Re-examining the Complexity of Popular Websites

Ran Tian and Reza Rejaie
{mty0083,reza}@cs.uoregon.edu
Department of Computer and Information Science
University of Oregon

***Abstract*—During the past decade, the Web has become increasingly more popular and thus more important for delivery of content and services over the Internet. At the same time, the number of requested objects, their size and delivery mechanisms for popular websites have become more complex. This in turn has various implications including the impact on page loading time that directly affects the experience of visiting users. Therefore, it is important to capture and characterize the complexity of popular web pages. An earlier study by Butkiewicz *et al.* [1] have characterized the complexity of 1700 popular pages in 2011.**

In this study, we adopt the methodology proposed by Butkiewicz *et al.*, develop the required tools and conduct a detailed measurement study to re-assess the complexity of 2000 popular web pages and present any observed trends in their complexity characteristics over the past four years. Our results show that the number of requested objects and contacted servers for each website has significantly increased. But a growing number of contacted servers are associated with third parties. Despite these changes, the page loading time remains rather unchanged and it is primarily affected by the same key parameters. Overall, our results sheds a useful light on trends in web site complexity and motivates a range of issues to be explored.

Keywords*—*Web page, Measurement

I. Introduction

During the past two decades, web has become increasingly more popular and thus more important for delivery of content and services to a wide range of users across the world. This in turn has caused these web pages to become increasingly more complex both in terms of the number and type of delivered objects as well as the number of contacted servers with different administrative domains. For example, the landing page of today's popular websites often consists of hundreds of objects including images, videos and scripts that are delivered not only from servers hosted by its provider but also from third party services (*e.g.,* advertising agencies) and content distribution networks (CDNs). The complexity of website could directly affect their loading times and thus impact the experience of visiting users. While different aspects of web traffic have been examined by prior studies [2], [3], the complexity of web pages have received little attention in recent years. Butkiewicz *et al.* [1] conducted a detailed measurement study on the complexity of 1700 popular websites and presented a range of their characteristics in 2011. The evolving nature of web pages and the associated services raise the following basic question: *Whether and how the basic complexity characteristics of popular web pages have changed during the past four years?"*

In this paper, we tackle this important question. Toward this end, we adopt the proposed methodology by Butkiewicz *et al.* [1] to characterize the complexity of 2000 popular websites across the Internet. We develop a crawler for automated browsing and related tools for data collection, parsing, cleaning and analysis. Using collected data from a few scattered vantage points across the world, we repeat the key analysis conducted by Butkiewicz *et al.* [1] and present the main changes in content and service complexity of popular websites over the past four years. In particular, we consider relative ranking and category of individual pages as well as MIME type of requested objects for each website in our analysis. Some of our main findings can be summarized as follows: *(i)* The number of requested objects for all web page have at least doubled over the past four years regardless of their type or rank. Furthermore, there is a clear gap in the number of requested objects by web pages with different ranks or categories that has widen over the past few years. *(ii)* The distribution of the number of object of certain MIME type exhibits only minor change over the past 4 years. *(iii)* The content associated with individual websites are fetched from a much larger number of servers but a larger fraction of these servers are managed by third parties. *(iv)* Despite the growing content and service complexity of websites, their loading times exhibits only a negligible increase. While the main factors that affect the loading time of individual web pages are still the number of request, servers and images, these factors have lower correlation with page load than four years ago. While these findings may not be surprising, our work is the first one that systematically quantifies these trends in web pages complexity over the past four years.

The rest of this paper is organized as follows: In Section II, we present an overview of our methodology. Section III describes our datasets. We present our content complexity analysis in Section IV and service complexity analysis in Section V. The implications of website complexity on page load time is discussed in Section VI. Finally, Section VII concludes the paper and sketches some of our future plans.

II. Methodology

This section presents the key aspects of our methodology for data collection, parsing, cleaning and analysis.

