The Complexity of Physicists

Varghese Jacob

CID: 01716176

Supervisor: Dr Tim Evans

Assessor: Dr Dave Clements

Word Count: 3293

Table of Contents

Declaration of Work Undertaken

This project was carried out solely by me, Varghese Jacob, so all results and developments were produced by myself with input from my supervisor, Dr Tim Evans. Before this project, I have not worked with my supervisor, and I am not enrolled in the Complexity and Networks module.

Abstract

A network was produced by creating nodes that correspond to individual physicists and adding edges between the nodes if a hyperlink was found between the Wikipedia pages of these physicists. Ranking these physicists by analysing centrality measures for the network, provided evidence to suggest that Albert Einstein and Niels Bohr are the most important physicists. This investigation also suggests that the hyperlinks between Wikipedia pages strongly suggest some form of influence between physicists.

Key Words: Networks, Physicists, Wikipedia, Centrality measures

The Complexity of Physicists

Physics has always been a subject for curious minds to discuss and collaborate on. While modern physics is heavily collaborative and large teams work to make steady progress, early physics was dominated by several people who made significant advances while also sharing their developments with each other. Inevitably, among all the people who work in the field of physics, there are a few people who were very important in shaping the scene of physics today. Other than surveying a very large sample of people to determine who the most important physicist is, there are no quantitative measures to determine importance. In this paper, we attempt to describe the process of quantifying the importance of a physicist and produce a definitive answer for who carries this title. Rather than surveying people, we will discuss what useful data is present on a physicist's Wikipedia page to achieve this goal and employ various network analysis techniques to identify who the most important physicist is and describe the associated uncertainty.

Methods

The information required to form a network of physicists was extracted from the open-source biographies contained within the Wikipedia pages of individual physicists. The list of physicists was produced by reading a Wikipedia page that lists the names of 1021 physicists, along with their nationality, date of birth and date of death [1]. These pages were downloaded as HTML (HyperText Markup Language) files and then read using a python package, called BeautifulSoup [2], which parses these files and can be used to identify specific components within them such as hyperlinks. By loading each file as a 'BeautifulSoup' object, the entire file could be read without formatting, i.e., as HTML code; the hyperlinks could then be extracted into a list by appending all items that were contained within <a> tags, e.g., Augustin-Jean Fresnel, which is a hyperlink to the Wikipedia page of Augustin-Jean Fresnel found on Émile Verdet's page. By cycling through each file and compiling lists of hyperlinks for each page, a dictionary could be formed in which the keys are the names of the physicists, and the values are the lists of hyperlinks found on each page.

A network is defined as a set of objects, called "nodes" or "vertices", that are connected by "links" or "edges" [3]. The edges can be directed or undirected, meaning a link between two nodes can either be one-way or two-way. In this project, each physicist is associated with a node and the undirected edges represent a relation between two physicists' work. The python package called NetworkX is the tool that was used to create the nodes and edges of the network that was worked on [4]. To begin forming the network, all 1021 nodes were produced and assigned the names of each physicist in the list. However, before they were produced, the list of file names had to be processed to make them readable. In general, this only involved removing the '.html' extension, but the main

processing occurs when a name contains an accented character. Due to the encoding of characters in URLs (Uniform Resource Locator), accented characters are translated into basic characters, e.g., É becomes %C3%89 [5]; the name of a physicist such as Émile Verdet needs to be extracted from the Wikipedia file called %C3%89mile_Verdet.html. After creating and labelling all the nodes, the edges had to be formed. It was decided that a hyperlink to a physicist's Wikipedia page found on another physicist's page strongly suggests there was an influence, whether mutual or not, on each other. This became the criteria for forming an edge between two nodes. However, we cannot determine that finding a hyperlink in physicist A's Wikipedia page to physicist B's page means that the influence is directed. Depending on the biases of the writers of these pages, they may have deemed it unnecessary to include the hyperlink to A's page on B's page and vice versa. The mention of another physicist may have also been in the context of describing that the mentioned physicist influenced the physicist on whose page the hyperlink was found. Due to the large variety of possible reasons for including a hyperlink to another physicist's page, adding an undirected edge seemed to be the only suitable solution that would allow for a tolerant depiction of the relation between two physicists' work. By comparing the dictionary consisting of 1021 unique lists of hyperlinks found on Wikipedia pages, with the list of physicist's names that was previously constructed, each hyperlink could be searched for the name of a physicist; if the link contained the name of another physicist, an undirected edge could be added between the corresponding nodes. By completing this process for the entire array, the full network representation of all the Wikipedia pages had been produced.

