Y3 Project: Using Network and Natural Language Processing to Analyse the Wikipedia Physicists

Author: Kaiyu Hu
CID: 01720985
Supervisor: Dr Time Evans
Assessor: Dr Dave Clements
Word count: 5481

Declaration of work undertaken:
This project consists of two main parts: clustering physicists by natural language processing and using network centrality measures to determine the rank of physicists. The former part is completed by the project partner Chao Fan. The author Kaiyu Hu has done the latter part.

Contents

Abstract

In this project, we use network centrality measures and natural language processing to analyse 1019 physicists' English Wikipedia web pages. A rank of physicists is produced based on our ranking scheme (ranking in order of average mark which is the mean of five rescaled centrality measures). It suggests that Albert Einstein is the most famous physicist, followed by Niels Bohr, Enrico Fermi and Werner Heisenberg. The robustness of this scheme is tested by our noise model which simulates the action of the edition of the Wikipedia website, and the overall rank is not influenced by noises significantly. We also analyse the text in the biographies of physicists. K-means clustering is applied to divide them into five clusters. The feature of each cluster can be summarised as: 'German', 'Quantum Theory', 'Nobel', 'Atomic and Nuclear', 'Generally Well-known'. We combine the clustering with our network to obtain a rank in each group. From comparisons within and between clusters, we find that many influential physicists work on Quantum, Atomic and Nuclear Physics. If physicists do not have significantly great achievements, nationality or winning a Nobel Prize will become one of their mean features. We also determine that the average mark of a physicist is linearly positively correlated with the length of the biography in log-log space. As anyone is able to edit the English Wikipedia, there is a bias in favour of English-speaking Physicists, which means our ranks and clusters will be different if we analyse Wikipedia in other languages.

1 Introduction

From the history of Physics, it can be seen that ideas are passed between physicists in different countries and eras. Theories and experiments inspire new ones. By looking into the biographies of physicists, we can find how the research topics are carried out and linked together. With the development of the internet, people can get access to information about physicists easily. It is common that a few questions may come up in the daily life:

1. Who is the most famous/important physicist?
2. Are there any ranks of physicists? If so, are they reasonable?
3. Do physicists have some general features?
4. Can we divide physicists into some groups? And How?

There are different answers to the questions determined by qualitative or quantitative methods[1][2]. In this project, we investigate the English Wikipedia of the biographies of physicists. Natural language processing and network centrality measures are used for clustering and ranking. The outcome of this project is a rank of physicists with uncertainties, which answers the first two questions. We also allocate physicists in five clusters with corresponding keywords based on the text in their biographies, and this gives answers to the last two questions.

2 Data and Methods

2.1 Extracting the Biographies of Physics & Data cleaning

The term 'physicists' is defined by a list of 1019 physicists on the English Wikipedia website that contains hyperlinks to their biographies. Request (a Python package) is used to read the URLs and download all HTML files to the local drive only once. All the following modifications are applied to the local files. The Python package Beautifulsoup is used to analyse these HTML documents.

We only consider the text in the main body of the website (we call this main text) for Natural Language Processing (NLP) and the hyperlinks in the main body for network analysis. The term 'main body' is the sections that remain after the processes of data cleaning, and Fig.1 shows the removed sections. The trial and error method was implied to find all of them. Regions A and B are toolbars; C is the sentence that appears on every Wiki website; D is the contents. These regions clearly do not contain significant information about the physicist.

However, region E in Fig.1, which consists of sidebars, does contain some key information (age, gender, college and etc.). Unfortunately, some sidebars also contain a list of names of scientists who have done great contributions to a field of Physics (e.g. Statistical Physics, Quantum Mechanics), and this results in an overwhelmingly large contribution to the degree of a physicist when doing centrality measures (see definitions in Section 2.3). It cannot be denied that physicists in the list have connections as contributors to a subject, but not all of them have the lists in the sidebar. For example, On Yang's website, the name Einstein can be seen in the sidebar 'Modern Physics', but on Einstein's website, there is no such a sidebar. Therefore, we decide to delete all the sidebars to ensure a fair comparison.

In the References section (region F in Fig.1), the text consists of article titles, authors, journals, publishers, doi/ISBN/page numbers and etc. As references have a certain format, some terms like 'doi' have a high frequency, which disturbs our text analysis. Not all the titles in References belong to scientific articles, so some of them are subjective or have commercial purposes, which is not suitable. One subtle point here is that Wiki uses different words to entitle this section, such as sources, notes, citations and etc., so we develop an algorithm that can detect

all of these words and remove the section. We also make sure that the section title is followed by '[edited]' as the same word may appear in the main text.

The last section in the Wiki website is usually External links (Region G in Fig.1) that contains several lists of hyperlinks of names (e.g. Nobel laureates, Presidents of American Physical Society), which causes the same issue as the sidebar (large degree). This is also unfair to the early physicists (e.g., Newton) when the Nobel Prize and other organisations were not established.

Although those sections that are cut off may contain some useful information, it must be mentioned in the main text as well if it is truly significant. Words with less than three letters are deleted as they are usually prepositions, pronouns, abbreviations of names that have no significance. In our cleaned data frame, we have the name of physicists with hyperlinks on their websites and the reduced main text of the biography.

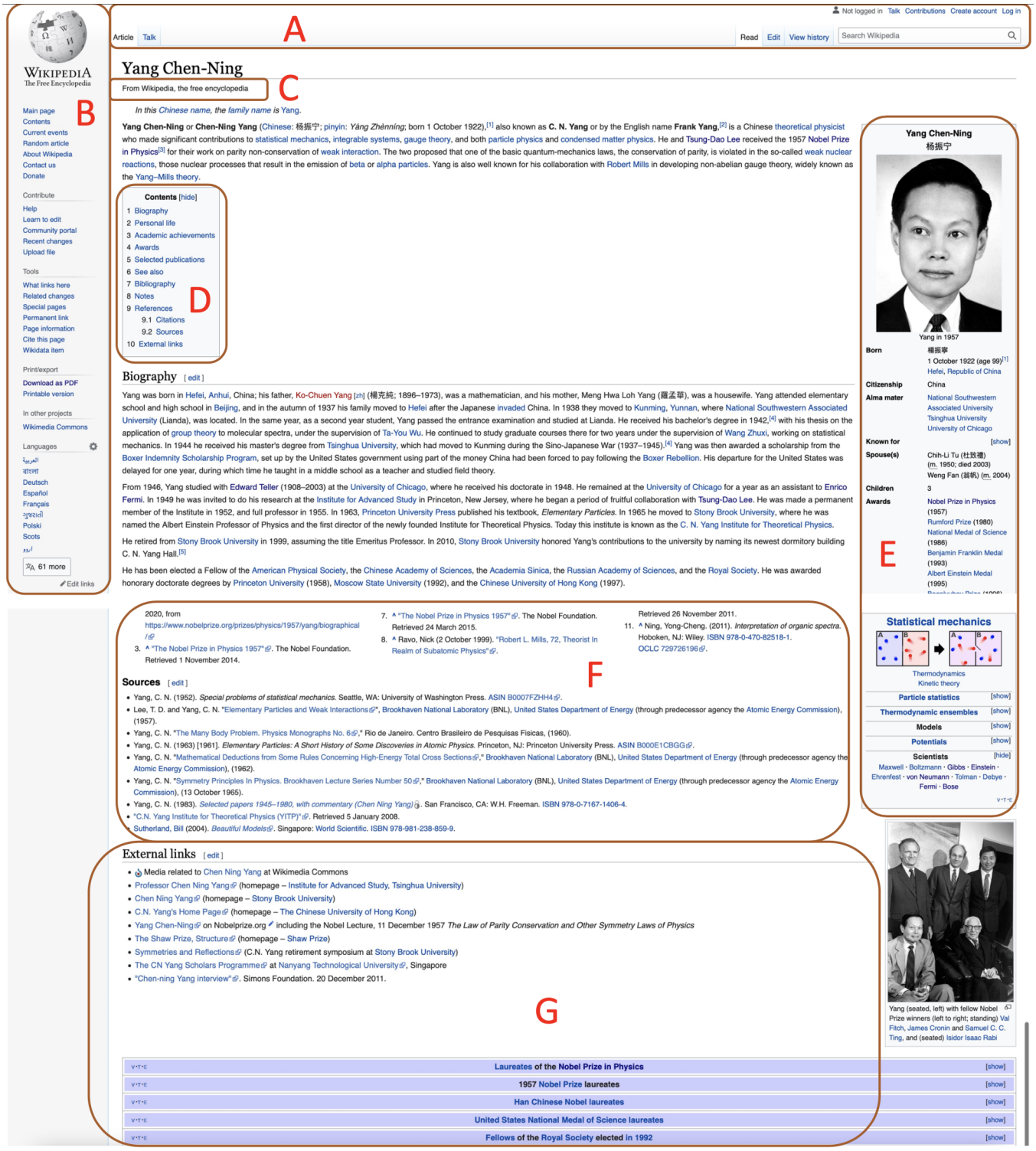

Figure 1: The screenshot of Yang Chen-Ning's Wiki website (`https://en.wikipedia.org/wiki/Yang_Chen-Ning`) to illustrate data cleaning. The regions (A-G) enclosed by red line are cut off. A,B: toolbars; C: 'From Wikipedia, the free encyclopaedia'; D: contents; E: sidebars; F: references; G: external links

