A Simple Approach to the Effect of the Wikipedia Network Structure on Pageviews

Candidate Number: 1033256

January 14, 2019

1 Introduction

Wikipedia is one of the most visited websites in the world, containing articles in almost 300 languages with an extensive coverage of a wide array of topics. Even in light of the fact that its contents are written by thousands of volunteers, its accuracy has been considered close to the precision of the Encyclopædia Britannica (Giles, 2005). It is, therefore, no surprise that Wikipedia have been used for multiple research projects. In 2019, Google Scholar returns more than 16,000 articles that contain the word Wikipedia in the title. The variety of these work ranges from computing semantic relatedness (e.g. Strube, & Ponzetto, 2006; Gabrilovich, & Markovitch, 2007) to predicting movie box office success (Mestyán, Yasseri, & Kertész, 2013).

The goal of this experiment is to develop an initial approximation about how Wikipedia pageviews relate to the structure of the network. This topic has been much more deeply explored by Gildersleve and Yasseri (2018), who studied the navigational behavior in Wikipedia using aggregated clickstream data (i.e. referer-resource pairs). In this short approximation, however, we will follow a much simpler approach analyzing the pageviews of some articles, which are simply a measure of how many people have visited an article during a given period of time.

First, consider there are three main ways to enter a Wikipedia article. The first is searching explicitly for it; the second is using the Wikipedia "random article" option; and the third is following an hyper-link that leads to a particular article. We will focus on the third one, and, in particular, only in the hyper-links between Wikipedia articles. If every article contains hyper-links to other ones, we can see the Wikipedia as a directed network in which the nodes are the articles and the edges correspond to the hyper-links between them. However, in this network not all edges have the same weight: it is common for an article to have many hyper-links directed to closely related concepts, while having only one or a couple for articles tangentially connected. For instance, the Wikipedia page about the Bitcoin has fourteen hyper-links to the Cryptocurrency article, while it only contains one to the University of Pittsburg. Therefore, the number of hyper-links between article A and

article B is a very simple measurement of the connection strength between an article A and an article B.

Pageviews are potentially related to the Wikipedia structure in two ways: directly or indirectly. Pageviews will be directly related if they are, in fact, a product of the hyper-link structure. That is, if fewer people visit article A, the hyper-links in it directing to article B will obviously receive less clicks, which in turn implies that article B will receive fewer visits. Note that this creates a loop by which the pageviews to both articles decrease. However, a correlation between pageviews could happen even if the articles did not contain hyper-links between them, as long as they entail some relationship. Namely, this will happen if the interest about article A and article B increased (or decreased) at the same time, causing people to search for article A and article B more during some periods and less in others. For instance, we could expect that the pageviews of articles about biochemistry and articles about cultural studies will be correlated. However, this could occur not because these two topics contain many hyper-links between each other, but because students tend to search for these articles more during exam periods and less in summer break.

Our research question asks about the relationship between the network structure and pageviews. In more concrete terms, we will test if the actual pageviews of one article can be predicted based on the weighted mean of the pageviews of the hyper-linked articles to that original article.

2 Methods

We will obtain two datasets. One is a simplified version of an article's network, and the other is consist of the pageviews of this articles as well as the articles in its network.

2.1 Structure

The first step is to obtain the contents of the article of interest using the Special:Export feature in Wikipedia. With that, we can discover the articles that are hyper-linked from it. As we noted in the introduction, some articles are hyper-linked many times while others are just hyper-linked once or twice. For instance, we could define as the article of interest the one about Neymar, a famous Brazilian football player. With this particular example, we will soon discover that just counting the number of hyper-links to the different articles will produce sub-optimal estimations of the connection between the article about Neymar and the hyper-linked articles. For example, the Neymar article contains only two links to the Brazil national football team, in which he currently plays, exactly the same number of links directed to the article about Flamengo, another Brazilian team against which he played in 2011, according to the own article.

A simple way to create a better estimate is to take into account the times that the Brazilian football team and Flamengo are mentioned. If we do it manually, we will

soon understand that there are many mentions to the Brazilian football team (42) while just a couple to Flamingo. However, note that counting the number of times that certain concept is mentioned in the original article is a simple task for a human, but can be tricky for a computer. The main reason is that the Brazilian football selection is called by various names in different parts the article.