Rank Range	From Quantcast	From Alexa
1-500	500	344
500-1000	300	11
1,000-5,000	300	54
5,000-10,000	400	37
10,000-20,000	500	44
Total	2000	500

TABLE I. SELECTION OF 2000 TARGET WEBSITES FROM THE LIST OF TOP 20K QUANTCAST AND TOP 500 ALEXA WEBSITES.

Dataset	Location	Collected	Redirected	Unresolved	Real
US-5-2	Eugene, OR	90%	1%	2%	86%
US-2-09	Durham, NC	90%	1%	1%	87%
BRA-2-15	Brazil	97%	1%	2%	93%
FRA-2-09	France	94%	1%	2%	91%
SPA-2-12	Spain	96%	1%	2%	92%
CHN-2-15	China	84%	5%	12%	66%

TABLE II. DETAILS OF COLLECTED DATA FROM ALL VANTAGE POINTS

Further details about our methodology can be found in our related technical report [4].

Website Selection: To conduct our study, we need to identify a collection of websites that properly represent different levels of popularity and various categories. We use two online resources to identify these websites as follows: First, *alexa.com*, a subsidiary of Amazon, provides a ranked list of the top 500 most popular websites along with their category (*e.g.,* business, sport, shopping). Second, *quantcast.com* offers a ranked list of roughly one million websites without any category information. Given the skewed distribution of websites popularity, we only consider the top 20K websites from Quantcast list. To leverage both the category information from Alexa as well as the longer lists from Quantcast, we use the ranked lists from these two sources to sample 2000 target websites as follows: We consider Quantcast list as our reference because of its longer length, use the ranking in this list as the ranking of each website for our analysis and focus only on its top 20K websites. We then divide this 20K website into 5 groups based on their ranks as shown in the first column of Table I. Our goal is to spread our 2000 samples across these 5 rank-groups to ensure adequate representation of all websites with different ranks while having a better coverage of higher ranked groups. The second column of this table presents the division of 2000 target websites across the five rank-groups. We simply refer to these rank-groups as groups in the rest of this paper. To identify the specified number of samples from each group, we select all the 500 websites from the Alexa list to have a plenty of samples with the category information in order to conduct category-based analysis. The third column in Table I shows how these 500 websites from Alexa are mapped to different groups based on the Quantcast ranking. The rest of samples for each group are randomly selected from the unselected websites in that group. *In summary, the selected 2000 sample websites include all the top 500 Alexa as well as random samples of websites with different range of ranking from the Quantcast list. We refer to these websites as target websites.*

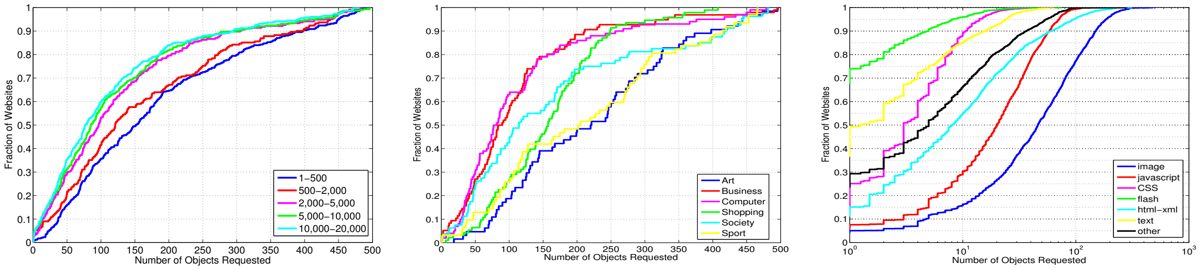
Data Collection: We developed a crawler for automated browsing of target websites. Our crawler uses a library to operate JAVA based scripts and manage the built-in version of Firefox [5] provided by Selenium WebDriver [6]. We then use the Firebug extension (version 2.0.2) [7] in addition to Firestarter (version 0.1a6) [8] and NetExport (version 0.9b6) [8] to produce the HTTP archive record (HAR) files [9]. These HAR files follow JSON format (a popular NoSQL data format) and contain all the details of requests/responses between the browser and each target website (*e.g.,* timing information and MIME type of each requested object). Using our crawler, we download the main (*i.e.,* landing) page of all the target websites that results in a separate HAR file for each website. Finally, we use our parser to extract all the desired information from the HAR file for our analysis.