2.2 Clustering by NLP

We now have a collection of documents (biographies of physicists). To divide them into clusters based on the feature of the text, we firstly tokenise the raw text, splitting it into small chunks called tokens. N-gram is used for tokenisation, which means we consider not only the single word but also the contiguous sequence of n terms (e.g. 'American Physical Society' is also a token). lemmatisation is performed to convert the word into its base form. We also set a list of stop words that could not indicate the feature of a physicist (e.g. the word 'Physics', 'equation', which occurs nearly every page). Tfidf vectorizer from Natural Language Toolkit (NLTK) is used to convert a collection of raw documents to a matrix of TF-IDF features. K-means clustering from sklearn is used to determine the clusters, and the optimal number of clusters is obtained by the Elbow method. Our output is five clusters with a list of weighted words showing the features. Word cloud is produced for better illustrations.

2.2.1 Lemmatisation

Lemmatizer from NLTK is used to convert the inflected forms of a word into its base form, called lemma (e.g. 'walk' has several inflected forms: 'walks', 'walked', 'walking'). We also tried to use Stemmer from NLTK, which makes the clustering algorithm run faster, but the results turned out inaccurate (e.g. the verb 'elect' and the noun 'electron' are both identified as 'elect'), so we rejected it. Unlike stemming, lemmatisation can select the appropriate lemma of a word.

2.2.2 Stop words

The NTLK package has a built-in library of stop words. We also add extra ones manually. In Table 1, there are some popular English and French names, which are meaningless since we cannot find out which one belongs to whom. Other words like 'Physics' and 'physicists' do not have any significant information to the analysis of the physicist's biography. We keep extending the stop-word list until all the keywords produced by our program are informative in representing the content of a text.

James	Robert	John	George	William	Prize
der	des	van	von	research	professor
Biography	Sciences	also	Institute	Physics	physicist

Table 1: Part of stop words added for clustering

2.2.3 TF-IDF Matrix

TF-IDF means term-frequnecy (TF) times inverse document-frequency (IDF). TF is the number of times a term t occurs in a given document, and IDF is computed as:

$$IDF(t) = log\frac{1+n}{1+DF(t)} + 1\,, \tag{1}$$

where n is the total number of documents in the collection, and DF(t) is the number of documents in the collection that contain the term t. The raw of the TF-IDF matrix shows each the lemma (base form of the word) in our document set and the column represents the index of the document (Fig.2). The element of the matrix is the TF-IDF value of a lemma in a certain document that determines how informative it is, therefore lemmas with relatively high TF-IDF values, compared with those in other documents and the other lemmas in the same document, demonstrate the features of the physicist.

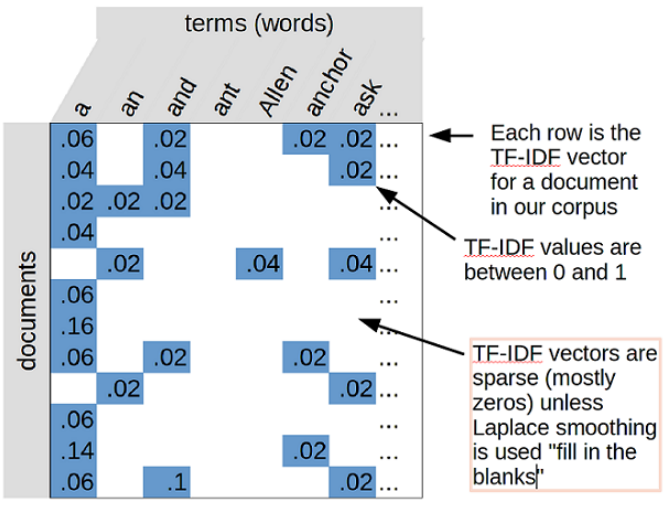

Figure 2: The diagram to illustrate the TF-IDF Matrix (not the one in this project). Figure reproduced from [3]

2.2.4 K-means and Elbow Method

K-means clustering is an unsupervised machine learning algorithm from sklearn which runs fast (the only advantage)[4]. Given a collection of vectors $(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n})$, each of which contains the TF-IDF values of lemmas in a document, K-means clustering split the n vectors (corresponding to n documents) into k sets $\mathbf{S} = \{S_1, S_2, ..., S_k\}$, minimising the within-cluster sum of squares (WCSS). The K-means algorithm aims to find:

$$\underset{\mathbf{S}}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathbf{S_i}} ||\mathbf{x} - \boldsymbol{\mu}_\mathbf{i}||^\mathbf{2}, \tag{2}$$

where $\boldsymbol{\mu}_i$ is the mean of points in S_i. The steps of K-means clustering are[4]:

1. k points are selected arbitrarily as the initial cluster centers.
2. The data points are allocated to clusters based on the distance to the centers.
3. The centers are recomputed as the mean of points in the cluster.
4. Steps 2 and 3 are repeated until the WCSS is minimised.

One problem with K-means is that it required the number of clusters k to start with. The Elbow method is used to find the optimal cluster number by fitting the model with a range value of k (Fig.3). The point of inflection ('elbow') on the curve indicates the model fits best at that point. Five is determined to be the optimal number of clusters. As K-means is stochastic (it starts with random points as initial cluster centers), we repeat the Elbow method ten times. Every time the result shows five is the best number.

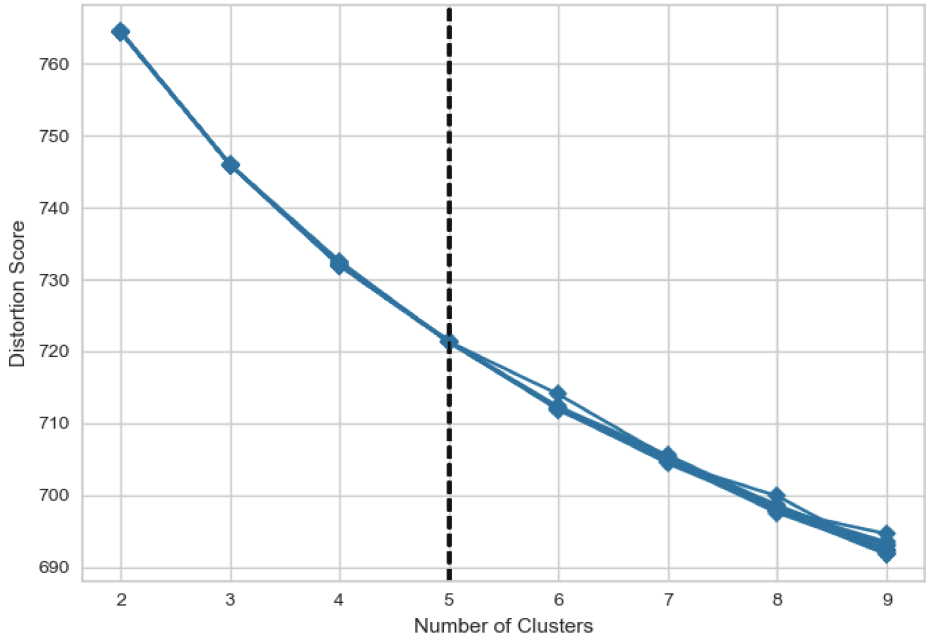

Figure 3: Distortion Score against number of clusters for ten runs. Elbow method indicates that five is the optimal number of clusters. The fluctuations show that clusters are different for each run due the stochastic nature of K-means.

2.2.5 Other methods to determine the optimal number of clusters

Apart from the Elbow method, we have tried the Average silhouette method and the Gap statistic method [5] to determine the optimal number of clusters for K-means clustering. Both of them are rejected as they suggest the more clusters we have, the better clustering we can achieve (i.e. a smaller WCSS). It is trivial that the WCSS is zero if each cluster has one person only, but that misses the point of clustering.

