1) Distribution of third-party HTTP requests

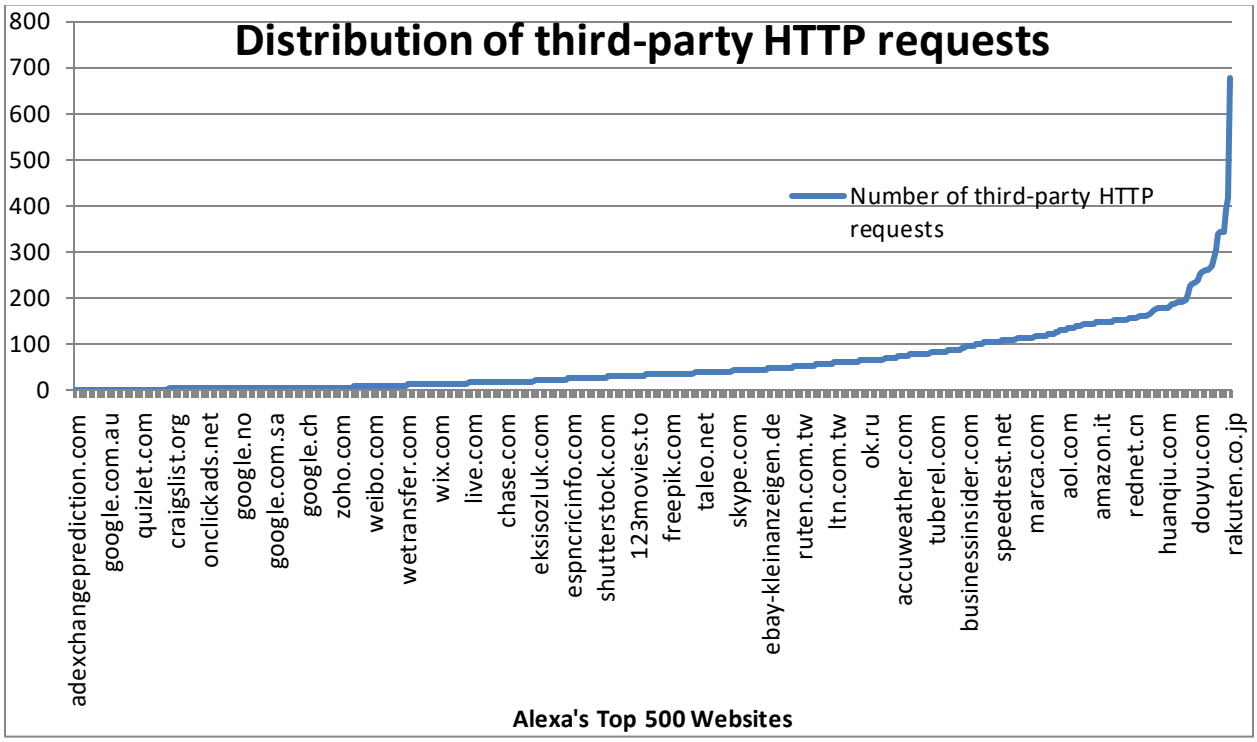

Fig 1. The graph shows distribution of third-party HTTP requests by Alexa's top 500 websites

Top 10 Websites with the most third-party HTTP requests are[1]:

Websites	Number of third-party HTTP requests
rakuten.co	676
azlyrics.com	418
extratorrent.cc	390
cnn.com	344
ettoday.net	342
sabah.com.tr	337
qq.com	301
indiatimes.com	268
sohu.com	262
nytimes.com	260

Table 1: The table shows top 10 websites (from Alexa's top 500 sites) with highest number of third-party http requests.

