Retrospective Study of Third-party Web Tracking

Tim Wambach and Katharina Bräunlich
Institute for IS Research, University Koblenz-Landau, Universitätsstr. 1, 56070 Koblenz, Germany

Keywords: Web Tracking, Tracker, Privacy, Data Leakage, Data Protection, Web Archive, Security, Retrospective Analysis, History of Web Tracking.

Abstract: Web tracking has seen a remarkable usage increase during the last years. Unfortunately, an overview of how web tracking evolved within the last ~15 years is missing. In this paper we present a retrospective analysis using archived data to quantify the usage and distribution of web tracking and how it changed throughout the last decade. We identify a more than five fold increase in external requests between 2005 and 2014. About half of the analyzed websites have a web tracking based inclusion today (2015). As web tracking is often associated with a risk of privacy loss, we also outline the security implications of monopolized ubiquitous tracking.

1 INTRODUCTION

Being tracked has become a part of modern life. About half the websites we visit include some kind of tracking mechanism ((Gelbmann, 2012)). A disadvantage associated with web tracking is the potential loss of privacy for end users – (Mayer and Mitchell, 2012) illustrate possible risks for consumers. Collecting data about their web surfing behavior can be used for business, marketing, or other purposes. From our point of view, the security implications for organizations and enterprises are still underestimated. For example, let us consider an employee working in a development department that uses the web to gain further information about products they are working on. The browsing history would usually be kept undisclosed, but (Englehardt et al., 2015) shows how an adversary on the web can reconstruct 62-73% of a typical user's browsing history. Information about current development within a company could be revealed by their web activities. The usage of company wide web proxies that aggregate requests do not solve the issue. Hiding internal IP address information or removing HTTP referrers do not prevent web tracking, because a wide range of different technologies exists that implement web tracking by other means.

Monitoring internet routers might be possible for internet providers or intelligence services, but they should not be able to analyze the content of TLS *(Transport Layer Security)* encrypted requests. This does not apply to web tracking mechanisms that create a separate (encrypted or unencrypted) request to the tracking provider and inform about activities. This could also include very detailed usage information about something like mouse movement.

Web tracking must also be considered if privacy protection on the web is intended by an end user. Network layer anonymization (e.g. using a TOR network (TorProject, 2015)) is not effective if the browser assists the user recognition. It is well known that cookies can be used for this. However, a strong cookie policy might not be an effective protection: (Kamkar, 2010) shows how other browser technologies can be used as cookie replacements or circumvent cookie deletion. How these so-called evercookies are used to facilitate web tracking is shown in (Mcdonald et al., 2011).

This makes web tracking a possible threat for both, the privacy of end users and the security of companies. However, web tracking as a privacy threat and a potential data leakage seams to be underestimated in IT security research. One reason could be that there is a wide range of tracking and advertising providers. Due to the variety of different tracking providers, it is not feasible to grasp the whole picture of a specific company or person. Another reason for this underestimation might be the novelty of this threat. 10 years ago third-party web tracking was relatively rare but is now growing into a serious problem.

Currently, no overview on how web tracking has grown over the last 10-16 years exists (cf. Section 5). Our goal is to analyze how third-party web tracking

Wambach, T. and Bräunlich, K.
Retrospective Study of Third-party Web Tracking.
DOI: 10.5220/0005741301380145
In *Proceedings of the 2nd International Conference on Information Systems Security and Privacy (ICISSP 2016)*, pages 138-145
ISBN: 978-989-758-167-0

has changed over the last 16 years. Our methods are explained in Section 2. The required software implementations are described in Section 3. However, the results in Section 4 show a remarkable increase in external connections that can mostly be linked to web tracking. A trend that was already partly shown in other work.

An assumption was, that a diversity of tracking providers exist so that a single tracker might not be able to gain the whole picture of a specific person (as explained above). It is based on the fact that 10 years ago, third-party web tracking was relatively rare compared to today. In this paper we show that few providers currently track the majority of users visiting the most popular websites. This is the first analysis that covers a more than 5 year range and (graphically) demonstrates how this tracking network has grown over the last decade.

2 METHODOLOGY

The questions this paper strives to answer are: How many trackers have exist historically, how the number of trackers has changed throughout the last years, how trackers have been distributed, and how their distribution has changed over time.

