A LONGITUDINAL ASSESSMENT OF WEBSITE COMPLEXITY

by

SEYED HOOMAN MOSTAFAVI

A THESIS

Presented to the Department of Computer and Information Science
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Master of Science

June 2018

THESIS APPROVAL PAGE

Student: Seyed Hooman Mostafavi

Title: A Longitudinal Assessment of Website Complexity

This thesis has been accepted and approved in partial fulfillment of the requirements for the Master of Science degree in the Department of Computer and Information Science by:

Reza Rejaie Advisor

and

Sara D. Hodges Interim Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2018

© 2018 Seyed Hooman Mostafavi

THESIS ABSTRACT

Seyed Hooman Mostafavi

Master of Science

Department of Computer and Information Science

June 2018

Title: A Longitudinal Assessment of Website Complexity

Nowadays, most people use several websites on a daily basis for various purposes like social networking, shopping, reading news, etc. which shows the significance of these websites in our lives. Due to this phenomenon, businesses can make a lot of profit by designing high quality websites to attract more people. An important aspect of a good website is its page load time. There has been a lot of studies which analyzed this aspect of the websites from different perspectives. In this thesis, we characterize and examine the complexity of a wide range of popular websites in order to discover the trends in their complexity metrics, like their number, size and type of the objects and number and type of the contacted servers for delivering the objects, over the past six years. Moreover, we analyze the correlation between these metrics and the page load times.

CURRICULUM VITAE

NAME OF AUTHOR: Seyed Hooman Mostafavi

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
University of Tehran, Tehran, Iran

DEGREES AWARDED:

Master of Science, Computer and Information Science, 2018, University of Oregon
Bachelor of Science, Software Engineering, 2016, University of Tehran

AREAS OF SPECIAL INTEREST:

Artificial Intelligence
Machine Learning
Data Analysis

PROFESSIONAL EXPERIENCE:

Web Administrator and Data Manager, Graduate School, University of Oregon, June 2017 - June 2018

Teaching Assistant, University of Oregon, Sep 2016 - June 2017

ACKNOWLEDGMENTS

I wish to express my sincere gratitude to my advisor Professors Reza Rejaie for his continuous support and assistance. This thesis would not have been possible without his guidance and immense knowledge. In addition, special thanks to the people who helped us in collecting our data from different locations and made significant contribution to our research; Andrew Hill from University of Oregon, Makan Taghavi from New York University, Heda Wang from China, Thiago Guarnieri form Brazil, and Yonas Kassa from Spain. My sincere thanks also go to Bahador Yeganeh from University of Oregon who always provided valuable resources and comments that helped in improving this study.

I would like to express my very profound gratitude to my parents, my brother, and my sister for their unfailing support and continuous encouragement throughout all years of my studies. This accomplishment would not have been possible without them.

Finally, my heartfelt thanks to my friends Sarah Shodja and Sina Abbasgholipour whose untiring help and support have always been there for me when I needed it the most.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

Figure Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

These days, due to the widespread use of internet, webpages are evolving with a fast pace and business owners try to design the content of their websites in a way to attract more users. There are many types of objects which could be used on the websites to make them look more appealing and easier to use. At the same time, the content design must not have negative effect on the performance of the website, specifically its page load time, which is a very influential factor on the user quality of experience.

There has been a lot of work on examining the performance of websites and identifying ways to improve the page load times. For instance, some studies (Wang, Balasubramanian, Krishnamurthy, & Wetherall, 2013) (Netravali, Goyal, Mickens, & Balakrishnan, 2016) focus on the dependencies within the page load process and optimizing the page load time by using the dependency graph. Some studies (Butkiewicz, Wang, Wu, Madhyastha, & Sekar, 2015) (Kelton, Ryoo, Balasubramanian, & Das, 2017) use content prioritization based on the user preferences in order to load more important objects earlier and enhance the user experience. Moreover, there are some online tools (Google-Developers, 2018) (WebpageTest, 2018) that assess the performance of websites and help developers to design and write more efficient webpages.

Another way to evaluate the performance of a website is by analyzing its complexity in terms of number and size of the objects, types of the objects, number of the contacted servers etc, since according to the results reported by (Butkiewicz, Madhyastha, & Sekar, 2011) (Tian & Rejaie, 2015), the page load time is affected by these metrics. In this study, we adopt a similar methodology as (Butkiewicz et al., 2011) and aim to conduct a longitudinal assessment of websites' complexity. we repeat and extend the analyses in our previous study (Tian & Rejaie, 2015) on around 2000 popular websites and also analyze the trends in complexity metrics and page load times over the last 6 years. Moreover, we take a closer look at the role of the contacted servers in providing the objects and affecting the load time of the page. Also, we explore other key factors that influence the page load time.

In the rest of this paper, first we explain our methodology in selecting a representative set of target websites and collecting and parsing the data. Then, we describe our dataset and the categories of websites in more details. Next, we examine the content complexity and service complexity of our target websites and present our results and comparisons. After that, we analyze the correlation between different complexity metrics and the page load time and use machine learning models to further explore these correlations. Finally, we summarize our results and demonstrate our main findings.

CHAPTER II

METHODOLOGY

This section describes the methodology that we used to collect the data and extract desired information from it. This methodology is similar to the one that is used in our previous paper (Tian & Rejaie, 2015).

Website Selection

We aim to select our target websites in a way to be representative of different levels of popularity and various categories. To fulfill this purpose, we use two online resources which provide a list of the most popular websites based on their number of visitors and number of page views. The first resource is Alexa.com which offers a ranked list of top 500 sites on the web along with their categories (e.g., art, business, shopping). The second resource is Quantcast.com that provides over half a million ranked websites according to their popularity, but without their categories. We only consider the top 20k websites from Quantcast list. Our method for selecting the 2000 target websites from these two resources is as follows: Since we want to have as many categorized website as possible, we take all the websites from Alexa list into account. To take advantage of the wide range of popular websites from Quantcast, we partition the list into 5 different rank groups and then randomly choose a number of websites from each rank group. Table 1 shows the 5 rank groups and the number of the websites selected from each of them (including our previous results (Tian & Rejaie, 2015)).

As it is more important for us to consider more websites with higher ranks, we select all the websites from the first rank group and a some of the websites from the other rank groups. The number of websites in other rank groups is proportional to how high the rank range is. It should be noted that in the Quantcast list, in each rank group, the name of the some of the websites is not available and it is referred to as Hidden Profile which is the reason of having a total of 418 websites from the first rank group. Moreover, in the third column of the table, the number of websites from Alexa list that map to each rank group is demonstrated. The rest of the websites from Alexa list (290 websites) are not among any of the defined rank groups. In the rest of this paper, we simply refer to rank groups as groups and to the

Rank Range	Quantcast	Alexa	Total
1 - 500	332 (Tian: 500)	86 (Tian: 344)	418 (Tian: 500)
500 - 1000	233 (Tian: 300)	15 (Tian: 11)	248 (Tian: 300)
1000 - 5000	196 (Tian: 300)	52 (Tian: 54)	248 (Tian: 300)
5000 - 10000	315 (Tian: 400)	33 (Tian: 37)	348 (Tian: 400)
10000 - 20000	424 (Tian: 500)	24 (Tian: 44)	448 (Tian: 500)
-	-	290 (Tian: 0)	290 (Tian: 0)
Total	1500	500	2000

Table 1: Selection of 2000 target websites from the list of top 20k websites of Quantcast and top 500 websites of Alexa

selected websites as target websites.

Data Collection

We developed a crawler that automatically browses target websites and collects a single HTTP Archived Record (HAR) file (Odvarko, 2017) for each website. When we refer to a HAR file of a website, we mean the HAR file that is exported from the homepage of the website. The crawler uses Firefox (Firefox, 2017) browser and is implemented with Selenium WebDriver (Selenium-WebDriver, 2017) in Java. It also uses HAR Export Trigger extension (HAR-Export-Trigger, 2017) (version 0.5.0-beta.7) for Firefox in order to export the HAR files. The HAR file is a JSON-formatted file that logs all the data related to the browser-website interaction. This data contains the details about the requests and responses such as the destination URL, timing values, response status code and size and type of the returned objects. After collecting the HAR files, we use the Haralyzer module (Haralyzer-Module, 2017) in Python to parse the files and extract the desired information from them. Each HAR file will be automatically exported when the page is completely loaded. However, we need to set two timeouts to be able to identify the completion of the load time of the page. The first timeout specifies the amount of time that the auto-exporter should wait after the last finished request before exporting the HAR file. This timeout is set to 2.5 seconds. The second timeout indicates how long we should wait for a single website to be loaded.

According to our experiments, some of the websites can have a very long load time, even more than 15 minutes, since the website keeps sending some objects to the browser even after the main content of the page is downloaded. To avoid allowing these kind of websites to stall our crawler for a long time, we empirically set a timeout of 3.5 minutes on the waiting time for each website. If the page is not loaded in this time, we do not export its corresponding HAR file. Instead, we add the website's URL to the list of not loaded websites. In different geographical locations, clients can experience dissimilar load times due to having different relative connectivity to the local servers. Furthermore, locally customized version of the websites might have different content, like different advertisements. To address these issues, we run our crawler from five geographically scattered vantage points. It is also worth mentioning that sometimes a request to a specific website might be redirected to another domain based on the location of the client. In this case, we collect the HAR file from the final domain address.

CHAPTER III

DATA SET

Among the five vantage points that we used in our study, two of them are in the US, one in the east coast (New York) and one in the west coast (Eugene); one of them is in Brazil, one of them is in Spain, and the other one is in China. Table 2 shows the details of the collected data from each vantage point and table 3 shows these details in our previous results (Tian & Rejaie, 2015).

Location	Collected	Not Loaded	Broken	Redirected
Eugene, OR	93.2%	1.8%	5%	1.4%
New York, NY	91.6%	2.9%	5.5%	1.5%
Brazil	90.3%	5.7%	4%	1.1%
Spain	89%	4.5%	6.5%	1.7%
China	80.5%	13.1%	6.4%	1.5%

Table 2: Details of collected data from each vantage point

Location	Collected	Unresolved	Redirected	Real
Eugene, OR	90%	2%	1%	86%
Durham, NC	90%	1%	1%	87%
Brazil	97%	2%	1%	93%
France	94%	2%	1%	91%
Spain	96%	2%	1%	92%
China	84%	12%	5%	66%

Table 3: (Tian & Rejaie, 2015) - details of collected data from each vantage point

The first column indicates the percentage of the HAR files that were successfully collected and considered in our analysis. The corresponding column in our previous results (Tian & Rejaie, 2015) is the ”Real” column. Second column represents the percentage of the websites that did not load completely in the specified time limit. Third column shows the fraction of the websites that were broken at the time of the crawling, meaning that the URL of these websites led to

no up and running web page. The similar column in our earlier results (Tian & Rejaie, 2015) is the ”Unresolved” column which is the percentage of the HAR files that caused parsing error. Last column demonstrates the percentage of the redirected websites. Note that the HAR files of these websites are included in the collected ones. Overall, we were able to collect and use the HAR files from a high percentage of the websites (more than 80% for all the vantage points). Comparing to our previous results (Tian & Rejaie, 2015), this percentage is higher for the vantage points in the US and China and is a little lower for Brazil and Spain. Moreover, The smaller rate of collected HAR files from China is due to the filtering of certain websites and restricted access of the clients. These blocked websites were mostly categorized as not loaded websites after running our crawler. It should be noted that our analysis is biased based on the reachable websites.