[1] For complete list please look at dist_tp_request.csv file under project1_1_materials folder

1.2) Most referred third-party websites by Alexa's top 500 websites

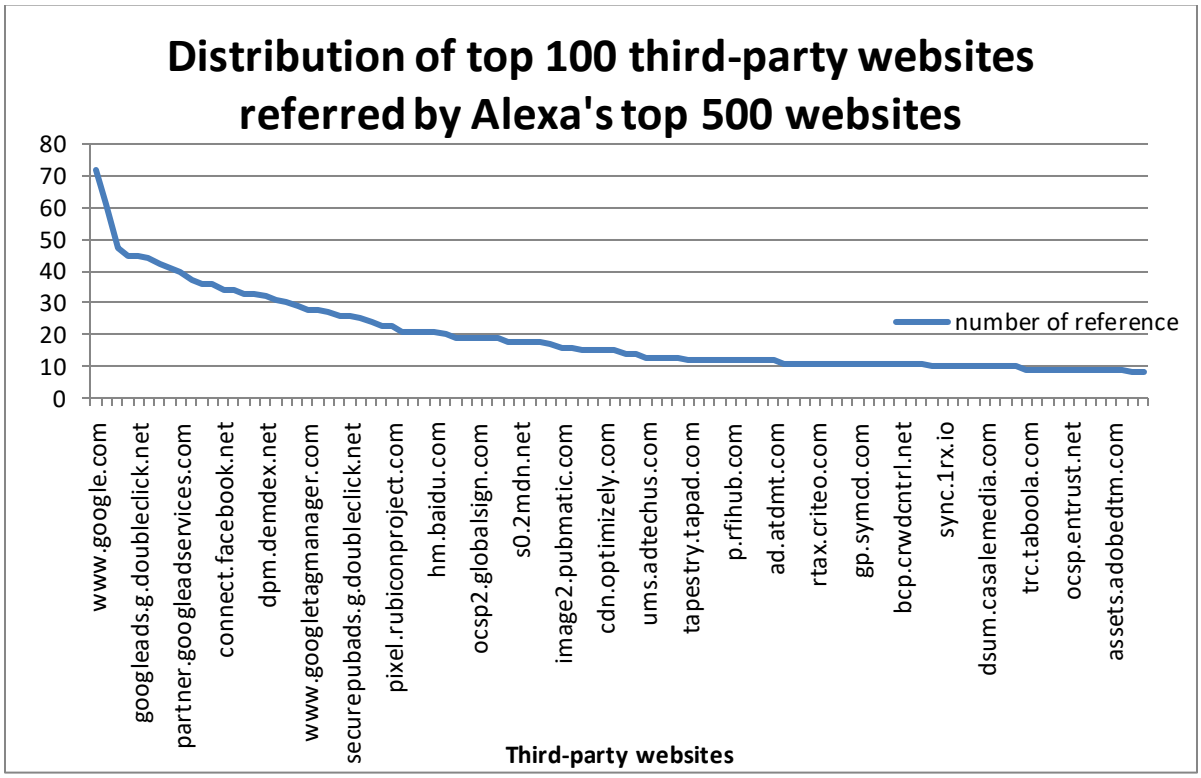

Figure 2: The graph shows distribution of top 100 third-party website referred by Alexa's top 500 websites

Top 10 popular third-party sites referred to by Alexa's top 500 websites are[2]:

Referred third-party sites	Number first-party that referred to the site (Referrals)
www.google.com	72
www.facebook.com	61
cm.g.doubleclick.net	47
www.googletagservices.com	45
googleads.g.doubleclick.net	45
ocsp.digicert.com	44
ss.symcd.com	42
fonts.googleapis.com	41
partner.googleadservices.com	40
clients1.google.com	37

Table 2: The table shows top 10 third-party websites referred by number of first-party websites that is in Alexa's top 500 websites

[2] For complete list please look at dist_referred_tp.csv file under project1_1_materials folder

Discuss:

1.1) Distribution of third-party HTTP Requests on Alexa's top 500 websites

Total number of first-party HTTP requests is 3,175 while that for third-party HTTP requests is 26,260.

Website that contains the most number of third-party HTTP requests is rakuten.co.jp which does 676 times third-party http requests. Rakuten is a Japanese electronic commerce and Internet company based in Tokyo, Japan. The e-commerce is the largest e-commerce site in Japan and among the world's largest by sales[3].