Section 2.1 provides a general overview of web tracking mechanisms. Section 2.2 explains how we detect trackers on websites. This method differs from privacy enhancing technologies but fits our special needs for such a retrospective analysis. For such an analysis, we need a large amount of website snapshots over the past years. We use website snapshots from *archive.org* and perform an analysis of web tracking on them – further explained in Section 2.3.

2.1 Tracker Mechanism

At the beginning of the world wide web, website analytics were performed by an analysis of the web server log files. The website operator collected information about the IP addresses of the visitors, possible locations and most requested websites. An analysis, performed by the web server's owner, can be classified as first-party analysis.

Embedding resources is a common technique for web tracking. HTML allows embedding content from local and remote servers. During the parsing and interpretation process, the browser automatically loads content from any location specified in the HTML code. The server notices a resource request from the visitor's browser. A picture (so-called web bug) embedded in a web site allows a third-party to track visitors over different domains. Such a request contains the user's IP address and additional meta information (e.g. HTTP header). Additionally, cookies can facilitate the recognition of visitors.

JavaScript, developed in 1995 and first provided in 1996 by Netscape Navigator 2.0 and Internet Explorer 3.0, introduced the possibility to execute scripts on the client side. This technique can also be used to perform web tracking to gain fine-grained usage information. Data collected on the client-side can afterwards be sent to the server. The *Same-Origin Policy* implemented in browsers since 1996 lead to a separation of data from different websites. To achieve cross domain tracking, a shared third party tracking provider is required.

In active web tracking, the browser is part of the tracking mechanism. Tracking from a third-party that is supported by the browser in any way (e.g. creating remote connections, performing scripts, etc.) is known as active third-party web tracking. Today, different techniques are used to track visitors across domains. Unlike the passive type of web tracking, like the use of browser fingerprinting ((Eckersley, 2010), (Boda et al., 2012)), the active part can be detected. (Mayer and Mitchell, 2012) explains how modern web tracking works.

2.2 Tracker Detection

Most currently available tracker recognition techniques are focused on detecting trackers on existing and active websites. For example by blacklisting known tracker patterns (e.g. Ghostery (Ghostery, 2015)) or by behavioral analysis (e.g. TrackingObserver (TrackingObserver, 2012)).

(Roesner et al., 2012) presents a classification framework and shows five different classes of typical tracker behavior. However, we cannot use existing tracking recognition tools to find tracking mechanisms from *archive.org* websites due to the fact that the web server of the tracker might not exist anymore or might show different behavior as in the past. If a visited website (A) includes an image or script that causes the browser to open a connection to the tracking domain (B), (B) might set a cookie for tracking purposes. This active part – setting a cookie – is not preserved by *archive.org*. But what we can see in the archived data of (A) is the fact that there is an active part (e.g. image, script, etc.) that causes a request by the browser to open a connection to (B).

Embedding content is not necessarily associated with tracking. Examples include embedding a video player or third-party requests caused by the server infrastructure (Content Delivery Network). By creating a new connection from the browser to another domain, a transmission of source/location information (IP address) is caused and may also reveal further protocol specific information like HTTP referrers. (Eckersley, 2010) shows how this kind of information can be used for passive web tracking. From our point of view today, disclosing IP address information must be classified as personal data transfer that can be used for tracking purposes. In this paper, tracking is defined as a connection to an external (third party) host, that is not part of the visited/requested website. It cannot be proven whether the third party uses personal data for tracking purposes or not. However, it is clear that the data could be used to track.

The results are similar to the ones generated by the Firefox add-on Lightbeam ((Lightbeam, 2015)) that also provides a graphical overview about third party connections. Due to the fact that we need the ability to block connections not directed to *archive.org*, further development was necessary. In Section 3 we describe how a development framework (PyQt) was modified to obtain all external requests that occur during a web request. As soon as a website is fully parsed, all network requests are saved in a list and can be processed further.

2.3 Retrospective Analysis

Founded in 1996, *archive.org* became well known as an internet library by preserving the state of popular websites. If not disabled by the website owner, *archive.org* stores the current state of public internet websites several times a year (Day, 2006), (Olston and Najork, 2010). Information about new popular websites are donated by the *Alexa.org* database. This information will be used for a retrospective analysis of the websites with a focus on third-party connections.

We can now obtain all requests for a given *archive.org* website. We also need to restrict our analysis to a set of websites. We decided to use the 10,000 most popular websites according to the *Alexa.org* database (as of March 2015). Unfortunately, *Alexa.org* was not able to provide the most visited websites for the years before 2007. Other databases, like Netcraft or *archive.org*, were not able to provide this either. Therefore, our analysis is based on the most popular 10,000 websites today.