Our request to a particular website might be redirected to another website. One scenario for such a redirection is where a website has a separate domain name for each country. For example, a request for *google.com* from a client in Spain is redirected to *google.es*. Since these URLs often represent separate websites in our list, we consider the captured characteristics for assessing the final website, *i.e., google.es* in the above example. It is worth noting that our crawling technique works on primitive pages that do not contain any credentials or confidential information, as an example *facebook.com* is not a primitive page.The download time of a website might slightly vary with time or location of a client. For example, clients at different location may have a different relative connectivity to the local server or their customized version have different content (*e.g.,* different local ads). To explore these issues, we run our crawler multiple times at six geographically scattered vantage points (VPs).

One basic challenge for automated crawling of targeted web pages is to determine that the page download has completed. For example, some web pages contain javascripts that periodically send new requests and receive update from the servers. Therefore page load may not conclude. In our crawler, the Firebug requires a timeout (τ) to determine the completion of a page download. If the browser does not send any request for τ seconds after the last response from the server, the Firebug concludes the completion of page loading and outputs the HAR file. Using a long timeout value may result in a long (or infinite) download time whereas small timeout value could result in a premature termination of data collection. In the absence of any known good practice for setting the timeout value, we empirically set the timeout value to 10 seconds.

III. Data Set

We perform our data collection from six geographically scattered vantage points (VPs), two in the US (east and west coast), two in Europe (Spain and Germany), one in



(a) Grouping web pages by rank (b) Grouping web pages by category (c) Object of each MIME type per page

Fig. 1. Content complexity (number of objects and their MIME type) across target web pages.

China and one in Brazil. Table II summarizes the main characteristics of collected datasets. Four rounds of measurements were conducted from our main VP (at Eugene, OR) and a single round from all other VPs. Each VP runs the crawler with the same list of target websites but using a different random order to visit the websites. Different columns in Table II show the (average) percentage of pages for which a HAR file was generated (labeled Collected), the percentage of pages that were redirected to another existing page in the list and thus excluded (labeled Redirected), and the percentage of pages whose HAR file resulted in parsing error (labeled Unresolved). The last column shows the percentage of web pages that were successfully crawled and considered for our analysis. Excluding the results from VP in China, the success rate from VPs inside and outside of the US is more than 86% and 91%, respectively. The success rate from our VP in China is much lower (66%). We believe that this drop in the success rate is due to the known blocking of access to certain web pages from China. Overall, our analysis is clearly biased by focusing on reachable web pages.

Categories: As we stated earlier, we collect the category information for a subset of target websites from Alexa. Alexa has 17 different categories and provides the top 500 most popular websites for each category in addition to the overall top 500 pages and their categories. We collect all these 18 lists from Alexa. We check each target web pages in our study against these 18 lists to determine their category whenever the information is available. Some pages are associated with more than one category by Alexa. For these pages, we used their most popular category in order to focus our analysis around a smaller number of more popular categories. The top 6 categories among the target web pages that we use for our analysis are as follows: Art, Business, Computer, Shopping, Society, and Sport.

IV. Content Complexity

In this section, we characterize the complexity of requested objects by target websites that represent a rather wide range of ranks (1-20K) and 6 popular categories. Furthermore, we compare our results with the result of similar analysis that were performed by Butkiewicz *et al.* in 2011 [1] to illustrate the trend in different aspects of complexity over the past four years.