The dataset of interest was sourced from many Wikipedia pages. While Wikipedia pages undergo moderation and special pages, such as those of important people, are locked and only edited by administrators [6], it is still a crowd-sourced dataset with a lot of room for error. However, these errors are hard to quantify as they are not numerical, and there is no correct answer for the biography of a physicist, aside from ensuring any facts presented are accurate. By having a large sample size, the effect of these errors is minimised, however, it is still necessary to have a sense of the uncertainty present in the ranking of influential physicists. Rather than estimating the effect of these errors, introducing noise into the network and then identifying the effect on the trends identified in the original network provides a thorough understanding of the stability of the ranking. The noise is introduced by randomly selecting a percentage of the existing edges, 5% in this case, and rearranging them, i.e., taking the edges that connect node A to B and C to D and swapping them, so A connects to D and B connects to C. This method produces an artificial network that is very similar to the original network but is slightly rearranged, representing potential erroneous connections produced by the writer of the page. By creating multiple distinct artificial networks that are each slightly altered, and comparing the multiple instances, the difference in rankings across the artificial networks depicts how stable the rankings are to slight perturbations introduced by errors.

Results

To be able to rank physicists by their influence on others, we need a way to quantify their importance in the network. In network analysis, a way to do this is to rank them according to various centrality measures. A centrality measure is a quantifiable definition of the importance of a node within a network, i.e., a property of a node that describes how central it is [7]. Many centrality measures could be used to produce a ranking of physicists, however here we utilise degree centrality, closeness centrality and betweenness centrality. These measures were calculated for each node in the original network, providing a ranking in terms of these measures. The same measures can be calculated for nodes in the artificial networks, allowing for a comparison of the rankings between networks.