Categories

To categorize the target websites, as we mentioned earlier, we use the category information from the Alexa website. Alexa provides the list of 500 most popular websites in 17 different categories. Moreover, in each category, there were a number of layers of subcategories which contained further number of websites. In order to extract the most extensive list of websites in each category, we automatically crawled the list of categories and their subcategories. Since in some cases there were many subcategories under each category, and subcategories with a small number of websites usually contained lots of duplicate websites, we only considered the subcategories with more than 25 websites. In this way, we were able to collect more than 30,000 websites in each category except the Games category which has about 10,000 websites. After obtaining the categories, we compared the list of target websites with the list of websites in each category. Table 4 shows the number of target websites in each category (including these numbers from our previous results (Tian & Rejaie, 2015)). In order to focus our analysis on the most popular categories, we only consider the top 6 categories which are: Art, Business, Computers, Shopping, Society, and Sports.

Category	Number of Websites
Computers	112 (Tian: 100)
Business	90 (Tian: 96)
Art	82 (Tian: 64)
Shopping	81 (Tian: 93)
Society	58 (Tian: 80)
Sports	25 (Tian: 62)
Total	448 (Tian: 495)

Table 4: Number of target websites in each category (including (Tian & Rejaie, 2015))

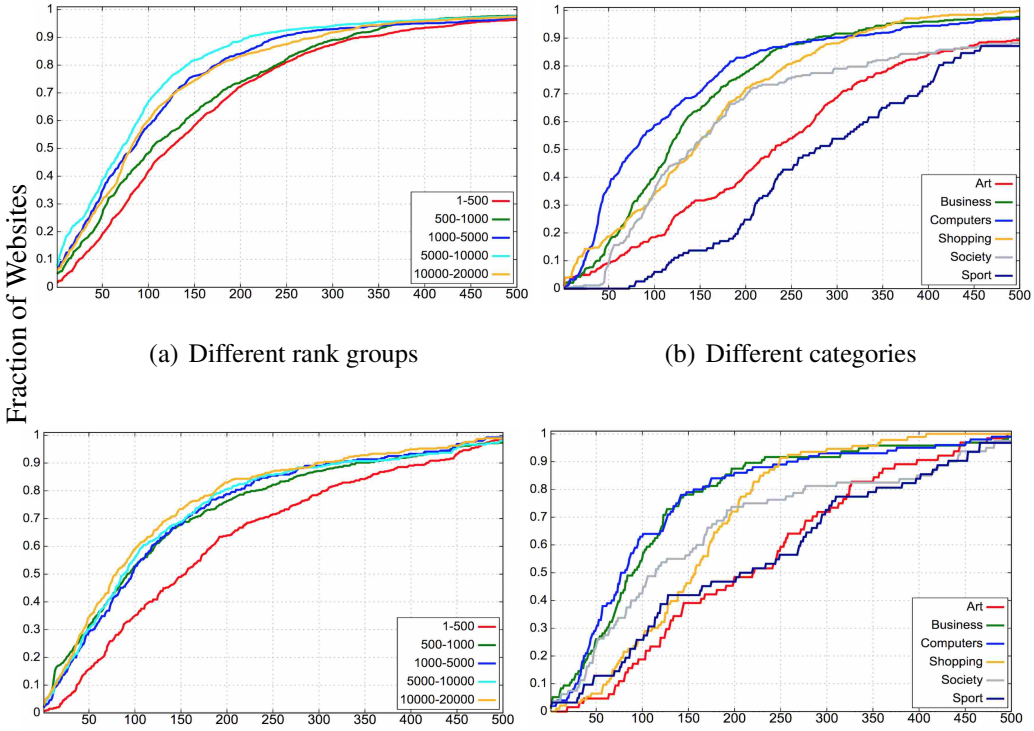
CHAPTER IV

CONTENT COMPLEXITY

In this part, we present our analyses regarding the number of the objects and different types of the objects of our target websites. These analyses comprise the measurements for all the target websites in all five vantage points. Moreover, we compare our results with our previous results (Tian & Rejaie, 2015) and the analysis that were conducted by (Butkiewicz et al., 2011) to demonstrate the trend of content complexity over the course of past six years.

Number of Requested Objects

First, we illustrate our findings about the number of the requested objects. Figure 1 show the cumulative distribution function (CDF) of the requested objects based on different rank groups and categories of the target websites.



(a) Different rank groups

(b) Different categories

(c) (Tian & Rejaie, 2015) - different rank groups

(d) (Tian & Rejaie, 2015) - different categories

Figure 1: CDF of number of the requested objects

These results show that the rank group and the category of the target websites correlate with their number of objects. Figure 1(a) demonstrates that the websites in

higher rank groups contain more number of objects except for the two lowest rank groups. More specifically, the median number of objects in the two top rank groups is 152 and 134, respectively, but for the other rank groups, the median number is in the range of 80 to 100. This is similar to our previous results (Tian & Rejaie, 2015), which are shown in Figure 1(c), meaning that on average, the number of the requested objects by websites in different rank groups has stayed more or less the same since two years ago. However, for the top 20 percent of the websites in the highest rank group (ranks between 1-500) the median number of objects is more than 450, but for the ranks between 5000-10000 the median is around 250. This analysis indicates around 50% increase in the number of the objects for the top 20 percent of the websites in comparison to our previous results (Tian & Rejaie, 2015). Moreover, another difference is that there is a smaller gap between the number of the objects of the websites in the top two rank groups.

Furthermore, Similar to (Tian & Rejaie, 2015) findings, comparing these results with (Butkiewicz et al., 2011) results, the main distinctions are: different order of categories in terms of having more objects, much more number of requested objects for all the websites in the current results (at least doubled for all rank groups and categories), and a larger gap between the number of the objects for different ranks and categories of websites.

Figure 1(b) demonstrates that there is an apparent distinction between the number of the objects for websites in different categories. According to this Figure, websites in the Sport and Art categories have much more objects than websites in other categories. In particular, the median number of objects in these two categories is around 250, however, websites in the Shopping and Society categories typically include around 140 objects. Websites in the Business category have less number of objects with a median around 120 and websites in the Computers category contain the fewest number of objects and have a median of around 70. These results are similar to our previous analysis (Tian & Rejaie, 2015) in terms of the order of the categories that have the most number of objects. The only difference is that in our current results, there is more gap between the number of the objects for different categories.

MIME Types

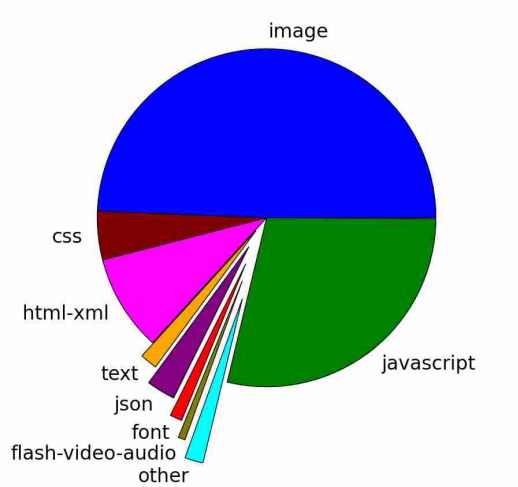
To examine various types of delivered objects by the target websites, we consider the 11 most common MIME types in our analyses. Table 5 shows these MIME types.

Name	Template
Javascript	*/javascript, */x-javascript
Image	image/*
HTML	*/html
CSS	*/css
Json	*/json
Text	text/plain
Flash	*/x-shockwave-flash, */x-flv
XML	*/xml
Font	font/*
Audio	audio/*
Video	video/*
Other	Other templates

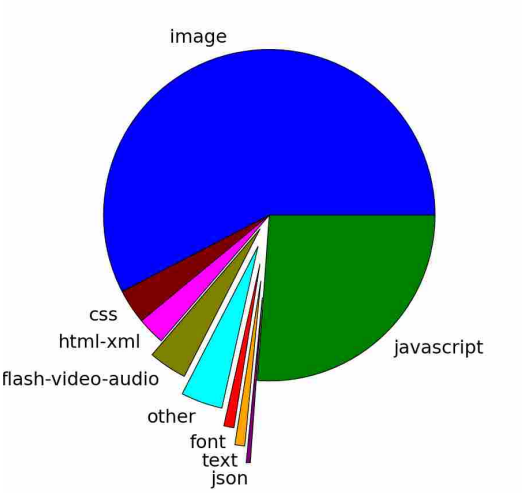
Table 5: Different MIME types

We draw the pie charts in Figure 2 in order to understand how the distribution of these MIME types is among all the delivered objects. Note that the returned objects from target websites in all the vantage points are considered. These pie charts indicate the distribution of number and size of the the various MIME types. Similar to our previous results (Tian & Rejaie, 2015), shown in Figure 2(c) and 2(d), image objects are responsible for almost half of the number of objects. However, in our new results, image objects compose more than half of the size of objects. The MIME type that has the second most contribution to the number and size of the delivered objects is javascript. Javascript objects also show a slight increase in their number and a little decrease in their size comparing to our previous findings (Tian & Rejaie, 2015); meaning that popular websites are using more javascript objects, which are smaller in size, than two years ago. In addition, like

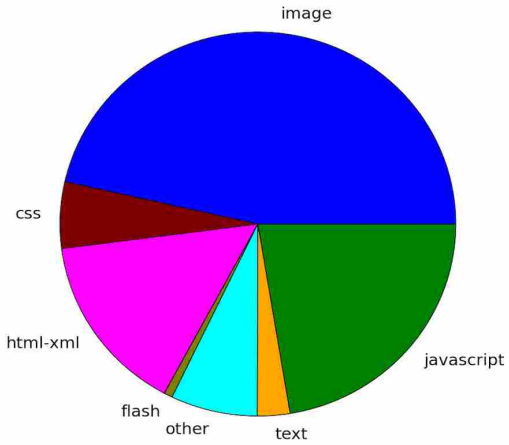
our previous results (Tian & Rejaie, 2015), some MIME types such as html-xml and text have less contribution to the size of the objects than the number of the objects while some other objects like flash, video, and audio are the opposite. This is due to the fact that html-xml and text objects have a smaller size in comparison to other objects such as flash, video, and audio.



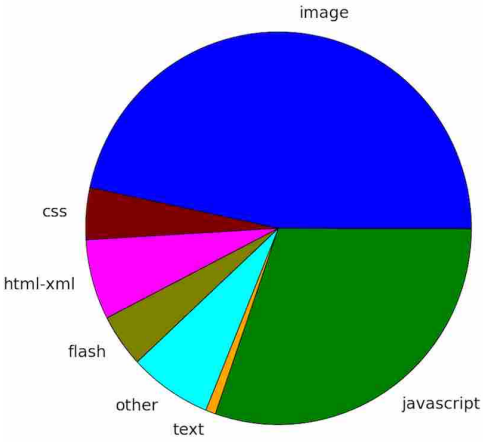
(a) Number of objects



(b) Size of objects



(c) (Tian & Rejaie, 2015) - number of objects



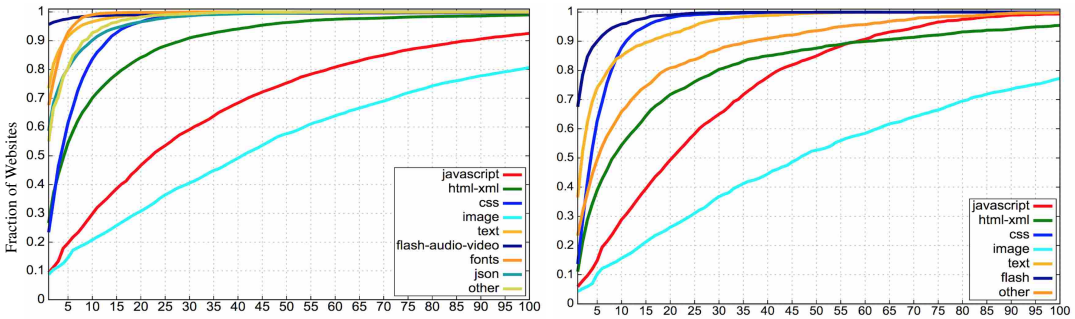
(d) (Tian & Rejaie, 2015) - size of objects

Figure 2: Distribution of number and size of the objects across different MIME types

If we classify the delivered objects by their MIME types, we can see that the number of the returned objects varies based on their type. Figure 3(a), which shows the CDF of the returned objects, having different MIME types, by target websites,

proves this statement. As mentioned in the previous paragraph, on average, each website delivers more number of image and javascript objects that any other type of object. Particularly, the median number of the image and javascript objects across different websites is 41 and 23 respectively; while this number is less than 10 for all other types of objects. The distribution of the objects with different MIME types is similar to our previous results (Tian & Rejaie, 2015), which is plotted in Figure 3(b), nevertheless, there are some differences that are worth mentioning. First, the median number of the image objects is slightly decreased, but the median size of the image objects is almost 50% increased. Moreover, both the median number and median size of the javascript objects is increased by roughly 20%.

