Network Analysis: Wikipedia Map of Science

Group 46: Daniel Castro 87644 João Tiago Aparício 97155 Miguel Trinca 86490

I. Introduction

In this project we are analysing a network which shows the similarities among different branches of science based on Wikipedia pages in outline of natural, formal, social and applied science. We analysed the graph and collected our metrics using Python, powerlaw: A Python Package for Analysis of Heavy-Tailed Distributions[1] and NetworkX, the package for network analysis.

II. Goals

With this experiment we intend to:

- Acquire a better understanding of the NetworkX package
- Interpret and compute the following metrics:
  - Degree Distribution (P(k))
  - Centrality or Importance
  - Cluster Coefficient (CC)
  - Average Path Length (APL)
- Understand how the network grows
- Infer and discuss the results of said metrics

In addition to Python and its packages we will be also using Gephi to visualize our network.

III. Dataset

The dataset is based on the one extracted by Alberto Calderone in 2020 [2]. As previously mentioned, this dataset depicts the different branches of science on Wikipedia. Nodes represent Wikipedia pages of a science field. An edge between two nodes is their page similarity above a 0.3 threshold according to cosine similarity. In other words, if the pages of node i and node j have a cosine similarity above 0.3, then there is an edge between node i and j. The graph is undirected and has a total of 687 nodes and 6523 edges with an average degree of 18.9898, which means that each field is somewhat similar to 18.9898 other fields on average. To model the information in each node we used the following data structure:

```
(0, {'name': 'Accounting', 'class':
'Applied', 'url': 'https://...'})
```

Each page has a class that can be either Social, Formal, Natural and Applied. The percentage of each is about 33.33% (Pink), 28.09% (Green), 24.89% (Orange) and 13.68% (Blue) respectively. The graph is represented in Fig. 1 using the Fruchterman Reingold algorithm.

By analysing the graph, we encountered five single connected components. Four of those were either pairs or triplets of nodes. These were sciences such as Coastal geography (As both natural and social sciences) and Hydrography; Oceanography and Marine Biology; And Food Science and Nutrition;

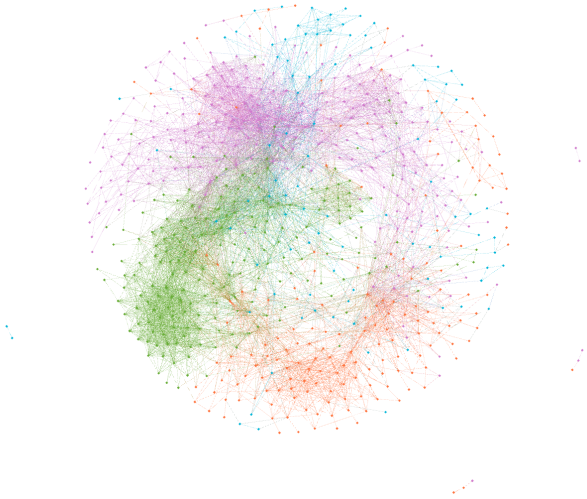

Fig. 1. Graph of the Wikipedia Map of Science, color-coded by Science Class.

IV. Degree Distribution

The **degree of a node** is the number of edges that the node has. The **Degree Distribution** is the probability distribution of these degrees over the whole network.

After collecting the degree of each node, we made an histogram with the results. We can observe that the majority of the nodes have only one connection. In practice, this says that each Science Wikipedia page is most likely to only have one other science page with high similarity.

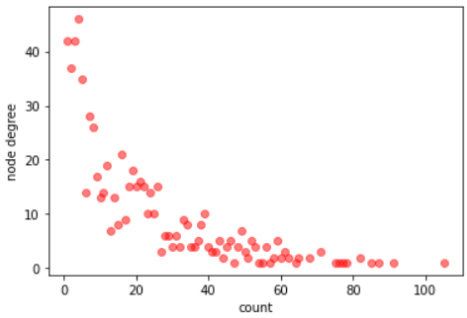

Fig. 2. Histogram of the degree distribution

We believe that this dataset also follows a power law distribution, since it is based on the number of Wikipedia pages and they usually have a rapid growth in terms of cardinality and reference. So we tested this hypothesis, to **understand how the network grows**.

A. Power Law

An important thing that we thought to analyse in our dataset was the question of, theoretically, how would our network grow?

A first intuitive interpretation of our problem: when a new Science arises, the tendency is to have new branches or derivations of that same science. Which makes a probable power growth in the number of Wikipedia Science pages. As they as well follow the creation of new science fields.