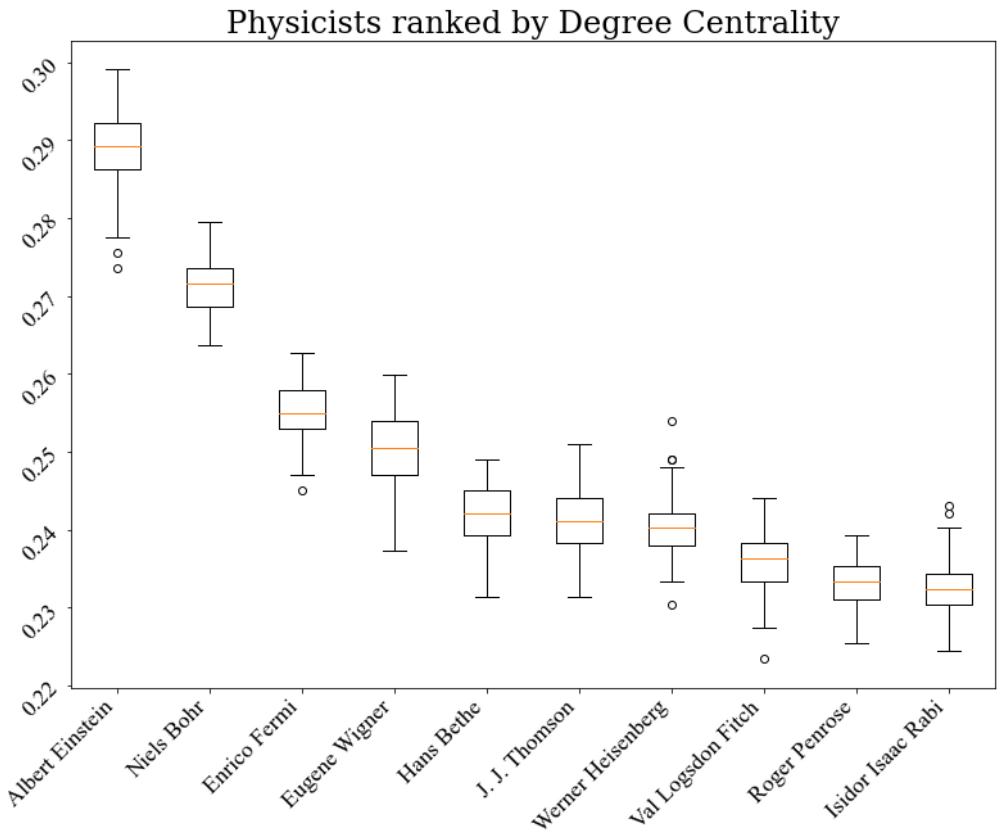

Figure 1: The ranking of physicists by decreasing degree centrality from left to right. The names of the physicists on the horizontal axis are in the order obtained from the original network and the box plots represent the error in the ranking displayed as obtained from the 10 artificial networks produced by introducing noise to the original network. The orange line is the median and the circles are outliers defined as being greater than 1.5 times the interquartile range from the median.

Degree centrality is the first measure of interest in ranking physicists by importance. The degree centrality is simply the number of edges to other nodes that one node has. This number is normalised by the number of possible edges that could branch from that node, which is the number of nodes in the network, n, barring the node of interest, so $n-1$ [7]. This provides a concise measure of how popular a physicist and their work is, as having the most connections suggests they have

produced work that is elementary to a field and have directly influenced many other physicists. For the original network, the degree centrality was calculated for each node, and this allowed the ranking to be formed as shown in the horizontal axis of Fig. 1. The box plots on this graph were produced by collating the degree centrality of each physicist from each of the 10 artificial networks that were produced with noise introduced.