Also, considering these results and the (Butkiewicz et al., 2011) results in 2011, we see similar distributions of objects across various MIME types with a couple of significant distinctions. First, the contribution of the image objects to the total number of bytes has increased from nearly 10% to more than 50% while their contribution to the number of the objects is just slightly increased. Second, the number of the javascript objects has been almost doubled when their contribution to the number of bytes has experienced a little decrease.



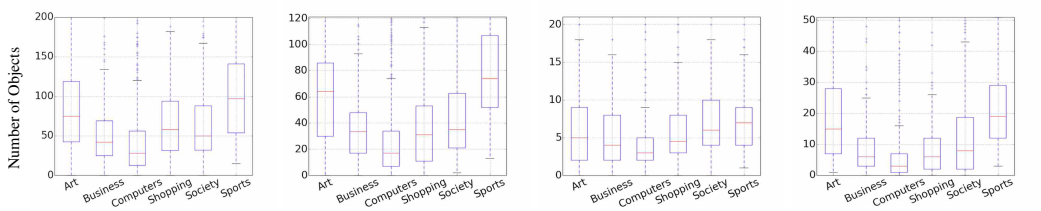
(a) Different MIME types

(b) (Tian & Rejaie, 2015) - Different MIME types

Figure 3: CDF of number of different types of objects

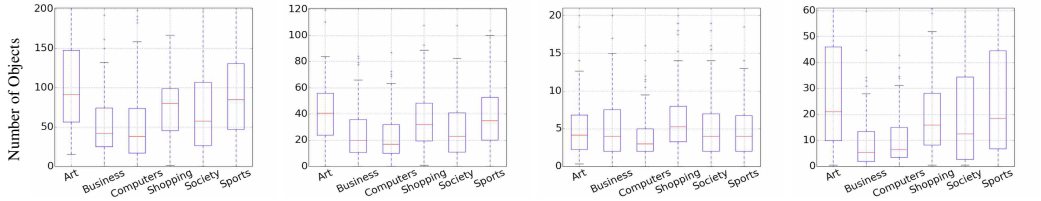
In the rest of this section, we analyze the distribution of the number and size of the objects with four major MIME types (image, javascript, css, and html-xml) across different categories and rank groups considering all the target websites in all the vantage points. Figure 4 demonstrates the distribution of the number of the delivered objects for different categories. We observe that websites in Art and Sport

categories have the most number of objects for all the MIME types except css. For css objects, websites in the society category surpass the websites in other categories. In our previous analyses (Tian & Rejaie, 2015), Figure 5, the results were similar for all the major MIME types except css which Business and Shopping websites had the most number of css objects.



(a) Image (b) Javascript (c) CSS (d) HTML-XML

Figure 4: Distribution of the number of objects with four major MIME types across different categories

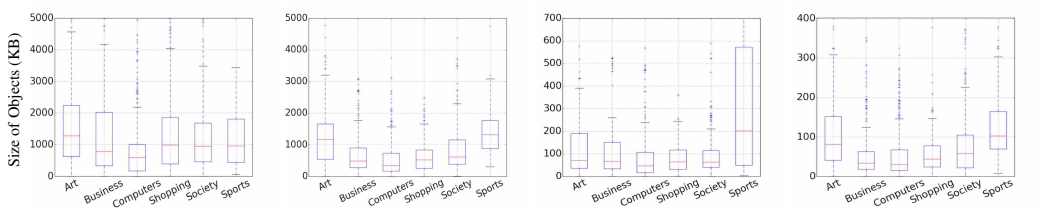


(a) Image (b) Javascript (c) CSS (d) HTML-XML

Figure 5: (Tian & Rejaie, 2015) - distribution of the number of objects with four major MIME types across different categories

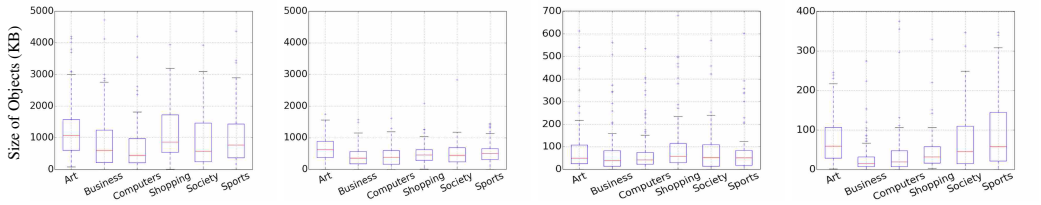
Figure 6 shows how the size of the objects with four major types is dispersed among the six categories. For javascript and HTML-XML objects, the distribution is similar to the distribution of the number of the objects, meaning that Art and Sport categories deliver the most number of bytes. With respect to the image objects, Art category still has the most contribution, but the Business category returns more bytes of data than the Sport category. Considering css objects, websites in the Sport category have much larger objects than other categories. These results regarding the image and css objects are different from our previous results (Tian & Rejaie, 2015), shown in Figure 7, in which Shopping category delivers the most bytes of image objects and Sport websites deliver smaller css

objects than some of the other categories such as Art, Shopping, and Society. Generally, these were the results that we expected to get based on our earlier findings in Figure 1(b) which denoted the fact that websites in the Art and Sport categories contain much more number of objects than other categories.



(a) Image (b) Javascript (c) CSS (d) HTML-XML

Figure 6: Distribution of the size of objects with four major MIME types across different categories

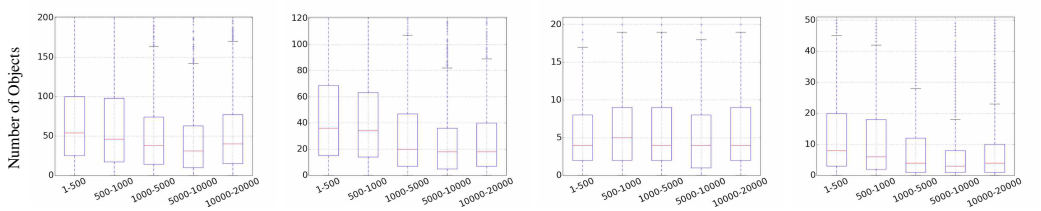


(a) Image (b) Javascript (c) CSS (d) HTML-XML

Figure 7: (Tian & Rejaie, 2015) - distribution of the size of objects with four major MIME types across different categories

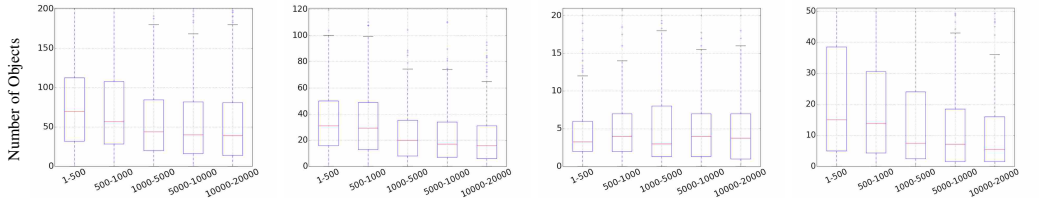
Now, we look into the distribution of the objects with the four major MIME types across different rank groups. Figures 8 and 10 demonstrate the distribution of the number and size of the returned objects by websites in different rank groups. Same as our previous results (Tian & Rejaie, 2015) (Figures 9 and 11), websites in higher ranks contain more number of image, javascript, and HTML-XML objects. The only distinction in our new results is that websites in the 10000-20000 rank group have more number of objects with these three MIME types than websites in the 5000-10000 rank group. Moreover, like our past analysis (Tian & Rejaie, 2015), although websites in higher rank groups include more number of images, the size of the image objects is about the same for all rank groups, which implies that higher ranked websites have more number of smaller sized images. However, size of the

javascript and html-xml objects has the same distribution as the number of these objects across different rank groups. Another interesting point about the distribution of objects in our current results and also previous results (Tian & Rejaie, 2015) is that the css objects have an almost uniform distribution over different rank groups in terms of both number and size of the objects.



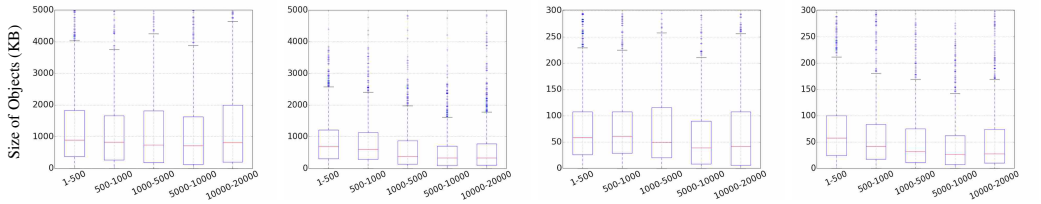
(a) Image (b) Javascript (c) CSS (d) HTML-XML

Figure 8: Distribution of the number of objects having four major MIME types across different rank groups



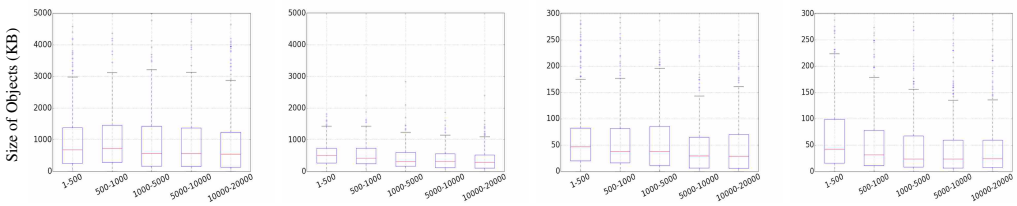
(a) Image (b) Javascript (c) CSS (d) HTML-XML

Figure 9: (Tian & Rejaie, 2015) - distribution of the number of objects having four major MIME types across different rank groups



(a) Image (b) Javascript (c) CSS (d) HTML-XML

Figure 10: Distribution of the size of objects having four major MIME types across different rank groups



