Exercise Sheet 0x00

PSI-AdvaSP-M: Advanced Security and Privacy

Privacy and Security in Information Systems Group

By Isabell Sailer (1863490),
Tobias Schwartz (1738195),
Barbara Hoffmann (1759786),
Sascha Riechel (1740803)

Design of our experiment

Which websites should be crawled? Random sites? If yes, how do we specify what a random site is? Restriction to sites that fulfil certain criteria? Which?

We decided to restrict the research area and only crawl websites that belong to a certain topic. The topic we chose is insurances. The reason for choosing only one field of reference is to guarantee the comparability of the results. When comparing an important website like Facebook with thousands of clicks per minute to a blog with only two entries and two clicks in the last year there is no real point of measurement given. We took insurances as a field of reference because they store lots of sensitive data. Not only for this data pool there should be a high protection but also the insurances websites should provide cover for the visitors' data, e. g. send requests to Google Analytics only with an anonymized IP address. As a basis we took Wikipedia's top list of German insurances.

Our list of at least 10 (we will use 11) websites:

```
https://de.wikipedia.org/wiki/Liste_der_gr%C3%B6%C3%9Ften_Versicherungen_in_Deutschland_nach_Beitragseinnahmen_im_Jahr_2009

www.allianz.com
www.munichre.com
www.talanx.com
www.generali.de
www.ruv.de
www.axa.com
www.debeka.de
www.vkb.de
www.zurich.de
www.signal-iduna.de
www.huk.de
```

The reasons why we chose these sites are already mentioned above.

Websites may send several requests to Google Analytics to collect statistics. Some anonymize the IP, some not. Handle do we handle this?

Several requests:
The crawler will reveal if there are several requests sent. We will count the number of requests and show them in a statistic.

Partial anonymized IP addresses:
By dint of the crawler we determine whether the website anonymizes the IP address or not. This could be found out by looking at a certain parameter that could be set via JavaScript and by looking the point of time a request is sent to Google Analytics. The results will be shown in a statistic.

Various pieces of information regarding third parties can be collected. Which information do we find interesting?

We are interested in the following aspects regarding included third parties besides Google Analytics:

- Which kind of third parties are included and what is their main aim? (placing advertisement e. g.)
- How many different websites are included as third parties?

Note that including a third party does not necessarily mean that the third party is reachable, so the HTTP request to them can fail. Include all requests or only the successful ones and why?

We decided to include the failed HTTP requests in the statistics to guarantee the completeness. Apart from that no further consideration is made.

Explanation for plots?

We used bar diagrams to easily visualize the extent of usage of different third parties, but also the amount of different HTTP status codes to see how much of a threat is present on the crawled websites.