2.3 Definition of network & centrality measures

We define the node as the name of a physicist in the Wiki list. An unweighted edge will be added to a pair of physicists if there is at least one hyperlink in one's website that links to the other (the direction doesn't matter). For example, in Fig.4, if we find a hyperlink to Einstein's page on the website of Newton, there will be an edge between them, and vice versa. Our hypothesis is that this hyperlink shows a significant connection between two people (e.g. They worked together; They were students and tutors; One purposed a theory based on the research of others).

As we have the name of physicists and hyperlinks on their websites, we can determine people that are connected to a physicist by analysing the URLs of the hyperlinks. The URL of the physicist's Wiki page has a fixed format (https://en.wikipedia.org/wiki/Yang_Chen-Ning). We split the URL by '/', and the string after the last '/' is the name. It is converted to utf-8 so that all the names including non-English ones in the URL can be matched to the names in the list.

The network contains 1019 nodes, 3349 edges and 200 components. As some physicists have a fairly short biography, there are 191 components with a single node. The largest connected component (LCC) contains 812 nodes, and we apply the centrality measures (except for degree) to the LCC only. Centrality means the importance of a node in the network. Five centrality measures are used to give different aspects to evaluate the importance of a physicist. The following subsections explain the definitions of measures in detail.

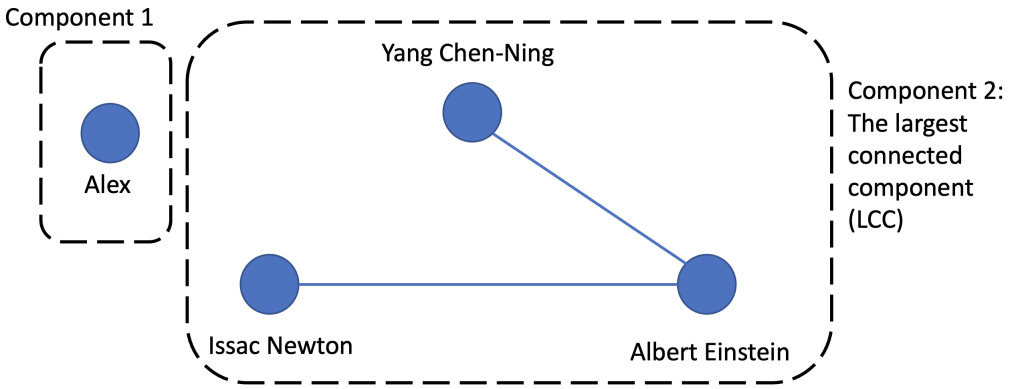

Figure 4: A simple network for illustration (not the one in this project), consisting of four nodes and two components. The LCC has three nodes

2.3.1 degree

The degree k of node i is simply the number of edges incident to it[6]. For instance, in Fig.4, the degree is 2 for Einstein and 0 for Alex. It is the simplest centrality measure, but the drawback of only considering the nearest neighbours makes it not sufficient enough to describe how famous a physicist is.

2.3.2 closeness

The closeness centrality c of node i is defined by:

$$c_i = \frac{n-1}{\sum_{v=1}^{n-1} d(v,i)}\,, \tag{3}$$

where n is the number of nodes in LCC; d(v, i) is the shortest-path distance between node v and i. Compared with degree centrality, the closeness takes all the nodes in the LCC into account, but it only considers the shortest paths as the important routes, and the other paths are ignored[7].

2.3.3 betweenness

For the betweenness centrality b of node i, we have:

$$b_i = \sum_{s,t \in LCC} \frac{\sigma(s,t|i)}{\sigma(s,t)}\,, \tag{4}$$

where s,t are nodes in LCC; $\sigma(s,t)$ is the number of shortest (s, t)-paths; $\sigma(s,t|i)$ is the number of those paths passing the node i; if s=t, $\sigma(s,t) = 1$; if i=s or i=t, $\sigma(s,t|i) = 0$. Like closeness, betweenness only uses the shortest path to measure the centrality.

2.3.4 eigenvector

The eigenvector centrality of node i is the i-th entry of the eigenvector of A associated with the largest eigenvalue, where A is the adjacency matrix that A_{ji} is one/zero if there is a link/no link from node i to j [8]. If the neighbours of a physicist have high eigenvector centralities, the physicist will have a high eigenvector centrality as well.

2.3.5 pagerank

The entry T_{ji} of a transfer matrix T represents the probability of a random walker moving from node i to j at the next time step:

$$T_{ji} = \frac{1}{s_i^{(out)}} A_{ji} \qquad s_i^{(out)} = \sum_j A_{ji} \tag{5}$$

The random walker has a probability α to follow the link chosen at random, while there is a probability of $(1-\alpha)$ that the current walk is stopped. This process can be described by the Markovian matrix G:

$$G_{ji} = \alpha T_{ji} + (1-\alpha)\frac{1}{N}\,, \tag{6}$$

where N is the number of nodes in LCC; the damping factor $\alpha = 0.85$. Pagerank of node i is proportional to the probability that a random walker stays at this vertex in the long time limit. It equals the i-th entry of the eigenvector of G with the largest eigenvalue.[1]

2.4 Noise Model

The edges are not definite as the Wikipedia pages can be edited and have been edited many times during the past years. To simulate the action of the edition, we build a noise model that removes 5% of edges randomly from the original network and adds them back randomly for one simulation. The edges are modified randomly in the edge list. We have done 100 simulations and calculated the standard deviation for each centrality measure and the average mark as the uncertainty. In principle, we should look into the historical data to determine how many edges are changed and then obtain the ratio. Here we assume 5% of hyperlinks are altered.

2.5 Ranking scheme

The ranking scheme is to sum up all the five rescaled centrality measures and take the average. The physicists are ranked in order of average mark. Rescaled measure C_i' for physicist i is defined by:

$$C_i' = \frac{C_i}{\mathrm{Max}(C_i|i \in V)} \times 100\,, \tag{7}$$

where C_i is the original centrality measure and V is the largest connected component. Therefore, every physicist has a mark between 0 and 100 for each measure, and the average mark reflects popularity.

There are many other ranking schemes (e.g. calculate the rank of each centrality measure and then rank physicists by the average rank), and there is no such 'best' one. Our scheme is good for demonstration as all five centrality measures are considered and the differences between physicists can be observed clearly.

3 Results and Discussion

3.1 Degree Distribution

In Fig.5, it is good to see that the degree distribution of our network is fat-tailed, as this is expected for many networks[9][10][11][12][13]. The straight line in the log-log scale indicates that the probability of degree follows the power law[14]. This graph helped us develop the data cleaning processes (in Section 2.1). When we kept the external links in the network analysis, some anomalous data appeared in the right region which means our network had an unexpectedly large number of people with a high value of degree. As mentioned, this is caused by the lists of Nobel laureates or presidents of the American Physical Society in the External Links section. After deleting that section, not all anomalous data were removed, and hence we spotted that the problem was caused by the sidebars, which contain the names of physicists who made great contributions to a specific field.

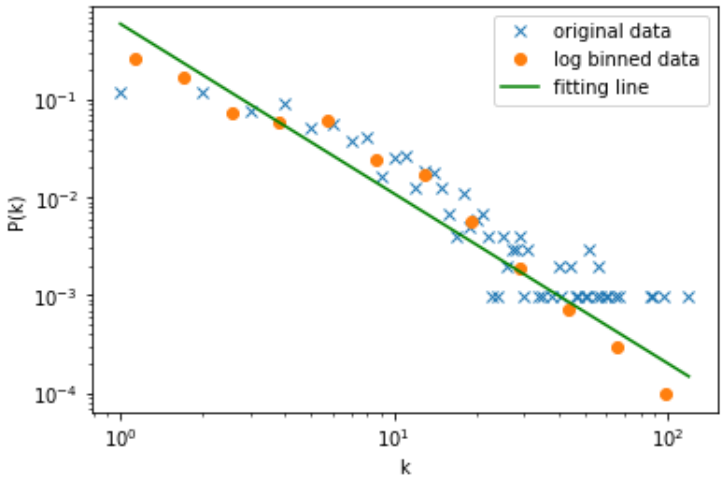

Figure 5: Degree distribution of physicists hyperlink network in log-log scale.The orange dot represents the same data binned with logarithmic bins (the ratio of consecutive bin edges is set to be 1.5). The slope of the line of best fit is 1.7 ± 0.1.