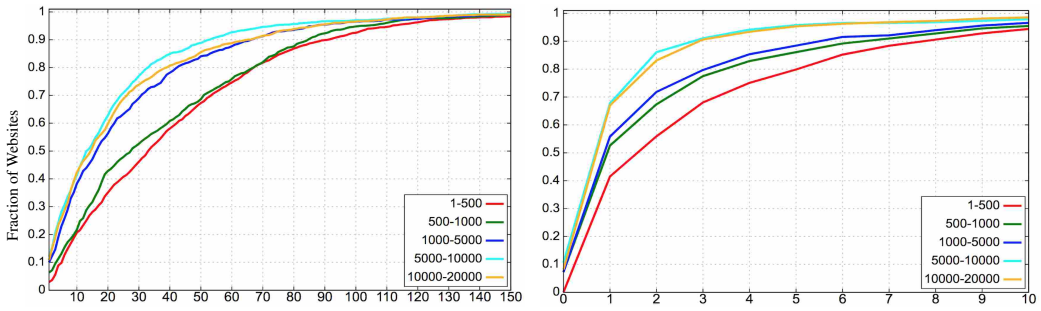
(a) Image (b) Javascript (c) CSS (d) HTML-XML

Figure 11: (Tian & Rejaie, 2015) - distribution of the size of objects having four major MIME types across different rank groups

CHAPTER V

SERVICE COMPLEXITY

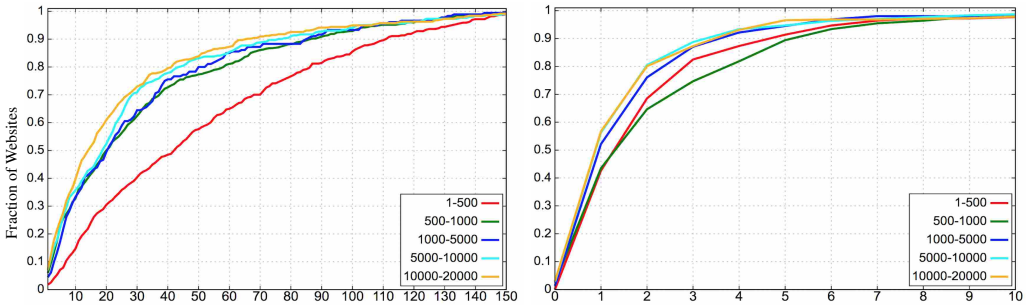
In this part, we analyze the number and role of the contacted servers which provide the objects of a website. Most of the time, each website contacts more than one server in order to fetch the objects and respond to the incoming requests from the browser. Moreover, usually more fraction of these contacted servers do not have the same domain name as the website's domain. Figure 12(a) shows the number of the contacted servers by websites in different rank groups. This Figure indicates that websites in higher rank groups usually contact more number of servers. The only exception is the websites in the 10000-20000 rank group which contact more servers than their upper rank group, which is 5000-10000. More specifically, websites in the two highest rank groups usually contact around 30 servers while the median number of contacted servers for other rank groups is 15.



(a) All servers (b) Origin servers

Figure 12: Number of contacted servers

In comparison to our previous results (Tian & Rejaie, 2015), websites in the highest rank group, which have a ranking between 1 to 500, contact roughly 20% fewer servers and considering all the websites, the median number of the contacted servers has been decreased a little (around 15%) over the past two years. Figure 13(a) demonstrates these findings. However, comparing with (Butkiewicz et al., 2011), the median number of the servers has become twice as large as their number in 2011.



(a) All servers (b) Origin servers

Figure 13: (Tian & Rejaie, 2015) - number of contacted servers

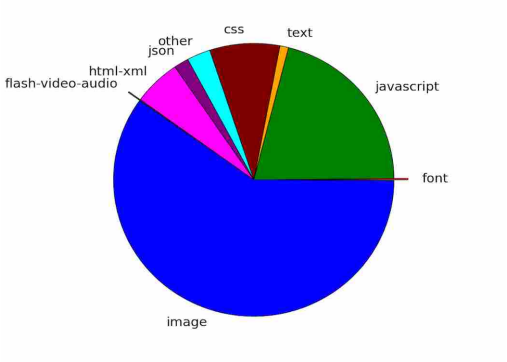
Origin and Non-Origin Servers

We categorize the contacted servers into origin and non-origin servers based on their authoritative nameservers. Origin servers are the ones that have the same set (or a subset) of authoritative nameservers as the target website. Other servers are considered as non-origin servers. For example, many of the websites use non-origin servers such as content delivery network (CDN) servers in order to fetch their objects or Google servers in order to show advertisements. Our methodology in identifying the origin and non-origin servers is using the ”dig” command which provides the list of authoritative nameservers for a website. Figure 12(b) demonstrates the number of contacted origin servers for websites in different rank groups. It shows that as we saw in the case of all the contacted servers, websites with higher ranks contact more number of origin servers. Furthermore, we observe that the number of origin servers is much smaller than the number of all the servers which means that non-origin servers constitute the most number of contacted servers. For 80% of all the target websites, the number of the origin servers is between 2 to 5, which is almost the same as our previous results (Tian & Rejaie, 2015), Figure 13(b). Thus, the small decrease in the number of the servers in the past two years is mostly due to the decrease in the number of the non-origin servers. Moreover, these results indicate that non-origin servers have continued to play an important role in providing the content of the websites since two years ago; Considering the fact that in the earliest results reported by (Butkiewicz et al., 2011), the number of the origin servers was larger than our current results but the total number of the contacted servers was much smaller.

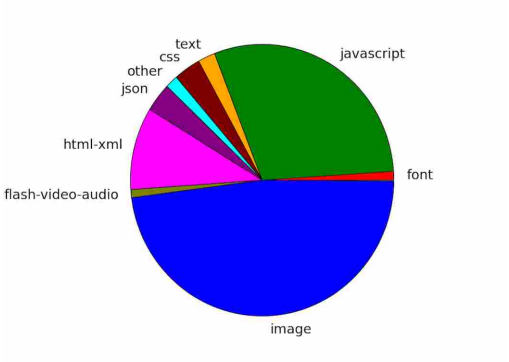
Another important aspect regarding the role of origin and non-origin servers is the fraction of returned objects by each group of servers and the distribution of different MIME types across those objects. Based on our analysis, considering all the target websites in all the vantage points, almost half of the objects and half of the bytes is delivered by origin servers and the other half is delivered by non-origin servers. More specifically, very similar to another study by (Ludin, 2017), for each website, median number of returned objects by non-origin servers is about 50 and median size of these objects is about 500 KB. This is different from our previous results (Tian & Rejaie, 2015), in which origin servers delivered roughly two-third of all the objects and bytes of data. It means that although the number of the non-origin servers has decreased slightly over the last two years, they deliver more number of objects and more bytes of data than before. Furthermore, the pie charts in Figures 14(a) and 14(b) present the distribution of MIME types across the objects returned by origin and non-origin servers. According to this Figure, for both group of servers, the most number of returned objects belongs to image and javascript MIME types. Moreover, same as our previous analysis (Tian & Rejaie, 2015), which is displayed in Figures 14(c) and 14(d), non-origin servers deliver higher percentage of objects of most of the MIME types except image and css objects. These results are also the same as the (Butkiewicz et al., 2011) results. In the next part, we further investigate the non-origin servers and types of the services that they provide.

Analysis of Non-Origin Servers

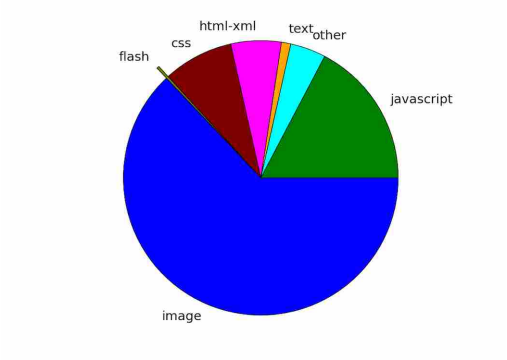
Now that the significance of non-origin servers has been cleared, we explore the type of the services that are offered by these servers. Since there are many non-origin servers that only appear in a few number of our target websites, we focus our analysis on the most frequent non-origin servers. For measuring the frequency of a non-origin server, we count the number of the websites which contact that server to receive one or more objects. In this way, we find the top 300 most frequent non-origin servers in each vantage point. Then, we use McAfee TrustedSource Web Database (McAfee, 2018) to get the service type or category of these non-origin servers.



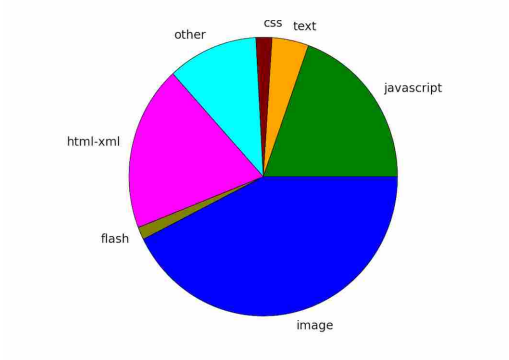
(a) Origin servers



(b) Non-Origin servers



(c) (Tian & Rejaie, 2015) - origin servers



(d) (Tian & Rejaie, 2015) - non-origin servers

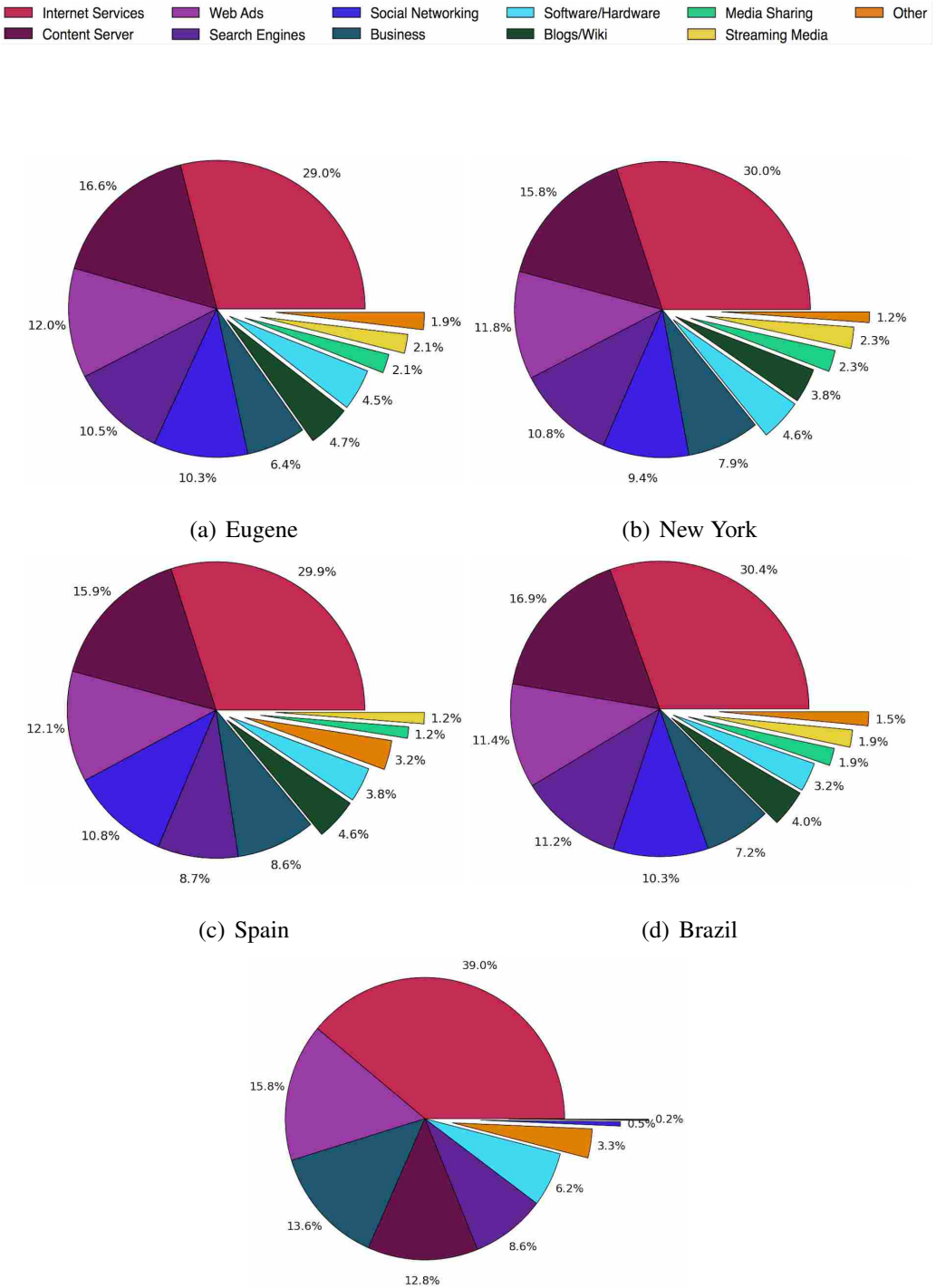
Figure 14: Distribution of different MIME types across the delivered objects by origin and non-origin servers