A simple approach to solve this problem comes with the idea that each internal link in Wikipedia is characterized by the name of the article to which it links, which we will call target, and the name that appears on the text, which we will call label. For example, we mentioned that there were two links to the Brazil football national team, but in this case, they were under two different labels: one was 'Brazil national team' and the other was 'national senior team'. To better estimate the relationship between Neymar and the Brazilian national team we will count the number of times that these labels appear on the text and sum it to the number of links. In particular, we found that there were 14 references to the 'Brazil national team' or the 'national senior team' apart from the two mentioned links, leaving a total of 16 mentions. Repeating the same procedure for all the links in the original article allows us to obtain a list of linked articles with the corresponding weights. This can be represented as a directed network, as shown below.

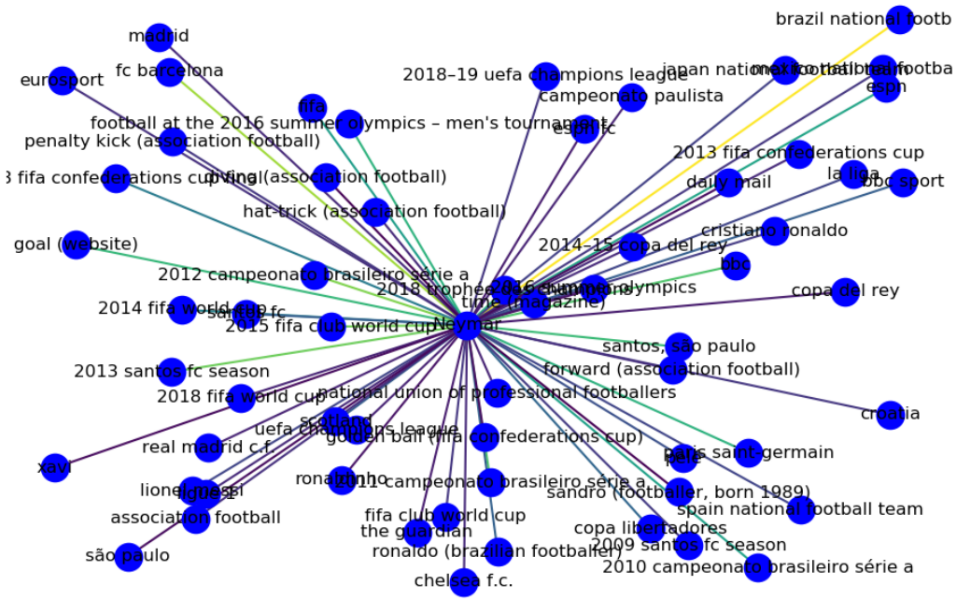

Figure 1: Network representing the articles hyper-linked from the article about Neymar

However, note that even counting the number of times that article B is mentioned in article A leaves a very simple measurement of the connection between the two. More advanced approaches will also include the number of hyper-links from B to A, and even other characteristics about the structure of the network (e.g. the number of triangles in which article A and article B both appear). However, for this initial approach to this topic we will use this very simple measure.

2.2 Pageviews

Pageviews can be interpreted as a measure of interest over time. That is, more people will visit the Neymar Wikipedia article when the general public interest about this football player increases. However, it should be noted that pageviews do not reflect the number of unique visitors (i.e. number of visitors that have requested that article, regardless of how often they visit), but the quantity of requests to load the HTML file containing a given article. Therefore, if the same person visited a Wikipedia article twice during a given period of time it will be counted as two visits.

In this case, we will study the Wikipedia pageviews between the 24/11/2015 and the 24/11/2018, using a day as the standard unit of time. This will offer enough granularity to take into account the impact of specific events that could have impacted the daily number of pageviews.