3.2 Overall ranking

Table 2 shows the best thirty physicists based on our ranking scheme (in Section 2.5). It suggests that Albert Einstein is the most popular physicist, followed by Niels Bohr, Enrico Fermi and Werner Heisenberg. The average mark of Einstein is 100, which means he has the highest score in every centrality measure. It is also noticeable that the uncertainties of all centrality measures except for eigenvector are 0, which indicates that the noises have almost no effects on his centrality. The large gap in average marks between Einstein and the others shows he had made an extraordinary influence on Physics. The 2nd, 3rd, and 4th physicists (Bohr, Fermi and Heisenberg), who are famous for Quantum Physics, have an average mark between 70-80. The gap between the 4th and 5th is relatively large as well, but after that, the change in average mark becomes more smooth. Most people have an average mark below 35.

Looking into the uncertainties, we find that there are several groups of physicists with similar average marks (e.g. Teller and Rutherford, Schrödinger and Maxwell), where the difference in average is within the range of uncertainty and therefore we cannot determine who is better than the other. However, the overall rank is not affected by the noises significantly.

It can be seen that the maximum value of Einstein's average mark is over 100. This is because the values of centrality measures and the average mark are determined from the original network (no edges are changed). The uncertainty is generated by our noise model, which is the standard deviation of 100 simulations. We didn't choose the mean as the value of the average mark since it cannot be obtained from the internet now, and we believe the value calculated from the original network gives a more precise rank of physicists based on the available information online.

Rank	Name	Degree	Closeness	Betweeness	Eigenvector	Pagerank	Average
1	Albert Einstein	100.00±0.00	100.00±0.00	100.00±0.00	100.00±1.42	100.00±0.00	100.00±0.28
2	Niels Bohr	80.83±1.56	93.47±0.87	39.89±3.57	99.28±2.74	75.78±1.77	77.85±1.47
3	Enrico Fermi	72.50±1.80	94.09±0.85	51.60±3.83	87.15±3.38	70.46±1.87	75.16±1.88
4	Werner Heisenberg	73.33±1.22	92.43±0.79	28.51±3.52	93.54±2.86	65.39±1.36	70.64±1.27
5	Max Born	55.83±1.19	89.19±0.79	16.70±2.46	77.81±2.89	51.10±1.25	58.13±1.20
6	Hans Bethe	53.33±0.64	90.65±0.73	27.49±2.79	67.71±2.29	51.43±0.97	58.12±0.81
7	Arnold Sommerfeld	50.83±1.51	86.70±0.83	32.28±2.24	56.96±2.62	57.45±1.86	56.85±1.41
8	Paul Dirac	46.67±1.05	93.66±0.87	31.36±2.57	64.12±2.68	44.76±1.17	56.11±1.19
9	Isaac Newton	50.00±1.24	89.15±0.79	51.42±3.14	28.01±1.66	54.93±1.51	54.70±1.30
10	Lise Meitner	47.50±1.02	89.11±0.80	17.97±2.10	68.12±2.75	42.28±1.08	52.99±1.12
11	Richard Feynman	45.83±1.08	88.72±0.84	32.15±2.79	48.25±2.32	49.51±1.23	52.89±1.18
12	Wolfgang Pauli	43.33±1.05	89.06±0.76	18.56±2.10	65.96±2.64	42.36±1.27	51.85±1.15
13	Edward Teller	43.33±1.17	89.37±0.98	20.49±2.16	57.34±2.27	43.06±1.28	50.72±1.26
14	Ernest Rutherford	46.67±1.27	88.76±1.00	26.24±2.28	44.29±2.22	46.15±1.35	50.42±1.27
15	Max Planck	42.50±0.94	87.66±0.76	13.64±2.10	58.64±2.64	39.02±1.05	48.29±0.96
16	J. Robert Oppenheimer	41.67±0.67	88.72±0.80	17.52±2.32	53.65±1.99	39.67±0.99	48.25±0.82
17	Erwin Schrödinger	39.17±1.00	89.02±0.71	12.69±1.62	57.46±2.35	36.21±1.02	46.91±0.92
18	James Clerk Maxwell	43.33±1.46	85.53±0.79	28.92±2.00	30.57±1.73	44.64±1.48	46.60±1.18
19	Isidor Isaac Rabi	38.33±0.82	86.54±0.81	11.37±1.66	54.96±2.22	35.33±0.91	45.31±0.83
20	Rudolf Peierls	36.67±0.81	84.89±0.67	11.62±1.77	53.32±2.15	35.31±1.05	44.36±0.80
21	Eugene Wigner	36.67±0.68	86.62±0.91	9.68±1.59	53.67±2.35	33.96±0.74	44.12±0.78
22	John von Neumann	33.33±0.72	83.92±0.73	12.39±1.62	38.57±1.92	34.64±0.94	40.57±0.76
23	Otto Hahn	33.33±0.53	82.12±0.70	4.65±1.84	48.31±2.05	29.42±0.79	39.57±0.60
24	J. J. Thomson	34.17±0.77	83.69±0.91	16.25±1.96	26.65±1.80	36.65±1.02	39.48±0.85
25	Stephen Hawking	31.67±0.99	80.07±0.75	33.06±2.59	9.34±0.65	40.73±1.34	38.97±0.98
26	Lev Landau	25.83±1.24	85.61±0.92	20.78±2.03	30.86±2.15	29.57±1.36	38.53±1.25
27	James Chadwick	29.17±0.94	84.23±0.71	4.76±0.95	43.61±2.26	26.44±0.89	37.64±0.80
28	George Gamow	28.33±0.90	86.82±0.79	11.22±1.16	33.38±1.85	28.09±0.87	37.57±0.83
29	Subrahmanyan Chandrasekhar	24.17±0.54	84.00±0.78	13.96±1.28	25.57±1.55	27.81±0.87	35.10±0.64
30	Arthur Compton	25.83±0.92	81.39±0.80	6.39±1.34	34.20±2.10	24.82±1.11	34.52±0.82

Table 2: Top thirty physicists ranked in order of the average of the five rescaled centrality measures.The uncertainty is determined by the noise model (in Section 2.4). Results are given in 2 d.p. for simplicity.

In Fig.6, the variation in the ranks of centrality measures for each physicist increases with the rank of average mark (especially for betweenness and eigenvector). There is an outlier (the red hexagon for eigenvector centrality) corresponding to Stephen Hawking. A relatively low eigenvector centrality means he is a distance away from the center of the network. As he is one of the latest physicists in the top thirty, and his research about black holes is cutting edge, the strong connections between him and earlier physicists haven't been built up yet.

Another way to demonstrate the rank of physicists is a directed acyclic graph (Fig.7) with transitive reduction[1]. The vertical level of physicist illustrates different tiers of importance. Although the effects of uncertainty are ignored, the graph produces a neat visualisation of our data. The various colors in the graph represent the cluster that the physicist belongs to, and the results of clustering will be shown in the next section.

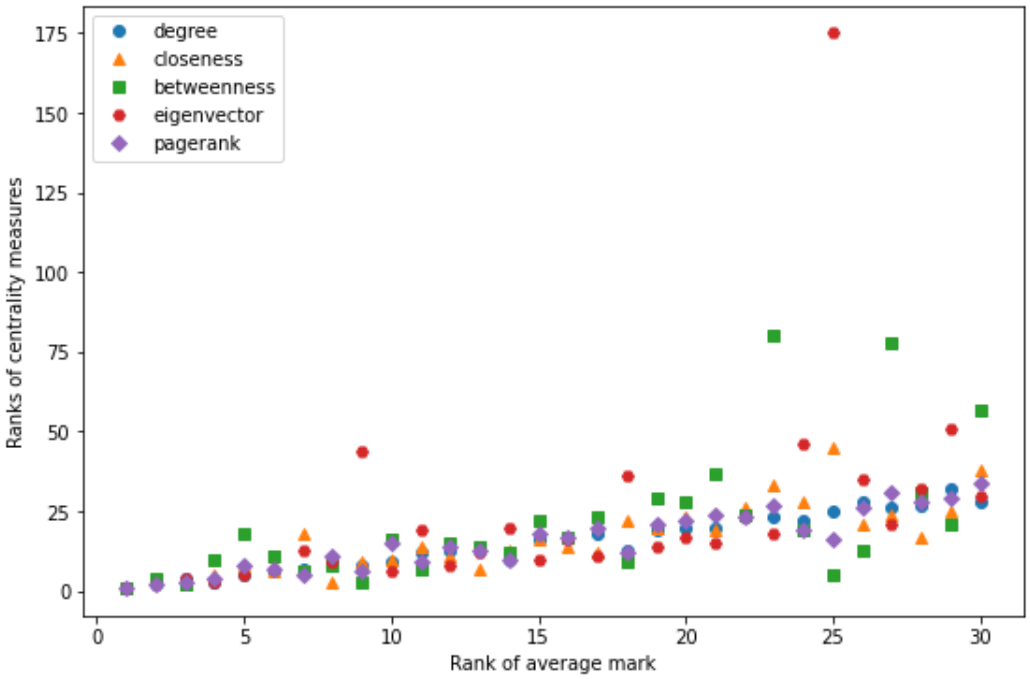

Figure 6: Comparison of rank of each centrality measure of the top thirty physicists ranked in order of the average mark