The top 10 websites with maximum number of third-party HTTP requests falls in the following categories

Categories	**Top 10 websites with maximum number of third party HTTP requests**
Ecommerce	rakuten.co.jp
News (current world affairs, domain specific eg. business and finance)	cnn.com, ettoday.net, sabah.com.tr, indiatimes.com, nytimes.com, caijing.com.cn
Entertainment (music, movies, online games)	azlyrics.com, ettoday.net, sohu.com
BitTorrent System (communication protocol P2P)	extratorrent.cc
Instant messaging software	qq.com
Search engine	sohu.com

Note: a websites may fall in more than 1 category

We see that 6 out of 10 websites fall into 'News' category followed by 'Entertainment' category with 3 out 10 websites falling into this group.

The type of categories the top 10 websites fall into are used by all people in general except for BitTorrent System. Hence, with more third-party HTTP requests being made, these top 10 websites have more potential to impact large number of people.

It is also interesting to see that 6 out of the 10 websites are from Asia – Japan-1, India-1, China-3, and Taiwan-1.

The figure 1 shows that about 50% of the top 500 Alexa websites has less than 20-25 third-party HTTP requests. For the rest 50% there is a steady increase and for a small section its extreme.

[3] Wikipedia contributors, "Rakuten.", *Wikipedia, The Free Encyclopedia,* https://en.wikipedia.org/w/index.php?title=Rakuten&oldid=741503792 (accessed October 15, 2016)

1.2) Most referred third-party domains

The third-party website that is referenced the most is google.com. It is referred by 72 different websites in Alexa's top 500 websites.

The top 10 websites that is referred the most falls in the following categories:

Categories	**Top 10 websites that is referred the most by Alexa's top 500 websites**
Internet related services and products	google.com
Online social media and networking service	facebook.com
Online ad serving services	cm.g.doubleclick.net, googletagservices.com, googleleads.g.doubleclick.net,
Security of server and network services	ocsp.digicert.com, ss.symcd.com ,
Front-end web design	fonts.googleapis.com
Tracker	partner.googleadservices.com , fonts.googleapis.com
Search suggestion	clients1.google.com

Note: a websites may fall in more than 1 category

It is very interesting to see that the top two most referred sites are within top 3 sites of Alexa's top 500 websites – google.com 1st and facebook.com 3rd.

As we see from the categories, the most type of third party referred is for online advertising services, tracking, suggestions and top websites (8 out of 10 referred websites). However, with the presence of references to security checking protocol (2 out of 10 referred websites) we can assume that there are many sites in Alexa's top 500 websites that keeps check on security and security compliance.

The figure 2 shows that top 100 referred third-party websites are referred by at least 10 Alexa's top 500 websites.