For this reason, we decided to analyze the dataset, using the powerlaw package [1], and the follow up article on it, Statistical Analyses Support Power Law Distributions Found in Neuronal Avalanches [3]. The Fig 3, shows us the distribution of degree probability in our network and and in a power law. On the y axis the p(k) - degree distribution and on the x axis the k, denoting number of nodes in the network.

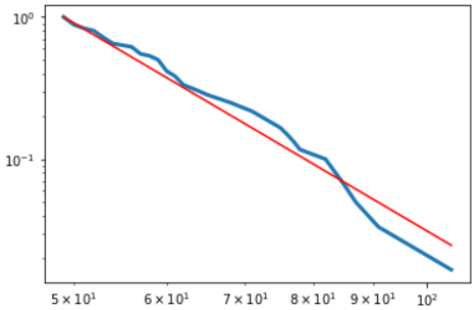

Fig. 3. Power law fit: we can see the degree distribution of the power law in blue and the page network in red.

The slope of our growth is $\gamma = 5.852082324422884$, following the power law formula:

$$p(x) \propto x^{-\gamma} \tag{1}$$

According to Klaus et al. [3], such a visual approach suffers from a lack of statistical rigor in identifying a significant difference between a power law and an exponential or other alternative model distributions. This problem is worsened when the availability of the data is limited, for example, for small sample sizes or when the range of values over which the distribution is analyzed is narrow.

This means that, although our initial reasoning makes complete sense with the network growth, we have the notion that this may be an overfit because of the relatively small size of our network.

In addition, typically the scaling parameter lies in the range $2 < \gamma < 3$, however there are occasional exceptions, which we believe to be in one of them. In practise, few empirical phenomena obey power laws for all values of x. More often the power law applies only for values greater than some minimum $Xmin$. In such cases we say that the tail of the distribution follows a power law. [4]

Given the γ value, which larger than 3, we may conclude that the tail of the distribution follows a power law, and the network can be approximated to a random network.

V. Centrality or Importance

How do we classify a branch of science as the most similar to every other science? Does the amount of links between other branches (in this case class) make it versatile, or is the relation between other fields? These are the type of questions we are trying to answer when computing the centrality of the nodes. By intuition, we may conjecture that the amount of links to other nodes would classify the page as important. Due to an impact on a higher number of fields, since they have more similarity and hence more influence on the concepts shared between one another, regardless of the importance of the concepts themselves.

A. Degree Centrality

To find the Hub in this network, we will start by comparing different centrality measures. Firstly we calculated the **degree centrality** to understand which science has a page with the highest degree of similarity to the remaining pages. In this case the hub is **School Psychology, as a social science** (https://en.wikipedia.org/wiki/School_psychology). This probably happens because this page has references to many other sciences, including biology, for the studying the of human thought process.

B. Eigenvector Centrality

To know the science that has a higher page similarity to the sciences with the highest degree of page similarity, we use the eigenvector centrality, which **yields the same result as degree centrality**, has similarity to the most prestigious sciences, or with the sciences with the highest similarity degree.

C. Betweenness Centrality

To understand which of the science pages is a bridge between different parts of the network, we use the **betweenness centrality**. In this case this tells us the hub on this centrality measure is the node that connects the most different science pages to each other. In this network that science page is **Population Biology, as a natural science** https://en.wikipedia.org/wiki/Population_biology.

We might hypothesize that this happens because this is a field that derives from the intersection of formal sciences and natural science. The second highest betweenness centrality is **Algorithm, as a formal science** https://en.wikipedia.org/wiki/Algorithm. This is intuitive because they can be applied in the context of any other science. Another science with a high degree of betweenness centrality that can be easily guessed is **Economics** (https://en.wikipedia.org/wiki/Economics) (actually the third highest betweenness centrality). This happens because it is a strong link between the social sciences and the formal sciences.

VI. Cluster Coefficient and Transitivity

Usually science fields have similar aspects between one another. By using cluster coefficients we may compute how likely that two neighbor nodes have a connection with each other. In other words, if they create triangles. This metric allows for the depiction of the network by clusters, where nodes from a cluster have similar properties.

To analyze this we computed, using networkX, transitivity, clustering and average clustering. The first one computes the fraction of all possible triangles present in the graph (transitivity). The second computes the local cluster coefficient. And finally, the last one computes the global clustering coefficient.