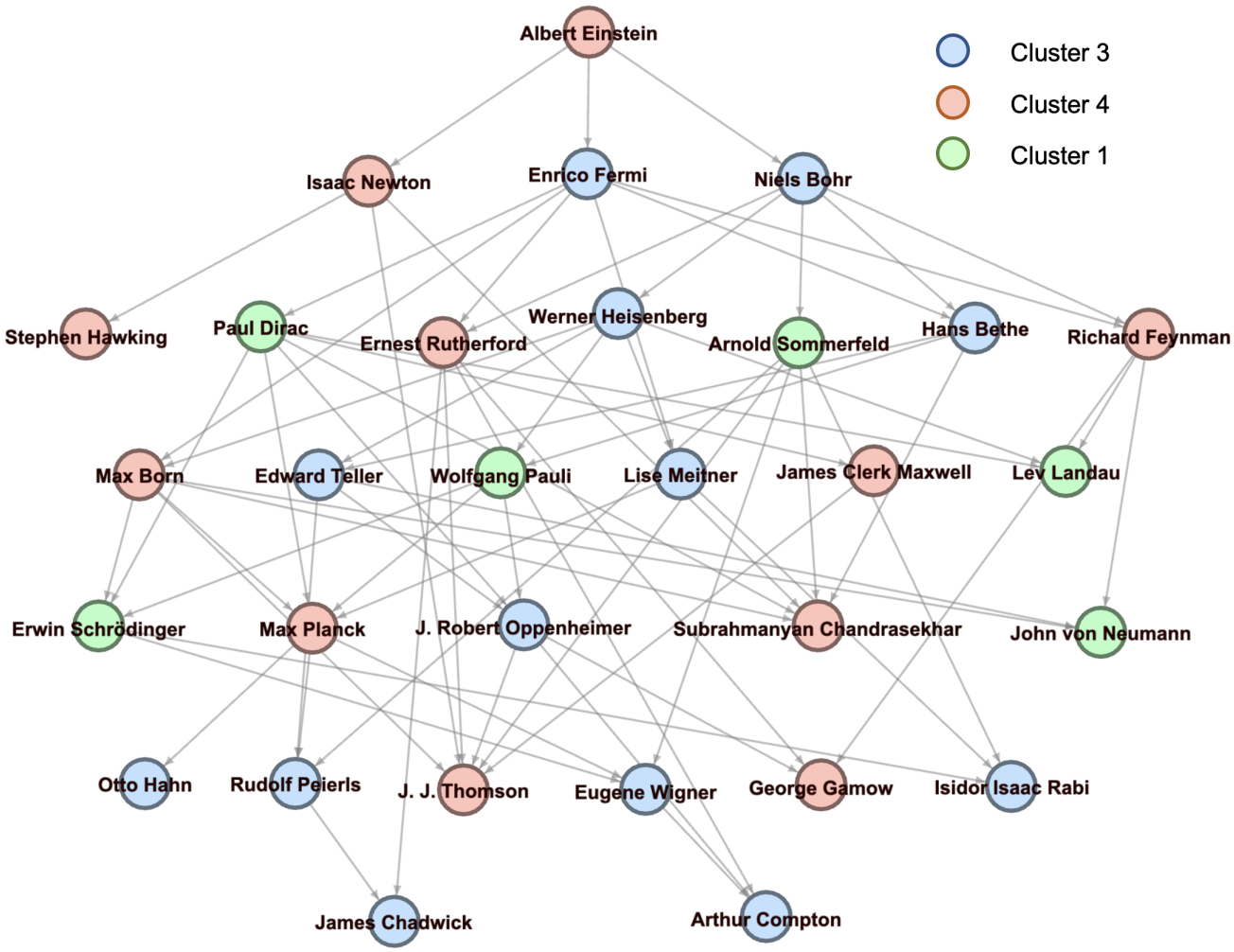

Figure 7: The directed acyclic graph of the top thirty physicists after transitive reduction. Colours indicate the cluster (see more details in Section 3.3). If physicist 'A' has a high value in every centrality measure than physicist 'B', an arrow is drawn from 'A' to 'B'. The vertical level of the physicist is defined by the longest path to the top node (i.e., Albert Einstein). Transitive reduction minimises the number of edges while still keeping the same reachability as the original graph, which makes the graph easier for visualisation.

3.3 Properties of clusters

Using the methods mentioned in Section 2.1, we divide physicists into five clusters. The number of people in each cluster is 141, 153, 176, 120, and 222. We then combine the clustering with the network we built. It can be seen that we have three relatively large clusters (4,3,1) and two small ones (2,0) in terms of the size of the nodes (i.e. the importance of the physicist). From the DAG (Fig.7), we can see that the top thirty physicists are in the three large clusters.

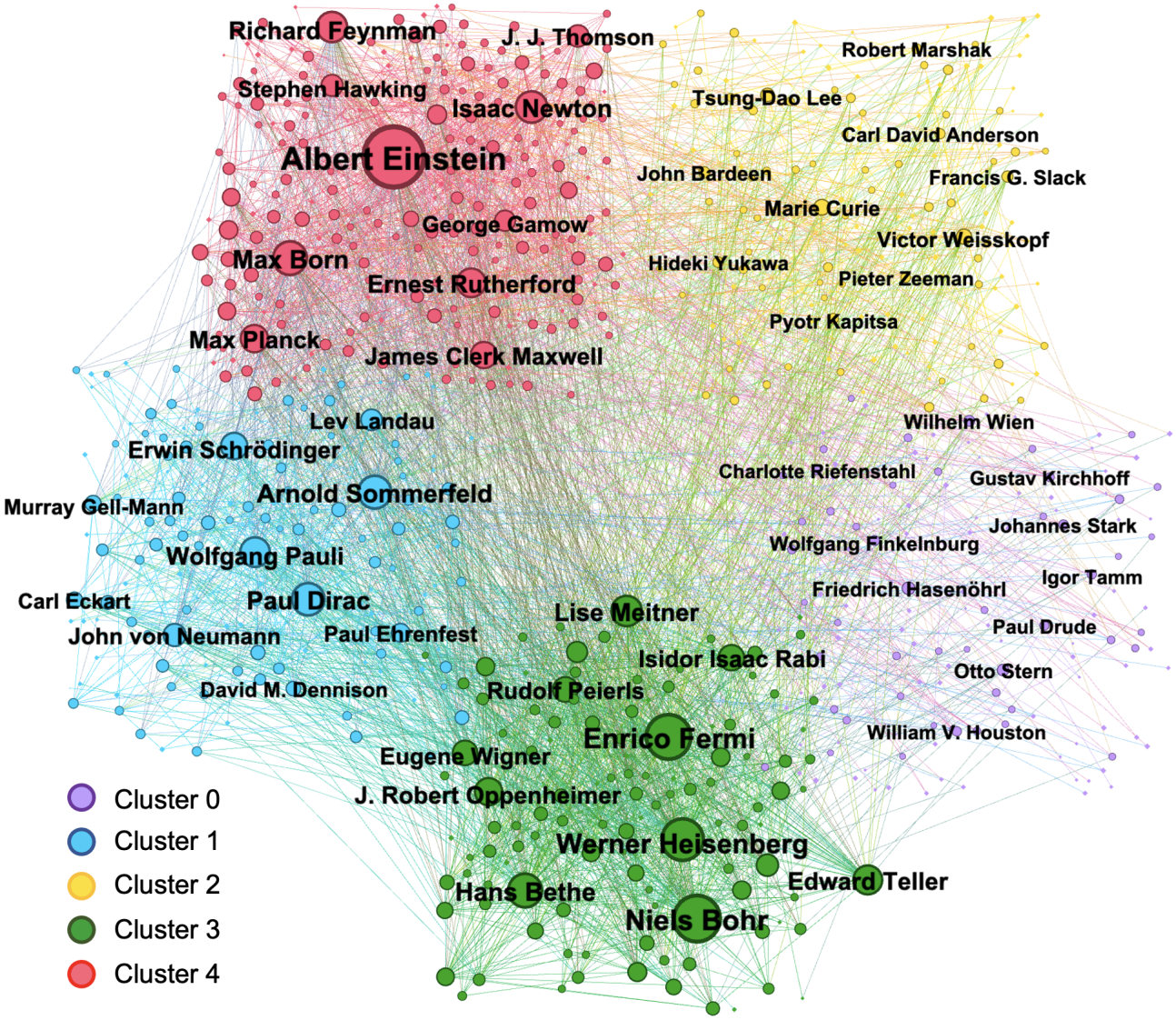

Figure 8: The network of physicists. Colours indicate the cluster. The names of top ten physicists (according to our ranking scheme) in each cluster are shown. The size of node is proportional to the average mark.

