48 links
0 confirmed trackers

job completed succesfully in 63 seconds

start processing ./data/eu.cyprus

2368 headers
73 nodes
126 links
24 confirmed trackers

job completed succesfully in 83 seconds

start processing ./data/eu.czech-TDS

3817 headers
54 nodes
108 links
1 confirmed trackers

job completed succesfully in 99 seconds

start processing ./data/eu.czech

4910 headers
89 nodes
233 links
17 confirmed trackers

job completed succesfully in 167 seconds

start processing ./data/eu.denmark-TDS

7810 headers
97 nodes
209 links
2 confirmed trackers

job completed succesfully in 148 seconds

start processing ./data/eu.denmark

15828 headers
149 nodes
468 links
16 confirmed trackers

job completed succesfully in 319 seconds

start processing ./data/eu.estonia-TDS

2973 headers
35 nodes
56 links
0 confirmed trackers
```

job completed succesfully in 64 seconds

start processing ./data/eu.estonia

3945 headers

54 nodes

130 links

4 confirmed trackers

job completed succesfully in 96 seconds

start processing ./data/eu.eu-TDS

1187 headers

19 nodes

22 links

0 confirmed trackers

job completed succesfully in 57 seconds

start processing ./data/eu.eu

1569 headers

48 nodes

85 links

14 confirmed trackers

job completed succesfully in 69 seconds

start processing ./data/eu.finland-TDS

10170 headers

132 nodes

353 links

0 confirmed trackers

job completed succesfully in 245 seconds

start processing ./data/eu.finland

13484 headers

177 nodes

676 links

11 confirmed trackers

job completed succesfully in 307 seconds

start processing ./data/eu.france-TDS

9171 headers

96 nodes

153 links

2 confirmed trackers

job completed succesfully in 176 seconds

```
start processing ./data/eu.france

12078 headers
187 nodes
507 links
43 confirmed trackers

job completed succesfully in 276 seconds

start processing ./data/eu.germany-TDS

19297 headers
127 nodes
212 links
4 confirmed trackers

job completed succesfully in 294 seconds

start processing ./data/eu.germany

17558 headers
232 nodes
770 links
57 confirmed trackers

job completed succesfully in 340 seconds

start processing ./data/eu.greece-TDS

4853 headers
55 nodes
84 links
0 confirmed trackers

job completed succesfully in 90 seconds

start processing ./data/eu.greece

3830 headers
80 nodes
177 links
10 confirmed trackers

job completed succesfully in 110 seconds

start processing ./data/eu.hungary-TDS

6869 headers
115 nodes
204 links
1 confirmed trackers

job completed succesfully in 131 seconds

start processing ./data/eu.hungary

7731 headers
```

```
148 nodes
345 links
11 confirmed trackers

job completed succesfully in 164 seconds

start processing ./data/eu.ireland-TDS

2848 headers
58 nodes
107 links
0 confirmed trackers

job completed succesfully in 83 seconds

start processing ./data/eu.ireland

3969 headers
91 nodes
293 links
18 confirmed trackers

job completed succesfully in 132 seconds

start processing ./data/eu.italy-TDS

10813 headers
86 nodes
147 links
1 confirmed trackers

job completed succesfully in 185 seconds

start processing ./data/eu.italy

13254 headers
137 nodes
408 links
21 confirmed trackers

job completed succesfully in 300 seconds

start processing ./data/eu.latvia-TDS

3158 headers
37 nodes
76 links
0 confirmed trackers

job completed succesfully in 87 seconds

start processing ./data/eu.latvia

3576 headers
50 nodes
122 links
2 confirmed trackers
```



job completed succesfully in 101 seconds

start processing ./data/eu.lithuania-TDS

7362 headers

91 nodes

165 links

1 confirmed trackers

job completed succesfully in 130 seconds

start processing ./data/eu.lithuania

9316 headers

159 nodes

362 links

17 confirmed trackers

job completed succesfully in 197 seconds

start processing ./data/eu.luxemburg-TDS

1051 headers

10 nodes

12 links

0 confirmed trackers

job completed succesfully in 57 seconds

start processing ./data/eu.luxemburg

1181 headers

23 nodes

30 links

4 confirmed trackers

job completed succesfully in 64 seconds

start processing ./data/eu.malta-TDS

1170 headers

20 nodes

23 links

0 confirmed trackers

job completed succesfully in 56 seconds

start processing ./data/eu.malta

1399 headers

26 nodes

45 links

1 confirmed trackers

job completed succesfully in 76 seconds

start processing ./data/eu.netherlands-TDS

6681 headers  
121 nodes  
225 links  
1 confirmed trackers

job completed succesfully in 140 seconds

start processing ./data/eu.netherlands

9792 headers  
230 nodes  
717 links  
57 confirmed trackers

job completed succesfully in 259 seconds

start processing ./data/eu.norway-TDS

11290 headers  
119 nodes  
227 links  
1 confirmed trackers

job completed succesfully in 208 seconds

start processing ./data/eu.norway

14939 headers  
168 nodes  
555 links  
20 confirmed trackers

job completed succesfully in 287 seconds

start processing ./data/eu.poland-TDS

7266 headers  
103 nodes  
219 links  
0 confirmed trackers

job completed succesfully in 149 seconds

start processing ./data/eu.poland

9122 headers  
134 nodes  
415 links  
13 confirmed trackers

job completed succesfully in 220 seconds

start processing ./data/eu.portugal-TDS

4572 headers

```
57 nodes
99 links
2 confirmed trackers

job completed succesfully in 119 seconds

start processing ./data/eu.portugal

2916 headers
79 nodes
173 links
15 confirmed trackers

job completed succesfully in 92 seconds

start processing ./data/eu.romania-TDS

9065 headers
139 nodes
294 links
1 confirmed trackers

job completed succesfully in 171 seconds

start processing ./data/eu.romania

12968 headers
211 nodes
596 links
43 confirmed trackers

job completed succesfully in 279 seconds

start processing ./data/eu.slovakia-TDS

2865 headers
41 nodes
66 links
0 confirmed trackers

job completed succesfully in 73 seconds

start processing ./data/eu.slovakia

3654 headers
65 nodes
143 links
7 confirmed trackers

job completed succesfully in 142 seconds

start processing ./data/eu.slovenia-TDS

3366 headers
40 nodes
60 links
1 confirmed trackers
```

```
job completed succesfully in 84 seconds

start processing ./data/eu.slovenia

4014 headers
66 nodes
139 links
9 confirmed trackers

job completed succesfully in 105 seconds

start processing ./data/eu.spain-TDS

21234 headers
135 nodes
241 links
2 confirmed trackers

job completed succesfully in 352 seconds

start processing ./data/eu.spain

29219 headers
264 nodes
971 links
67 confirmed trackers

job completed succesfully in 536 seconds

start processing ./data/eu.sweden-TDS

9845 headers
106 nodes
216 links
3 confirmed trackers

job completed succesfully in 187 seconds

start processing ./data/eu.sweden

13385 headers
182 nodes
602 links
43 confirmed trackers

job completed succesfully in 294 seconds

start processing ./data/eu.uk-TDS

11000 headers
145 nodes
280 links
3 confirmed trackers

job completed succesfully in 188 seconds
```

start processing ./data/eu.uk

16993 headers

299 nodes

1084 links

90 confirmed trackers

job completed succesfully in 381 seconds

## C Tracking Detection System (TDS)

If you have come here to start using the Web Tracking Detection System, this is the right place. In this appendix we will see how the rule set with protective regular expressions can be installed. We will also see how the protective opt-out cookies can be installed.

The TDS is based upon three risk controls:

- appear to be a first time visitor.
- limit the collection of data.
- limit the use of data.

Updating will be done automatically. The rule set with regular expressions has an update interval of five days. The opt-out cookies will auto update when a new release of the BeefTaco is available. Firefox takes care of this automatically when searching for updates for the installed add-ons Adblock Plus and BeefTaco.

So in fact, we could add a fourth risk control:

- set and forget.

## Tracking Detection System (TDS)

Risk management strategies to reduce tracking.

Version 1.0, 23 juli 2011

Rob van Eijk

1. [REDUCTION] appear to be a first time visitor.
2. [REDUCTION] limit the collection of data.
3. [REDUCTION] limit the use of data.

- Adjust the default browser settings (Tools|Options|Privacy tab) to custom setting with 'Accept third-party cookies' and 'Clear history when firefox closes'.

- Install [Beef Taco with IAB Europe opt-out cookies](#) | [View Patch](#)

- Install [Adblock Plus](#)

- Install [Tracking Detection System \(TDS\) for Adblock](#) | [View Ruleset](#)

The purpose of TDS is to reduce the privacy risk of collection and use of unique identifiers that can track you across different websites. A granular approach is needed in order to allow 3rd party content without being tracked. Blocking ads is not part of the scope. However an important side effect of limiting the collection of data with TDS, is less ads appearing on webpages you visit. In order to achieve a granular control for the user to reduce tracking, the following strategies are suggested:

- Reduce data flying in and out of the browser by using regular expressions. The Tracking Detection System (TDS) for Adblock contains an effective set of regular expressions. **Note:** do not mix with other rulesets because other rulesets can overrule the regular expressions with filter and allow rules resulting in a false sense of tracking protection.
- Reduce the use and sharing of data which are not blocked by regular expressions by sending opt-out cookies. The patched Beef Taco Add On keeps your opt-outs and sets a long expiration period. Cookies set through [youronlinechoices.com](#) are inconsistent on the expiration period. Furthermore, [youronlinechoices.com](#) tracks your opt-out behavior.
- Finally, it is important to accept 3rd party cookies in Firefox, otherwise opt-out cookies will not be send.
- Note: with TDS it is not necessary to use Privacy or Ghostery (Evidon) or Trackerblock (Privacy Choice) or the likes. Keep things simple and open source.

An example of a visualization displaying the visit to a set of websites can be found here: without TDS, and the same set with TDS.

Currently TDS works for [Firefox version 5 and 6](#) and [Firefox Portable Edition](#).



## D Source code: TDS cookies

The patch in this appendix, with cookies for the EU context derived from the industry supplies opt-out website [youronlinechoices.com](http://youronlinechoices.com), has been added to the release on July 23, 2011. Version 1.3.6 and higher of the Firefox Ad-on BeefTaco<sup>14</sup> contain the custom cookies.

All the experiments in this thesis have been using the opt-out cookie set from version 3.6. Since the opt-out cookie repository is a community effort, the number of opt-out cookies is growing over time.

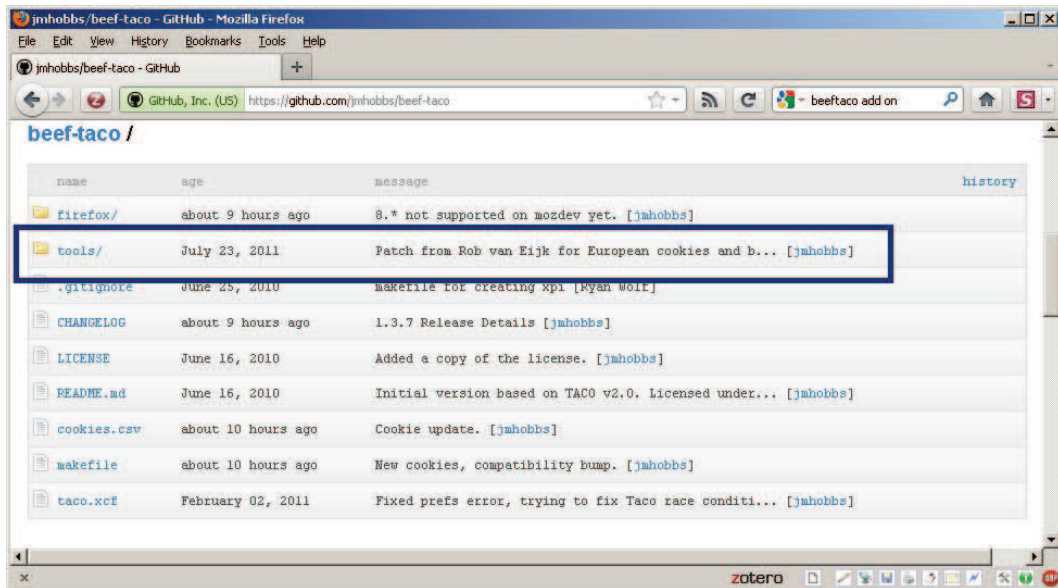


Figure 18: BeefTaco source code repository.

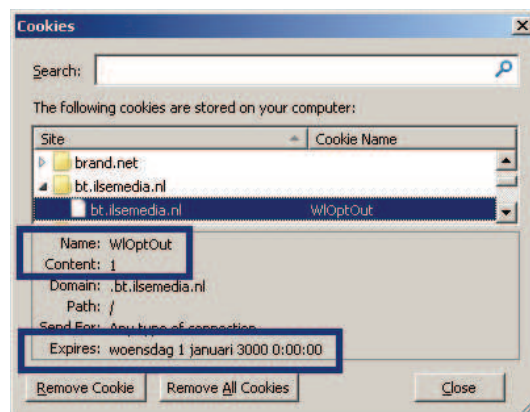


Figure 19: Example of BeefTaco cookie with a custom period. The browser extension enables persistent reloading of all opt-out cookies at launch of the Firefox web browser.

<sup>14</sup><http://jmhobbs.github.com/beef-taco/>