The *archive.org* JSON API[1] allows us to check how many snapshots are available and where they can be found. For each of the 10,000 websites and for each year between the years 2000 and 2015, we request a snapshot overview from *archive.org*. The result is a list of snapshots that can be processed. This processing results in up to 16 lists of resources (for each year) that the browser loads after visiting the website in the archive. Finally, we perform an analysis of what kind of trackers were used historically and how tracker usage changed in the last years.

As already stated in Section 2.2, an ideal analysis of web tracking cannot be performed due to the fact that active parts (web servers) sometimes do not exist anymore or do not show the same behavior. Furthermore, redirections to content on other websites cannot be followed if they are not preserved by *archive.org*. For example if an advertising spot was sold by the website owner and filled with different content for each request. This could generate much more external requests if visited multiple times. Due to this, the results of this analysis must be interpreted as a minimum of tracking, but could be higher.

3 IMPLEMENTATION

For our analysis, it is necessary to identify external connections from an archived website. A static analysis, like using regular expression to find external resources within HTML source code, has been shown to be insufficient. A reason for this is code obfuscation that looks like this:

```
var src = (document.location.protocol ===
’https:’ ? ’https:/’ : ’http:/’)
+ ’/imagesrv.adition.com/js/srp.js’;

document.write(’<scr’ + ’ipt type=
"text/javascript" src="’ + src + ’"
charset="utf-8"></scr’ + ’ipt>’);
```

In this code, the address of the tracker (adition.com) is obfuscated in a simple form, but good enough to defeat an automatic URL search in the source code.

Thus, a more dynamic analysis of websites is required. PyQt is a library that connects the Qt C++ cross-platform application framework with the interpreted language Python. Qt is a toolkit that includes a web browser widget that supports all modern web techniques (JavaScript, CSS, AJAX etc.) according to their whitepaper (Riverbank, 2013). When this browser widget is parsing a website, there are various points where resources (images, scripts, etc.) are requested. We identified the `PyQt4.QtNetwork.QNetworkAccessManager` class where all network-based requests come together. If a resource must be loaded, the method `createRequest` is called and contains the full address (URL) of the resource. We overwrote this class, so that:

[1] JSON API for *archive.org* services and metadata, https://archive.org/help/json.php

- all requests on the network are stored permanently, and
- requests not directed to *archive.org* are blocked using an empty dummy request.

A dummy request allows the library to continue the parsing process of the website without this requested/blocked resource – so this approach can unfortunately not reveal requests resulting from resources (e.g. scripts) that are not available any more.

Using this modified library, we obtain a list of all connections that are created by the browser during the parsing process. If the requested resource is available on *archive.org*, it is loaded and processed. All other requests outside the archive are blocked. It must be clarified that only the main website was initially requested – the browser does not try to ”click” on hyperlinks if it is not forced to do so by the content. However, we assume that most trackers will be loaded by the main page.

In (Krishnamurthy and Wills, 2009), third-party domains are identified by their DNS record. Due to the fact that we do not have DNS records from the last 16 years, this method cannot be applied here. Therefore, we define a tracker as a loading process from an external domain. We define an external domain as a host, where the second-level domain differs from the requested. For example, if ”example.com” is requested, ”web.example.com” and ”example.net” count as internal resources, while ”notexample.com” counts as external. In Table 1 we provide an overview of request types and how they are handled. Each external domain is counted once. For external domains, we also take the top level domain into account, so that ”facebook.com” and ”facebook.net” are different resources.

For graphical representation, we use networkx (Hagberg et al., 2008). All data was obtained in April 2015.

4 RESULTS

For our analysis, we determined how many years of history of the Alexa Top 10,000 Domains (as of April 2015) we can find on *archive.org*. For 896 domains, no history is available – website owners are able to block[2] being archived. For 3,042 domains, *archive.org* only provides a ≤ 5 year history. For the following analysis, a history of ≥ 10 years must be available, which includes 4,833 domains. How the snapshots are distributed over the 10-16 years is not relevant for our analysis. For 1,426 domains, we have the full 16 year history. In 123 cases, we had browser errors and removed these domains from the analysis. In conclusion, we identified 4,710 domains with at least a 10 year history for further analysis.