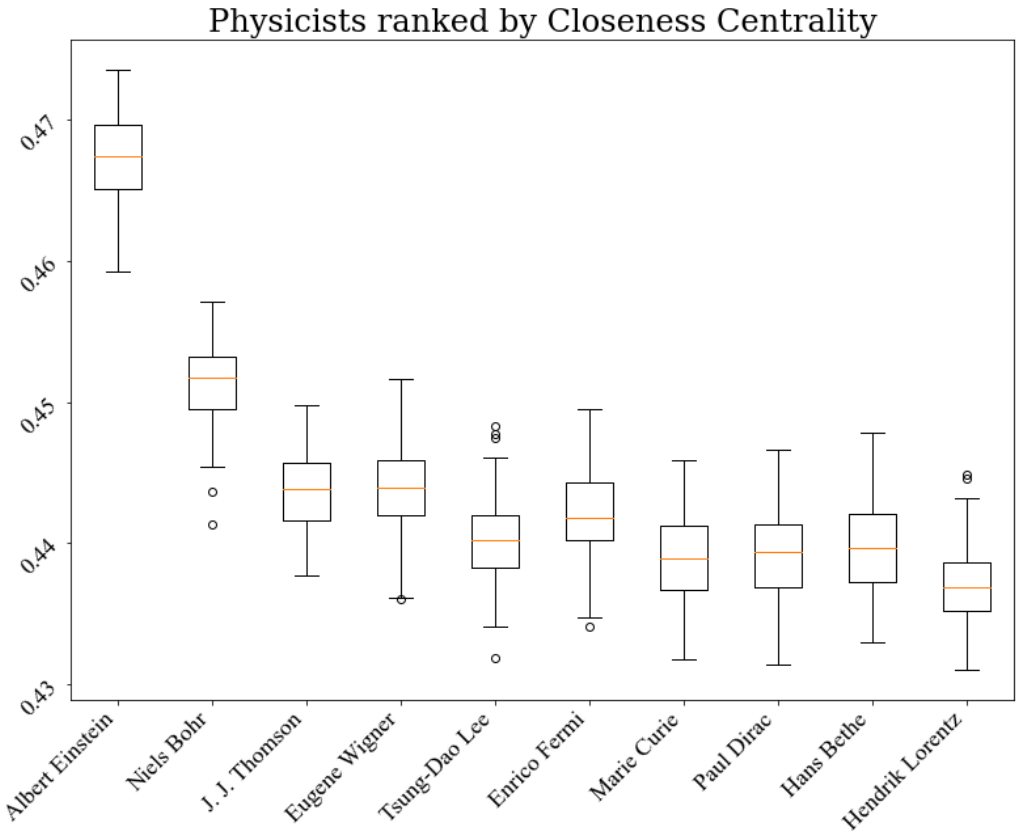

Figure 2: The ranking of physicists by decreasing closeness centrality from left to right. The names of the physicists on the horizontal axis are in the order obtained from the original network and the box plots represent the error in the ranking displayed as obtained from the 10 artificial networks produced by introducing noise to the original network. The orange line is the median and the circles are outliers defined as being greater than 1.5 times the interquartile range from the median.

Closeness centrality is the inverse of the average length of the shortest path from one node to all other nodes in the network. This measure is expressed as,

$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)}, \quad (1)$$

where *C(u)* is the closeness centrality of node *u*, *n-1* is the number of nodes in the network excluding *u* and the expression in the denominator is the sum of the shortest paths between the node *u* and all other nodes. This measure essentially describes how close a physicist is to all the other physicists in the network [8], which is different to degree centrality in the sense that it considers connections to all

other physicists, as opposed to those only directly connected to a specific node. Considering these extra connections allows for a greater understanding of a physicist's sphere of influence and reinforces the idea that a physicist with a high closeness centrality produced developments of great importance that could affect the connections between other physicists. The horizontal axis of Fig. 2 shows the ranking of physicists based on closeness centrality and the box plots represent the uncertainty in this ranking as obtained from the artificial networks.