```

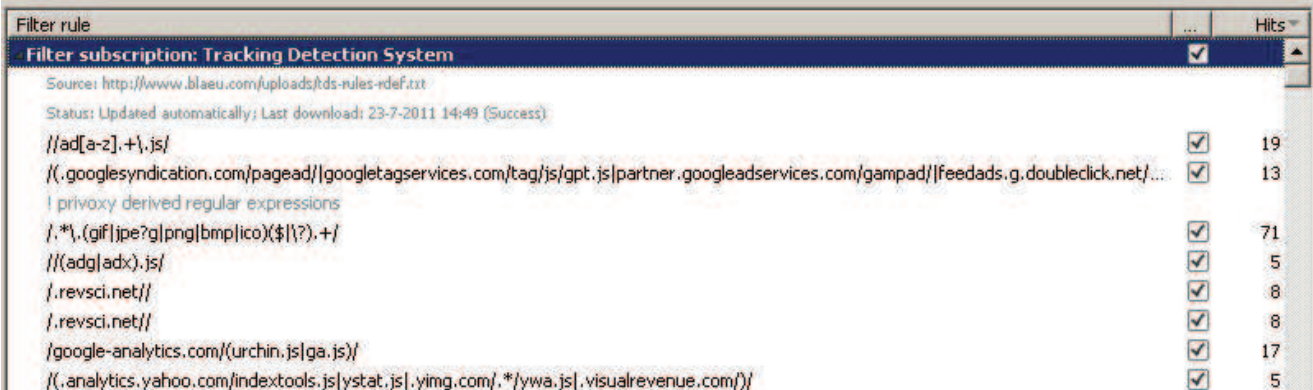
--- cookies.csv 2011-07-23 11:50:15.892800000 +0200
+++ cookies including IAB v5.csv 2011-07-24 11:18:44.961800000 +0200
@@ -258,3 +258,36 @@
, "vitamine.networldmedia.net", "/", "OPTOUT", 1, "2010-11-16 00:00:123", "Thanks to Jim Brock at
PrivacyChoice",
, "www.acxiom.com", "/", "CP", "null*", "2010-11-16 00:00:124", "Thanks to Jim Brock at
PrivacyChoice",
, "www.inadcoads.com", "/", "_iad_vsid", "99999999-9999-9999-9999-999999999999", "2010-11-16
00:00:125", "Thanks to Jim Brock at PrivacyChoice",
+ "Adform", ".adform.net", "/", "C", 3, "2011-07-23
12:00:00", "http://www.youronlinechoices.com/nl/uw-advertentie-voorkeuren", "http://track.adform
.net/serving/opt/iab/opt-out/?token="
+ "adGENIE", ".adgenie.co.uknet", "/", "_ngtid", "OPTOUT", "2011-07-23
12:00:00", "http://www.youronlinechoices.com/nl/uw-advertentie-voorkeuren", "http://adverts.adge
nie.co.uk/optout/optout.php?token="
+ "blinkx", ".blinkx.com", "/", "up", "off", "2011-07-23
12:00:00", "http://www.youronlinechoices.com/nl/uw-advertentie-voorkeuren", "http://www.blinkx.c
om/thirdparty/iab/opt-out?token="
+ "Crimtan", ".delivery.ctasnet.com", "/", "RTC8", "a_", "2011-07-23
12:00:00", "http://www.youronlinechoices.com/nl/uw-advertentie-voorkeuren", "http://www.ctasnet.
com/optout.html?token="
+ "Google/DoubleClick", ".doubleclick.net", "/", "_drt_", "OPT_OUT", "2011-07-23
12:00:00", "http://www.youronlinechoices.com/nl/uw-advertentie-voorkeuren", "http://googleads.g.
doubleclick.net/ads/preferences/iaboptout/optout?token="
+ "Microsoft Advertising", ".microsoft.com", "/", "TOptOut", "1", "2011-07-23
12:00:00", "http://www.youronlinechoices.com/nl/uw-advertentie-voorkeuren", "http://g.live.com/A
IPRIV/IABEU/optout"
+ "Sanoma Media Netherlands", ".bt.ilsemedia.nl", "/", "WlOptOut", "1", "2011-07-23
12:00:00", "http://www.youronlinechoices.com/nl/uw-advertentie-voorkeuren", "http://rc.bt.ilseme
dia.nl/iab/opt-out.php?token="
+ "Unanimis", ".unanimis.co.uk", "/", "OPTOUT", "True", "2011-07-23
12:00:00", "http://www.youronlinechoices.com/nl/uw-advertentie-voorkeuren", "http://www.unanimis
.co.uk/iab_optout/optout.php?redirect=true&token="
+ "YD", ".254a.com", "/", "ydmk[set]", "false", "2011-07-23
12:00:00", "http://www.youronlinechoices.com/nl/uw-advertentie-voorkeuren", "http://d.254a.com/y
ddb?type=yocout&token="
+ "Yell", ".servedby.yell.com", "/", "ytrmain", "-", "2011-07-23
12:00:00", "http://www.youronlinechoices.com/nl/uw-advertentie-voorkeuren", "http://standard.ser
vedby.yell.com/tr/iab-optout.php?token="
+ "Adtech", ".adtech.com", "/", "OptOut", "we will not set any more cookies", "2011-07-23
12:00:00", "http://www.youronlinechoices.com/nl/uw-advertentie-voorkeuren",
+, ".atdmt.com", "/", "TOptOut", "1", "2011-07-23
12:00:00", "http://www.youronlinechoices.com/nl/uw-advertentie-voorkeuren",
+, ".atwola.com", "/", "atdses", "O", "2011-07-23
12:00:00", "http://www.youronlinechoices.com/nl/uw-advertentie-voorkeuren",
+ "Admeld", ".tag.admeld.com", "/", "admeld_opt_out", "true", "2011-07-24
10:00:00", "http://www.networkadvertising.org/managing/opt_out.asp",
+ "Batanga (DoubleClick)", ".doubleclick.net", "/", "id", "OPT_OUT", "2011-07-24
10:00:00", "http://www.networkadvertising.org/managing/opt_out.asp",
+ "Buysight", ".pulsemgr.com", "/", "p", "OPTOUT", "2011-07-24
10:00:00", "http://www.networkadvertising.org/managing/opt_out.asp",
+ "Cognitive Match", ".cognitivematch.com", "/", "naiOptout", "cm", "2011-07-24
10:00:00", "http://www.networkadvertising.org/managing/opt_out.asp",
+ "Cognitive Match", ".cmadsasia.com", "/", "naiOptout", "cm", "2011-07-24
10:00:00", "http://www.networkadvertising.org/managing/opt_out.asp",
+ "Cognitive Match", ".cmads.com.tw", "/", "naiOptout", "cm", "2011-07-24
10:00:00", "http://www.networkadvertising.org/managing/opt_out.asp",

```