2.1) Distribution of number of third-party cookies on Alexa's top 500 websites

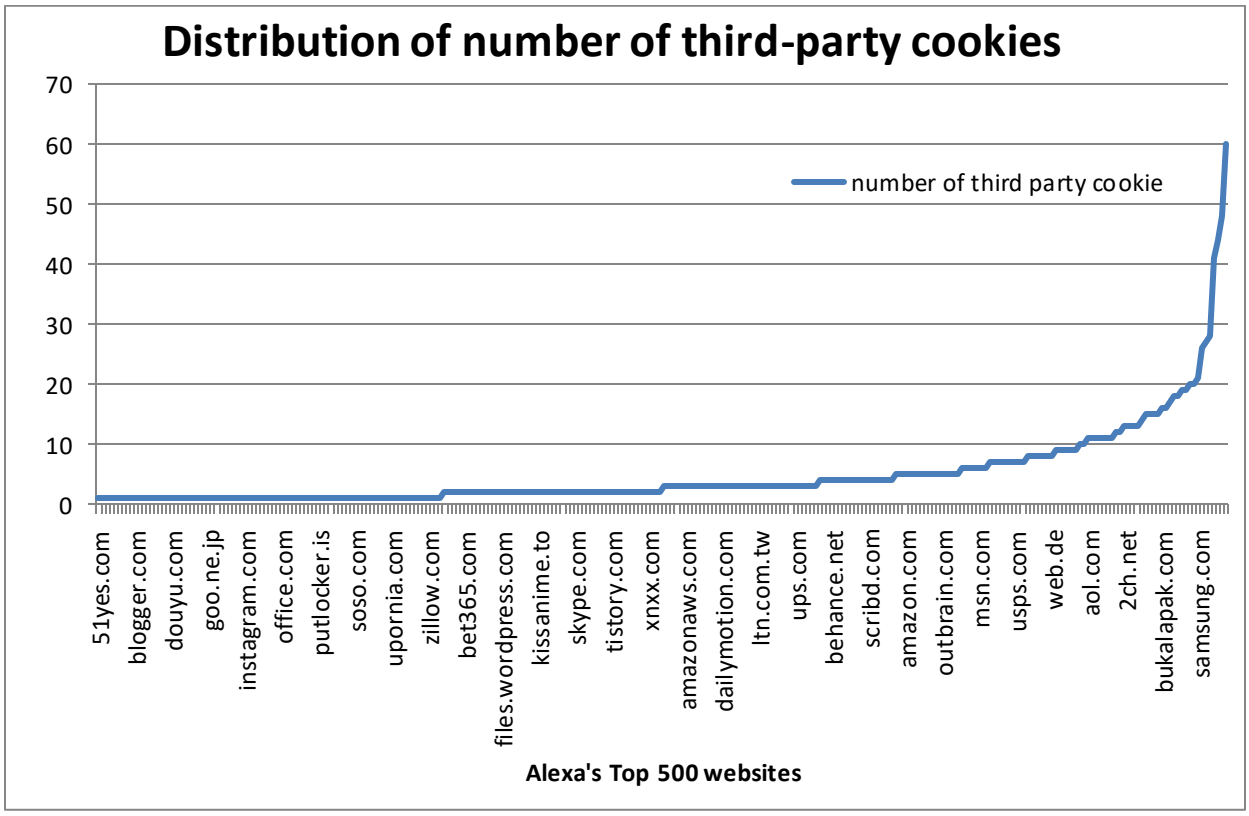

Figure 3: The graph shows distribution of third-party HTTP requests on Alexa's top 500 websites

Top 10 websites where the most third-party cookies are set[4]:

Alexa Websites	Number of third party cookie set
elpais.com	60
indiatimes.com	48
capitalone.com	44
sogou.com	41
weebly.com	28
bloomberg.com	27
booking.com	26
samsung.com	21
subscene.com	20
foxnews.com	20

Table 3: The table shows top 10 websites with the most third-party cookies set

[4] For a complete list please look at dist_tp_cookie.csv file project1_2_materials folder

2.2) Distribution of third-party websites that set cookie in different Alexa's top 500 websites