Once we have the most frequent non-origin servers and their type of services, in order to examine the importance of each service type (category), we obtain the target websites whose non-origin servers only include one or more of the top 300 frequent servers. After that, for these websites, we compute the fraction of the number of the objects and size of the objects which are returned by each category of servers. Figures 15 and 16 demonstrate these results for different vantage points. According to these pie charts, we have very similar results in all the vantage points except China. Due to the fact that many servers are blocked in China, the contribution of categories such as Social Networking and Blogs/Wiki are much less than other vantage points and also, Media Sharing and Streaming Media categories provide no objects for target websites. Furthermore, for all the vantage points, non-origin servers in the ”Internet Services” category have the most contribution to the number of the objects and for all the vantage points except Eugene and New York, they have the most contribution to the size of the objects. As stated in the

category description of the McAfee TrustedSource Web Database (McAfee, 2018), Internet Services category includes services for publication and maintenance of websites such as web design, statistics and access logs, domain registration, internet service providers and broadband and telecommunications companies that provide web services.

Taking the number of the objects into account, the next most important categories are Content Servers and Web Ads. However, with regard to the size of the objects, categories like Social Networking, Blogs/Wiki, Media Sharing, and Streaming Media become more influential. More specifically, servers offering Media Sharing and Streaming Media services have a large contribution to the size of the objects in spite of their small contribution to the number of the objects. These results is different from the findings of (Butkiewicz et al., 2011) in a couple of ways. First, content servers provide more number of objects than advertising servers (except for China). Second, Social Networking servers constitute larger fraction of objects and bytes of data. Third, Content Servers do not dominate the number of bytes anymore and on average, Internet Services provide about the same fraction of the number of bytes.

It should be noted that considering the top 300 frequent non-origin servers, the exact list of these servers and their distribution across various categories is slightly different for each vantage point. Tables 6 through 10 indicate the number of the frequent servers that belong to each category as well as the number of the target websites that contact at least one of the frequent servers in that category. According to these tables, the results are very similar in all the vantage points except China, in which there are much fewer number of non-origin servers having categories such as Social Networking and Search Engines. However, the major categories across all the vantage points are the same which are: Internet Services, Content Server, Web Ads, Business, Software/Hardware, Search Engines, and Social Networking. These results are similar to the (Butkiewicz et al., 2011) findings in 2011. Based on their results, the major type of services that non-origin servers provide were: Analytics, Advertising, Tracking Cookies, Services/Widgets, CDN, Social Networking, and Programming API. In our results, Analytics and Services/Widgets services are mostly included in the Internet Services category and Tracking Cookies and



(a) Eugene

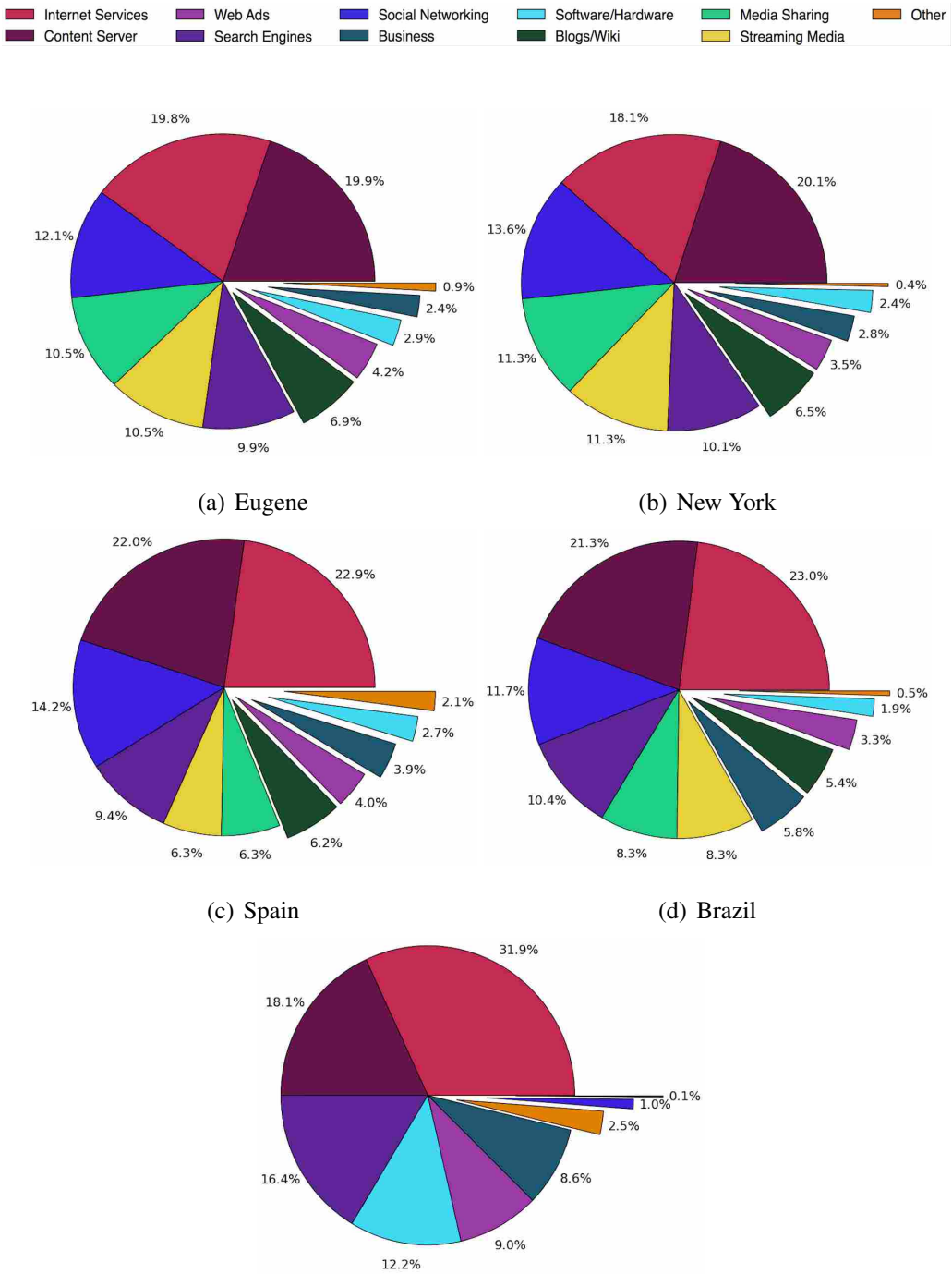
(b) New York

(c) Spain

(d) Brazil

(e) China

Figure 15: Distribution of number of the objects across different categories of non-origin servers



(a) Eugene

(b) New York

(c) Spain

(d) Brazil

(e) China

Figure 16: Distribution of size of the objects across different categories of non-origin servers

Programming API fall under the Software/Hardware and Business categories. An interesting point about the Search Engines servers is that despite their small number, they appear in many of the target websites. Some examples of these servers are: Google.com, Bing.com, and Baidu.com.

Category	Number of non-origin servers	Number of target websites
Internet Services	102	297
Business	69	84
Content Server	46	208
Web Ads	45	152
Software/Hardware	39	86
Social Networking	10	86
Blogs/Wiki	6	43
Search Engines	5	150
Online Shopping	3	5
Marketing/Merchandising	3	12
Streaming Media	3	28
Media Sharing	2	28
Portal Sites	2	20
Professional Networking	2	3
Interactive Web Applications	1	1
Total	**300**	**313**

Table 6: Eugene - Categories of Non-Origin Servers

Category	Number of non-origin servers	Number of target websites
Internet Services	105	285
Business	65	79
Content Server	44	203
Web Ads	43	152
Software/Hardware	38	81
Social Networking	10	82
Blogs/Wiki	6	38
Search Engines	5	146
Online Shopping	5	7
Streaming Media	4	25
Marketing/Merchandising	3	8
Media Sharing	2	25
Portal Sites	3	22
Professional Networking	2	2
Interactive Web Applications	1	3
Total	**300**	**299**

Table 7: New York - Categories of Non-Origin Servers

Category	Number of non-origin servers	Number of target websites
Internet Services	104	300
Business	68	95
Content Server	42	234
Web Ads	39	164
Software/Hardware	38	86
Social Networking	10	100
Blogs/Wiki	6	50
Search Engines	6	172
Marketing/Merchandising	6	27
Online Shopping	5	7
Portal Sites	3	16
Streaming Media	2	25
Media Sharing	2	25
Professional Networking	2	1
Games	1	5
Interactive Web Applications	1	2
General News	1	1
Total	**300**	**319**

Table 8: Spain - Categories of Non-Origin Servers

Category	Number of non-origin servers	Number of target websites
Internet Services	103	318
Business	68	94
Content Server	40	243
Web Ads	40	164
Software/Hardware	40	85
Social Networking	10	95
Blogs/Wiki	6	48
Search Engines	6	177
Streaming Media	4	30
Marketing/Merchandising	4	11
Media Sharing	4	30
Online Shopping	4	8
Portal Sites	2	19
Professional Networking	2	3
Games	1	4
Interactive Web Applications	1	3
Pornography	1	3
Total	**300**	**346**

Table 9: Brazil - Categories of Non-Origin Servers

Moreover, we find the URLs of the top 10 most frequent non-origin servers to understand what these servers are exactly and how frequent they are among target websites in each vantage point. We also look into the type of the services which these servers provide. Tables 11 through 15 represent these results. As it can be seen in these tables, the top 10 most frequent non-origin servers are almost the same

Category	Number of non-origin servers	Number of target websites
Internet Services	109	186
Business	71	69
Web Ads	45	109
Software/Hardware	39	68
Content Server	33	124
Online Shopping	5	11
Marketing/Merchandising	4	15
Portal Sites	4	14
Social Networking	3	20
Search Engines	3	61
Blogs/Wiki	3	5
Professional Networking	2	1
Games	1	4
Interactive Web Applications	1	3
Forum/Bulletin Boards	1	1
Entertainment	1	1
Total	**300**	**200**

Table 10: China - Categories of non-origin servers

in all the vantage points. Again, the only exception is China for which servers such as ”facebook.com”, ”facebook.net”, and ”google.com” are not among the top frequent non-origin servers probably due to the access restrictions. Furthermore, servers in ”Internet Services” category compose 50% of the top 10 most frequent servers. Comparing these results with the results from (Butkiewicz et al., 2011), half of the top 10 most popular non-origin servers has stayed the same which are: google-analytics.com, doubleclick.net, quantserve.com/scorecardresearch.com, facebook.com, googleapis.com (all of our vantage points except Spain have quantserve.com in their list of top 10 non-origin servers, however, Spain has scorecardresearch.com instead). Considering the other half, googleadservices.com, scorecardresearch.com, and 2mdn.net are still among the top 40 servers; atdmt.com is among the top 150 servers; and yieldmanager.com is not among the top 300 servers anymore.