Number of Requested Objects: We start by examining the number of requested objects associated with the front page of each targeted website. Figure 1(a) and 1(b) show the CDF of the number of objects associated with individual target pages that are grouped based on their rank and category, respectively. While the number of requested objects exhibits a rather wide range, these distributions based on page rank and category are clearly separated. Figure 1(a) reveals that the number of requested objects generally increases with the popularity (*i.e.,* rank) of a site. In particular, the median number of objects for the top two groups is 120-150 while for other groups drops to around 70-100. The top 20 percentile of pages in high (low) rank groups request more than 270-300 (180-200) objects. Examination of the number of requested objects across pages with the same category in 1(b) indicates that pages related to Sport and Art typically contain the largest number of objects (200). Web pages related to Shopping typically request 150 objects while pages related to Society, Business and Computer often include a smaller number of objects. Comparing this result with similar analysis by Butkiewicz *et al.* reveals two interesting points: *First, the top categories in our analysis are different from those reported by Butkiewicz* et al.. *Furthermore, the number of requested objects for all groups have at least doubled over the past four years. Second, the gap in the number of requested objects by different groups (based on both rank or category) have also widens in this period.*

MIME Types: We now take a closer look at the requested objects based on their MIME type. Table III shows the definition of different MIME types according to the official registration by IANA. We note that there are more than 100 different official MIME types in addition to experimental ones. These two pie charts in Figure 2 present the fraction (*i.e.,* normalized composition) of all delivered objects and bytes for all websites across different MIME types, respectively. These charts show two interesting points as follows: First roughly half of all objects and bytes are associated with images. The next largest contribution is from javascripts that make up roughly 20% of delivered objects and bytes. Second, we observe that the fraction of bytes associated with text-xml and html objects are much smaller than their fraction of objects while the situation for flash and other objects are the opposite. This simply

Type Name	MIME label in HTML
image	$image/*$
javascript	$*/javascript$
CSS	$*/css$
Flash	$*/x-flv$ or $*/x-shockwave-flash$
html-xml	$application/xml$
text	$text/*$
other	$Unknown$

TABLE III. MIME TYPE SPLIT AND NOTES

indicates that text-xml and html objects are much smaller than flash and other objects [1].

Figure 1(c) presents the CDF of the number of requested objects of certain MIME types across target websites using log scale for x axis. In essence, this figure shows a "page level" view of collected objects. The number of objects of certain MIME types widely varies across different websites. The median number of images (javascripts) are 50 (20) while this number for other object types is less than 10. Examination of object size indicates that the main deriving factors for the size is the number of requested objects of each type. Contrasting our results with the earlier study shows a few interesting differences and similarities: (i) *The distribution of the number of object of certain MIME type exhibits only minor change over the past 4 years. However, the total size of requested images per website has increased by more than 2 orders of magnitude.* (ii) *Both the number of objects and the aggregate size of flash and javascript objects in each website have grown roughly by an order of magnitude.*

We have also examined the distribution of the number and total size of objects with the four major MIME types (image, javascript, css, flash) across different rank-group and categories to identify any major relationship between web page rank/category and the MIME type of their objects. For example, each plot in Figure 3 shows the number of requested objects for individual web pages that are of certain MIME type among web pages of specific category. We can see that Web pages related to Sport, Art and Shopping contain more images and javascripts than other categories. Interesting, the fraction of image and javascript objects across all categories are similar. Some of our other results from this examination are as follows: *(i)* Higher ranked web pages generally have more images and javascripts. Despite larger number of image objects for higher rank websites, the total size of images across different ranks are similar (*i.e.,* higher rank pages have a larger number of smaller-sized images). However, more javascript in those pages results in larger size in higher rank websites. These results are available in the related technical report [4]

V. SERVICE COMPLEXITY

We turn our attention to service complexity by examining the number and role of servers contacted for delivered content to each website. We note that requested objects from a website are often delivered from multiple servers. Some of these servers are related to the same domain name and are used to ensure the scalability and robustness of its provided services. Figure 4(a) shows the distribution of the number of contacted servers by individual websites grouped by rank. We observe that the number of contacted servers for websites with different ranks are clearly different. For instance, the typical number of contacted servers by the two highest ranked groups is around 30-40. Interestingly, comparing these results with Butkiewicz *et al.* [1] illustrates that the number of contacted servers for all groups have more than doubled over the past four years.