```
+ "Cognitive Match", ".cmadseu.com", "/", "naiOptout", "cm", "2011-07-24
10:00:00", "http://www.networkadvertising.org/managing/opt_out.asp",
+ "Cross Pixel Media", ".crosspixel.net", "/", "OPTOUT", "1", "2011-07-24
10:00:00", "http://www.networkadvertising.org/managing/opt_out.asp",
+ "Dapper", ".admonkey.dapper.net", "/", "DAPPEROPTOUT2", "OPT-OUT", "2011-07-24
10:00:00", "http://www.networkadvertising.org/managing/opt_out.asp",
+ "Glamcam", ".glam.com", "/", "optout", "1", "2011-07-24
10:00:00", "http://www.networkadvertising.org/managing/opt_out.asp",
+ "Media Innovation
Group", ".mookie1.com", "/", "%2emookie1%2ecom/%2f/1/o", "0/cookie", "2011-07-24
10:00:00", "http://www.networkadvertising.org/managing/opt_out.asp",
+ "Media Innovation
Group", ".decideinteractive.com", "/", "%2edecideinteractive%2ecom/%2f/1/o", "0/cookie", "2011-07-24
4 10:00:00", "http://www.networkadvertising.org/managing/opt_out.asp",
+ "Media Innovation Group", ".decdna.net", "/", "%2edecdna%2enet/%2f/1/o", "0/cookie", "2011-07-24
10:00:00", "http://www.networkadvertising.org/managing/opt_out.asp",
+ "Media Innovation Group", ".pm14.com", "/", "%2epm14%2ecom/%2f/1/o", "0/cookie", "2011-07-24
10:00:00", "http://www.networkadvertising.org/managing/opt_out.asp",
+ "Pulse360", ".pulse360.com", "/", "pulse360-opt-out", "1", "2011-07-24
10:00:00", "http://www.networkadvertising.org/managing/opt_out.asp",
+ "Red Aril", ".raasnet.com", "/", "o", "0", "2011-07-24
10:00:00", "http://www.networkadvertising.org/managing/opt_out.asp",
+ "Rocket Fuel", ".rfihub.com", "/", "k", "", "2011-07-24
10:00:00", "http://www.networkadvertising.org/managing/opt_out.asp",
+ "Wall Street on Demand", ".wsod.com", "/", "ub", "OPT_OUT", "2011-07-24
10:00:00", "http://www.networkadvertising.org/managing/opt_out.asp",
+ "Wall Street on Demand", ".ad.wsod.com", "/", "u", "OPT_OUT", "2011-07-24
10:00:00", "http://www.networkadvertising.org/managing/opt_out.asp",
+ "[x+1]", ".ru4.com", "/", "X1ID", "OO-0000000000000000", "2011-07-24
10:00:00", "http://www.networkadvertising.org/managing/opt_out.asp",
\ No newline at end of file
```

## E Source code: TDS regular expressions

In this appendix we find the regular expressions that have been used in the experiments in this thesis. The latest rule set can be downloaded from the TDS project homepage<sup>15</sup>. For instructions on how to install and use the TDS regular expressions, see Appendix C.



Filter rule	...	Hits
<b>Filter subscription: Tracking Detection System</b>		
Source: <a href="http://www.blaeu.com/uploads/tds-rules-rdef.txt">http://www.blaeu.com/uploads/tds-rules-rdef.txt</a>		
Status: Updated automatically; Last download: 23-7-2011 14:49 (Success)		
//ad[a-z].+\.js/	<input checked="" type="checkbox"/>	19
(/.google syndication.com/pagead/ google tagservices.com/tag/js/gpt.js partner.googleadservices.com/gampad/ feedads.g.doubleclick.net/ ...)	<input checked="" type="checkbox"/>	13
! privoxy derived regular expressions		
/.*\.(gif jpe?g png bmp ico)(\? \?)+/	<input checked="" type="checkbox"/>	71
/(adg adx).js/	<input checked="" type="checkbox"/>	5
/.revsci.net/	<input checked="" type="checkbox"/>	8
/.revsci.net/	<input checked="" type="checkbox"/>	8
/google-analytics.com/(urchin.js ga.js)/	<input checked="" type="checkbox"/>	17
(/.analytics.yahoo.com/indextools.js ystat.js .yimg.com/.*/ywa.js .visualrevenue.com/)/	<input checked="" type="checkbox"/>	5

Figure 20: Adblock Add-on with TDS rule set in action.

<sup>15</sup><http://www.blaeu.com/uploads/tds-rules-rdef.html>

**[Adblock]**

```
! Checksum: 8AGtKIyqcQsSSH8MOC6+fg
! Fetched from: http://www.blaeu.com/uploads/tds-rules-rdef.txt
! Expires: 5 days (update frequency)
! -----
! Name       : tds-rules-rdef.txt
! Author      : Rob van Eijk <rob@blaeu.com>
! Version     : Aug 14, 2011 12:57 CEST
! Copyright   : MPL 1.1/GPL 2.0/LGPL 2.1, see bottom of this file
! Description : Tracking Detection System (TDS)
! Usage       : http://www.blaeu.com/uploads/tds-rules-rdef.html
! -----
/.(barneveldsekrant|penthion).nl/scripts/edoris.+/
/.cognitivematch.com/
/.cmadsasia.com/
/.cmads.com.tw/
/.cmadseu.com/
/.crosspixel.net/
/.decideinteractive.com/
/.decdna.net/
/.pml4.com/
/.pulse360.com/
/weborama.fr/
/mint/\?js/
/mint/js/
/mintjs/
/.adgenie.co.uknet/
/.blinkx.com/
/.delivery.ctasnet.com/
/.bt.ilsemedia.nl/
/.unanimis.co.uk/
/.254a.com/
/.servedby.yell.com/
/escape.insites.eu/
/data.rtl.nl/system/track/
/ad-emea.doubleclick.net/adj/
/ad.aim4media.com/st\?.+/
/adserver3.adremedy.com/ttj\?.+/
/images.webads.nl/
/adonline.infobel.com/openx/www/delivery/spcjs.php\?.+/
/smartinit.webads.nl/Bin/SmartInit.dll\?.+/
/tracking.quisma.com/c.cfs\?.+/
/ad-serving.unibet.com/renderImage.aspx\?.+/
/www.geld.nl/WebResource.axd\?.+/
/proximis.fr/(app|affiliate)/
/ad.360yield.com/(ad|adj|adi)\?.+/
/(events|publishing).kalooga.com/overloadWidget/
/rc.bt.ilsemedia.nl/(Tag|Get)/ilsemedia/JS/
/csi.gstatic.com/csi\?.+/
/mediate.eu/(eas|media|EAS_tag)/
/(omniture|wanalytics|mbx|hbx|omniunih)(.*)?.js/
//.jpg/
//xtclicks.js/
/(a|d).ligatus.com/
//TextAds.js/
//googleads.html\?.+/
//ga-links.js/
```

```
! -----
! privoxy inspired regular expressions
! -----
/*\.(gif|jpe?g|png|bmp|ico)($|\?)\./
//ad[a-z]+\.\js/
/ads.php\?\.+/
/popup.js\?\.+/
/www.linkedin.com/analytics/noauthtracker\?+/
! -----
! Ghostery inspired regular expressions
! -----
/(pub.lookery.com/js/|lookery.com/look.js|/j/pub/look.js)/
/google-analytics.com/(urchin.js|ga.js)/
/.mybloglog.com//
/(.quantserve.com/|quant.js)/
/indextools.js/
/sitemeter.com/(js/counter.js|meter.asp)/
/www.lijit.com/informers/wijits/
/cetrk.com//
/(shots.snap.com/snap_shots.js|spa.snap.com/snap_preview_anywhere.js)/
/.statcounter.com/counter/(counter[0-9a-zA-Z_]*|frames).js/
//piwik.js/
//mint/\?js/
/facebook.com/beacon//
/www.typepad.com/t/stats/
/(stats.wordpress.com/|s.stats.wordpress.com/w.js)/
//salog.js.aspx/
/(.analytics.yahoo.com/indextools.js|ystat.js|.yimg.com/.*/ywa.js|.visualrevenue.com/)/
/otracking.com/js/
/analytics.engagd.com/archin-std.js/
/.nuggad.net/bk/
/static.crowdsience.com/start(-.*)\?\.js/
/.fmpub.net//
/(bid|d[0-9]\?)\.openx.(org|net)//
/www.assoc-amazon.(com|ca|co.uk|de|jp)/(e/ir|s/(ads.js|asw.js|link-enhancer))/
/(feeds.feedburner.com/~[fs]|/feedproxy.google.com/~fc)/
/clustrmaps.com/counter//
/(feedjit.com/serve/|feedjit.com/map//)
/(.googlesyndication.com/pagead/|googletagservices.com/tag/js/gpt.js|partner.googleadservices.com/gampad/|feedads.g.doubleclick.net/~at)/
/.hittail.com/mlt.js/
/friendfeed.com/embed/widget//
//woopra(.v(2|3|4))\?\.js/
/static.scribfire.com/ads.js/
/.imrworldwide.com//
/ad(-apac)\?\.([a-z][a-z])\?doubleclick.net//
/(Tacoda_AMS_DDC_Header.js|an.tacoda.net/|an.secure.tacoda.net//)
/(ad.yieldmanager.com/|optimizedby.rmxads.com|e.yieldmanager.net/script.js)/
/(content.dl-rms.com/|.dlqm.net/|.questionmarket.com//)
/trackingTags_v1.1.js/
/.xiti.com/(hit.xiti|get.at)/
/(/share-this.php|w.sharethis.com//)
/(/seesmic_topposters_v2.js|seesmic-wp.js)/
/static.addtoany.com/menu/(feed|page).js/
//addthis_widget.(js|php)/
/.revsci.net//
/(ads|clients|container).pointroll.com/
```





```
/.crwdcntrl.net//
/advertising.cpxinteractive.com//
/lypn.com/lp//
/revelations.trovus.co.uk/tracker//
/touchclarity/
/.insightexpressai.com//
/.kanoodle.com//
/(tags.bluekai.com|bkrtx.com/js)//
/(tr-metrics.loomia.com|assets.loomia.com/js)//
/otheronline.com/*/[a-z0-9]+.js/
/twittercounter.com/remote//
/rate.thummit.com/js//
/(.dtmpub.com|login.dotomi.com/ucm/ucmcontroller)/
/scripts.chitika.net//
/ad.spot200.com//
/.hitslink.com//
/www.w3counter.com/tracker.js/
/awstats_misc_tracker.js/
/stat.onestat.com//
/twitter.com/(javascripts/[0-9a-z]+.js|statuses/user_timeline)//
/.bmmatrix.com//
/include.reinvigorate.net//
/api.postrank.com//
/service.collarity.com//
/.smrtlks.com//
/www.tumblr.com/dashboard/iframe/
/blogrollr.com/embed.js/
/.casalemedia.com//
/track.blogcounter.de//
/api.widgetbucks.com/script/ads.js/
/www.nooked.com/javascripts/clearspring.js/
/(.mediaplex.com|.fastclick.net)//
/(www.haloscan.com/load|js-kit.com/[0-9a-z/]+.js)/
/buzzster.com/widget//
//trackalyze.js/
/(.burstbeacon.com|.burstnet.com)//
/.metricsdirect.com//
/(bspixel.bidsystem.com|bidsystem.adknowledge.com)//
/.nebuadvertising.com//
/.media6degrees.com//
//functionalTrends.js/
/.nuconomy.com/n.js/
/(.adrevolver.com|ads.bluelithium.com)//
/.glam.com/app/site/affiliate/viewChannelModule.act/
/storage.trafic.ro/js/trafic.js/
/.clicktracks.com//
/.enquiste.com/log.js/
/.extreme-dm.com//
/analytics.live.com//
/.sweepery.com/javascripts/*/[0-9a-zA-Z_]*.js/
/.socialtwist.com/
/(tracking.percentmobile.com|/percent_mobile.js)/
/stat.netmonitor.fi/js//
/munchkin.marketo.net//
/(api|leads).demandbase.com//
/pixel.fetchback.com//
/gw-services.vtrenz.net//
```