Rank	Name	Fraction of Websites	Service Type
1	google-analytics.com	0.66	Internet Services
2	doubleclick.net	0.56	Web Ads
3	googleapis.com	0.49	Internet Services
4	gstatic.com	0.46	Content Server-Internet Services
5	facebook.com	0.45	Social Networking
6	google.com	0.44	Search Engines
7	facebook.net	0.41	Social Networking
8	googlesyndication.com	0.33	Search Engines
9	googletagservices.com	0.30	Internet Services
10	quantserve.com	0.28	Internet Services

Table 11: Eugene - most frequent non-origin servers

Rank	Name	Fraction of Websites	Service Type
1	google-analytics.com	0.65	Internet Services
2	doubleclick.net	0.56	Web Ads
3	googleapis.com	0.48	Internet Services
4	gstatic.com	0.46	Content Server-Internet Services
5	facebook.com	0.44	Social Networking
6	google.com	0.44	Search Engines
7	facebook.net	0.41	Social Networking
8	googlesyndication.com	0.33	Search Engines
9	googletagservices.com	0.29	Internet Services
10	quantserve.com	0.27	Internet Services

Table 12: New York - most frequent non-origin servers

Rank	Name	Fraction of Websites	Service Type
1	google-analytics.com	0.67	Internet Services
2	doubleclick.net	0.58	Web Ads
3	googleapis.com	0.46	Internet Services
4	facebook.com	0.46	Social Networking
5	gstatic.com	0.45	Content Server-Internet Services
6	facebook.net	0.42	Social Networking
7	googlesyndication.com	0.33	Search Engines
7	google.es	0.32	Search Engines
9	googletagservices.com	0.30	Internet Services
10	scorecardresearch.com	0.27	Business

Table 13: Spain - most frequent non-origin servers

Rank	**Name**	**Fraction of Websites**	**Service Type**
1	google-analytics.com	0.64	Internet Services
2	doubleclick.net	0.55	Web Ads
3	googleapis.com	0.48	Internet Services
4	gstatic.com	0.47	Content Server-Internet Services
5	facebook.com	0.43	Social Networking
6	facebook.net	0.39	Social Networking
7	googlesyndication.com	0.31	Search Engines
7	google.com.br	0.31	Search Engines
9	googletagservices.com	0.28	Internet Services
10	quantserve.com	0.26	Internet Services

Table 14: Brazil - most frequent non-origin servers

Rank	Name	Fraction of Websites	Service Type
1	google-analytics.com	0.63	Internet Services
2	doubleclick.net	0.59	Web Ads
3	googleapis.com	0.36	Internet Services
4	gstatic.com	0.33	Content Server-Internet Services
5	googlesyndication.com	0.33	Search Engines
6	quantserve.com	0.30	Internet Services
7	googletagservices.com	0.30	Internet Services
8	scorecardresearch.com	0.26	Business
9	quantcount.com	0.24	Internet Services
10	cloudfront.net	0.24	Content Server

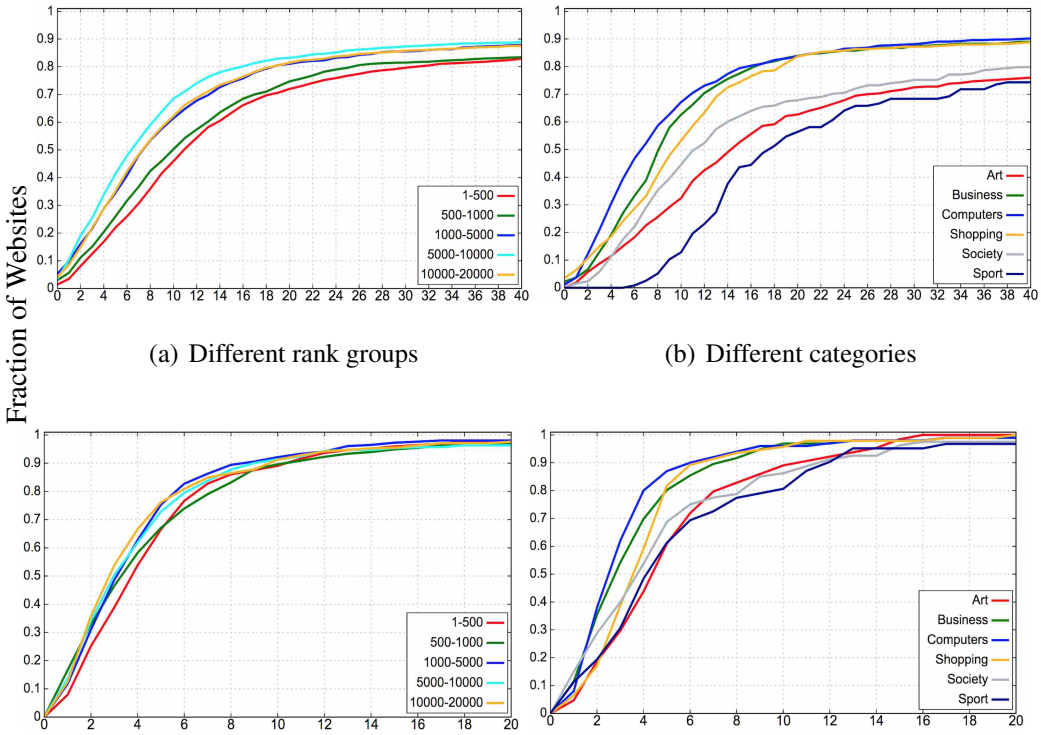
Table 15: China - most frequent non-origin servers

CHAPTER VI

PAGE LOAD TIME ANALYSIS

In this section, we examine how the complexity of the target websites affects their load time, which is an important factor that influences the user experience. In order to measure the page load time, we use the information in the HAR files. In the HAR files, we have the start time and the execution time of each request. Therefore, we obtain the load time of the page by computing the time difference between the start time of the first request and the finish time (start time plus the execution time) of the last request. Moreover, before calculating page load times, we conduct further data cleaning. The reason is that, although the HAR export trigger extension on Firefox is supposed to export the HAR file when there is no request for 2.5 seconds after the last finished request, in a number of HAR files, this is not the case. Fraction of these files is less than 6% in all the vantage points except China. However, there is a significant number of these files in China which compose around 45% of all the files. In these HAR files, we find the request that is sent more than 2.5 seconds after the last finished request and then, in the page load time computation, we ignore that request and all the proceeding ones. Figures 17(a) and 17(b) show the CDF of the page load times for different ranks and categories of websites. According to these Figures, almost 80% of the target websites in different rank groups have a load time less than 20 seconds. This load time is more than two times larger than our previous results (Tian & Rejaie, 2015) and also the (Butkiewicz et al., 2011) results, although in the last two years, the number of the requested objects and the number of the contacted servers have stayed near the same. Our previous results (Tian & Rejaie, 2015) are demonstrated in Figures 17(c) and 17(d).

An important point is that what fraction of objects and bytes are downloaded within the first few seconds of the page load time. Considering all the target websites across all the vantage points, Figures 18(a) and 18(b) indicate the percentage of number of the objects and size of the objects which are loaded within the first 10 seconds of the page load time. Based on these findings, for 77% of the websites, all the objects are loaded within 10 seconds. For the remaining websites,



(a) Different rank groups

(b) Different categories

(c) (Tian & Rejaie, 2015) - different rank groups

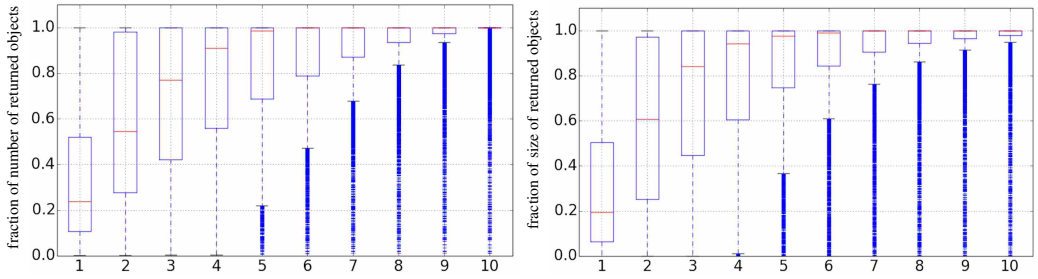
(d) (Tian & Rejaie, 2015) - different categories

Figure 17: Page load times (seconds)

on average, almost 80% of their objects are loaded within the first 10 seconds. Regarding the size of the objects, for 70% of the websites, all the bytes of data are loaded in 10 seconds and for the rest of them, on average, again almost 80% of the bytes are loaded within 10 seconds. These results imply that the long load times of some of the websites would barely be experienced by the users because they can access the majority of the content of a website during the first few seconds after visiting the page. In the following section, we investigate the reasons behind having some websites with a very long load time.

Websites With Long Page Load Times

In this part, we look into the websites with a page load time larger than 10 seconds, which we will refer to as problematic websites in the rest of this paper. These websites constitute approximately 38% of the target websites across all the vantage points and table 16 shows the fraction of these websites in each vantage point. Based on our further investigation, there are three main reasons that contribute to the long load times of these websites. The first one is the execution



(a) First 10 seconds (b) First 10 seconds

Figure 18: Fraction of number of returned objects and size of returned objects within the first 10 seconds of the page load time

time of the requests, the second one is the total number of the returned objects, and the third one is the total number of the contacted servers.

The execution time of each request comprises the following values:

- Blocked time: The time that the request spends in a queue waiting for a network
- DNS time: The DNS resolution time
- Connect time: The required time to create TCP connection
- Send time: The required time to send the request to the server
- Wait time: The waiting time for a response from the server
- Receive time: The required time to read the entire response from the server

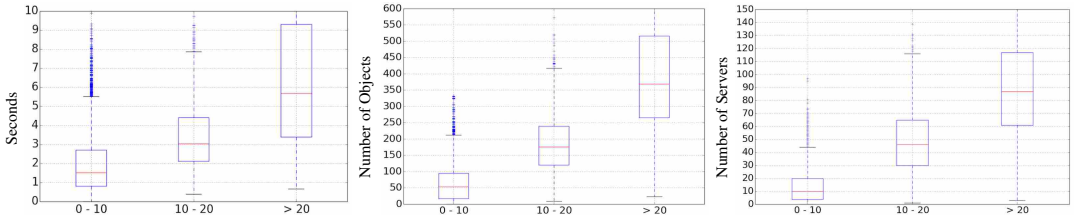
A lengthy delay in any of these timing values will increase the execution time of a request and subsequently, the page load time. We compute the ratio of the page load time over the maximum execution delay among all the requested objects in that page. If this value is less than 2, it implies that the maximum delay is probably the main reason for the long page load time. Moreover, it should be noted that the number of the websites having this criterion is different in each vantage point, but nearly 12% of all the target websites in all the vantage points have this condition. Table 16 presents the following information for each VP: The percentage of the problematic websites, the percentage of websites with a very long execution delay in executing at least one of the requests, and number of the times that each timing parameter was the main reason for the long execution delay. The parameter that contributes the most to the maximum delay (has the largest delay among all parameters) is identified as the main reason for the delay.