Fig.8 shows that cluster 4 is led by Einstein who has the highest average mark, while cluster 3 is the group with the largest median (Fig.9), which can be interpreted as the most influential group. The page length, defined as the number of words in the main text (text after data cleaning), is calculated for each cluster (Fig.10). It can be seen that clusters 0,1,2 have a small mean page length. Clusters 3 and 4 have a large average page length with a relatively small standard deviation. These will be useful when looking into each cluster in detail.

For each cluster, a word cloud is produced (Fig.11). The size of the word shows how important it is (i.e. its frequency), and different colours are used for better visualisation. It can be seen that the majority are single words, and there are also some terms with more than one word (e.g. 'quantum mechanics' in cluster 2) due to the implication of n-gram (in section 2.2). The following subsections will show the details of each cluster.

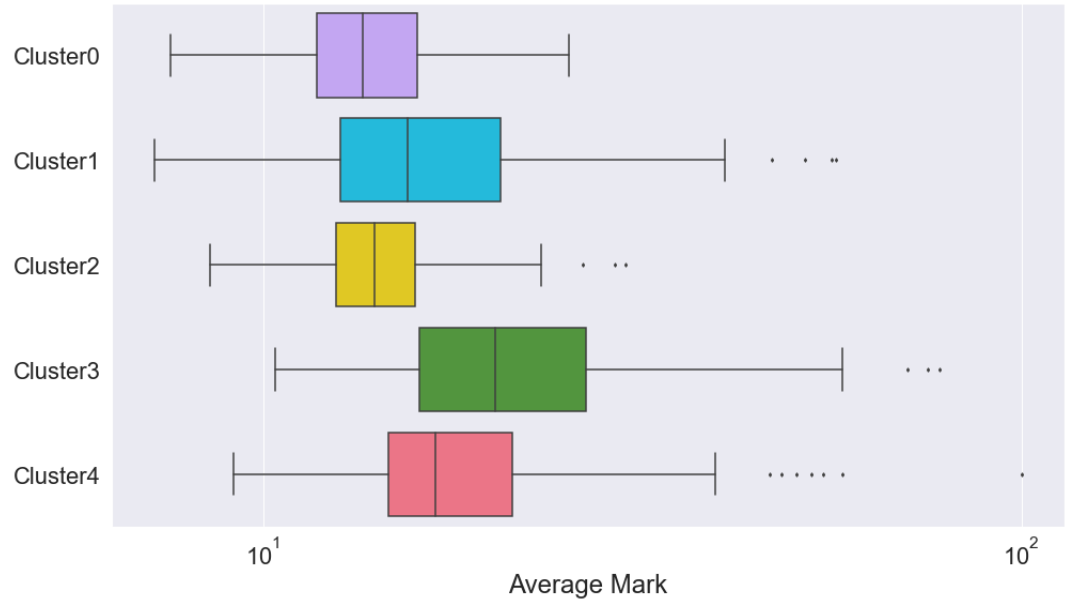

Figure 9: The box-plot of the average mark of five clusters, showing minimum, first quartile, median, third quartile, maximum, and outliers. Maximum/minimum is defined as third/first quartile ± IQR. IQR is the difference between third and first quartile.

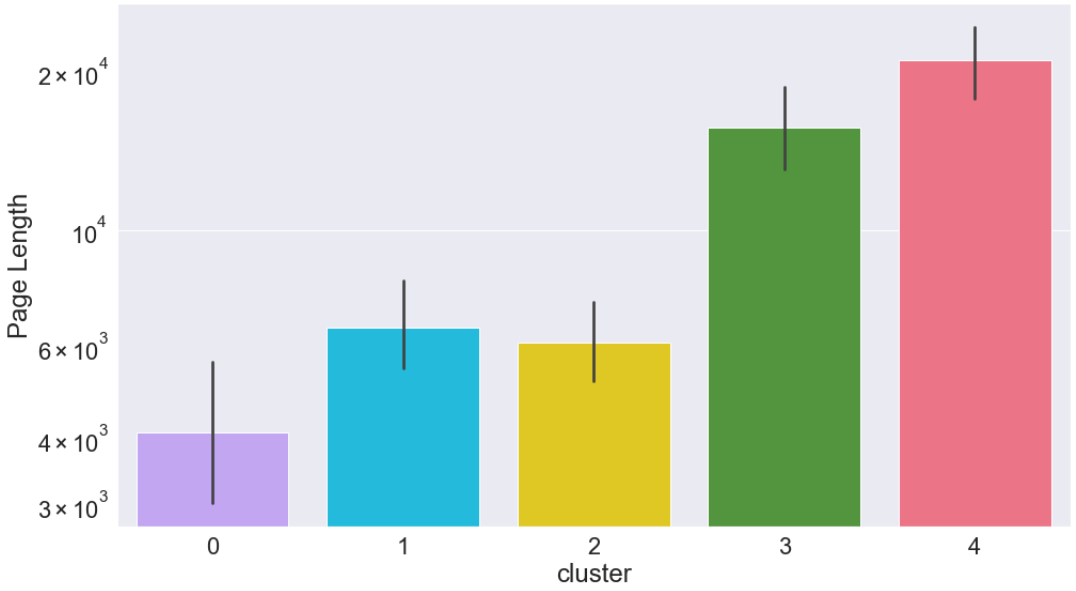

Figure 10: The bar-plot of the average page length of five clusters. The vertical black line in each bar represents the standard deviation. Page length is the number of words in the main text (see definitions in Section 2.1)

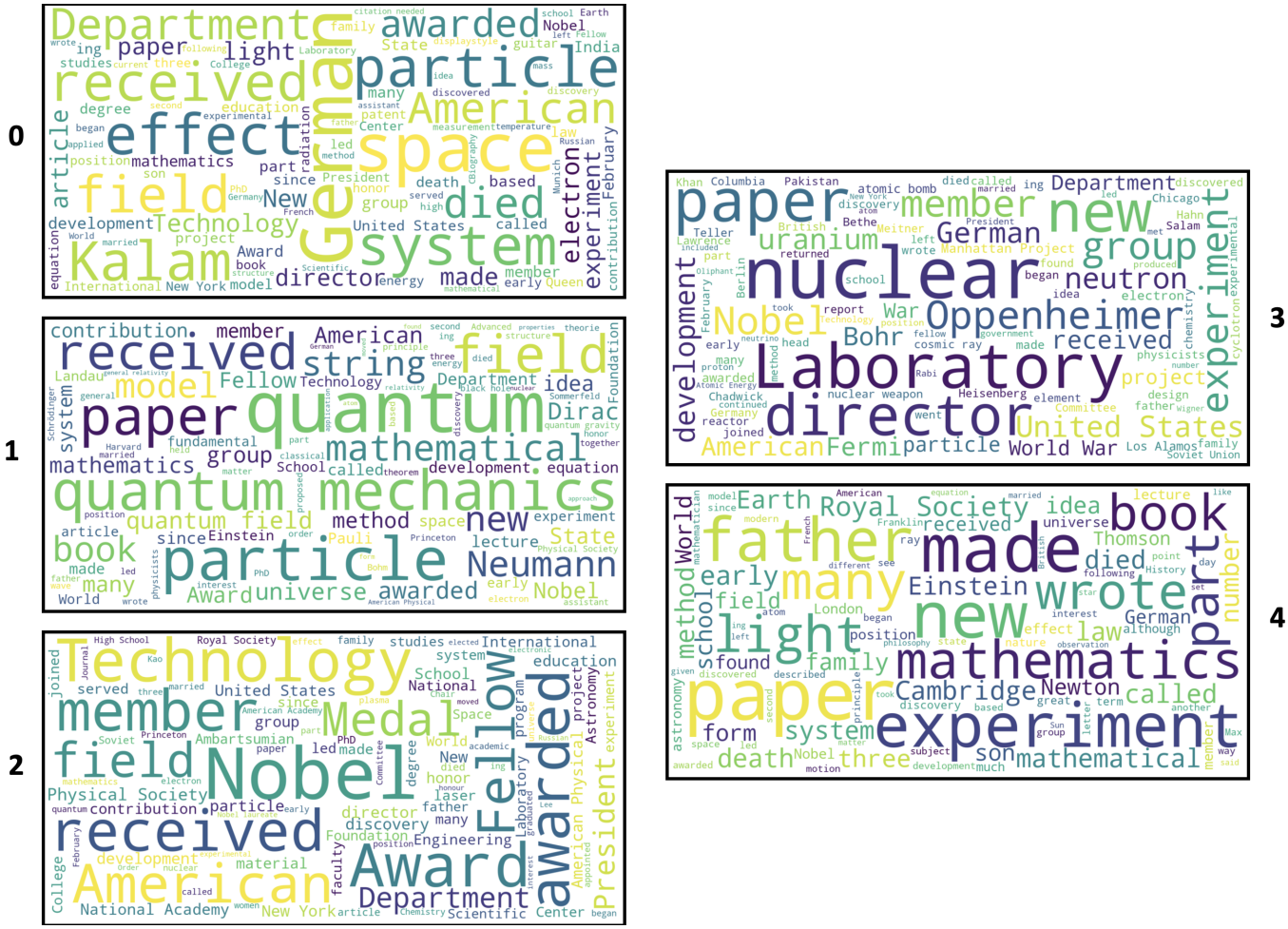

Figure 11: The word clouds of five clusters. The size of word is proportional to its frequency in the cluster.