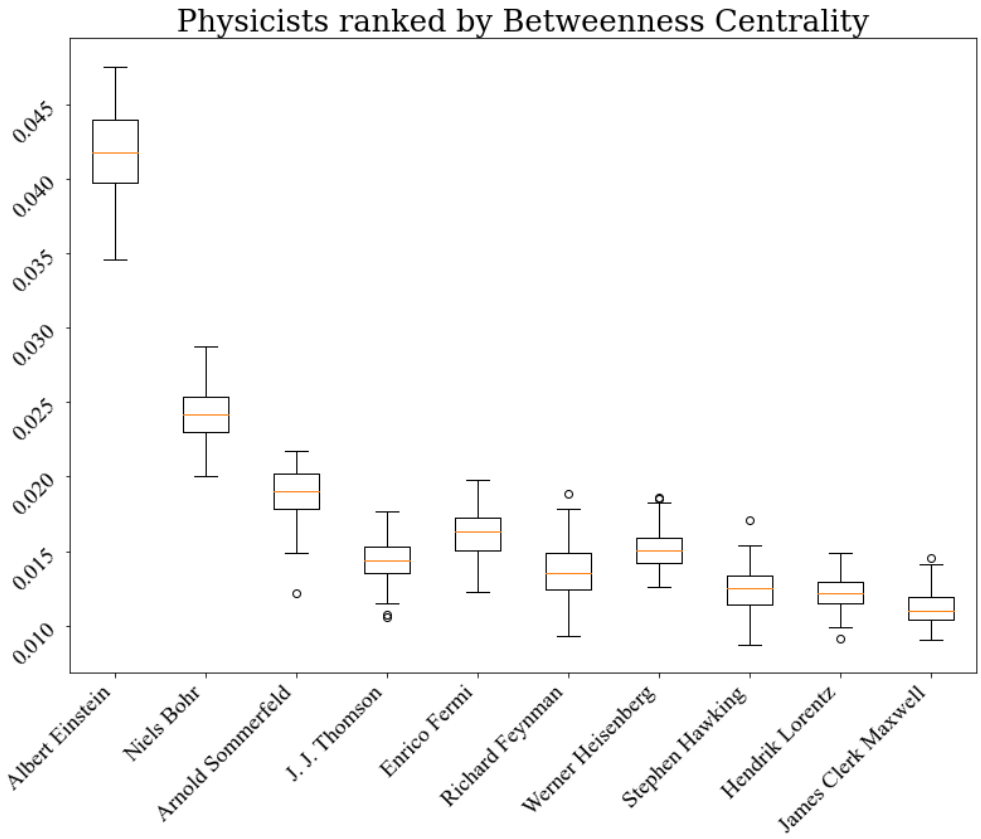

Figure 3: The ranking of physicists by decreasing betweenness centrality from left to right. The names of the physicists on the horizontal axis are in the order obtained from the original network and the box plots represent the error in the ranking displayed as obtained from the 10 artificial networks produced by introducing noise to the original network. The orange line is the median and the circles are outliers defined as being greater than 1.5 times the interquartile range from the median.

Betweenness centrality is the ratio of the sum of shortest paths between all pairs of nodes in the network that pass through a node against the total number of possible shortest paths in the network. This is expressed as,

$$B(u) = \sum_{x,y \in N} \frac{\sigma(x,y|u)}{\sigma(x,y)} , \quad (2)$$

where *B(u)* is the betweenness centrality of a node *u*, *N* is the number of nodes and the numerator and denominator are the number of shortest paths between two nodes, *x* and *y*, that pass through *u* and the number of shortest paths between all pairs of nodes, *x* and *y*, respectively. This measure is very useful

as it quantifies how instrumental a physicist is in connecting various physicists. It again suggests that a physicist may have produced significant developments or laid the groundwork for a field if they have a high betweenness score and can bridge the gap between various nodes [8]. The horizontal axis of Fig. 3 shows the ranking of physicists based on betweenness centrality and the box plots represent the uncertainty in this ranking as obtained from the artificial networks.

Discussion

From the results of the network analysis, there is strong evidence to suggest that Albert Einstein and Niels Bohr are the most influential physicists in the network of Wikipedia pages. By observing the uncertainty produced by the imposed noise as represented by the box plots on the graphs, it becomes clear that Albert Einstein and Niels Bohr hold positions in all three centrality measures that are very stable under the noise model. From the third position onwards in each of the figures, the whiskers of the box plots have considerable overlap, and in many cases, the medians are very similar. Both Fig. 2 and Fig. 3 contain examples of the rank of a physicist from the original network being different to the rank obtained by observing the median of a given centrality measure. One of these examples is the betweenness centrality of J. J. Thompson in Fig. 3; the ranking obtained from the original network suggests he holds the fourth position in terms of betweenness centrality however in comparison to the median obtained from the box plot, he would be placed nearer the seventh position. The data obtained from the noise model suggests that physicists that are below the second rank on the graph have a very similar influence on other physicists across centrality measures. The significant overlap in the error bars provides further evidence that these physicists have a similar importance as they show that, depending on the artificial network being examined, the order of rankings changes dramatically at lower levels, as the nodes share similar values of centrality and a small change due to noise will have a significant effect on the rank of a physicist. Extending the investigation beyond the top ten physicists shows a similar effect among the lower ranks, however, the overall rankings are the same since there are levels of influence that many physicists share, i.e., a physicist might move down by 30 positions in an artificial network but out of the 1021 ranks and in comparison with physicists of similar ranking, their level of influence in the original network is accurately quantified.

The criteria for a person's Wikipedia page to be added to the list of physicists is unclear, however, considering that Wikipedia is a set of open-access pages produced by many independent authors, the definition of a physicist can be someone who has contributed enough to physics that they are recognised by society and the physics community for their contributions to their field. For example, some may consider Roger Penrose to be a mathematician rather than a physicist, but his contributions are primarily to the mathematical physics of general relativity and cosmology. While this is a succinct definition that provides an intuitive representation of what it means to be a physicist, it is not comprehensive in the context of our network. There are several names that were expected to