VP	Problematic Websites	Large Delay in Processing a Request	Blocked	DNS	Connect	Send	Wait	Receive
Eugene	31%	3%	0%	15%	28%	7%	43%	7%
New York	34%	14%	17%	2%	14%	0%	47%	20%
Spain	28%	3%	20%	6%	0%	0%	45%	29%
Brazil	48%	16%	27%	3%	1%	0%	31%	38%
China	47%	26%	69%	13%	1%	0%	3%	14%
All VPs	**38%**	**12%**	**27%**	**8%**	**9%**	**1%**	**34%**	**22%**

Table 16: Fraction of problematic websites and websites with large delay in processing a single request along with the contributing parameters to the delay

Now, putting aside this 12% and considering the rest of the problematic websites, we can observe that in addition to the execution delay of some of the requests, the total number of the returned objects and contacted servers definitely contribute to the long load time. The box plots in Figures 19(a), 19(b), and 19(c) demonstrate the distribution of the maximum execution delay, number of the returned objects, and number of the contacted servers across all the websites in all the vantage points, respectively. In these box plots, we categorize the target websites into three different groups based on their load times: less than 10 seconds, between 10 and 20 seconds, and more than 20 seconds. Moreover, we include all the websites except the websites mentioned in the previous paragraph since we already know that the main reason for their lengthy load time is the execution delay of one or some of the requests. As it is shown in these plots, considering the websites with a load time longer than 20 seconds, the median number of the returned objects and median number of the contacted servers are approximately 7 and 9 times larger than the corresponding medians for the websites with a load time less than 10 seconds. Additionally, similar relationship exists between the medians of the maximum execution delays with the difference that the ratio is almost 4 to 1. Therefore, besides the execution delay of requests, number of the returned objects and number of the contacted servers have significant influence on the long load times of the websites which take more than 20 seconds to load.

In order to examine the mutual influence of these parameters, we group these websites based on their computed load time over maximum execution delay value (between 2 and 3, 3 and 4, and so on up to 8 and 9) and plot the box plots of the number of the returned objects and number of the contacted servers for these



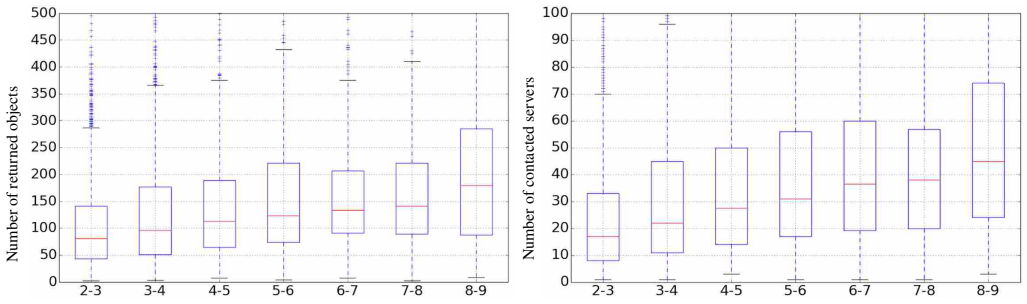
(a) Maximum delay

(b) Number of returned objects

(c) Number of contacted servers

Figure 19: Contributing factors to the long page load times

groups. Figure 20 shows these box plots which indicate that as the page load time over the maximum execution delay becomes larger, the number of the returned objects and contacted servers increases. It implies that when the max delay gets relatively smaller, there are more returned objects and contacted servers which cause the load time to still be large. Hence, we can conclude that the combination of these three factors is a very important reason for such long page load times.



(a) Different groups of websites based on ”load_time/max_delay” - distribution of returned objects

(b) Different groups of websites based on ”load_time/max_delay” - distribution of contacted servers

Figure 20: Box plots of number of the returned objects and contacted servers for different groups of websites based on their ”load_time/max_delay”

Correlation

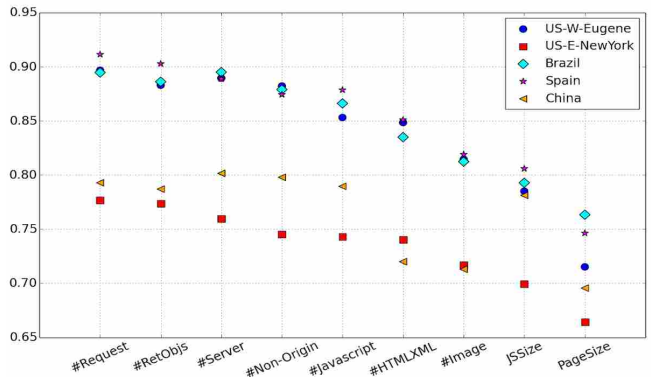
In the previous part, we saw the role of number of the returned objects and contacted servers in increasing the page load times. In this part, we aim to further analyze the correlation between various attributes of a website with its page load time. To accomplish this goal, first we compute the Spearman’s Correlation Coefficients which represent the correlation ratios between the page load time and

each attribute. In this analysis, we do not include the websites mentioned in the second column of table 16 since the main reason for their long load time seems to be the long delay in executing one of the requests. Figure 21(a) shows these coefficients for each of the vantage points. Considering all the vantage points, the most correlated metrics are: number of the requested and returned objects, number of the servers, and number of the javascript objects. Furthermore, Figure 21(b) indicates our previous results (Tian & Rejaie, 2015). Although the correlation ratios are larger in our new results, the most correlated attributes are very similar to our earlier results (Tian & Rejaie, 2015). A significant difference is that number of javascript objects and number of servers have become more correlated than number of image objects. Moreover, number of the returned objects, which is a new attribute defined in our study, has a high correlation with the page load time similar to number of the requested objects. Returned objects are a subset of the requests (on average, around 80% of the requests for each website) that have a status code of 200 which shows that the browser successfully received the object. Also, in the results reported by (Butkiewicz et al., 2011) in 2011, the correlation ratios were close to our new findings and the most correlated metrics were total number of objects, number of javascript objects, and the total page size. Thus, the only difference is that in the new results, number of the servers has become more correlated and the total page size has become less correlated.

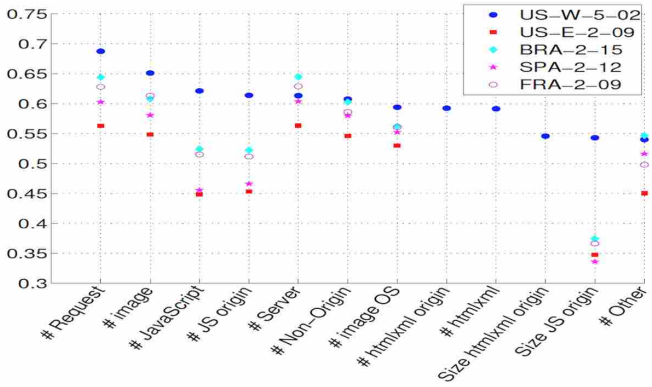
Moreover, in Figure 22, we take a closer look at the relationship between the number of returned objects and the page load time by grouping the websites based on their number of returned objects and drawing the box plots of the page load time values for each group.

Then, based on the Spearman's Correlation Coefficients, we select the features with the most correlation as the input features for our regression model. These features are:

- Number of requested objects
- Number of returned objects
- Number of non-origin servers
- Number of all contacted servers
- Number of javascript objects



(a) Current results



(b) (Tian & Rejaie, 2015)

Figure 21: Spearman's Correlation Coefficients

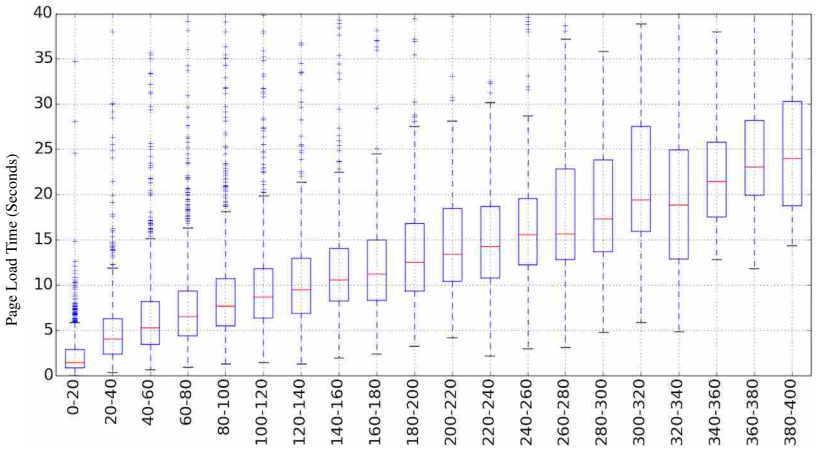


Figure 22: Correlation between number of returned objects and page load time

- Number of image objects

- Number of HTML-Xml objects

- Size of javascript objects

- Total page size

The purpose of training the regression models is to predict the page load times with a decent accuracy and then use the importance of different features among the models to find out whether the more correlated features according to the Spearman's Correlation Coefficients are also more important and effective in predicting the page load times using the regression models.

Regression

Now that we have gained some insight into how different website metrics are correlated with the page load time, we can use those metrics in training a regression model to predict the page load times. In order to get good results in predicting the load times, we take the following steps:

- Removing Outliers: After examining the information of each website in our dataset, we realized that there are some data points that the relationship between their load time and other metrics does not comply with the general data. These websites are the ones explained in table 16, whose delay in receiving one or a few number of objects causes their load time to be very long. Hence, none of the considered features in our regression models really affects their load time. Therefore, we removed these websites (which have a "load_time/max_delay" value of less than 2) from our dataset before training the regression model.

- Scaling Data: Since the distribution of the page load times and all the features are skewed, we need to scale the data to get more accurate prediction results. A good scaling method is using the logarithmic transformation both on the features and the load times. In this way, the data will shrink and the skewed values will have less influence on the final prediction error. It should be noted that predicting the logarithm of page load times does not affect the validity of our results since in predicting the page load times, we care about the relative error between the predicted and real values not the exact difference.

- Feature Engineering and Training the Models: Although we selected the initial features based on the Spearman's Correlation Coefficients, we further

examined different combinations of features by taking the other complexity metrics into account. For instance, we tried adding features like: number of the css objects, percentage of objects returned by origin or non-origin servers, and also nonlinear terms like the squared of existing features. Nevertheless, still the initial set of features gave us the best accuracy. Moreover, we used a number of different regression models such as: Linear Regression, Ridge and Lasso Regression, Elastic Net Regression, and Random Forest Regressor. The reason behind selecting linear models is that after looking at the scatter plots of page load time against different features, we observed a linear like relationship between the load time and the features. Moreover, as we do not have a very large dataset, more complex nonlinear models did not perform well on our dataset because of overfitting. In addition, we selected Random Forest Regressor as it has a good accuracy on our dataset and can be used for identifying important features.

After training the models and testing them on our dataset, all of them had a very similar accuracy. Table 17 shows the R-squared score and the Root Mean Square Error (RSME) for each model. As we can see in this table, the Lasso Regression performs slightly better than the other models. Figure 23 shows the scatter plot of the predicted load times (using Lasso Regression) against the actual load times and illustrates how the linear model fits the data.