```

/(eluminate.js|data.cmcore.com/imp|data.coremetrics.com)/
/www.dialogmgr.com/tag/lib.js/
/tracking.fathomseo.com//
/(now.eloqua.com|elqcfg(xml)\?.js|elqimg.js)/
/.imiclk.com//
/.mmismm.com//
/rt.trafficfacts.com//
/.adnxs.com//
/.pro-market.net//
/.collective-media.net//
/.exelator.com//
/.fimserve.com//
/.interclick.com//
/.nexac.com//
/.trafficmp.com//
/.turn.com//
/(.realmedia.com/|realmedia/ads/|.247realmedia.com/realmedia/ads)/
/code.etracker.com//
/.(scorerresearch|securestudies|scorecardresearch).com//
/(.bizographics.com/|ad.bizo.com/pixel)/
/.snoobi.com/snoop.php/
/.rfihub.com//
/.shinystat.(com|it)//
/(sniff|stats).visistat.com//
/.sitestat.com//
/(.tynt.com/ti.js|.tynt.com/javascripts/tracer.js)/
/.i-stats.com/js/icounter.js/
/tracking.summitmedia.co.uk/js//
/.yandex.ru/(resource|metrika)/watch.js/
/ad.adriver.ru//
/.spylog.(com|ru)//
/.conversiondashboard.com//
/(spruce.rapleaf.com|.rlcdn.com)/
/static.zemanta.com//
/.alexametrics.com//
/vizisense.komli.net/pixel.js/
//phpmyvisites.js/
/cdn.doubleverify.com/[0-9a-zA-Z_-]*.js/
/one.statsit.com//
/.leadforcel.com/bf/bf.js/
/widgets.backtype.com//
/.iperceptions.com//
/.searchforce.net//
/tweetboard.com/tb.js/
/(tweetmeme.com/i/scripts/button.js|zulu.tweetmeme.com/button_ajax.js)/
/.zendesk.com/external/zenbox/overlay.js/
/.ivwbox.de//
/(media|recs).richrelevance.com//
/counter.yadro.ru//
/vistrac.com/static/vt.js/
/.blvdstatus.com/js/initBlvdJS.php/
/clixpy.com/clixpy.js/
/logdy.com/scripts/script.js/
/widgetserver.com/syndication/subscriber/
/widgets.clearspring.com//
/.google-analytics.com/siteopt.js/
/lt.navegg.com/lt.js/

```

```
.rsvpgenius.com//
/tracker.wordstream.com//
.kissmetrics.com//
/api.mixpanel.com//
.adtegrity.net//
/(.unica.com/|ntpagetag)/
/rover.ebay.com//
.inq.com//
.clixmetrix.com//
//liveball_api.js/
.marinsm.com//
/(cid|pi).pardot.com//
/js.stormiq.com/sid[0-9]*_[0-9]*.[0-9]*.js/
.histats.com//
/(widgets.amung.us/*.js|whos.amung.us/widget//)
/data.gosquared.com/
/www.apture.com/js/apture.js/
/c.compete.com/bootstrap/./bootstrap.js/
/(s3.amazonaws.com/wingify/vis_opt.js|dev.visualwebsiteoptimizer.com/deploy/js_visitor_setting
s.php.*|server.wingify.com/app/js/code/wg_consolidated.js)/
/mashlogic.com/(loader.min.js|brands/embed//)
/s3.buysellads.com/
/cim.meebo.com/cim/
/(c1|beta).web-visor.com/c.js/
/stags.peer39.net//
.eproof.com/js/*.js/
/(autocontext|o).begun.ru//
/foresee-(trigger(.*)\?|alive|analytics(.*)\?).js/
.quintelligence.com/quint.js/
/3dstats.com/cgi-bin/3dstrack(ssl)\?.cgi/
/addfreestats.com/cgi-bin/afstrack.cgi/
/(.webtrekk.net|webtrekk.js)/
/channelintelligence.com//
/ads.doclix.com/adserver/serve//
/bdv.bidvertiser.com//
/view.binlayer.com/ad/
/get.mirando.de//
/ads.adtiger.de//
/js.adscale.de//
/dsa.csdata1.com/
/(js|tag).admeld.com/
/cdn.wibiya.com/(toolbars|loaders)/
//adam/(cm8[0-9a-z_]+.js|detect)/
/www.actonsoftware.com/acton/bn//
/(.res-x.com/ws/r2/resonance|resxcls.js)/
/(/gomez.+\.js|.rt].axf8.net//)
/(cdn.mercent.com/js/tracker.js|link.mercent.com//)
/(.content.ru4.com/images/|.edge.ru4.com/smartserve/|.xp1.ru4.com/|ad.xplusone.com/|/xplus1/xp
1.js)/
/www.googleadservices.com/pagead/conversion/
/s.clickability.com/s/
/(xslt.alexa.com/site_stats/js/s/|widgets.alexa.com/traffic/javascript//)
/ad.retargeter.com/seg/
/tags.mediaforge.com/if/[0-9]+/
.visualdna.com//
/pixel.33across.com/
/tracking.searchmarketing.com//
```

```
/saas.intelligencefocus.com/sensor//
/www.domodomain.com/domodomain/sensor//
/.optimost.com//
/(html|image|js).ng//
/t.p.mybuys.com/js/mybuys3.js/
/track.roiservice.com/track//
/(bh|tag|ds).contextweb.com/
/.dmtracker.com//
/cdn.triggertag.gorillanation.com/js/triggertag.js/
/nr7.us/apps//
/searchignite.com/si/cm/tracking//
/.wa.marketingsolutions.yahoo.com/script/scriptservlet/
/.doubleclick.net/activityi/
/prof.estat.com/js//
/(adstat.4u.pl/s.js|stat.4u.pl/cgi-bin)//
/.hit.gemius.pl/
/jlinks.industrybrains.com/jsct/
//z(i|a)g.(js|gif)/
/bs.serving-sys.com/burstingpipe/(activityserver|adserver).bs/
/segment-pixel.invitemedia.com/
/assets.newsinc.com/(ndn.2.js|analyticsprovider.svc)//
/.predictad.com/scripts/(molosky|publishers)//
/(track|files).netshelter.net/
/.iesnare.com/
/.(brcdn|brsrvr).com//
/(beacon|js).clিকেequations.net/
/mct.rkdms.com/sid.gif/
/server[2-4].web-stat.com/
//internal/jscript/dwanalytics.js/
/a.giantrealm.com/
/tags.dashboardad.net/
/.oewabox.at/
/.amgdgt.com/(ads|base)//
/img.pulsemgr.com/
/(ads|image2).pubmatic.com/adserver//
/ad.fed.adecn.com/
/adelixir.com/(webpages/scripts/ne_roi_tracking.js|neroitrack)/
/keywordmax.com/tracking//
/amadesa.com/static/client_js/engine/amadesajs.js/
/.srtk.net/www/delivery//
/(tns-counter.ru|tns-counter.js|.tns-cs.net|statistik-gallup.net)/
/[a|c].adroll.com/
/hints.netflame.cc/service/script//
/.tynt.com/ts.js/
/ad.xtendmedia.com/
/newstogram.com/(.*)/(histogram|toolbar).js/
/ad.adlegend.com/
/a.mouseflow.com/
/(rotator.ad juggler.com|banners/ajtg.js|servlet/ajrotator)/
/picadmedia.com/js//
/(m|js|api|cdn).viglink.com/
/sageanalyst.net/
/svlu.net/
/w55c.net/
/(cdn|crosspixel).demdex.net/
/adserver.(adtechus.com|adtech.de)/
/.r.msn.com/scripts/microsoft_adcenterconversion.js/
```

```

(dw|adlog).com.com//
/netmng.com//
/px.owneriq.net/
/app.insightgrit.com/1//
/ads.bridgetrack.com//
/hits.convergetrack.com//
/(.dt07.net|mg.dt00.net/(u)\?js)/
/(publishers|.hat).halogennetwork.com//
/app.phonalytics.com/track/
/cn.clickable.net/js/cct.js/
/s0b\?.bluestreak.com/ix.e/
/tracking.dsmmadvantage.com/clients//
/(cdn.nprove.com/npcore.js|go.cpmadvisors.com//)
/.intermundomedia.com//
/vtracker.com/(counter|stats|tr.x|ts|tss|mvlive|digits)/
/(scripts|tm).verticalacuity.com/vat/mon/vt.js/
/(impression|ca).clickinc.com//
/.skimresources.com/js/
/ad.metanetwork.com//
/rt.legolas-media.com//
/cdn.krxd.net/krux.js/
/(clickserve.cc-dt.com/link/|gan.doubleclick.net/gan)/
/.fwmrm.net/(ad|g)//
/.ibpxl.com//
/pmetrics.performancing.com/(js|in.php|[0-9]*.js)/
/(tracking.conversionlab.it|conversionlab.trackset.com/track//)
/visualpath[0-9].trackset.it//
/.adblade.com//
/ads.undertone.com/
/ads.lucidmedia.com/clicksense/pixel/
/track.did-it.com//
/tag.didit.com/(didit|js)//
/(content|track).pulse360.com//
/.adgear.com//
/j.clickdensity.com/cr.js/
/visitorville.com/js/plgtrafic.js.php/
/ads.affbuzzads.com//
/cts.vresp.com/s.gif/
/.linksynergy.com/
/websitealive[0-9].com//
/b.monetate.net/js//
/(adserver|int).teracent.net//
/html.aggregateknowledge.com/iframe/
/.tellopart.com/crumb/
/vitamine.networldmedia.net/bts//
/projectwonderful.com/(ad_display.js|gen.php)/
/voicefive.com/*.pli/
/(econda.*.js|www.econda-monitor.de/els/logging)/
/adspeed.(com|net)/ad.php/
/live1.netupdater.info/live.php/
/.ic-live.com/(goat.php|[0-9][0-9][0-9][0-9].js)/
/adreadytractions.com/rt//
/ad.yieldmanager.com/pixel/
/(adcode|conv).adengage.com/
/server.cpmstar.com/
/tracker.financialcontent.com//
/www.csm-secure.com/scripts/clicktracking.js/

```