Table 2 shows an overview of the results. The second column (column #) shows the number of analyzed websites that were available for a particular year. The columns x_{min} and x_{max} show the minimum and maximum number of external requests from a single website. $\bar{x}_{med}$ is the median, $x_{Q0.25}$ is the first and $x_{Q0.75}$ the third quartile. Column $\bar{x}$ is the sample mean, σ is the standard deviation, and $\sigma^2 = Var(x)$.

During the years 2000 to 2004, the number of analyzed websites is below 4,000 and therefore difficult to compare with the rest. As of the year 2005, the number is between 4,000 and 4,500. Due to the fact that the analysis was executed in the beginning of 2015, not all websites were available at this time for the year 2015. Therefore, the best range for a comparison is between 2005 and 2014. Within this range the number of web inclusions increased more than five fold. The median in 2015 shows that half of the websites have at least 5 different external resources included and a quarter at least 2. The fact that in the year 2005 75% have no inclusions lead us to the question how these external resources are distributed.

4.1 Most Used Trackers

In this section, we show which domains are most prominently included throughout our 16 years of analysis. For each year, we analyzed how often a domain showed up in our results. The five most used domains can be seen in Table 3. For the year 2000, the most included domain was *doubleclick.net* which is still a popular web tracking domain. But in our 2,677 analyzed websites it only occurred 177 times. In comparison: the most included domain in 2014 was *google-analytics.com* with 2,094 occurrences on 4,274 analyzed domains.

It needs to be noted that domains are not grouped by company. For example, *Akamai Technologies* is still in the top 20 of the year 2015 with 346 occurrences, but currently known under the domain *akamaihd.net*.

As explained in the previous section, the range most suitable for comparison are the years 2005 to 2014. In 2005, *doubleclick.net* is the most included domain but with 209 of 4,049 websites it is only present in about 5% of the websites. In the year 2014, *google-analytics.com* is the most included domain with 2,094 occurrences. On nearly every second website, the visitor is tracked by *Google Analyt-*

[2] Removing Documents From the Wayback Machine, https://archive.org/about/exclude.php

Table 1: Overview of the request types. Protocol: http or https.

Type	Request	Counted	Blocked
I	`web.archive.org/web/<time>/<int domain>/<resource>`	no	no
II	`web.archive.org/web/<time>/<ext domain>/<resource>`	yes	no
III	`<int domain>/<resource>`	no	yes
IV	`<ext domain>/<resource>`	yes	yes

Table 2: Statistical report of external requests on websites for the years 2000 to 2015.

Year	#	x_{min}	$x_{Q0.25}$	$\bar{x}_{med}$	$x_{Q0.75}$	x_{max}	$\bar{x}$	σ	$Var(x)$
2000	2677	0	0	0	1	22	0.71	1.54	2.37
2001	3117	0	0	0	1	22	0.86	1.69	2.87
2002	3277	0	0	0	1	22	0.88	1.64	2.7
2003	3534	0	0	0	1	17	0.91	1.66	2.74
2004	3861	0	0	0	1	30	15.0	1.99	3.98
2005	4049	0	0	0	2	42	1.17	2.11	4.47
2006	4251	0	0	1	2	37	1.42	2.23	4.97
2007	4327	0	0	1	3	26	1.88	2.48	6.16
2008	4457	0	0	1	3	29	2.13	2.69	7.24
2009	4328	0	0	2	4	26	2.6	2.95	8.7
2010	4328	0	1	2	5	31	3.1	3.34	11.14
2011	4412	0	1	3	6	46	42.0	4.15	17.18
2012	4414	0	1	4	7	52	4.8	4.74	22.43
2013	4410	0	1	4	8	39	5.24	4.92	24.25
2014	4274	0	2	5	9	47	5.91	5.37	28.86
2015	4031	0	2	5	9	52	6.15	5.53	30.6

Table 3: Top 5 of the most included external domains. Legend: A: doubleclick.net, B: akamai.net, C: rambler.ru, D: bfast.com, E: imgis.com, F: spylog.com, G: list.ru, H: imrworldwide.com, I: googlesyndication.com, K: google-analytics.com, L: google.com, M: quantserve.com, N: scorecardresearch.com, O: facebook.com, P: googleapis.com, R: fbcdn.net, S: facebook.net.