3.3.1 cluster 0

Table 3 shows the top ten physicists based on our ranking scheme in this cluster. This group mainly contains people who have (had) the nationality of German as the term 'German' is shown in Fig.11-0. On average, they have a relatively short biography and a low average mark. Compared with other factors (e.g. theories, achievements), nationality becomes the main feature of this cluster.

Rank in cluster	Overall rank	Name	Average mark	Page length
1	82	Friedrich Hasenöhrl	25.30±0.70	8,185
2	112	Otto Stern	23.55±0.76	3,387
3	131	Wilhelm Wien	22.43±0.66	3,634
4	167	Wolfgang Finkelnburg	20.77±0.69	3,891
5	172	Johannes Stark	20.60±0.45	6,141
6	183	William V. Houston	20.18±0.43	5,931
7	191	Igor Tamm	19.97±0.87	5,702
8	194	Paul Drude	19.89±0.54	2,272
9	207	Gustav Kirchhoff	19.49±0.44	5,853
10	223	Charlotte Riefenstahl	19.06±0.41	2,370

Table 3: Top ten physicists ranked in order of the average mark in cluster 0

3.3.2 cluster 1

In Fig.11-1, the terms 'quantum' and 'quantum mechanics' suggest that people in this cluster are mainly the ones who have significant contributions to Quantum Physics and Quantum Mechanics in the first half of the 20th century. It can be seen that in Table 4 there are Sommerfeld, Dirac, Pauli, Schrödinger and etc., who are famous quantum physicists.

Rank in cluster	Overall rank	Name	Average mark	Page length
1	7	Arnold Sommerfeld	56.85±1.41	18,248
2	8	Paul Dirac	56.11±1.19	48,554
3	12	Wolfgang Pauli	51.85±1.15	13,890
4	17	Erwin Schrödinger	46.91±0.92	20,534
5	22	John von Neumann	40.57±0.76	79,726
6	26	Lev Landau	38.53±1.25	14,058
7	31	Paul Ehrenfest	34.20±0.76	10,711
8	42	Murray Gell-Mann	30.35±0.89	12,147
9	45	Carl Eckart	30.28±0.85	8,721
10	54	David M. Dennison	28.06±0.92	5,866

Table 4: Top ten physicists ranked in order of the average mark in cluster 1

3.3.3 cluster 2

This group is similar to cluster 0. Since people have a relatively small page length and low average mark, winning the Nobel Prize becomes the main feature of this cluster (the term 'Nobel' can be seen in Fig.11-2). Another relatively big word 'American' is in the word cloud as well, because the laureates are mainly American. Different from cluster 0 which mainly consists of German, cluster 2 contains multiple nationalities. In Table 5, we have Marie Curie (Polish-French), Victor Weisskopf (Austrian-born American), Tsung-Dao Lee (Chinese-American), Hideki Yukawa (Japanese), Pyotr Kapitsa (Union of Soviet Socialist Republics).

Rank in cluster	Overall rank	Name	Average mark	Page length
1	47	Marie Curie	30.04±0.59	77,249
2	52	Victor Weisskopf	29.12±0.41	3,502
3	76	Tsung-Dao Lee	26.44±0.82	10,830
4	115	John Bardeen	23.26±0.45	15,670
5	124	Francis G. Slack	22.85±0.63	6,518
6	138	Hideki Yukawa	22.05±0.58	4,634
7	152	Pyotr Kapitsa	21.50±0.47	7,086
8	157	Carl David Anderson	21.24±0.67	3,245
9	175	Pieter Zeeman	20.54±0.58	5,287
10	177	Robert Marshak	20.48±0.55	10,830

Table 5: Top ten physicists ranked in order of the average mark in cluster 2

3.3.4 cluster 3

This group is the one with the largest median of the average mark (Fig.9). Compared with cluster 1 which represents the physicists who founded Quantum Physics, people in this group applied the theory to the real world in the mid-20th century, and they worked more on Atomic, Nuclear and Particle Physics. This group contains many experimental physicists ('Laboratory' and 'experiment' are high-frequency words in Fig.11-3) and they are strongly connected by the Manhattan Project. Oppenheimer, the leader of the project and the 'father of atomic bomb', appears in the word cloud with other terms 'war', 'uranium', 'neutron' which describe how nuclear weapons were developed during World War II.

Rank in cluster	Overall rank	Name	Average mark	Page length
1	2	Niels Bohr	77.85±1.47	74,106
2	3	Enrico Fermi	75.16±1.88	44,635
3	4	Werner Heisenberg	70.64±1.27	55,598
4	6	Hans Bethe	58.12±0.81	30,634
5	10	Lise Meitner	52.99±1.12	56,324
6	13	Edward Teller	50.72±1.26	44,665
7	16	J. Robert Oppenheimer	48.25±0.82	103,516
8	19	Isidor Isaac Rabi	45.31±0.83	25,199
9	20	Rudolf Peierls	44.36±0.80	25,586
10	21	Eugene Wigner	44.12±0.78	19,081

Table 6: Top ten physicists ranked in order of the average mark in cluster 3

3.3.5 cluster 4

This cluster is led by Einstein and his name appears in the word cloud (Fig.11-4). Unlike people in clusters 1 and 3 that are linked by one topic (Quantum Physics or Atomic & Nuclear Physics), people in cluster 4 had made great impacts in a wide range of areas. For example, Einstein and Newton can be seen in Table 7. The words 'Royal',

'Society', 'Cambridge' indicate that this group contains many British Physicists. In the history of science, the British scientists dominate before the 20th century and had great contributions to fundamental subjects such as Mechanics and Electromagnetism. Therefore, we identify this group as the generally famous physicists who are well-known to common people because of their equations, experiments and discoveries.

Rank in cluster	Overall rank	Name	Average mark	Page length
1	1	Albert Einstein	100.00±0.28	134,551
2	5	Max Born	58.13±1.20	31,317
3	9	Isaac Newton	54.70±1.30	104,927
4	11	Richard Feynman	52.89±1.18	76,219
5	14	Ernest Rutherford	50.42±1.27	48,031
6	15	Max Planck	48.29±0.96	24,704
7	18	James Clerk Maxwell	46.60±1.18	54,997
8	24	J. J. Thomson	39.48±0.85	23,735
9	25	Stephen Hawking	38.97±0.98	119,179
10	28	George Gamow	37.57±0.83	19,531

Table 7: Top ten physicists ranked in order of the average mark in cluster 4

3.4 Correlation between page length and average mark

We have also studied the correlation between page length and average mark to determine whether a famous physicist has a longer biography. The answer is yes, unsurprisingly, and we find the relationship is a power law as we obtain a Pearson r of 0.58 in the log-log space. The gradient of the line in Fig.12 is determined to be 0.20±0.01, which means the average mark is proportional to the page length to the power of 0.2. It is worth noting that the page length is not simply the number of the word in a document but the main text that carries the significant information. The Pearson r is defined by:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{8}$$

where x,y is the variable; n is the number of the variable; $\bar{x}$,$\bar{y}$ is the mean of the variable.