Model	R-Squared	RMSE
Linear Regression	0.7757	0.4091
Ridge Regression	0.7760	0.4024
Lasso Regression	0.7762	0.4020
Elastic Net Regression	0.7759	0.4022
Random Forest Regressor	0.7753	0.4096

Table 17: Performance of different regression models

Now that we have the trained models, to identify the more important features, we can look at the coefficients of the features, in Linear Regression, Ridge and Lasso Regression, and Elastic Net Regression, and also the feature importance values given by the Random Forest model. It should be noted that since the Elastic Net model is a combination of the Ridge and Lasso models, its coefficients is very

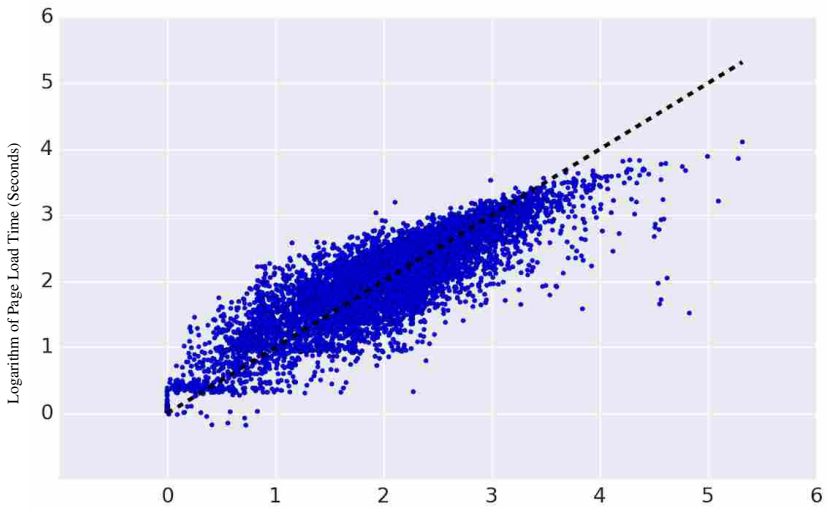
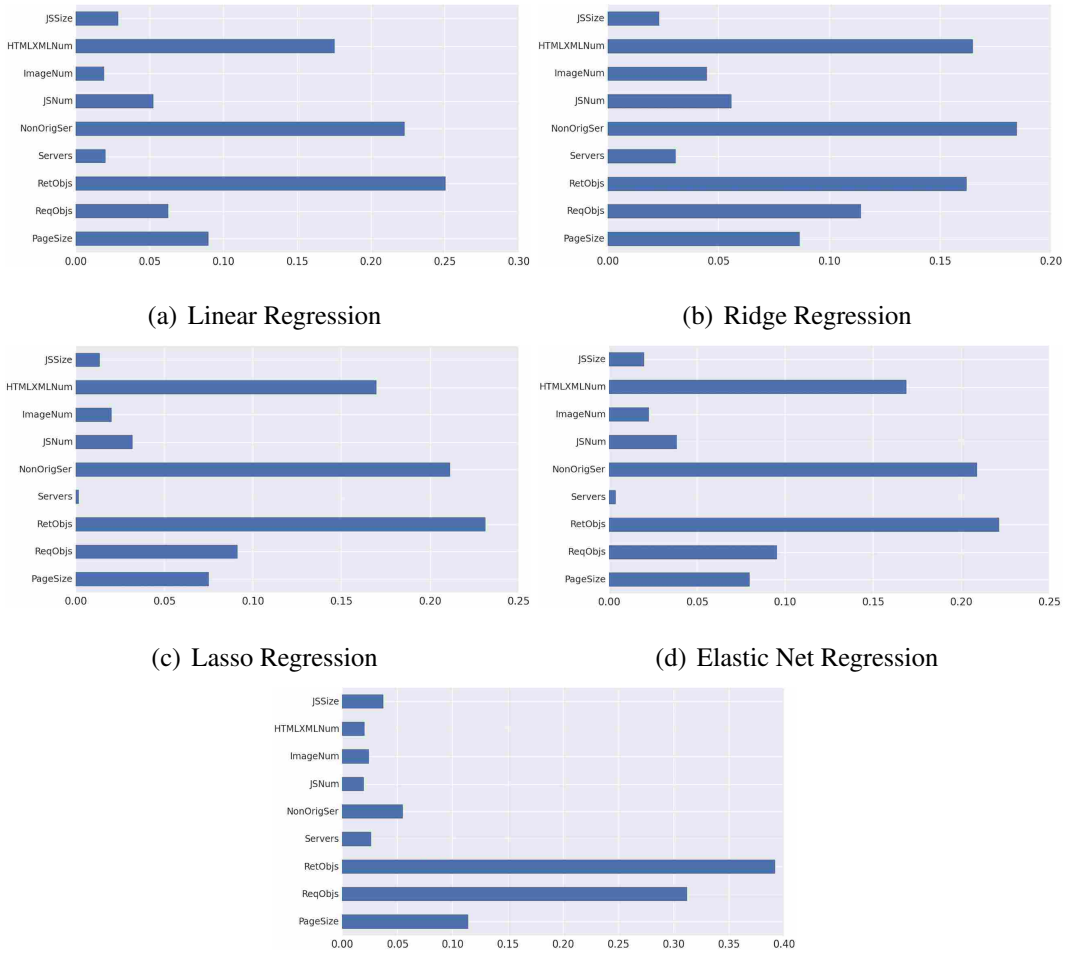


Figure 23: Predicted load times against actual load times

similar to those models. Figure 24 demonstrates the feature importances. Considering all the models except the Random Forest, the three most important features are: number of returned objects, number of non-origin servers, and number of HTML-XML objects. However, for the Random Forest model, these three features are: number of returned objects, number of requested objects, and the total page size.

Generally, these results confirms the correlations that we found in the previous part and also reveals a couple of interesting points. First, number of the HTML-XML objects are more important in our regression models (except Random Forest Regressor) than some metrics such as number of the javascript objects or number of the requested objects which have a larger correlation ratio than number of the HTML-XML objects. Second, although the Spearman's Correlation Coefficient of the total page size is smaller than all the other selected features, it is among the top most important features in all the models. Moreover, these findings prove our claims in the previous part regarding the main reasons that contribute to the long load time of the problematic websites. In fact, based on the feature importances, we realize that number of the non-origin servers is the key factor rather than the total number of the servers. This is also expected due to the fact that non-origin servers form the majority of the contacted servers by each website.

According to the (Butkiewicz et al., 2011) results, number of the requested objects and number of the servers were among the three top features in their



(a) Linear Regression

(b) Ridge Regression

(c) Lasso Regression

(d) Elastic Net Regression

(e) Random Forest Regressor

Figure 24: Feature importances of different regression models

regression models. Therefore, we see that these metrics are still very effective in determining the page load time. The main difference is that the number of the HTML-XML numbers and total page size has become more important than 6 years ago and number of the javascript objects has become less important. Moreover, (Butkiewicz et al., 2011) used a Lasso Regression model that performed 50% better than a naive estimator, which simply predicts the mean value of page load times, but our trained Lasso Regression model has an R-squared score of 0.77 which means it outperforms the naive estimator by 77%. Considering that most of the input features are the same in both studies, this improvement implies the importance of the number of the returned objects, which is a new feature in our study, as well as the more correlation between other selected features and the page load time than six years ago.

CHAPTER VII

COMPARISON AND CONCLUSION

In this study, we re-examined the complexity of a set of 2000 popular websites and compared the results with the earlier findings in 2011 (Butkiewicz et al., 2011) and 2015 (Tian & Rejaie, 2015). We also extended some of the analysis regarding the content complexity and service complexity of the target websites.

The most important finding is that although generally, number of the objects and number of the servers have not changed that much since two years ago, there is a significant increase in the page load time of some of the websites. Based on our further analysis, this increase is due to the large number of returned objects and contacted servers by these websites as well as the delay in processing some of the requests. In all the vantage points except China, in about 65% of the times, this delay in caused by the waiting time for the server response and the required time to read the response from the server. In China, in almost 70% of the times, this delay is induced by the time that the request is blocked and needs to wait for a network.

Another important finding is that in spite of the fact that number of the non-origin servers has even slightly decreased since two years ago, their contribution in providing the objects and bytes of data for the target websites has nearly doubled. They also still deliver more diverse object types than origin servers.

Additionally, we observed that number of the objects of a website and the page load time are still correlated with the rank and category of the website. Similar to our earlier analysis (Tian & Rejaie, 2015), higher ranked websites generally deliver more objects and have longer load times. Also, Art and Sport categories, have more objects and longer load times than other categories, same as our results from two years ago (Tian & Rejaie, 2015).

One more interesting finding is that the most correlated complexity metrics with the page load time have stayed almost the same during the last six years. There are only a couple of dissimilarities which are the more correlation of number of the servers (specifically, number of the non-origin servers) and number of HTML-XML objects and the less correlation of number of the image objects.

Furthermore, we discovered that using the data of our target websites about the

complexity features and page load times, it is possible to predict the page load times with better accuracy than six years ago (Butkiewicz et al., 2011) by training linear regression models like Ridge and Lasso regression.

Considering these results and findings, although number of the requested objects and number of the servers experienced a significant increase between 2011 and 2015 (Butkiewicz et al., 2011) (Tian & Rejaie, 2015), the former has remained nearly the same and the latter has experienced a 15% decrease since 2015. Moreover, fraction of number of the objects and number of the bytes that non-origin servers deliver has increased from one-third to one-half on average since 2015, which shows their crucial role in providing the objects of the target websites. With regard to the page load times, according to our previous study (Tian & Rejaie, 2015), they have stayed rather unchanged from 2011 to 2015, however, we observed that the median value of the page load times has been more than doubled since 2015. Although we found out that number of the objects and number of the non-origin servers are still the most effective metrics on the page load time, taking these trends in the websites' complexity and their page load times into account, an interesting future work would be to conduct additional analysis to realize whether the requests to non-origin servers are the main ones that are delaying the page load time or if there are specific non-origin servers which mostly act as the bottleneck for the page load time.

REFERENCES CITED

Butkiewicz, M., Madhyastha, H. V., & Sekar, V. (2011). Understanding website complexity: measurements, metrics, and implications. In *Proceedings of the 2011 acm sigcomm conference on internet measurement conference* (pp. 313–328).

Butkiewicz, M., Wang, D., Wu, Z., Madhyastha, H. V., & Sekar, V. (2015). Klotski: Reprioritizing web content to improve user experience on mobile devices. In *Nsdi* (Vol. 1, pp. 2–3).

Firefox. (2017). *Firefox browser.* Retrieved from `https://www.mozilla.org/en-US/firefox/`

Google-Developers. (2018). *Pagespeed insights.* Retrieved from `https://developers.google.com/speed/pagespeed/insights/`

Haralyzer-Module. (2017). *Haralyzer module in python.* Retrieved from `https://pypi.org/project/haralyzer/`

HAR-Export-Trigger. (2017). *Har export trigger extension.* Retrieved from `https://github.com/firebug/har-export-trigger`

Kelton, C., Ryoo, J., Balasubramanian, A., & Das, S. R. (2017). Improving user perceived page load times using gaze. In *Nsdi* (pp. 545–559).

Ludin, S. (2017). Measuring what is not ours: A tale of 3rd party performance. In *Passive and active measurement: 18th international conference, pam 2017, sydney, nsw, australia, march 30-31, 2017, proceedings* (Vol. 10176, p. 142).

McAfee. (2018). *Mcafee trustedsource web database.* Retrieved from `https://www.trustedsource.org/`

Netravali, R., Goyal, A., Mickens, J., & Balakrishnan, H. (2016). Polaris: Faster page loads using fine-grained dependency tracking. In *Nsdi* (pp. 123–136).

Odvarko, J. (2017). *Har 1.2 spec.* Retrieved from `http://www.softwareishard.com/blog/har-12-spec/`

Selenium-WebDriver. (2017). *Selenium webdriver.* Retrieved from `https://www.seleniumhq.org/projects/webdriver/`

Tian, R., & Rejaie, R. (2015). Re-examining the complexity of popular websites. In *Hot topics in web systems and technologies (hotweb), 2015 third ieee*

workshop on (pp. 25–30).

Wang, X. S., Balasubramanian, A., Krishnamurthy, A., & Wetherall, D. (2013). Demystifying page load performance with wprof. In *Nsdi* (pp. 473–485).

WebpageTest. (2018). *Webpagetest framework.* Retrieved from `http://www.webpagetest.org/`