	’00	’01	’02	’03	’04	’05	’06	’07	’08	’09	’10	’11	’12	’13	’14	’15
A	177	210	205	183	190	209	224	366	454	880	787	769	614	810	1202	1050
B	46	93	91	77	68	71	71	58	43	31	38	29	26	26	23	21
C	39	51	59	67	77	71	84	93	91	96	90	90	75	60	48	44
D	31															
E	31															
F		49	39	36	37	30	33	25	24	22						
G		42	47	53	56	57	69	67	62	48	39	31	26	18		
H				47	67	75	81	83	107	99	104	213	223	199	162	134
I					118	206	322	364	356	342	326	318	290	276	280	310
K						18	563	1182	1704	2001	2175	2394	2530	2417	2094	1971
L				16	21	43	64	651	146	219	266	468	742	769	805	735
M								44	223	388	422	436	448	418	368	328
N										72	304	462	549	567	602	588
O										38	254	667	698	583	510	427
P									18	119	273	543	763	1003	1170	1179
R										35	107	542	595	463	255	36
S											65	392	701	804	797	837

ics. Due to the fact that since 2007, *doubleclick.net* is part of *Google Inc.*, 4 of these top 5 included domains have been part of a *Google Inc.* service for the last years (since 2011). It is still unclear how data from Google services like *googleapis.com* are connected to other Google services.

A further analysis of the drop in requests (Table 3) between 2013 and 2014 of *Google Analytics* showed that 209 of them changed to *doubleclick.net* and 92 of these domains that removed *Google-Analytics* in 2013 added *googletagmanager.com* in 2014. So this could be part of a Google internal separation process. Further analysis of tracker removal can be found in Section 4.3.

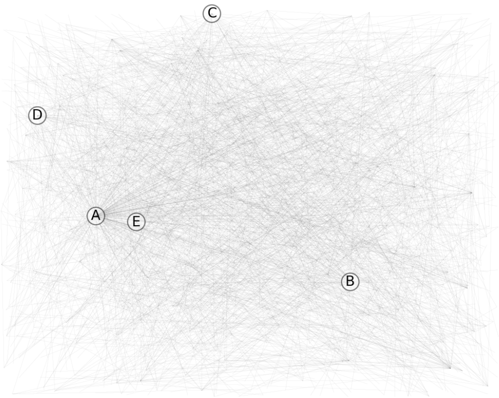

Figure 1: Year 2000.

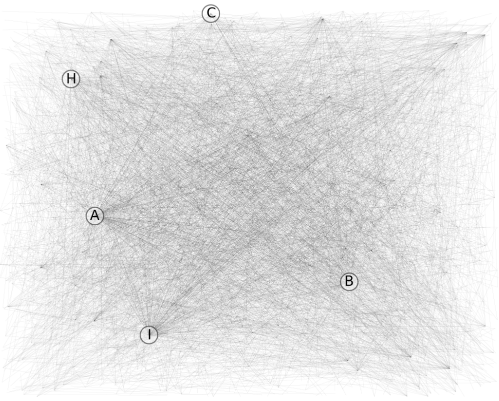

Figure 2: Year 2005.

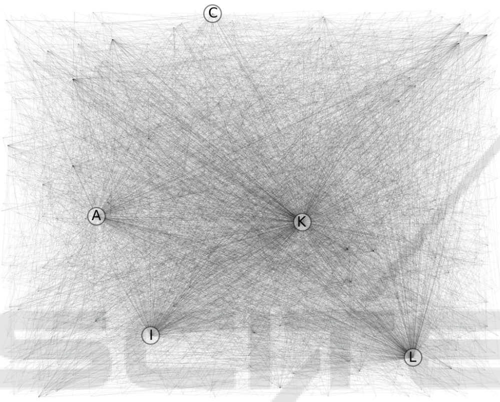

Figure 3: Year 2007.

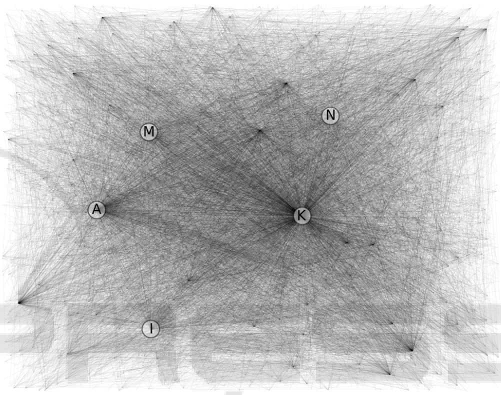

Figure 4: Year 2010.