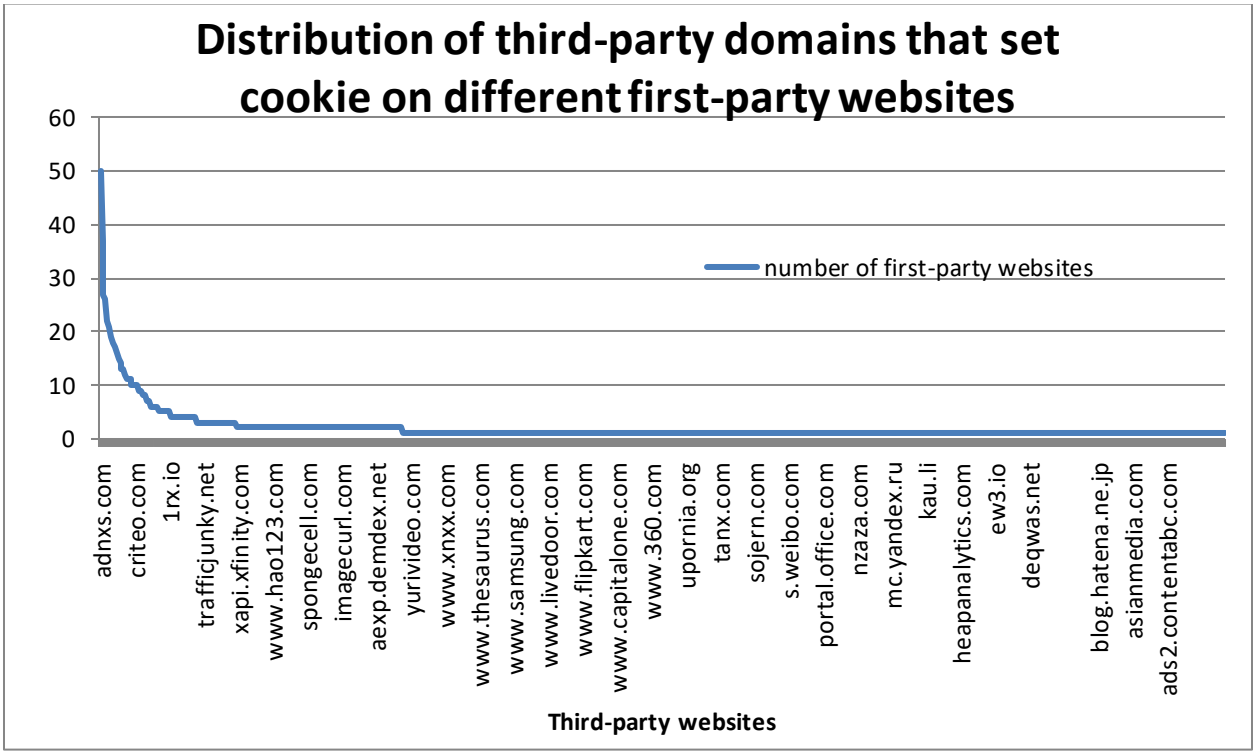

Figure 4: The graph shows number of different first-party domains where third-party domain sets cookie.

Top 10 third-party websites that sets cookie the most are[5]:

Third-Party website setting cookies	Number of different Alexa top 500 websites
adnxs.com	50
demdex.net	37
rubiconproject.com	27
rlcdn.com	26
casalemedia.com	22
pixel.rubiconproject.com	21
bluekai.com	19
pubmatic.com	18
dpm.demdex.net	18
optimizely.com	17

Table 4: The table shows top 10 third-party websites that set cookie on different Alexa top 500 websites

[5] For a complete list, please look at dist_tp_cookie_set_domain.csv file project1_2_materials folder

Discussion:

2.1: Distribution of number of third-party cookies on Alexa's top 500 websites

The website which has the most third-party cookies set is elpais.com. It is the highest-circulation daily newspaper in Spain[6]. It has 60 different cookies set. List of top 10 websites with the most third-party cookies is in table 3. In total there are 6,912 third-party and 3,377 first-party cookie set.

The top 10 websites that has the most third-party cookies set falls in following category:

Category	**Top 10 websites that has the most third-party cookie set**
News—current world affairs, content portals of different domains	elpais.com, indiatimes.com, Bloomberg.com, foxnews.com
Services -- banking, financial software, accommodation booking, Web-hosting,	capitolone.com, bloomberg.com, booking.com, weebly.com,
Search engine	sogou.com
Products – technological gadgets	samsung.com
Entertainment – Movie database	Subscene.com

The top 10 websites that has the most third-party cookies set falls into five main categories – news, services, search-engine, products and entertainment. Two of the categories are free of charge – news and search-engine. Rest three of the categories --- service, products and entertainment are paid products. One of the reasons for these websites to have the most third-party cookie set might be because these all websites have high unique visitors. Hence, in order to know more about behavior of people visiting their websites third-party cookie might be set.

The figure 3 shows that about 50% of top 500 Alexa websites have less than 5 cookies set. For the rest 50% there is a steady increase. For few websites there are high numbers of third-party cookies set.

[6] [6] Wikipedia contributors. "El Pais," *Wikipedia, The Free Encyclopedia,* *https://en.wikipedia.org/w/index.php?title=El_Pa%C3%ADs&oldid=739378313* *(accessed October 16, 2016).*

2.2: Most popular third-party domain that set cookies on most top 500 Alex's website

The most popular third-party domain that sets cookie on most top 500 Alexa websites is adnxs.com. It is a portal for publishers to the AppNexus—international internet advertising and ad management software, online auction exchange used to sell advertising space[7]. It sets cookie in 50 different Alexa websites.