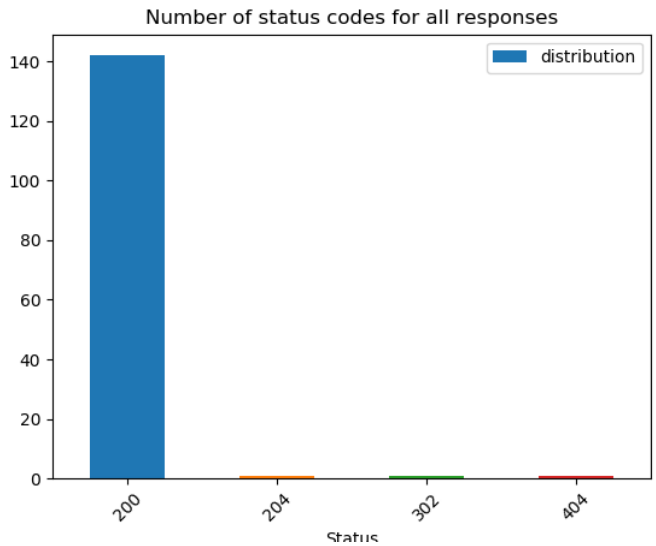
Evaluation of results

AIP Crawler

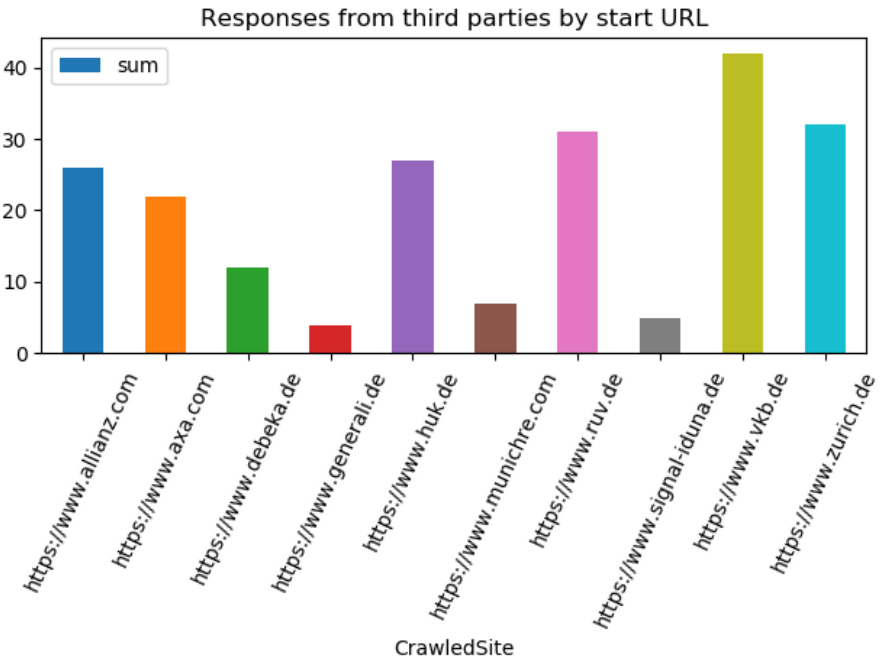
- From all 10 crawled websites, only 5 of them were using Google Analytics on their site
- From these 5 websites, only 3 were using the aip function correctly
- 1 website did not use aip at all!
- 1 website did not use aip correctly, they set a Boolean for aip, not “1”

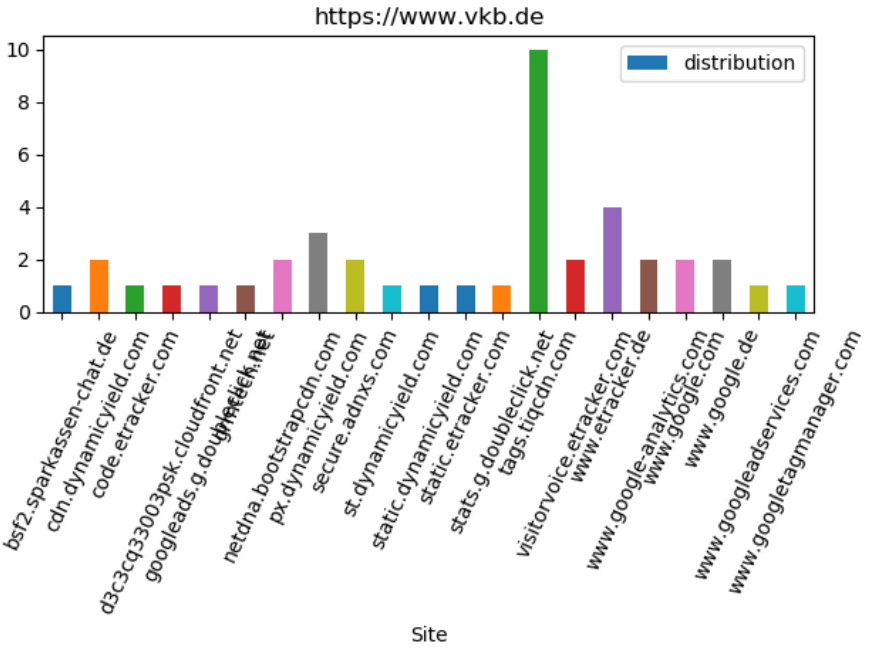
Third party crawler

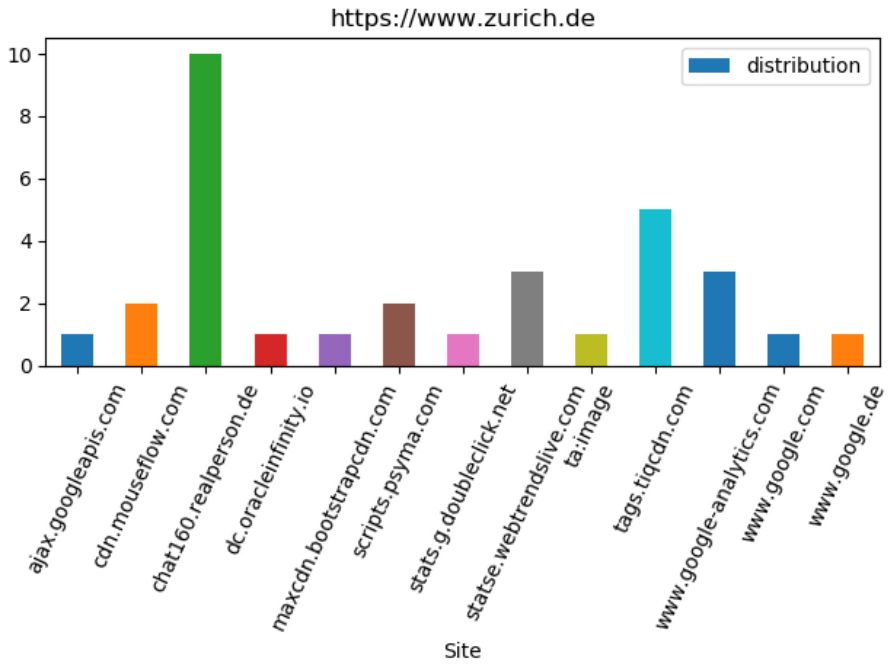
1. Response status
    - From all responses that were received together from all 10 pages, the status code “404” (website not reachable) was only returned once
    - 141 times, the status code was “200”, which says that everything is okay
    - 1 time, the status code was “204” which means that the site was processed correctly, but no content was returned
    - 1 time, the status code was “302” which is a temporary redirect
    - Altogether, only once a security problem arose with the “404” status code