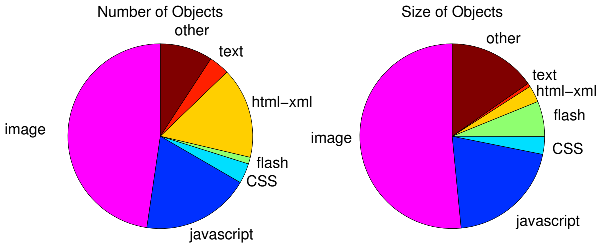


Fig. 2. Composition of MIME types according to all objects

Origin vs Non-Origin Servers: To gain more insights into the service complexity of target websites, we adopt the following methodology [10] to group involved servers into *origin* and *non-origin* based on their authoritative DNS servers:

- **Origin Servers**: Any server whose authoritative DNS server is the same as the target website, is considered an origin server of the website.

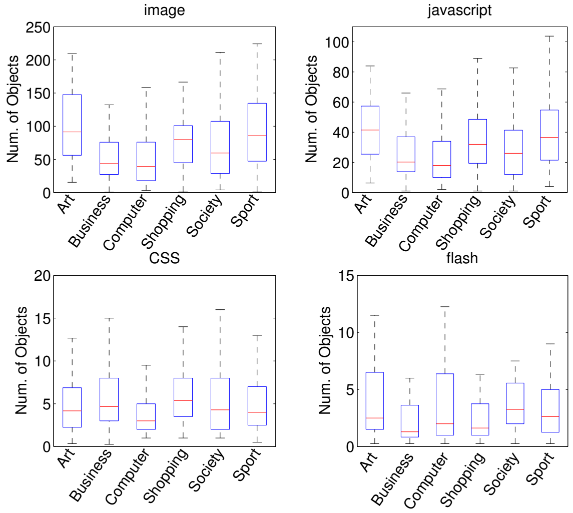
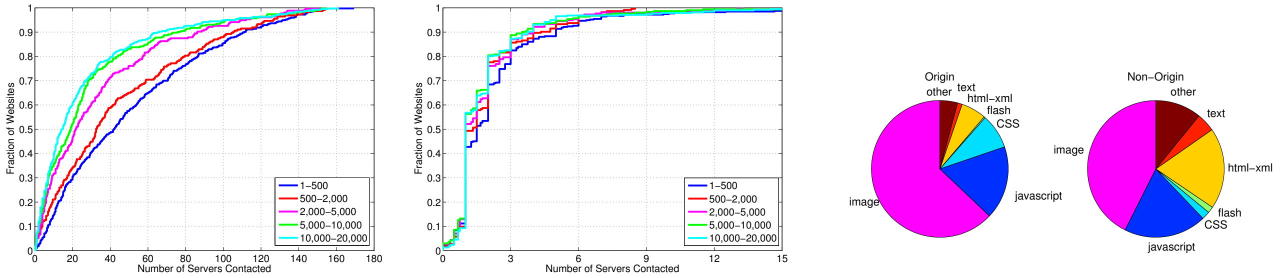


Fig. 3. The summary distribution of the number of requested objects of certain MIME type across web pages with different categories

[1] We note that some flash objects may be downloaded by other objects and thus are not captured by our crawler as we only observe the interaction between our browser/crawler and different web servers



(a) Contacted servers per page

(b) Contacted origin servers per page

(c) Delivered objects by origin/non-origin servers

Fig. 4. Characteristics of the number of contacted servers and its break down to origin and non-origin.

- **Non-origin Server**: Any server whose authoritative DNS server is different from the target website, is considered a third party and thus a non-origin server.

Toward this end, we use dig command to identify the authoritative DNS servers for all target websites. This approach allows us to identify and properly group multiple origin and non-origin servers that deliver the object associated with a target website. For example, contacting $qq.com$ results in the delivery of content from $qpic.cn$ and $gtimg.cn$ that share the same DNS server (and are co-owned by the same company) and thus are considered as origin servers for $qq.com$. As another example, many popular websites contain Google ads that are provided by Google servers and our methodology classifies them as non-origin servers for those websites.