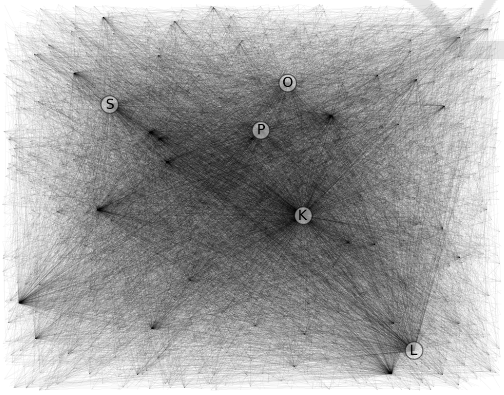

Figure 5: Year 2012.

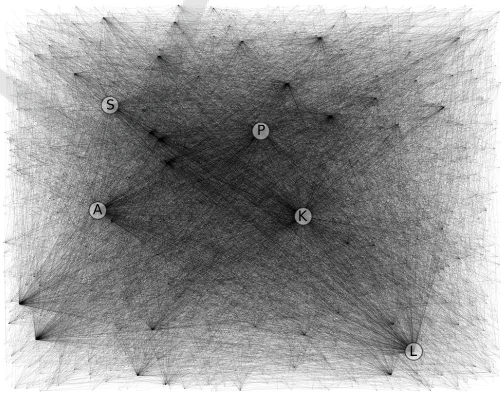

Figure 6: Year 2014.

4.2 Distribution

In this section we use a graph for a visual representation of the distribution of trackers. Figures 1 - 6 show undirected graphs where every edge stands for at least one load process from an external domain. The top 5 from Section 4.1 are marked with the corresponding letter as explained in Table 3. For the sake of clarity, nodes without any connection are not plotted. The position of each domain (node) is consistent for all graphs.

It must be noted that for 2005, nearly the same amount of websites were analyzed (4,049) as for 2014 (4,274). Thus, the differences between these graphs are mostly additional requests. In Figures 4 and 5 we can see a monopolization: while in earlier years we had a more equal distribution, the graph from 2010 shows few nodes with significantly more connections than the others. Especially *google-analytics.com* (K) is in the top 5 since 2006 and its growing influence on the web is easy to see.

4.3 Tracker Removal

As we can see in Table 3, some trackers were removed. In this section, we analyze what happened with the websites that removed a specific tracker. As one of the most included resources since 2006, we start this analysis with Google Analytics (GA). We checked how many websites used GA and removed it later. The analysis showed that 869 domains used GA and removed it. Between the years 2006 and 2011, on less than 50 domains per year, GA was removed. The highest amount was in the years 2013, 2014 and 2015 with 225, 369 and 239 removals respectively. In only a few cases (about 25 per year) the removal of GA resulted in an empty list of trackers. In 258 cases where GA was removed, *doubleclick.net* was added the year after or later.

5 RELATED WORK

Using data from *archive.org* in connection with the Alexa top 10,000 was also applied in (Stamm et al., 2010) for an analysis of web application vulnerabilities. Analysis of archived data in general is also performed by the LAWA project (Longitudinal Analytics of Web Archive data) (Spaniol and Weikum, 2012).

In (Krishnamurthy and Wills, 2009), a long-term study over 5 epochs between October 2005 and September 2008 was performed. The results showed a steadily decreasing number of entities where a handful of companies are able to track users' movement across almost all of the popular websites. Our results also show that *Google-Analytics* and *Doubleclick* are the most widely used trackers. (Chaabane et al., 2012) shows an analysis of the Alexa Top 10,000 regarding tracking mechanisms and social media plugins but is limited to October 2012.

A fully comparable overview of the history of web tracking is missing and existing analysis cover only a few years. Our analysis showed the trend of less diversity found in (Krishnamurthy and Wills, 2009) continued.

6 CONCLUSION

In our work, we showed how embedding external content has seen a usage increase in the 10,000 most popular websites today. Within the best comparable range between 2005 and 2014 we have shown a significant increase (five fold) of external requests. In the year 2014 we found an average of 5.9 external requests per website. This means at least 5 other hosts were informed about the visitation of a website. The most used external hosts could be connected with web tracking and so the request could be used to deduce even more information about a user, e.g. an analysis of the user's behavior. We presented an impression of the distribution and diversity in our graphs.