```

/sitecompass.com/(sc_cap/[ij]pixel)/
/link.ixsl.net/s/at/
/upsellit.com/
/accelerator-media.com/pixel/
/(haku|puma|cheetah).vizu.com/
/.adfox.ru/preparecode/
/(goku.brightcove.com|admin.brightcove.com/js)/
/.mathtag.com/
/(ad|ad2|counter).rambler.ru/
/lookup.bluecava.com/
/monitus(_tools)\?.js/
/service.optify.net//
/ad.reduxmedia.com//
/(servedby|geo).precisionclick.com//
/destinationurl.com/
/counters.gigya.com/
/webiqonline.com/
/radar.cedexis.(com|net)/
/as00.estara.com/as/initiatecall2.php/
/.atgsvcs.com/js/atgsvcs.js/
/.mookie1.com/
/stats.businessol.com/
/webtraxs.(js|com)/
/pixel.adbuyer.com/
/adadvisor.net/
/vertster.com/.*vswap.js/
/voice2page.com/naa_1x1.js/
/ad.z5x.net/
/data.resultlinks.com/
/eyewonder.com//
/ad(media)\?.wsod.com/
/ats.tumri.net/
/.visiblemeasures.com/log/
/fx.gtop(stats.com|.ro)/js/gtop.js/
/adsfac.(eu|us|sg|net)/
/pixazza.com/(static/)\?widget/
/(qnsr.com|thecounter.com/id)/
/(smp|leads).specificmedia.com/
/gmads.net/
/levexis.com/
/(mmcore.js|cg-global.maxymiser.com)/
/nexus.ensighten.com/
/uptrends.com/(aspx/uptime.aspx|images/uptrends.gif)/
/visitstreamer.com/vs.js/
/crm-metrix.com/
/utd.stratigent.com/
/facebook.com/(plugins|widgets)/.*.php/
/.chango.(ca|com)/
/cdna.tremormedia.com/
/admonkey.dapper.net/
/p-td.com/
/yieldoptimizer.com/
/displaymarketplace.com/
/ads.addynamix.com/category/
/(dd|ff).connextra.com/
/(api|apps).conduit.com/
/udmserve.net/

```

/pixel.adpredictive.com/  
 /px.steelhousemedia.com/  
 /tracking.godatafeed.com/  
 /.trumba.com/scripts/spuds.js/  
 /adspaces.ero-advertising.com/  
 /ads.adxpansion.com/  
 /clarity.adinsight.eu/static/adinsight/  
 //performable/pax/  
 /snapabug.appspot.com/snapabug/  
 /breathe.c3metrics.com/  
 /um.simpli.fi/ab\_match/  
 /dt.admission.net/retargeting/displaytracker.js/  
 /yumenetworks.com/dynamic/  
 /trk.vindicosuite.com/Tracking//  
 /creativeby2.unicast.com/(assets|script2)/  
 /(hs|sw).interpolls.com/  
 /rs.gwallet.com/r1/pixel/  
 /wtp101.com//  
 /.adshuffle.com//  
 /(servedby|stat|cdn|a).flashtalking.com//  
 /(servedby|ads|event).adxpose.com/  
 /utag.loader.js/  
 /(us.img|ads).e-planning.net/  
 /choices.truste.com/ca/  
 /btstatic.com/  
 /tldadserv.com/impopup.php/  
 /wiredminds.de/track/  
 /segs.btrll.com/v[0-9]/tpix//  
 /p.brilig.com/contact/bct/  
 /xcdn.xgraph.net/([0-9]|partner.js)/  
 /cdn.doubleverify.com/oba/  
 /beencounter.com/b.js/  
 /traveladvertising.com/live/tan/  
 /everestjs.net|pixel([0-9]\*)\?.everesttech.net/  
 /ctasnet.com/  
 /raasnet.com/  
 /eyereturn.com/  
 /rainbow.mythings.com/  
 /esml.net/  
 /adconnexa.com/  
 /.spongecell.com//  
 /((p(shared|files|thumbnails)|embed).5min.com//  
 /stat.yellowtracker.com/  
 /.backbeatmedia.com/  
 /ad.adserverplus.com/  
 /(amconf|core|adcontent).videoegg.com/(siteconf|eap|alternates|ads)//  
 /contaxe.com/go/  
 /ad.zanox.com/  
 /zbox.zanox.com/  
 /www.zanox-affiliate.de/ppv//  
 /rts.sparkstudios.com//  
 /adserver.juicyads.com/  
 /ads.pheedo.com/  
 /www.cbox.ws/box/  
 /yandex.ru/cycounter/  
 /ads.ad4game.com/  
 /a.ucoz.net/

/c.bigmir.net/  
 /.tradetracker.net/  
 /gwp.nuggad.net/  
 /tqlkg.com/image/  
 /ads.brand.net/  
 /js.geoads.com/  
 /certifica.js/  
 /certifica-js14.js/  
 /hits.e.cl/  
 /prima.certifica.com/  
 /ads.traffiq.com/  
 /livepass.conviva.com/  
 /surveybuilder.buzzlogic.com/  
 /cadreon.com/tags/defaultads/  
 /184.73.199.28/tracker/event/  
 /a.akncdn.com/  
 /c.betrad.com/geo/ba.js/  
 /beacon.dedicatednetworks.com/  
 /ads.dedicatedmedia.com/  
 /adserver.veruta.com/  
 /veruta.com/scripts/trackmerchant.js/  
 /anormal-tracker.de/tracker.js/  
 /anormal-tracker.de/countv2.php/  
 /roia.biz//  
 /(ads|sync).(adaptv|tidaltv).(tv|com)/  
 /specificmedia.com/otherassets/ad\_options\_icon.png/  
 /.keewurd.com/  
 /rt.liftdna.com/  
 /paid-to-promote.net/images/ptp.gif/  
 /777seo.com//  
 /.bridgetrack.com/track/  
 /.bridgetrack.com/a/s//  
 /m.webtrends.com/  
 /.adsrvr.org/  
 /hurra.com/ostracker.js/  
 /spl.convertro.com/  
 /domdex.(net|com)/  
 /qjex.net/  
 /mi.adinterax.com/(js|customer)/  
 /cdn.undertone.com/js/ajs.js/  
 /ad.adperium.com/  
 /adlily.adperium.com/  
 /.adperium.com/js/adframe.js/  
 /.adperium.com/abd.php/  
 //\_\_utm./  
 /tags.nabbr.com/  
 /.dmtry.com/  
 /perf.overture.com/  
 //ki.js//  
 /j.kissinsights.com/  
 /doug1izaerwt3.cloudfront.net/  
 /dnn506yrbagrg.cloudfront.net/  
 /cdn.optimizely.com/js//  
 /webgozar.ir/c.aspx/  
 /webgozar.com/counter/  
 /r.i.ua/  
 /hotlog.ru/cgi-bin/hotlog/



/qksz.net/  
/dlqpxklwfeh8v1.cloudfront.net/  
/cn01.dwstat.cn/  
/(gopjn|pjatr|pjtra|pntra|pntrac|pntrs).com//  
/ftjcfx.com/  
/tqlkg.com/  
/yceml.net/  
/analytics.matchbin.com/  
/ads.matchbin.com/  
/shareasale.com/  
/pages.etology.com//  
/s.thebrighttag.com/  
/ads2\?.smowtion.com/  
/imgsrv.nextag.com/imagefiles/includes/roitrack.js/  
/rcm-ca.amazon.ca/e/cm/  
/rcm-uk.amazon.co.uk/e/cm/  
/rcm-de.amazon.de/e/cm/  
/fls.doubleclick.net/  
/pixel.adsafeprotected.com/  
/fw.adsafeprotected.com/  
/.tradedoubler.com/  
/.streamray.com/  
/.pop6.com/  
/.cams.com/  
/.nostringsattached.com/  
/.getiton.com/  
/.adultfriendfinder.com/  
/.double-check.com/  
/.facebookofsex.com/  
/mediacdn.disqus.com/  
/widgets.digg.com/  
/.lijit.com/delivery/  
/cdn.technoratimedia.com/  
/log.feedjit.com/  
/.smartadserver.com/  
/.sponsorads.de/  
/google.com/afsonline/show\_afs\_ads.js/  
/conduit-banners.com/  
/spotxchange.com/track/  
/adserver.advertisespace.com/  
/ads.advertisespace.com/  
/dinclinx.com/  
/widgets.twimg.com/j//  
/platform.twitter.com/widgets/  
/sitebro.net/track.js/  
/static.poll daddy.com/p/  
/contextlinks.netseer.com/  
/.adform.net/  
/ads.newtention.net/  
/trk.newtention.net/  
/adserving.cpxadroit.com/  
/choicesj.truste.com/ca/  
/google.com/adsense/search/ads.js/  
/eplayer.clipsyndicate.com//  
/.adition.com/  
/ad.clickotmedia.com/  
/badge.facebook.com/

