Summary Essay
Kumar Suyash (113277210)

High web-page load times are frustrating for users. Studies suggest about 49% of users would abandon a site or switch to a competitor upon facing performance issues. In 2009, Akamai found users typically wait for 3 seconds before navigating away from a webpage. With the rise in web usage, the complexity of websites has also increased. This increase in complexity can lead to an increase in the page load time.

Unfortunately, there hasn't been much study done into understanding how the complexity of the websites affects the page load times. This paper intends to fill this void by discerning the metrics characterising the complexity, quantifying them, and analysing them. The authors propose the study would help the web providers understand the factors most responsible for slowing down of their websites.

Hence, in order to characterize and quantify the complexity, the authors make use of around 2000 websites as test subjects. Firefox is used as the browser for firing the webpages, enabled with the Firebug extension, and Net: Export and Firestarter as add ons for exporting the log of all request and response involved in rendering a webpage. The measurements are performed on four distinct vantage points.

Post the study, the authors make two major claims - in terms of page render and load times - the number of objects requested is the most important factor while the number of servers is the major factor affecting the variability in load times. For the study, the complexity is divided into two aspects: content-level - complexity characterization based on the content fetched in rendering a webpage (metrics : number of objects fetched, size of objects and types of content) and service-level: characterization based on the services these websites are built upon. For analysing the number of objects required to load a webpage, across the ranks it is observed that more than 40 objects are fetched for loading the webpage in the median case and about 20% websites request more than 100 objects. Upon investigating these sites based on the category they belong to, it is observed, the news sites request far more objects than any other types. While, on observing the service-level complexity - it is observed that about from 25% to as high as 55% of websites might require the use of at least 10 distinct servers for loading their base page and about 60% of websites require the usage of more than 5 non-origin servers to load their content. 30% of the objects and 35% of bytes fetched are contributed by the non-origin servers but their contribution to page load time is comparatively low due to requests' parallelization by browsers. A significant portion of the content contributed by these non-origin servers is Javascript.

To assess the impact of these metrics on page load time, the load time is quantified with 'RenderEnd' measure - signifying the total time to fetch and render all contents on the webpage. The metrics having significant correlation with the load time measure are then found using Spearman correlation - it is observed that the parameters like total number of objects loaded, the number of Javascript objects and the total webpage size are heavily correlated with load time measure. Upon performing the regression analysis, the result confirms correlation's findings. The authenticity of these findings is further verified upon plotting the actual page load time against the load time predicted by the regression model. In order to analyze the factors impacting the variation in load time, the variability in load times of a website is defined as the difference between 75th and 25th percentile values of RenderEnd and the correlation of variability with the metrics is plotted. The correlation now is lower and the most dominant factor is the number of servers. These findings also successfully debunk the notion that a website's rank is a significant indicator of its complexity.

The paper thus, provided me with a unique data-driven insight into the world of web technologies. It was a revelation to learn there are various unknown analytics and ad services featuring on a fairly significant number of websites, delivering content as non-origins. I also learnt about tools like Firebug which can provide an entire log report of the requests sent while loading a page - something I was completely unaware of. Finally, I also learnt that modern browsers use parallelization to enhance their speed.

However, I did find a few places where the paper can be improved. The paper has used two load metrics - onContentLoad and onLoad however ignores other good metrics like “above-the-fold” time, time to first “paint,” etc. Adding these metrics can improve the performance. Secondly, the paper does not consider personalised webpages for the study - which is a considerable gap considering most websites now have the functionality of having a personalized feed. Lastly, the paper only considers client-side view but ignores the back-end infrastructure - another topic which can affect the results of the study.

As an extension - I would like to use the Lasso regression technique I learnt for my data science projects like predicting the prices of condominiums in specific localities of Miami given a set of attributes for automating the feature selection. As a developer, I can also use the Firestarter for debugging my website’s requests considering the report also gives the response data and status code.