We showed that the diversity of web tracking and content providers, that we assumed in the beginning, does not exist. This lead to security issues: if a globally acting company has an insight into every second visited domain of a specific user, it is clear that keeping company information secret is difficult. The usage of further services like search, mail, calendar, or translation services contributes to the problem. From our point of view, information security officers should be more sensitive about privacy and web tracking for two reasons: for employee privacy and for an effective protection of corporate information. The fact that fewer companies collecting more information about us with an upward tendency is a privacy concern for end users. The consequences of such a centralization of tracking, which allows the formation of an increasingly complete picture about a specific person, are difficult to predict.

We showed that a reason for the underestimation of third-party web tracking consequences could be due to the fact that it did not exist 10 years ago. From a security point of view, considering web tracking and the usage of PET (Privacy-Enhancing Technologies) should be a part of every corporate security policy. This is more important today than it was in the past.

Further research in security implications of web tracking is required. A further study about the motivation of web tracking usage to explain this increase could be performed in future work. With respect to the implementation, a deeper analysis of the embedded content would also be beneficial to increase our understanding of modern web tracking today.

ACKNOWLEDGEMENT

This study has been made possible by the participation of the interdisciplinary project ”Transformations of Privacy”[3] funded by VolkswagenStiftung[4].

REFERENCES

Boda, K., Földes, d., Gulyás, G., and Imre, S. (2012). User tracking on the web via cross-browser fingerprinting. In Laud, P., editor, *Information Security Technology for Applications*, volume 7161 of *Lecture Notes in Computer Science*, pages 31–46. Springer Berlin Heidelberg.

Chaabane, A., Kaafar, M. A., and Boreli, R. (2012). Big friend is watching you: Analyzing online social networks tracking capabilities. In *Proceedings of the 2012 ACM Workshop on Workshop on Online Social Networks*, WOSN ’12, pages 7–12, New York, NY, USA. ACM.

Day, M. (2006). The long-term preservation of web content. In *Web Archiving*, pages 177–199. Springer Berlin Heidelberg.

Eckersley, P. (2010). How unique is your web browser? In *Proceedings of the 10th International Conference on Privacy Enhancing Technologies*, PETS’10, pages 1–18, Berlin, Heidelberg. Springer-Verlag.

Englehardt, S., Reisman, D., Eubank, C., Zimmerman, P., Mayer, J., Narayanan, A., and Felten, E. W. (2015). Cookies that give you away: The surveillance implications of web tracking. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, pages 289–299.

Gelbmann, M. (2012). Google can’t track every single click of your web surfing. only most of them.

Ghostery, I. (2015). Ghostery - home page. https://www.ghostery.com/.

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA.

Kamkar, S. (2010). evercookie - virtually irrevocable persistent cookies.

Krishnamurthy, B. and Wills, C. E. (2009). Privacy diffusion on the web: A longitudinal perspective. In *In Procs World Wide Web Conference*, page 09.

Lightbeam (2015). Lightbeam addon for firefox. https://addons.mozilla.org/en-US/firefox/addon/lightbeam/.

Mayer, J. R. and Mitchell, J. C. (2012). Third-party web tracking: Policy and technology. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, SP ’12, pages 413–427, Washington, DC, USA. IEEE Computer Society.

Mcdonald, A. M., Cranor, L. F., Mcdonald, A. M., and Cranor, L. F. (2011). A survey of the use of adobe flash local shared objects to respawn http cookies.

Olston, C. and Najork, M. (2010). Web crawling. *Found. Trends Inf. Retr.*, 4(3):175–246.

Riverbank (2013). PyQt Whitepaper. http://www.riverbankcomputing.com/.

Roesner, F., Kohno, T., and Wetherall, D. (2012). Detecting and defending against third-party tracking on the web. In *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*, NSDI’12, pages 12–12, Berkeley, CA, USA. USENIX Association.

Spaniol, M. and Weikum, G. (2012). Tracking entities in web archives: The lawa project.

Stamm, S., Sterne, B., and Markham, G. (2010). Reining in the web with content security policy. In *Proceedings of the 19th International Conference on World Wide Web*, WWW ’10, pages 921–930, New York, NY, USA. ACM.

TorProject (2015). Tor: An anonymous Internet communication system. http://www.torproject.org/.

TrackingObserver (2012). A browser-based web tracking detection platform. http://trackingobserver.cs.washington.edu/.

[3] http://www.strukturwandeldesprivaten.de/

[4] https://www.volkswagenstiftung.de/en.html