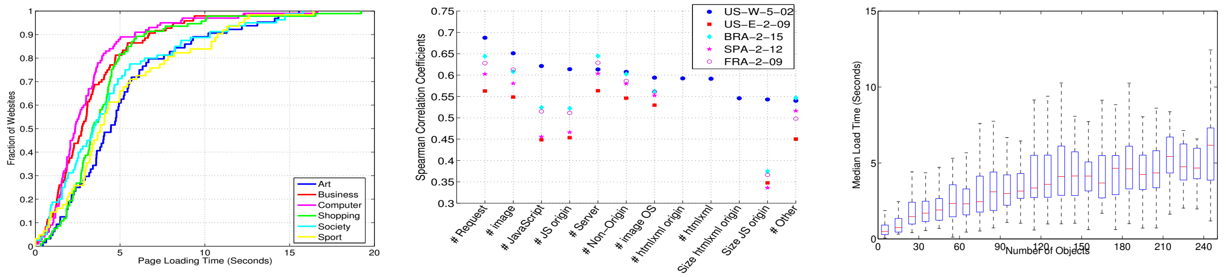
Figure 4(b) presents the distribution of the number of origin servers among websites for different rank-groups. It is interesting that the number of origin servers in more than 80% of websites is very small, namely less than 3-5 servers and in almost all cases is less than 12. The number of origin servers is much smaller than what was reported by Butkiewicz *et al.* [1] in 2011. *In summary, our results illustrate that the total number of contacted servers by individual website has significantly increased while the number of origin servers has decreased during the past 4 years. This clearly illustrates a dramatic growth in the involvement of non-origin servers to deliver objects for these popular websites.*

The next natural question is "how the contribution and MIME type for delivered objects and bytes from origin and non-origin servers differ?". Across all web pages, roughly two-third of all objects and the same fraction of all bytes are delivered by origin servers. To offer more insight, the pie charts in Figure 4(c) present the distribution of MIME types for delivered objects by origin and non-origin servers across all websites. We observe that images make up a significant majority of delivered objects by origin (>60%) and non-origin (>40%) servers. While the relative percentage of javascripts for both groups is the same, the fraction of delivered objects of other MIME types by non-origin servers (except for CSS) is larger than origin servers. The more balanced distribution of delivered objects across different MIME types by non-origin servers indicates the diverse roles that they play. A closer examination of the role of non-origin servers remains as a future work item.

VI. Implications on Page Download Time

We now turn our attention to the impact of website complexities on its download time as it is the key performance measure that affect the experience of their visitors and thus their popularity. In this analysis, we use the information in HAR files to calculate page download time which is the total time to fetch and render all the objects of the web page. In this analysis, we consider the mean page load time across all vantage points for each reachable website to obtain a more reliable estimate. Figure 5(a) shows the CDF of mean page load across websites in each category. This figure shows that websites in categories with more objects exhibit longer load time. Overall, 80% of websites from various categories are downloaded within 4-7 seconds. Comparing this result with the earlier study in 2011, we observe that the page load time for 80% of pages remains rather unchanged over the past 4 years despite the significant growth in the number (and size) of delivered objects and the number of involved servers associated with each web page. This is plausible given the significant increase in the bandwidth connectivity of end users. However, the page load time for the top 20% of web pages have clearly increased especially for categories such as Art and Business.