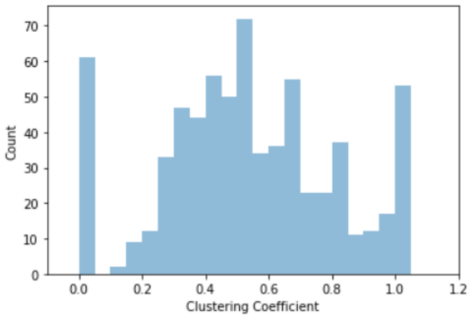

Fig. 4. Cluster coefficient per node.

The Global Clustering Coefficient is about 0.5302 and the transitivity is approximately 0.4692.

As we can observe, the majority of nodes have a cluster coefficient between 0.4 and 0.8, and others are really close to either 1 or 0. This will result in a global clustering coefficient and transitivity of approximately 0.53 and 0.47, respectively, which are both huge. These results imply that our network is very related, which makes sense since we are talking about branches of science, which have inter-class relations, whether it is math, physics, philosophy, etc. Nodes like "Automated reasoning", "Statistical theory", "Time series", have the highest values (approx. 0.984, 0.990, 0,994 respectively), which means that they can be correlated to every branch of science.

Looking at concrete examples and cross referencing with results of betweenness centrality (BC), we observe that pages that bridge different communities - i.e. have the highest BC value - such as "population biology", "algorithm", "economics" also have a very low local clustering coefficient (approx. 0.1328, 0.2387, 0.2664, respectively). This was expected, since these nodes are bridging different communities causing similarity to diminish.

Another interesting aspect is that nodes which are close to the bridges have also a high local clustering coefficient, for instance "population ecology" (cc approx. 0.9) which is close to "population biology" as seen above. This would suggest that these nodes are in fact bridging communities.

VII. Average Path Length (APL)

In this case, the average shortest path length is the smallest amount of hops I have to take, on average, to be on another science field page, within Wikipedia. We could hypothesize that **the smaller the APL the more similar** (in terms of cosine similarity) **all science fields are**, since they are more interrelated with one another.

To calculate this we only used one of the single connected components and did not use the four pairs and triplets that were disconnected from the remaining massive (relatively speaking) component with 677 nodes. The average shortest path length, approximately 3.433. Since each link represents a higher similarity between 2 pages, this means that on average the science pages are fairly indirectly similar with every other page, with the exception of the disconnected components. In other words, we can change subjects/classes with only 3.4 hops on average, meaning that pages are fairly similar.

The small-world property is obtained [5] by having a small APL and a high clustering coefficient, which is the case of our network. This implies that the growth of the APL scales in a logarithmic way with the growth of the network.

$$L \propto logN \tag{2}$$

In other words, by adding more pages to our network the average path length would not differ much from what it is now. In addition, these pages would also be highly correlated to the pages already in the network since the global clustering coefficient is also high.

VIII. Conclusions

Summarizing the main ideas on this analysis, we can conclude that:

It is probable that a new science page being added to the graph would have a high probability to have a similarity to only one other science page. This means that the degree distribution has a very high probability with k = 1.

Given the results of the power law analysis, we can conclude that this network behaves approximately like a random network, and is a small-word network. Which can be corroborated by the results from the cluster coefficient and average path length analysis.

The analysis of the centrality of our network were interesting and surprising. The School Psychology has the highest similarity and with most similar neighbors to other science Wikipedia pages. The nodes that have the highest betweenness centrality, i.e., are important bridges between the science pages, do not have a high degree centrality but they have key edges to those who have. Some science pages with this characteristic are Population Biology, Economics and Algorithms.

Also, we got a high clustering coefficient (≈ 0.5), which means that sciences neighbors are very likely to be connected.

Combining this with the relatively small APL, creates the small-world property that shows us the slow growth of the APL when the graph increases its size.

References

[1] J. Alstott, E. Bullmore, and D. Plenz, "Powerlaw: A python package for analysis of heavy-tailed distributions," *PloS one*, vol. 9, no. 1, e85777, 2014.

[2] A. Calderone, *A wikipedia based map of science*, Jan. 2020. DOI: 10.6084/m9.figshare.11638932.v5. [Online]. Available: https://figshare.com/articles/dataset/A_Wikipedia_Based_Map_of_Science/11638932/5.

[3] A. Klaus, S. Yu, and D. Plenz, "Statistical analyses support power law distributions found in neuronal avalanches," *PloS one*, vol. 6, no. 5, e19779, 2011.

[4] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[5] J. Davidsen, H. Ebel, and S. Bornholdt, "Emergence of a small world from local interactions: Modeling acquaintance networks," *Physical review letters*, vol. 88, no. 12, p. 128 701, 2002.