The top 10 third-party domain that sets cookie on most top 500 Alexa websites fall in the following categories.

Category	**Top 10 third-party domains that set cookie in most sites in Alexa top 500 websites**
Online auction for selling ad space	adnxs.com, rubiconproject.com,casalemedia.com, pubmatic.com
Website tester for adware, spam, scams, email-practices	demdex.net,
Scam site	rlcdn.com, pixel.rubiconproject.com, bluekai.com, dpm.demdex.net
Unknown	dpm.demdex.net , rlcdn.com, pixel.rubiconproject.com,
Optimizing experience in different platforms – mobile, desktop	optimizely.com

Note: a websites may fall in more than 1 category

The top 10 third-party domains that set cookie in most sites in Alexa top 500 websites fall into 4 categories – online auction for selling ad space, website tester for malwares, scam sites unknown and optimization of user experience in different platforms.

4 out of top 10 third-party domains that set cookie in most Alexa 500 sites falls under online ad selling. 1 fall under website tester for detecting any malware, 1 fall under enhancing user experience and 1 (bluekai.com) fall under scamming site. Result for rest 3 – dpm.demdex.net, rlcdn.com and pixel.rubicon.com, is unknown or is a scam. Google search on these three websites did not display their profile and some were blocked by robot.txt file. There is a high possibility that these 3 sites are scam sites since they have domain name similar to legit sites (demdex.net, rubiconproject.com). For rlcdn.com the 'cdn' string may be used for hiding its true nature. Upon more google search there were many threads linking virus and these 3 domains.

[7] https://www.adnxs.com/ (accessed October 16, 2016)

The figure 4 shows that more than 75% of third-party domain sets cookie in more than one first-party sites. However there is a steady rise for rest 25% which peaks for few last %. Considering that there are 876M web sites, the few % of 876M is a huge number of websites that has potential affect many people.

It is also very interesting to see that first-party domain with maximum number of third-party cookie set falls under news, services, search-engine and entertainment categories. Correspondingly third-party domains that set cookie on different first-party domains are mixed bag of ad sellers, website security checkers and malwares.

3) Visualizing using graphing tools – D3[8]

Here is an algorithm for creating graph:

```
1) create a json file for proj1_1_materials/fp_values.txt file. The fp_values.txt file contains a dictionary of first-party domain and it's respective third-party HTTP calls.The newly created json file will contain objects of first-party and its associated third-party values.

2) In the created new json file have objects with 'first_party' attribute to 'key' of the dictionary and 'third_party' attribute to 'values' of the dictionary.

3) In D3's canvas load this json file. Create nodes for all 'first_party' attribute's values and 'third_party' attribute's values. In 'third_party' values do not make duplicate nodes if a node with the domain is created.

4) Using an object's 'third_party' attribute, create a link between a first-party domain and its associated third-party domains.

5) The D3's code will be similar to that of link given in description of the project(https://bl.locks.org/mbostock/950642)
```

[8] Due to limited time, I just wrote an algorithm for graphing tracking graph

however with modification needed due to the structure of the json objects.

```

As inferring from two files (dist_tp_request.csv and dist_referred_tp.csv) and analysis part 1, the graph will not be fully connected since there is few number of domains being referred (in-degree) than number of third-party HTTP requests (out-degree) .

A vertex that will have the most out-degree is rakuten.co.jp since it has the maximum number of third-party HTTP request called. A vertex with the most in-degree will be google.com since it referred the most.

The vertex with both maximum number out-degree and in-degree will be the most central in the graph.

**Limitation of the project:**

1) Due to limited processing power of a laptop, not all data was taken in consideration. At least 75% of data is used in all part of analysis.

2) Given a little more time, I would have completed making graph for the project since I have computed data[9] for the graph.

**Improvements:**

1) Have a computer with better processing power when computing with large amount of data or use tools that can minimize processing of files such as sqlite in this project.

2) If writing scripts to process large amount of data, use good data structure to improve time and space efficiency.

[9] proj1_1_materials/fp_values.txt
```