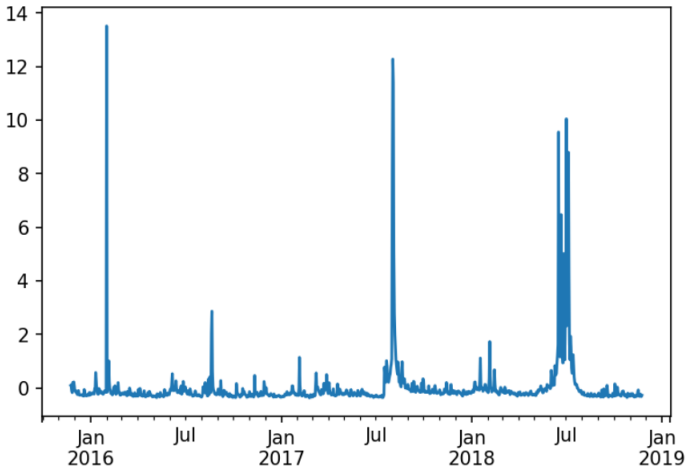

Figure 2: Pageviews of the Wikipedia article about Neymar

Note that the three most clear spikes in the pageviews reflect moments in which interest about this player in the general public increased, as the own Wikipedia article reflects. The first spike corresponds to an investigation over a potential tax fraud that appeared in the news in February 2016. The second spike corresponds with the mediatic transfer of this player to the Paris Saint-Germain. The third corresponds to the 2018 FIFA World Cup.

We used the same approach to extract the pageviews not only of the article about the original term, but also of all the hyper-linked articles from the original article. Note, however, that it is possible that some of these articles did not existed the 24/11/2015. In this case, we only obtained the pageviews after the article was created.

2.3 Study design

As we stated before, our main goal is to investigate if the pageviews of a certain article can be predicted based on the pageviews of its hyper-linked articles. This

involved calculating the correlation between the actual pageviews of the original article and an estimation that was done based on the hyper-linked articles. Although more complex approaches are possible, we defined this estimation as the average of the pageviews of the hyper-linked articles weighted by the amount of connection between each hyper-linked article and the original article. Summing up, our hypothesis states that the actual pageviews of an original article will correlate with the weighted average of its hyper-linked articles.

After defining our method, the last step was to choose a selection of original articles. In this case, we choose popular articles related to very different topics. In particular, we selected: (1) the article about Neymar, the football player mentioned before; (2) David Hume, a Scottish Enlightenment philosopher; (3) Ubuntu, a Linux distribution; (4) Pink Floyd, a British rock band; (5) Bitcoin, a cryptocurrency; and (6) Cubism, an 20th-century art movement.

3 Results

We find that, as expected, the weighted average of the hyper-linked articles pageviews correlates strongly with the actual pageviews of the original article. For Hume we obtained an r = 0.74; for Neymar r = 0.58; Ubuntu r = 0.91; Pink Floyd r = 0.7; Bitcoin r = 0.67; and Cubism r = 0.81. The results can be presented in a correlation plot.

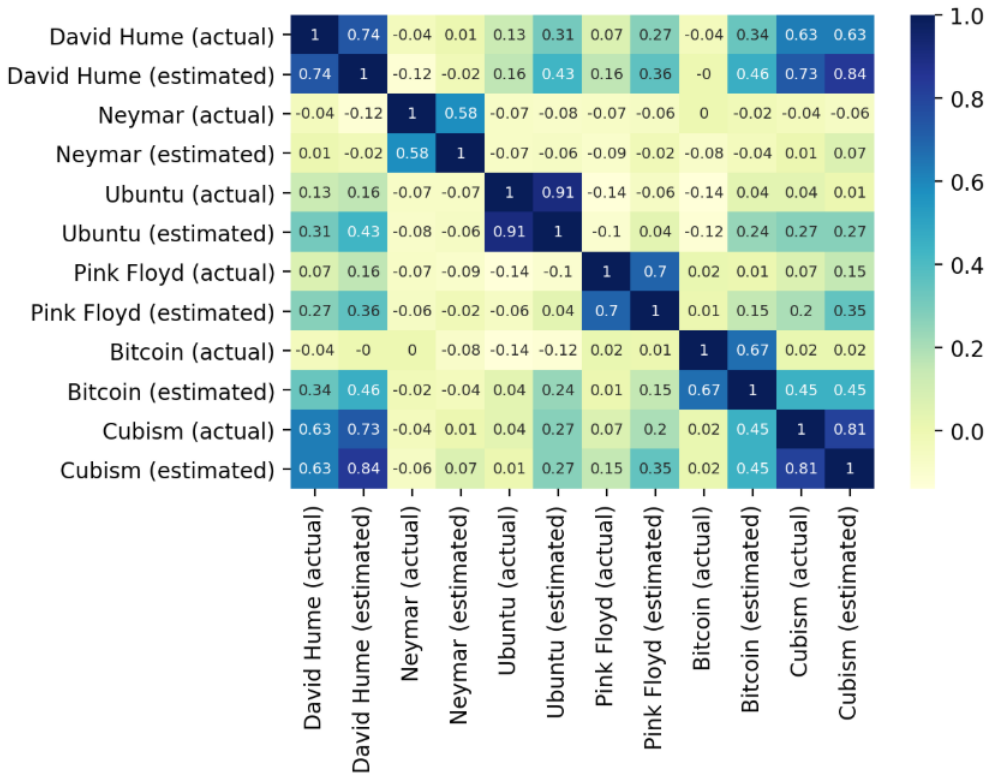

Figure 3: Correlation plot between the actual and estimated pageviews for different Wikipedia articles