We further examine the impact of several key variables associated with each page that are likely to affect its load time including number of objects, number of contacted servers, and number of images. In particular, for each variable, we calculate its Spearman's Correlation Coefficient with the load time across all pages to determine which variable is more correlated with the load time for a given vantage point. Figure 5(b) shows the Spearman's Correlation Coefficients for all these variables organized on the x axis based on their overall level of correlation across all vantage points. The results for each vantage point are shown with a different color. We observe that the number of requests, number of servers and number of images are the top 3 most correlated variables with page load time. To directly check the correlation between these



(a) Mean page load time for pages per category (b) Correlation with key variables (c) Correlation with number of objects

Fig. 5. Effect of page complexity on page load time

variables and page load time, we examine the trend in the summary distribution of page load time as a function of each variable. For instance, Figure **??** shows the summary distribution of page load time across different group of web sites where each group contains x objects. This figure clearly illustrates that the median download time generally grows with the number of objects in websites. Butkiewicz *et al.* [1] also reported the same three variable as having the highest correlations with the page load time. However, they observed a higher correlation coefficients (0.7-0.8) than our results (0.55-0.65). We believe that the increasing diversity of key variables leads to more diverse and subtle impact

VII. Conclusions & Future Work

In this paper, we assessed the complexity of 2000 popular web pages with respect to content and service complexity through systematic measurement. We compared our findings with a similar earlier study and presented the basic trends in the complexity of popular web pages over the past four years. In particular, we show that the number of requested objects and the number of contacted servers for downloading each page has significantly increased over the past four years but a much smaller number of them are origin servers. Furthermore, we examined the implications of main characteristics of individual web pages on their download times to identify the key factors that affect download times. Our study provides a valuable set of basic results on changing trends in the complexity of major web sites.

We plan to extend this preliminary work in a couple of directions as follows: First, given the clear decrease in the number of origin and major increase in the number of non-origin servers suggests a shift in the role of origin and non origin servers. We are exploring this issue by inferring the role of major non origin servers and their interactions with our client. Second, we take a closer look at the performance bottlenecks for different web sites at different VPs. In particular, we investigate whether origin or non-origin servers for various web sites affect the page load time, and in particular whether a specific non origin server may be the performance bottleneck. Other intriguing issues to explore are the effect of time of loading, VP's location, choice of browser and client platform (mobile vs desktop) on the performance. Along the same line, the causes for unreachable web sites and their higher fractions in the US (compared to Europe) requires further investigations.

VIII. Acknowledgements

We thankfully acknoweldge Juan Miguel Carrascosa Amigo from Spain, Balarishman Chandrasekaran from Duke, Roberto Gonzalez from France, Flavio Vinicius from Brazil and Xu Yuan from China who provided access to their local machines to server as VP for our data collection.

References

[1] M. Butkiewicz, H. V. Madhyastha, and V. Sekar, "Understanding website complexity: measurements, metrics, and implications," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, 2011, pp. 313–328.

[2] F. Schneider, S. Agarwal, T. Alpcan, and A. Feldmann, "The new web: Characterizing ajax traffic," in *Passive and Active Network Measurement*. Springer, 2008, pp. 31–40.

[3] F. Schneider, A. Feldmann, B. Krishnamurthy, and W. Willinger, "Understanding online social network usage from a network perspective," in *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*. ACM, 2009, pp. 35–48.

[4] R. Tian and R. Rejaie, "Re-examining the Complexity of Popular Websites: Measurement, Analysis and Implications," University of Oregon, Tech. Rep. CIS UO-TR-7-2, 2015. [Online]. Available: http://onrg.cs.uoregon.edu/pub/tr-7-1.pdf

[5] "Firefox," https://www.mozilla.org/en-US/firefox/new/. [Online]. Available: https://www.mozilla.org/en-US/firefox/new/

[6] "Selenium webdriver," http://www.seleniumhq.org/projects/webdriver/. [Online]. Available: www.seleniumhq.org/projects/webdriver/

[7] "Firebug," http://getfirebug.com/. [Online]. Available: http://getfirebug.com/

[8] "Firebug extensions," https://getfirebug.com/wiki/index.php/Firebug_Extensions. [Online]. Available: https://getfirebug.com/wiki/index.php/Firebug_Extensions

[9] "Har 1.2 spec," http://www.softwareishard.com/blog/har-12-spec/. [Online]. Available: http://www.softwareishard.com/blog/har-12-spec/

[10] B. Krishnamurthy and C. Willis, "Privacy diffusion on the web: A longitudinal perspective," in *WWW*, 2009.