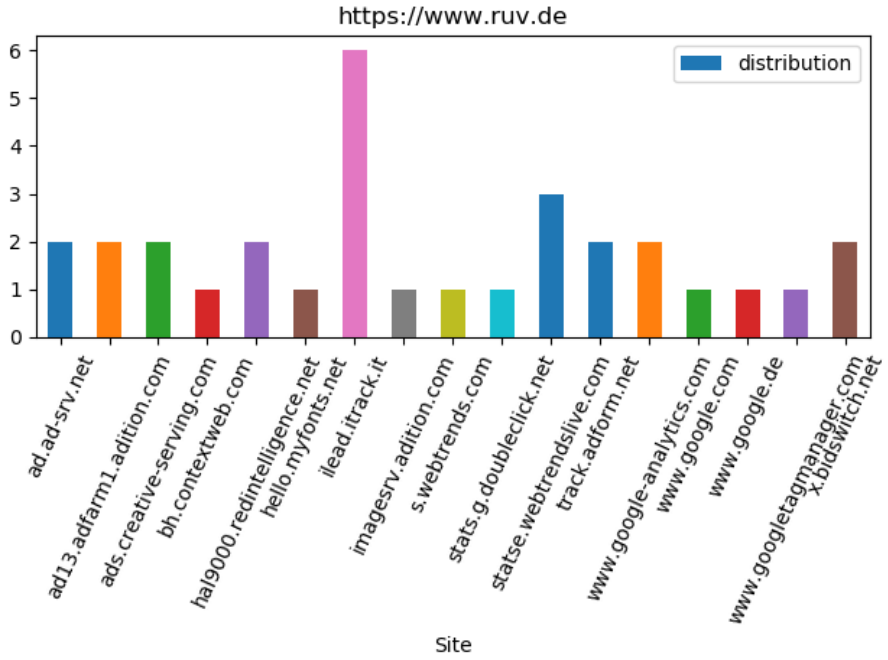


2. Overview of responses per third party
    - It is salient that 3 websites used very little third parties, namely Generali, Munichre and Signal-Iduna
    - The most third-party requests were received while loading the site of VKB, followed by Zurich, RUV and HUK
        i. Upon regarding the third parties of VKB, one can see that VKB uses many different trackers, but the mostly tiqcdn and etracker
        ii. Zurich uses a web chat client, but also tiqcdn like VKB
        iii. RUV also uses a lot of different trackers


Responses from third parties by start URL
sum
40
30
20
10
0
https://www.allianz.com
https://www.axa.com
https://www.debeka.de
https://www.generali.de
https://www.huk.de
https://www.munichre.com
https://www.ruv.de
https://www.signal-iduna.de
https://www.vkb.de
https://www.zurich.de
CrawledSite


https://www.vkb.de
distribution
10
8
6
4
2
0
bsf2.sparkassen-chat.de
cdn.dynamicyield.com
code.etracker.com
d3c3cq33003psk.cloudfront.net
googleads.g.doubleclick.net
netdna.bootstrapcdn.com
px.dynamicyield.com
secure.adnxs.com
st.dynamicyield.com
static.dynamicyield.com
static.etracker.com
stats.g.doubleclick.net
tags.tiqcdn.com
visitorvoice.etracker.com
www.etracker.de
www.google-analytics.com
www.google.com
www.google.de
www.googleadservices.com
www.googletagmanager.com
Site


https://www.zurich.de
distribution
10
8
6
4
2
0
ajax.googleapis.com
cdn.mouseflow.com
chat160.realperson.de
dc.oracleinfinity.io
maxcdn.bootstrapcdn.com
scripts.psyma.com
stats.g.doubleclick.net
statse.webtrendslive.com
ta:image
tags.tiqcdn.com
www.google-analytics.com
www.google.com
www.google.de
Site


https://www.ruv.de
distribution
6
5
4
3
2
1
0
ad.ad-srv.net
ad13.adfarm1.adition.com
ads.creative-serving.com
bh.contextweb.com
hal9000.redintelligence.net
hello.myfonts.net
ilead.itrack.it
imagesrv.adition.com
s.webtrends.com
stats.g.doubleclick.net
statse.webtrendslive.com
track.adform.net
www.google-analytics.com
www.google.com
www.google.de
www.googletagmanager.com
x.bidswitch.net
Site