It is also clear that the pageviews estimated with the mean of the hyper-linked articles of a certain original term do not correlate with the actual pageviews of the other search terms, apart from for one exception: there is a clear correlation between the estimated pageviews for David Hume and the actual pageviews for Cubism, as well as a similar effect between the estimated pageviews for Cubism and actual pageviews for David Hume.

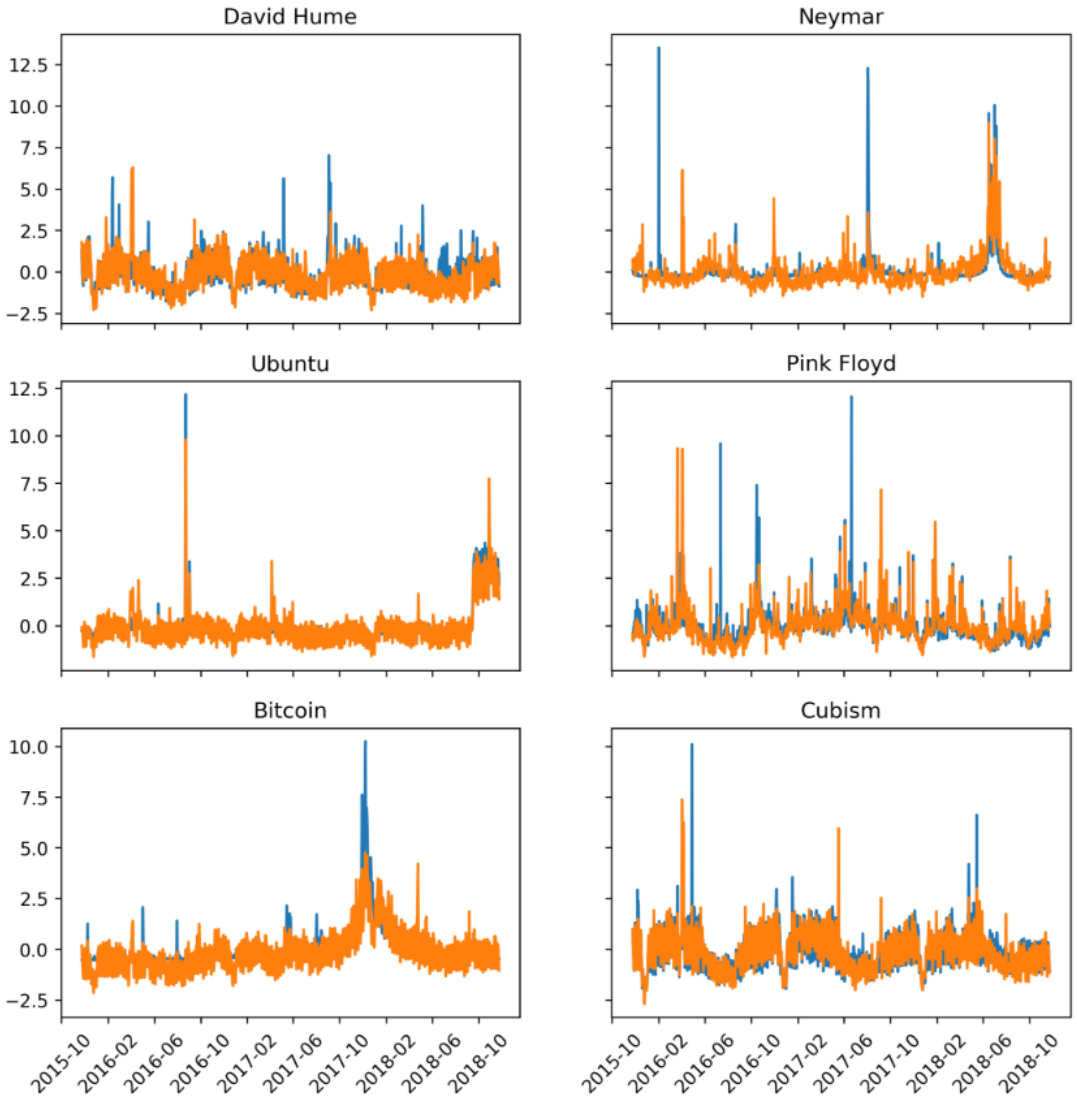

Figure 4: Actual (blue) and estimated (orange) pageviews for the Wikipedia articles that were considered