```

/dlros97qkrwjf5.cloudfront.net/
/apis.google.com/js/plusone.js/
/effectivemeasure.net/
/clicktale.pantherssl.com/
/srv.clickfuse.com/
/d1l6p2sc9645hc.cloudfront.net/
/.opentracker.net/
/(img|script).footprintlive.com/
/adreactor.com/
//xgemius.js/
/adocean.pl/
/adcentriconline.com/
/waterfrontmedia.com/
/customerconversio.com/
/facebook.com/connect/
/connect.facebook.net/
/static.ak.connect.facebook.com/*.js.php/
//fbconnect.js/
/fbcdn.net/connect.php/js/fb.share/
/adviva.net/
/.skimlinks.com/(api|js)//
/.1[12]2.2o7.net//
/hitbox.com/
/.omtrdc.net//
/(omniture|mbox|hbx|omniunih)(.*)\?.js/
/s(c)\?_code[0-9a-zA-Z_-]*(.[0-9a-zA-Z_-]*)\?.js/
/common.onset.freedom.com/fi/analytics/cms//
/c.statcounter.com//
//(adg|adx).js/
/(afr|ajs|avw).php/
/wms.assoc-amazon.com/
/rcm.amazon.com/e/cm/
/.doubleclick.net/pagead//
//webtrends(.*)\?.js/
/.webtrendslive.com/
//js_xiti.js/
//xtcore.js/
/.addthis.com/js/widget.(js|php)/
/l.addthiscdn.com/
/wunderloop.net//
/ad.targetingmarketplace.com//
/revsci.(.*)/gw.js/
/dis(.*)\?.criteo.com/
/adsyndication.msn.com/delivery/getads.js/
/o.aolcdn.com/js/mg2.js/
/(r1.ace|ace-tag|servedby|uac).advertising.com/
/.atwola.com//
/baynote.net/
/.visualdna-stats.com//
/netupdater[0-9].de/
//netupdater(_live)\?/
/.mail.ru/counter/
/.list.ru/counter/
/(l|b)ive.monitus.net/
/(do.am|at.ua)/stat//
/ucoz.(.*)/(stat/|main/\?a=ustat)/
/.meteorsolutions.com//

```

/freeonlineusers.com//  
 /.statistics.ro//  
 /.keymetric.net//  
 /.advertserve.com//  
 /.successfultgether.co.uk//  
 /d3pkntwtp2ukl5.cloudfront.net//  
 /t.unbounce.com//  
 /zopim.com//  
 /ads.brainient.com//  
 /.etracker.de//  
 /ad.103092804.com//  
 /.trafficrevenue.net//  
 /.adbull.com//  
 /.complexmedianetwork.com//  
 /.complex.com//  
 /pocketcents.com//  
 //hellobar.js/  
 /.ppctracking.net/  
 /expo-max.com/  
 /hitsniffer.com//  
 /.impressiondesk.com//  
 /objects.tremormedia.com/embed/js/  
 /.gigya.com/js/socialize.js/  
 /.atemda.com/  
 /lookery.com/  
 /google-analytics.com/  
 /mybloglog.com/  
 /quantserve.com|com.quantserve/  
 /sitemeter.com/  
 /lijit.com/  
 /(2o7.net|omtrdc.net) /  
 /cetrk.com/  
 /(shots.snap.com|spa.snap.com) /  
 /www.statcounter.com/  
 /www.typepad.com/  
 /stats.wordpress.com/  
 /(analytics.yahoo.com|.yimg.com) /  
 /otracking.com/  
 /analytics.engagd.com/  
 /nuggad.net/  
 /static.crowdscience.com/  
 /fmpub.net/  
 /openx.(org|net) /  
 /assoc-amazon.com/  
 /(feeds.feedburner.com|feedproxy.google.com) /  
 /clustrmaps.com/  
 /(feedjit.com) /  
 /(googlesyndication.com|googleadservices.com|2mdn.net) /  
 /hittail.com/  
 /friendfeed.com/  
 /static.scribefire.com/  
 /imrworldwide.com/  
 /doubleclick.net/  
 /tacoda.net/  
 /(ad.yieldmanager.com|optimizedby.rmxads.com|e.yieldmanager.net) /  
 /(dl-rms.com|dlqm.net|questionmarket.com) /  
 /webtrendslive.com/

/xiti.com/  
/sharethis.com/  
/addtoany.com/  
/addthis.com/  
/(revsci.net|targetingmarketplace.com) /  
/pointroll.com/  
/static.chartbeat.com/  
/static.getclicky.com/  
/rubiconproject.com/  
/lct.salesforce.com/  
/sphere.com/  
/criteo.(pro|com) /  
/(social.bidsystem.com|cubics.com) /  
/statisfy.net/  
/(adsl.msn.com|adsyndication.msn.com) /  
/outbrain.com/  
/specificclick.net/  
/(atdmt.com|adbureau.net) /  
/assets.skribit.com/  
/kona.kontera.com/  
/adbrite.com/  
/adultadworld.com/  
/gunggo.com/  
/doublepimp.com/  
/ads.sexinyourcity.com/  
/clicksor.com/  
/static.hubspot.com/  
/adsonar.com/  
/technorati.com/  
/xslt.alexa.com/  
/tribalfusion.com/  
/disqus.com/  
/ads.sixapart.com|saymedia.com/  
/ads.blogherads.com/  
/(o.aolcdn.com|advertising.com|atwola.com) /  
/leadback.advertising.com/  
/overture.com/  
/intensedebate.com/  
/connect.facebook.com/  
/btbuckets.com/  
/gumgum.com/  
/hook.yieldbuild.com/  
/d.yimg.com/  
/triggit.com/  
/digg.com/  
/cache.blogads.com/  
/zedo.com/  
/intellitxt.com/  
/afy11.net/  
/gmodules.com/  
/server.iad.liveperson.net/  
/clicktale.net/  
/(crwdcntrl.net|lotame.com) /  
/adserving.cpxinteractive.com/  
/lypn.com/  
/revelations.trovus.co.uk/  
/insightexpressai.com/

/kanoodle.com/  
 /bluekai.com/  
 /assets.loomia.com/  
 /otheronline.com/  
 /twittercounter.com/  
 /rate.thummit.com/  
 /dtmpub.com/  
 /chitika.net/  
 /ad.spot200.com/  
 /counter.hitslink.com/  
 /bmmatrix.com/  
 /include.reinvigorate.net/  
 /postrank.com/  
 /service.collarity.com/  
 /smrtlnks.com/  
 /www.tumblr.com/  
 /blogrollr.com/  
 /casalemedia.com/  
 /api.widgetbucks.com/  
 /mediaplex.com/  
 / (haloscan.com|js-kit.com) /  
 /buzzster.com/  
 / (burstbeacon.com|burstnet.com|burtsmedia.com) /  
 /metricsdirect.com/  
 / (bidsystem.com|adknowledge.com) /  
 /nebuadvertising.com/  
 /media6degrees.com/  
 /nuconomy.com/  
 /adrevolver.com/  
 /glam.com/  
 /clicktracks.com/  
 / (enquisite.com|eightfoldlogic.com) /  
 /extreme-dm.com/  
 /analytics.live.com/  
 /sweepery.com/  
 /socialtwist.com/  
 /tracking.percentmobile.com/  
 /munchkin.marketo.net/  
 /demandbase.com/  
 /fetchback.com/  
 /gw-services.vtrenz.net/  
 /dialogmgr.com/  
 /tracking.fathomseo.com/  
 / (imiclk.com|abmr.net) /  
 /mmismm.com/  
 /rt.trafficfacts.com/  
 /adnxs.com/  
 /pro-market.net/  
 / (collective-media.net|collective.com) /  
 /exelator. (com|net|biz) /  
 /fimserve.com/  
 /interclick.com/  
 / (nexac.com|nextaction.net) /  
 /trafficmp.com/  
 /turn.com/  
 / (247)\?realmedia.com/  
 /code.etracker.com/

/ (scorerresearch|securestudies|scorecardresearch) .com/  
 / (bizographics.com|bizo.com) /  
 /snoobi.com/  
 /rfihub.com/  
 /visistat.com/  
 /sitestat.com/  
 /tynt.com/  
 /i-stats.com/  
 /tracking.summitmedia.co.uk/  
 /adriver.ru/  
 /spylog.(com|ru) /  
 /conversiondashboard.com/  
 / (rapleaf.com|rcldn.com) /  
 /static.zemanta.com/  
 /alexametrics.com/  
 /vizisense.komli.net/  
 /doubleverify.com/  
 /one.statsit.com/  
 /leadforcel.com/  
 /widgets.backtype.com/  
 /iperceptions.com/  
 /searchforce.net/  
 /tweetboard.com/  
 /tweetmeme.com/  
 /zendesk.com/  
 /ivwbox.de/  
 /richrelevance.com/  
 /counter.yadro.ru/  
 /vistrac.com/  
 / (blvdstatus.com|seoq.com) /  
 /clixpy.com/  
 /logdy.com/  
 /widgetserver.com/  
 / (clearspring.com|connectedads.com) /  
 /navegg.com/  
 /rsvpgenius.com/  
 /tracker.wordstream.com/  
 /kissmetrics.com/  
 /api.mixpanel.com/  
 /adtegrity.net/  
 /unica.com/  
 /rover.ebay.com/  
 /inq.com/  
 /clixmetrix.com/  
 /marinsm.com/  
 /pardot.com/  
 /c.compete.com/  
 / (wingify.com|visualwebsiteoptimizer.com) /  
 /webtrekk.net/  
 / (link|cdn) .mercent.com/  
 /retargeter.com/  
 / (dizzads.com|mediaforge.com) /  
 /visualdna.com/  
 /tracking.searchmarketing.com/  
 /dmtracker.com/  
 /nr7.us/  
 /searchignite.com/

/serving-sys.com/  
 /iesnare.com/  
 /rkdms.com/  
 /web-stat.com/  
 /a.giantrealm.com/  
 /pubmatic.com/  
 /keywordmax.com/  
 /xtendmedia.com/  
 /viglink.com/  
 /w55c.net/  
 /bridgetrack.com/  
 /mvtracker.com/  
 /ad.metanetwork.com/  
 /legolas-media.com/  
 /fwrm.net/  
 /adblade.com/  
 /undertone.com/  
 /lucidmedia.com/  
 /clickdensity.com/  
 /websitealive[0-9].com/  
 /(smtad.net|teracent.net|ytasa.net)/  
 /aggregateknowledge.com/  
 /(bvmedia.ca|networldmedia.net)/  
 /adspeed.(com|net)/  
 /adreadytractions.com/  
 /adengage.com/  
 /cpmstar.com/  
 /sitecompass.com/  
 /accelerator-media.com/  
 /vizu.com/  
 /tumri.net/  
 /mathtag.com/  
 /bluecava.com/  
 /service.optify.net/  
 /reduxmedia.com/  
 /precisionclick.com/  
 /mookie1.com/  
 /adbuyer.com/  
 /z5x.net/  
 /ad(media)\?.wsod.com/  
 /(dw|adlog).com.com/  
 /visiblemeasures.com/  
 /adsfac.(eu|us|sg|net)/  
 /qnsr.com/  
 /(adviva.net|specificmedia.com)/  
 /gmads.net/  
 /chango.com/  
 /p-td.com/  
 /yieldoptimizer.com/  
 /displaymarketplace.com/  
 /ic-live.com/  
 /ads.addynamix.com/  
 /connextra.com/  
 /conduit.com/  
 /udmserve.net/  
 /adpredictive.com/  
 /steelhousemedia.com/

/simpli.fi/  
/yumenetworks.com/  
/admission.net/  
/vindicosuite.com/  
/unicast.com/  
/adadvisor.net/  
/interpolls.com/  
/gwallet.com/  
/(adnetik|wtp101).com/  
/adshuffle.com/  
/flashtalking.com/  
/adxpose.com/  
/ru4.com/  
/bluestreak.com/  
/cmcore.com/  
/hitbox.com/  
/adtech(us)\?.(com|de) /  
/adlegend.com/  
/33across.com/  
/amgdgt.com/  
/contextweb.com/  
/dotomi.com/  
/gigya.com/  
/demdex.net/  
/adjuggler.com/  
/admeld.com/  
/adroll.com/  
/owneriq.net/  
/pulsemgr.com/  
/admonkey.dapper.net/  
/intermundomedia.com/  
/fastclick.net/  
/invitemedia.com/  
/netmng.com/  
/amadesa.com/  
/wunderloop.net/  
/rlcdn.com/  
/adgear.com/  
/tellapart.com/  
/predictad.com/  
/mybuys.com/  
/acxiom.com/  
/adcentriconline.com/  
/nspmotion.com/  
/btrll.com/  
/choicestream.com/  
/smartadserver.com/  
/spotexchange.com/  
/adinterax.com/  
/tealium.com/  
/ads.e-planning.net/  
/p.brilig.com/  
/xgraph.net/  
/proximic.com/  
/struq.com/  
/beencounter.com/  
/traveladvertising.com/



/everesttech.net/  
/adchemy.com/  
/ctasnet.com/  
/tidaltv.com/  
/raasnet.com/  
/eyereturn.com/  
/mythings.com/  
/esml.net/  
/domdex.com|qjex.net/  
/adv.adsbwm.com/  
/spongecell.com/  
/ads.heias.de/  
/ask.com/  
/wedorama.fr/  
/valueclick.net/  
/effectivemeasure.net/  
/tradedoubler.com/  
/brand.net/  
/scanscout.com/  
/tracking.quisma.com/  
/buzzlogic.com/  
/tracking.reedge.com/  
/veruta.com/  
/5min.com/  
/adinsight.eu/  
/videoegg.com/  
/ad.zanox.com/  
/tqlkg.com/  
/traffiq.com/  
/cadreon.com/  
/ads.dedicatedmedia.com/  
/adaptv.com/  
/.adsrvr.org/  
/adform.net/  
/newtention.net/  
/.checkm8.com/  
/channelintelligence.com/  
/clickfuse.com/  
/adreactor.com/  
/adocean.pl/

! -----  
! Version: MPL 1.1/GPL 2.0/LGPL 2.1  
!  
! The contents of this file are subject to the Mozilla Public License Version  
! 1.1 (the "License"); you may not use this file except in compliance with  
! the License. You may obtain a copy of the License at  
! <http://www.mozilla.org/MPL/>  
!  
! Software distributed under the License is distributed on an "AS IS" basis,  
! WITHOUT WARRANTY OF ANY KIND, either express or implied. See the License  
! for the specific language governing rights and limitations under the  
! License.  
!  
! The Initial Developer of the Original Code is Rob van Eijk.  
!  
! Portions created by the Initial Developer are Copyright (C) 2010  
! the Initial Developer. All Rights Reserved.

-1-

## F Source code: creating nodes and links

We find the PERL source code for processing the data in this appendix. Raw HTTP data that can be saved to a file with the Firefox Add-on LiveHTTPheaders<sup>16</sup>. In order to process the HTTP data, we can use the following command: `perl main.pl <file>`. The code has only been tested under cygwin, but should be platform independent. You can use your favorite PERL implementation. The resulting files of the data process are:

- log file, <file.log>
- output file in JSON format, <file.json>
- compressed archive of the sqlite database, <file.gz>

The log files are added to this thesis in Appendix B. The output files are stored in a JSON format. We can see a short example of the output layout to get a good view of the structure of the JSON format:

```
{"nodes":[{"name":"195.177.x.x","group":2},(...),{"name":"zie.nl","group":2}],
"links":[{"source":0,"target":48,"value":7},(...),{"source":50,"target":50,"value":7}]}
```

The archive of the sqlite database itself in plain ASCII hand has the following layout:

```
PRAGMA foreign_keys=OFF;
BEGIN TRANSACTION;
CREATE TABLE trackers (tracker, groups);
INSERT INTO "trackers" VALUES('aamulehti.fi','0');
(...)
CREATE TABLE distinct_groups (tracker PRIMARY KEY, groups);
INSERT INTO "distinct_groups" VALUES('247realmedia.com','1');
(...)
COMMIT;
```

---

<sup>16</sup><http://livehttpheaders.mozdev.org/>

```
#!/usr/bin/perl
# -----
# Name      : main.pl
# Author    : Rob van Eijk <rob@blaeu.com>
# Version   : 29 juli 2011
# Copyright : MPL 1.1/GPL 2.0/LGPL 2.1, see bottom of this file
# Description : convert LiveHTTPHeaders to D3 json
# Usage     : main.pl <file>
#
# Run the LiveHTTPHeaders Add On in Firefox. Visit websites and save the
# captured headers to a <file>
# This PERL script will convert the <file> into a JSON that can be used
# with the D3 JavaScript library to display tracking in a force-diagram.
# -----

use strict;
use warnings;
use DBI;
use Data::Dumper;
use Switch;
use URI;

# read the specified file
die "Usage: $^X $0 liveHTTP-headers\n" unless @ARGV;
my $file = $ARGV[0];

# -----
print "start processing '$file'\n";

# -----
my $start = time;
my $outputfile = $file . ".json";
my $logfile = $file . ".log";
my $dumpfile = $file . ".gz";
$dumpfile =~ s/ /\ \ /g;
my $log = "start processing $file\r\n";

my $http_headers = readFile($file);

# -----
print "[1/8] opening sqlite database\n";

# -----
my $dbh = DBI->connect( "dbi:SQLite:connections.sqlite",
    "", "", { RaiseError => 1, AutoCommit => 1 } );

eval {
    local $dbh->{PrintError} = 0;
    $dbh->do("DROP TABLE trackers");
    $dbh->do("DROP TABLE domains");
    $dbh->do("DROP TABLE names");
    $dbh->do("DROP TABLE distinct_groups");
    $dbh->do("DROP TABLE distinct_domains");
    $dbh->do("DROP TABLE distinct_names");
};

# (re)create table(s)
```

```
$dbh->do("CREATE TABLE trackers (tracker, groups)")
|| die "Could not create table TRACKERS";
$dbh->do("CREATE TABLE distinct_groups (tracker PRIMARY KEY, groups)")
|| die "Could not create table GROUPS";
$dbh->do("CREATE TABLE domains ( source_domain, target_domain )")
|| die "Could not create table DOMAINS";
$dbh->do(
"CREATE TABLE distinct_domains (id INTEGER PRIMARY KEY, source_domain, target_domain,
source_node INTEGER, target_node INTEGER, tot_value INTEGER)"
) || die "Could not create table DISTINCT_DOMAINS";
$dbh->do("CREATE TABLE names (id INTEGER PRIMARY KEY, domain )")
|| die "Could not create table NAMES";
$dbh->do(
"CREATE TABLE distinct_names (id INTEGER PRIMARY KEY, domain, groups INTEGER)"
) || die "Could not create table DISTINCT_NAMES";

# -----
# print "pre-loading preliminary websites...\n";

# -----
my $group = 0;
foreach my $websites (
qw(tvrgids.nl telegraaf.nl nos.nl rtl.nl sunweb.nl live.com fok.nl hyves.nl geenstijl.nl
dumpert.nl wehkamp.nl ad.nl nu.nl buienradar.nl linkedin.com ah.nl marktplaats.nl
volkskrant.nl funda.nl ing.nl bol.com facebook.com google.nl rabobank.nl tweakers.net
twitter.com youtube.com trouw.nl parool.nl nrc.nl nd.nl bndestem.nl agd.nl
gooieneemlander.nl frieschdagblad.nl lc.nl gic.nl leidschdagblad.nl dvhn.nl
brabantsdagblad.nl tetubantia.nl limburg.nl pzc.nl noordhollandsdagblad.nl
barneveldsekrant.nl ed.nl ijmuidercourant.nl haarlemsdagblad.nl refdag.nl
katholieknieuwsblad.nl metronieuws.nl spitsnieuws.nl destentor.nl)
)
{
    $dbh->do("INSERT INTO trackers VALUES ('$websites', '$group')");
}

# -----
print "[2/8] pre-loading confirmed tracking domains...\n";

# -----
$group = 1;
foreach my $tracker (
qw(adgenie.co.uk blinkx.com ilsemedia.nl unanimis.co.uk 254a.com yell.com adtech.com)
)
{
    $dbh->do("INSERT INTO trackers VALUES ('$tracker', '$group')");
}
$file = "trackers.json";
my $priv_choice = readFile($file);
my @track_headers = split /\./, $priv_choice;
foreach my $tracker (@track_headers) {
    if ( $tracker =~ m/domain/ ) {
        $tracker =~ s/\r\n \\"domain\": \\"//g;
        $tracker =~ s/\\"//g;
        $tracker =~ s/co\.uk/co_uk/g;
        my @subdomains = split /\./, $tracker;

        switch ( scalar @subdomains ) {
```

```

        case 2 { $tracker = $subdomains[0] . "." . $subdomains[1] }
        case 3 { $tracker = $subdomains[1] . "." . $subdomains[2] }
        case 4 {
            $tracker = $subdomains[2] . "." . $subdomains[3];
        }
    }
    $tracker =~ s/co_uk/co\.uk/g;
    $dbh->do("INSERT INTO trackers VALUES ('$tracker', '$group')");
}

undef $priv_choice;
undef @track_headers;

# -----
print "[3/8] reading liveHTTPheaders...\n";

# -----
my @live_headers = split /\r\n/, $http_headers;
my $target;
my $domainsource;
my $domaintarget;
my $sameframe = 0;
my $cnt_headers;

foreach (@live_headers) {
    $target = $_;

    if ( $target =~ m/Host:/ ) {
        $target =~ s/Host: //g;
        $target =~ s/co\.uk/co_uk/g;
        my @subdomains = split /\./, $target;

        my $test_ipadres = $subdomains[0] . ".";

        switch ( scalar @subdomains ) {
            case 2 { $domaintarget = $subdomains[0] . "." . $subdomains[1] }
            case 3 { $domaintarget = $subdomains[1] . "." . $subdomains[2] }
            case 4 {
                if ( $test_ipadres =~ m/\b\d{1,3}\./ ) {

                    $domaintarget =
                        $subdomains[0] . "." . $subdomains[1] . ".x.x";

                }
                else {

                    $domaintarget = $subdomains[2] . "." . $subdomains[3];

                }

            }

        }
        $sameframe = 0;
    }

    if ( $target =~ m/Referer:/ ) {
        $target =~ s/Referer: //g;
    }
}

```

```

my $url      = URI->new("$target");
my $domain   = $url->host;

$domain =~ s/co\.uk/co_uk/g;
my @subdomains = split /\./, $domain;

my $test_ipadres = $subdomains[0] . ".";

switch ( scalar @subdomains ) {
  case 2 { $domainsource = $subdomains[0] . "." . $subdomains[1] }
  case 3 { $domainsource = $subdomains[1] . "." . $subdomains[2] }
  case 4 {
    if ( $test_ipadres =~ m/\b\d{1,3}\./ ) {

      $domainsource =
        $subdomains[0] . "." . $subdomains[1] . ".x.x";

    }
    else {

      $domainsource = $subdomains[2] . "." . $subdomains[3];

    }
  }
}

$sameframe = 1;

if ( $target =~ m/-----/ ) {
  $cnt_headers++;
  switch ($sameframe) {
    case 0 {
      $domaintarget =~ s/co_uk/co\.uk/g;
      $dbh->do(
"INSERT INTO domains VALUES ('$domaintarget', '$domaintarget')"
      );
    }
    case 1 {
      $domainsource =~ s/co_uk/co\.uk/g;
      $domaintarget =~ s/co_uk/co\.uk/g;
      $dbh->do(
"INSERT INTO domains VALUES ('$domainsource', '$domaintarget')"
      );
    }
  }
  $sameframe = 0;
}

undef $http_headers;
undef @live_headers;

# distinct domains
my $sth = $dbh->prepare(
"select distinct source_domain, target_domain from domains order by domains.target_domain"
);
$sth->execute;
my $cnt_domains;

```

```

while ( ( my $source, my $target ) = $sth->fetchrow_array() ) {
    $dbh->do(
        "INSERT INTO distinct_domains VALUES (NULL, '$target', '$source', NULL, NULL, NULL)"
    );
    $dbh->do("INSERT INTO names VALUES (NULL, '$source')");
    $dbh->do("INSERT INTO names VALUES (NULL, '$target')");
    $cnt_domains++;
}
$sth->finish();

# distinct names
$sth = $dbh->prepare("select distinct domain from names order by names.domain");
$sth->execute;
my $cnt_names;
while ( ( my $domain ) = $sth->fetchrow_array() ) {
    $dbh->do("INSERT INTO distinct_names VALUES (NULL, '$domain', NULL)");
    $cnt_names++;
}
$sth->finish();

# distinct trackers
$sth = $dbh->prepare(
    "select distinct tracker, groups from trackers order by trackers.tracker");
$sth->execute;
while ( ( my $tracker, my $groups ) = $sth->fetchrow_array() ) {
    $dbh->do("INSERT INTO distinct_groups VALUES ('$tracker', '$groups')");
}
$sth->finish();

# -----
print "[4/8] linking nodes...\n";

# -----
# lookup distinct_names.groups in tabel distinct_groups
my @correct_group;
my $id;
$sth = $dbh->prepare("select * from distinct_names");
$sth->execute;
while ( ( $id, my $domain, my $groups1 ) = $sth->fetchrow_array() ) {
    my $sti = $dbh->prepare(
        "select groups from distinct_groups where tracker='$domain'");
    $sti->execute;
    if ( ( my $groups2 ) = $sti->fetchrow_array() ) {
        $correct_group[$id] = $groups2;
    }
    else {
        $correct_group[$id] = '2';    # other non confirmed trackers
    }
    $sti->finish();
}
$sth->finish();
for ( $id = 1 ; $id <= $cnt_names ; $id++ ) {
    $sth = $dbh->prepare(
        "UPDATE distinct_names SET groups='$correct_group[$id]' WHERE id='$id'"
    );
    $sth->execute;
    $sth->finish();
}

```



```

}

# lookup distinct_domains.source_node in tabel distinct_names
my @correct_source;
$sth = $dbh->prepare("select id, source_domain from distinct_domains");
$sth->execute;
while ( ( $id, my $source_domain ) = $sth->fetchrow_array() ) {
    my $sti = $dbh->prepare(
        "select id from distinct_names where domain='$source_domain'");
    $sti->execute;
    $correct_source[$id] = $sti->fetchrow_array();
    $correct_source[$id]--;
    $sti->finish();
}
$sth->finish();
for ( $id = 1 ; $id <= $cnt_domains ; $id++ ) {
    $sth = $dbh->prepare(
        "UPDATE distinct_domains SET source_node='$correct_source[$id]' WHERE id='$id'"
    );
    $sth->execute;
    $sth->finish();
}

