Exercise Sheet 0x01PSI-AdvaSP-M: Advanced Security and PrivacyPrivacy 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 a 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?
* Which cookies are set on the different websites? Maybe this could lead us to cookie synchronization
* Which versions of browsers are supported by the third parties? Do they support antediluvian versions which are not supported by the browser companies any longer? This could be a lack of security

**Note that including a third party does not necessarily mean that the third party is**

**actually 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.

TODO:

* Ausgabe der Crawler Informationen in einem file
* Zählen der Requests an google analytict (eine, mehrere, wie viele)
* Schauen, ob Google Analytics aufgefordert wird, die IP zu anonymisieren (AIP = 1) und schauen zu welchem Zeitpunkt dieser Parameter gesetzt wird. Wird er nach der GA-Anfrage gesetzt, ist er wirkungslos und die IP wird nicht anonymisiert (siehe dazu Kommentare im Pythonskript, JavaScriptException usw)
* Ergänzung der Dinge, die uns an third parties nebst GA interessieren plus entsprechende Auswertung derer
* Darstellung der Feststellungen in Graphen