appear among the top ranks that instead appeared at much lower rankings across the centrality measures. After further investigation of individual Wikipedia pages, the reason for this became apparent. There are several physicists whose importance is concisely represented by the importance of their discoveries. Many of these larger discoveries became entire fields, like Einstein's theory of general relativity, but some also have produced work that is named after them, such as Newtonian mechanics. This means that on many Wikipedia pages, the author had decided to mention the theorem or field that a physicist had developed and included the link to that page rather than the page of the physicist who produced it. One example of this is Hamiltonian mechanics, a well-known formalism that bridges the gap between classical and quantum mechanics. On many pages, there will be a link to the Wikipedia page for Hamiltonian mechanics but no link to William Hamilton's page. Other examples include Noether's theorem when talking about conservation laws and Newton's theory of gravitation in the context of comparison with general relativity.

It is important to note that the Wikipedia pages that were analysed were English language pages. This suggests that they were written primarily by writers from the Western hemisphere. This simple fact implies that they would most likely have been raised with western ideologies and been surrounded with information that also has this bias. It can be argued that since the writers likely do the majority of their research on the Internet, where information is easily accessible and can be curated from multiple international sources, this bias can be given less attention, however, it cannot be assumed that every writer is aware of this western bias while they write, so there will be an inherent focus on the achievements of western physicists, even if it is only slight. This suggests that while we have evidence to claim that Albert Einstein is the most important physicist, looking at other languages' Wikipedia pages may have a different focus and produce a different result. Another possible bias would be toward male physicists; while there is progress in the subject of being more inclusive to female and minority physicists, seeing this progressive attitude be represented in the pages of female physicists will take more time while these attitudes become the norm.

Conclusion

By searching through a list of over 1000 physicists' Wikipedia pages, a network was formed to represent these physicists and the connections between them. By employing some network analysis techniques, evidence was gathered to suggest that Albert Einstein is the most important physicist, followed by Niels Bohr. A possible extension of this project is utilising a python package called NLTK (Natural Language ToolKit) [10]. This package would allow us to read the actual text of the pages and extract extra information such as the nationality of a physicist, the fields that they mainly contributed to and the beginning and end dates of their careers. By attaching these attributes to each node, in-depth analysis of the most important physicist within a field, country or timeframe can be carried out. Another application of this package is identifying the context surrounding a hyperlink to another

physicist's page. This would allow for directed edges to be added to the network if the uncertainty of using NLTK to do this is within a threshold, giving a clearer picture of the direction of influence between physicists.

References

[1] List of physicists [Internet]. Wikipedia. Wikimedia Foundation; 2022 [cited 2022Apr24]. Available from: https://en.wikipedia.org/wiki/List_of_physicists
[2] Beautiful Soup documentation¶ [Internet]. Beautiful Soup Documentation - Beautiful Soup 4.9.0 documentation. [cited 2022Apr24]. Available from: https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[3] Evans T. Network notes. London: Physics Department, Imperial College London; 2022. (Complexity and Networks Course).
[4] NetworkX documentation [Internet]. NetworkX. [cited 2022May1]. Available from: https://networkx.org/
[5] Accented Characters and Ligatures in HTML and JavaScript [Internet]. Accented characters and ligatures in HTML and JavaScript. [cited 2022Apr30]. Available from: http://www.javascripter.net/faq/accentedcharacters.htm
[6] Administrators [Internet]. Wikipedia. Wikimedia Foundation; 2022 [cited 2022May7]. Available from: https://en.wikipedia.org/w/index.php?title=Wikipedia%3AAdministrators&oldid=1086361466
[7] Newman MEJ. 7.1 Centrality. In: Networks. Oxford: Oxford University Press; 2019.
[8] Andrew Disney 2nd J2020. Social network analysis: Understanding centrality measures [Internet]. Cambridge Intelligence. 2022 [cited 2022May9]. Available from: https://cambridge-intelligence.com/keylines-faqs-social-network-analysis/
[9] Chen B, Lin Z, Evans TS. Analysis of the Wikipedia network of Mathematicians [Internet]. arXiv.org. 2019 [cited 2022May10]. Available from: https://arxiv.org/abs/1902.07622
[10] NLTK. [cited 2022May10]. Available from: https://www.nltk.org/