# lookup distinct_domains.target_node in tabel distinct_names
my @correct_target;
$sth = $dbh->prepare("select id, target_domain from distinct_domains");
$sth->execute;
while ( ( $id, my $target_domain ) = $sth->fetchrow_array() ) {
    my $sti = $dbh->prepare(
        "select id from distinct_names where domain='$target_domain'");
    $sti->execute;
    $correct_target[$id] = $sti->fetchrow_array();
    $correct_target[$id]--;
    $sti->finish();
}
$sth->finish();
for ( $id = 1 ; $id <= $cnt_domains ; $id++ ) {
    $sth = $dbh->prepare(
        "UPDATE distinct_domains SET target_node='$correct_target[$id]' WHERE id='$id'"
    );
    $sth->execute;
    $sth->finish();
}

# lookup value
my @correct_value;
$sth = $dbh->prepare("select id, source_domain from distinct_domains");
$sth->execute;
while ( ( $id, my $source_domain ) = $sth->fetchrow_array() ) {
    my $sti = $dbh->prepare(
        "select groups from distinct_names where domain='$source_domain'");
    $sti->execute;
    $correct_value[$id] = $sti->fetchrow_array();
    $sti->finish();
}
$sth->finish();
for ( $id = 1 ; $id <= $cnt_domains ; $id++ ) {

```

```

my $calculation = 1 + $correct_value[$id] * 3;
$sth = $dbh->prepare(
    "UPDATE distinct_domains SET tot_value='$calculation' WHERE id='$id'");
$sth->execute;
$sth->finish();
}

# -----
print "[5/8] writing $ouputfile\n";

# -----
# create nodes
my $json = "{ \"nodes\": [";

$sth = $dbh->prepare("select * from distinct_names order by distinct_names.id");
$sth->execute;
while ( ( $id, my $tracker, my $grouping ) = $sth->fetchrow_array() ) {
    if ( $id == 1 ) {
        $json = $json . "{ \"name\": \"$tracker\", \"group\": $grouping }";
    }
    else {
        $json = $json . ", { \"name\": \"$tracker\", \"group\": $grouping }";
    }
}
$json = $json . " ]";
$sth->finish();

# create links
$json = $json . ", \"links\": [";
my $first_entry = 1;

$sth = $dbh->prepare(
    "select id, source_node, target_node, tot_value from distinct_domains");
$sth->execute;

while ( ( $id, my $source, my $target, my $tot_value ) =
    $sth->fetchrow_array() )
{
    if ( $first_entry == 1 ) {
        $json = $json
            . "{ \"source\": $source, \"target\": $target, \"value\": $tot_value }";
        $first_entry--;
    }
    else {
        $json = $json
            . ", { \"source\": $source, \"target\": $target, \"value\": $tot_value }";
    }
}
$json = $json . " ] }";
$sth->finish();
writeFile( $ouputfile, $json );

# query confirmed trackers
my $cnt_tracker;
$sth = $dbh->prepare(
    "select COUNT(groups) from distinct_names WHERE distinct_names.groups='1'");
$sth->execute;

```

```
$cnt_tracker = $sth->fetchrow_array();
$sth->finish();

# -----
print "[6/8] closing sqlite database\n";

# -----
$dbh->disconnect();

my $proc = time - $start;

if ($proc) {
    $log = $log . "\r\n$cnt_headers headers\r\n";
    $log = $log . "$cnt_names nodes\r\n";
    $log = $log . "$cnt_domains links\r\n";
    $log = $log . "$cnt_tracker confirmed trackers\r\n\r\n";
    $log = $log . "job completed succesfully in $proc seconds\r\n\r\n";
}

print "[7/8] writing log file\n";
writeFile( $logfile, $log );

print "[8/8] archiving tables...Done!\r\n\r\n";
exec("echo '.dump' | sqlite3 connections.sqlite | gzip -c >$dumpfile") || die "Could not
dump connections.sqlite to file '$file'";

exit 0;

# -----

sub readFile {
    my $file = shift;

    open( local *FILE, "<", $file ) || die "Could not read file '$file'";
    binmode(FILE);
    local $/;
    my $result = <FILE>;
    close(FILE);

    return $result;
}

sub writeFile {
    my ( $file, $contents ) = @_;

    open( local *FILE, ">", $file ) || die "Could not write file '$file'";
    binmode(FILE);
    print FILE $contents;
    close(FILE);
}

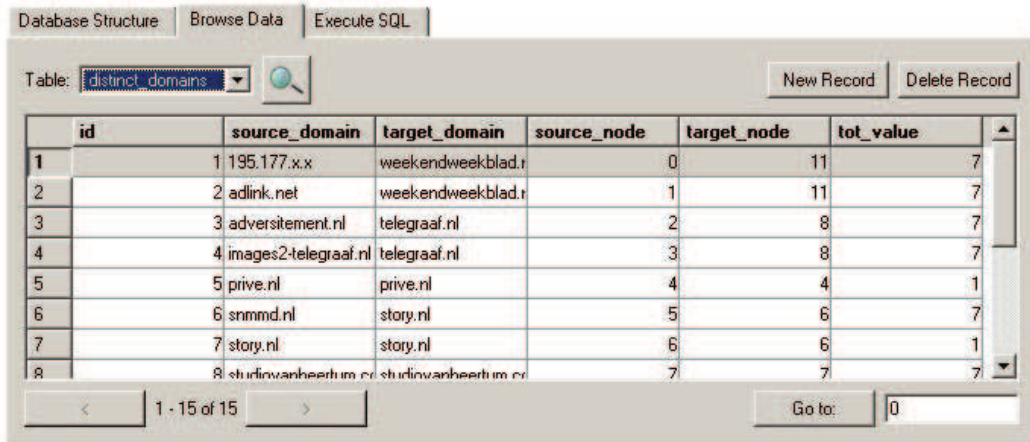
# -----
# Version: MPL 1.1/GPL 2.0/LGPL 2.1
#
# The contents of this file are subject to the Mozilla Public License Version
# 1.1 (the "License"); you may not use this file except in compliance with
# the License. You may obtain a copy of the License at
```

```
# http://www.mozilla.org/MPL/
#
# Software distributed under the License is distributed on an "AS IS" basis,
# WITHOUT WARRANTY OF ANY KIND, either express or implied. See the License
# for the specific language governing rights and limitations under the
# License.
#
# The Initial Developer of the Original Code is Rob van Eijk.
#
# Portions created by the Initial Developer are Copyright (C) 2010
# the Initial Developer. All Rights Reserved.
#
# Contributor(s):
#
# Alternatively, the contents of this file may be used under the terms of
# either the GNU General Public License Version 2 or later (the "GPL"), or
# the GNU Lesser General Public License Version 2.1 or later (the "LGPL"),
# in which case the provisions of the GPL or the LGPL are applicable instead
# of those above. If you wish to allow use of your version of this file only
# under the terms of either the GPL or the LGPL, and not to allow others to
# use your version of this file under the terms of the MPL, indicate your
# decision by deleting the provisions above and replace them with the notice
# and other provisions required by the GPL or the LGPL. If you do not delete
# the provisions above, a recipient may use your version of this file under
# the terms of any one of the MPL, the GPL or the LGPL.
# -----
```

## G Source code: visualizing links and nodes with connectivity color maps

In order to color the connectivity color maps, the xlink:href attribute value has been defined as a function, returning different image URLs based on the node data (d): `"".attr(xlink:href, function(d) { return my_nodecolor(d.group); })`. The `my_nodecolor` function handles the different colored bullets:

```
function my_nodecolor(a){
  switch(a)
  {
    case 0:
      return /bullet-green.png;
    break;
    (...)
  }
}
```



id	source_domain	target_domain	source_node	target_node	tot_value
1	195.177.x.x	weekendweekblad.nl	0	11	7
2	adlink.net	weekendweekblad.nl	1	11	7
3	adversitement.nl	telegraaf.nl	2	8	7
4	images2.telegraaf.nl	telegraaf.nl	3	8	7
5	prive.nl	prive.nl	4	4	1
6	snmmd.nl	story.nl	5	6	7
7	story.nl	story.nl	6	6	1
8	studiorvanheertum.nl	studiorvanheertum.nl	7	7	7

Figure 21: Data for a connectivity color map

Special care has to be taken in the order of nodes and groups. This is best handled in tables, which is the main reason to do the processing in SQL.

The number in column "tot\_value" is the cause for the evenly spreading out of the network map. The higher the value the stronger it's influence on the spreading pattern.

```

<!DOCTYPE html>
<html>
<head>
<script type="text/javascript" src="http://www.blaeu.com/d3.js"></script>
<script type="text/javascript" src="http://www.blaeu.com/d3.layout.js"></script>
<script type="text/javascript" src="http://www.blaeu.com/d3.geom.js"></script>
<style type="text/css">
.link { stroke: #ccc; }
.nodetext { pointer-events: none; font: 10px sans-serif; }
</style>
</head>
<body>
<script type="text/javascript">
function my_nodecolor(a) {
switch(a)
{
case 0:
return "/bullet-green.png";
break;
case 1:
return "/bullet-purple.png";
break;
case 2:
return "/bullet-blue.png";
break;
default:
return "/bullet-red.png";
}
}

var w = 1200,
    h = 900

var vis = d3.select("body").append("svg:svg")
    .attr("width", w)
    .attr("height", h);

d3.json("../json/eu.belgium.json", function(json) {
    var force = self.force = d3.layout.force()
        .nodes(json.nodes)
        .links(json.links)
        .gravity(.05)
        .distance(100)
        .charge(-100)
        .size([w, h])
        .start();

    var link = vis.selectAll("line.link")
        .data(json.links)
        .enter().append("svg:line")
        .attr("class", "link")
        .attr("x1", function(d) { return d.source.x; })
        .attr("y1", function(d) { return d.source.y; })
        .attr("x2", function(d) { return d.target.x; })
        .attr("y2", function(d) { return d.target.y; });

    var node = vis.selectAll("g.node")

```

```
.data(json.nodes)
.enter().append("svg:g")
  .attr("class", "node")
  .call(force.drag);

node.append("svg:image")
  .attr("class", "circle")
  .attr("xlink:href", function(d) { return my_nodecolor(d.group); })
  .attr("x", "-8px")
  .attr("y", "-8px")
  .attr("width", "16px")
  .attr("height", "16px");

node.append("svg:text")
  .attr("class", "nodetext")
  .attr("dx", 12)
  .attr("dy", ".35em")
  .text(function(d) { return d.name });

force.on("tick", function() {
  link.attr("x1", function(d) { return d.source.x; })
    .attr("y1", function(d) { return d.source.y; })
    .attr("x2", function(d) { return d.target.x; })
    .attr("y2", function(d) { return d.target.y; });

  node.attr("transform", function(d) { return "translate(" + d.x + "," + d.y + ")"; });
});
});
</script>
</body>
</html>
```