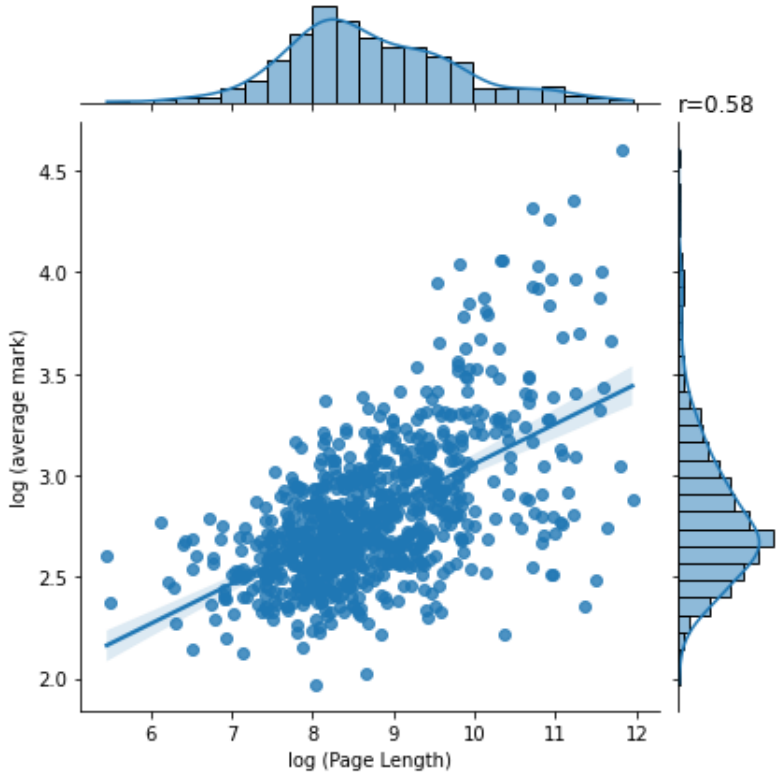

Figure 12: Correlation between page length and average mark in log-log scale with Pearson r = 0.58. Page length is defined as the number of words in the main text (see definitions in Section 2.1). The straight line of best fit has a slope of 0.20 ± 0.1 and y-intercept of 1.09 ± 0.02.

3.5 Bias of English Wikipedia

The main source of bias is language. In this project, all the data is extracted from English Wikipedia. It is clear that the majority of readers and editors are people from English-speaking countries. As anyone is able to edit the wiki page, the biography of English-speaking physicists is likely to have more details (including text and hyperlinks), because editors can access and understand their information more easily. That is to say, in English Wikipedia, English-speaking physicists have longer biographies with more details, which makes them have a higher centrality in the network and hence a higher rank, compared with those who speak other languages. Another issue is that there is information loss due to translation between languages, which means the results of our text analysis to the non-English-speaking physicists are less accurate, and that affects the clustering. We assume this is a minor effect as the content is mainly Physics, which is described by objective and rigorous words, but the effect will be much more significant if we study Poetry.

Gender is also a source of bias. Although the rights of women have been improved today, females including female physicists were treated unfairly in history. Some contributions they made may not be recorded properly, resulting in a lower rank from our network analysis. Nationality also causes bias. In some countries, textbooks for students will have more descriptions of the national scientists, so editors may be biased towards the physicists in their own country (adding more text and hyperlinks). The effects can hardly be evaluated as it depends on many factors, such as the population and education policy. Political events may also lead to bias of nationality (e.g. wars).

It is also noticeable that many physicists in cluster 5 (Table 7) actually are not experimentalists but in the word cloud (Fig.11-4) 'experiment' has a greater size than 'theory'. This suggests a bias in theory and experiment. As Wikipedia is open to the general public, it does not contain too much content in the academic aspect. Compared with complicated theories, experiments may be more attractive and understandable to the readers, so experiments may have longer and more detailed descriptions, resulting in a bias in our text analysis.

3.6 Robustness of K-means clustering

As mentioned in Section 2.2.4, we have to choose the number of clusters first for K-means clustering, and then the algorithm starts with the cluster centers chosen at random positions. Each point in the dataset is assigned to the closed cluster based on the distance between the points and centers[4]. The centers are recomputed if the within-cluster sum of squares (WCSS) is not minimised. After a number of iterations, the positions of the center become fixed and the clusters are obtained. The stochastic choices of starting point may yield different clustering results on different runs. To test the robustness, we iterated K-means five times. The top ten physicists in all five clusters did not change, and the main words in the word clouds remained the same. Therefore, in this project, the K-means clustering is robust, so we choose the outputs from one run to be the results. Although the results may not be repeatable and lack consistency, it has almost no effect on the analysis.

4 Conclusions & Improvements

In this project, network centrality measures and K-means clustering are used to analyse the Wikipedia pages of physicists. Several conclusions can be drawn:

1. Einstein is the most famous physicist, followed by Bohr, Fermi and Heisenberg.
2. Physicists can be divided into five clusters:

   Cluster 0: 'German'

   Cluster 1: 'Quantum Theory'

   Cluster 2: 'Nobel'

   Cluster 3: 'Atomic and Nuclear'

   Cluster 4: 'Generally well-known'
3. A large number of influential physicists work on Quantum, Atomic and Nuclear Physics, and many of them are connected by the Manhattan project.
4. If physicists do not have very great achievements, nationality or winning the Nobel Prize becomes one of their mean features.
5. The average mark of a physicist is proportional to the page length of the website to the power of 0.2.

For improvements, we cloud:

1. Use a better noise model. A node with higher degrees should have a greater probability to be removed or added edges. This matches the process of editing in real life as a long biography is more likely to be modified. We should also collect the data in previous years to determine how many edges were changed and then use the calculated fraction instead of 5%.
2. Test the robustness of K-means clustering quantitatively. For example, we could calculate the fraction of people changed in each cluster for each run, as well as the change in frequencies of the keywords.
3. Use different clustering methods. As mentioned in Sections 2.2.4 and 3.6, K-means has two drawbacks: it needs a given number of clusters and starts with random points, which makes the results unreproducible. It also performs poorly if there are many outliers in our dataset[4]. Other methods can be used to produce consistent results (Agglomerative Hierarchical Clustering and Expectation–Maximization EM Clustering)[15].
4. Investigate Wikipedia in different languages. As mentioned in Section 3.5, there is a bias in favour of English-speaking physicists. We cloud apply the same methods to analyse Wikipedia in other languages and then do comparisons to obtain a more objective picture of physicists.

5 Acknowledgement

The author would like to express many thanks to the project partner, who has made great contributions to the clustering part of this project, as well as the supervisor Dr Tim Evans, who guided the author and shared many ideas and methods which are fairly helpful for this project.

References

[1] B. Chen, Z. Lin, and T. S. Evans, "Analysis of the wikipedia network of mathematicians," *arXiv:1902.07622*, 2021.

[2] F.N.Silva, M.P.Viana, B.A.N.Travençolo, and L. da F.Costa, "Investigating relationships within and between category networks in wikipedia," *Journal of Informetrics*, vol. 5, pp. 431–438, 2011.

[3] H. Lane, C. Howard, and H. Hapke, *Natural Language Processing in Action.* Manning, 2019.

[4] K. Singh, imple Malik, and N. Sharma, "Evolving limitations in k-means algorithm in data mining and their removal," *International Journal of Computational Engineering Management*, vol. 12, 2011.

[5] M. J. Zaki and W. Meira, *Data Mining and Machine Learning: Fundamental Concepts and Algorithms.* Cambridge University Press, 2020.

[6] R. Diestel, *Graph theory.* Springer Publishing Company, Incorporated, 2018.

[7] M. Coscia, *The Atlas for the Aspiring Network Scientist.* Michele Coscia, 2021.

[8] M. Newman, *Networks: an introduction.* Oxford University Press, 2010.

[9] A.J.Lotka, "The frequency distribution of scientific productivity," *J.Washington Acad.Sci.*, vol. 16, p. 137, 1926.

[10] S.Redner, "How popular is your paper? an empirical study of the citation distribu- tion," *Eur.Phys.J.B*, vol. 4, p. 131, 1998.

[11] M.E.J.Newman, "Scientific collaboration networks: I. network construction and fundamental results," *Phys.Rev.E*, vol. 64, 2001.

[12] R.Albert, H.Jeong, and A.-L. asi, "The diameter of the world-wide web," *Nature*, vol. 401, p. 130, 1999.

[13] M.Faloutsos, P.Faloutsos, and C.Faloutsos, "On power-law relationships of the inter- net topology," *Comput.Commun.Rev.*, vol. 29, p. 251, 1999.

[14] T. S. Evans, "Complex networks," *Contemporary Physics*, vol. 45, pp. 455–474, 2004.

[15] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications.* Society for Industrial and Applied Mathematics, 2020.

Appendix A: GitHub Repository

The GitHub repository for this project can be accessed from here: `https://github.com/Kaiyu-cpu/BSc-Project-Complexity-of-Physicists`