Additionally, there is a noticeable trend in this data. The articles about popular topics (e.g. Neymar, Pink Floyd, Bitcoin) seem to have lower correlations than the apparently less popular articles (e.g. Ubuntu, Cubism, David Hume). We tested this relationship calculating the correlation between the popularity of these six articles (i.e. measured as the mean pageviews between the 24/11/2015 and the 24/11/2018) and the association between actual and estimated pageviews (i.e. the correlations mentioned above), obtaining a r = -0.721.

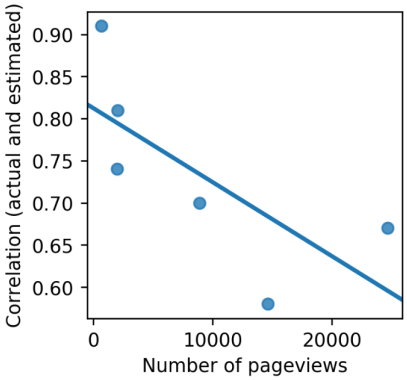

Figure 5: Negative relationship between the number of pageviews and the correlation between actual and estimated pageviews

4 Discussion

First and foremost, we have shown that the structure in the network has a clear relationship with the pageviews. We can estimate with high accuracy the pageviews of one Wikipedia article solely based on the pageviews of the articles hyper-linked from it. As we explained in the introduction, there are two different effects that could lead to this phenomenon. One is a direct impact of the structure of the network. For instance, when users are in an article about Linux they may click on Ubuntu, and therefore the pageviews on the former will be at some degree predictive of the pageviews on the latter. The second potential explanation is that related terms tend to be looked for in the same periods of time. For instance, during the last two years there have been periods of popular interest about bitcoin, cryptocurrency, and blockchain. During these periods of high interest, it is possible that people looked for the article about bitcoin, cryptocurrency, and blockchain directly. A clear example of this second explanation comes in the high correlations between the estimated and actual pageviews of David Hume and cubism. This effect cannot be justified with the first explanation because these two pages are not hyper-linked directly. However, looking at the time series it is apparent that the actual and estimates pageviews follow a cyclic trend that seems to overlap with the academic calendar.

We do not know how many pageviews are a product of the structure of the network and how many result from an indirect relationship. However, the negative trend between the number of pageviews and the predictability of the pageviews based on the hyper-linked articles gives us some information about these two factors. This evidence appears to show that the very popular articles (indicated by the number of pageviews) are the ones searched for in the first place, and this is the reason that their pageviews cannot be predicted with the same precision based on the information provided by their hyper-linked articles. On the other hand, less popular articles appear to receive more traffic from the hyper-linked articles, and their pageviews seem to be more influenced by the Wikipedia structure.

Even considering the important limitation that we only studied six articles and did not account for more details of the network, some evidence was found suggesting

that the network structure is highly relevant to explain the visits that each article of Wikipedia receives – especially the articles with fewer pageviews. This is a good starting point to continue studying how hyper-links may affect the navigation of users through Wikipedia.

5 References

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. *Proceedings of the 20th international joint conference on Artifical intelligence* (pp. 1606-1611).

Gildersleve, P., & Yasseri, T. (2017). Inspiration, Captivation, and Misdirection: Emergent Properties in Networks of Online Navigation. *arXiv preprint* arXiv:1710.03326.

Giles, J. (2005). Internet encyclopaedias go head to head. Nature, 438(7070), 900.

Mestyán, M., Yasseri, T., & Kertész, J. (2013). Early prediction of movie box office success based on Wikipedia activity big data. *PloS one*, 8(8), e71226.

Strube, M., & Ponzetto, S. P. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. *Proceedings of the 21st national conference on Artificial intelligence* (Vol. 6, pp. 1419-1424).

6 Code