Network of Scientific Concepts

Rajk College for Advanced Studies, Network Science

Held by: András Kárpáti

Hanga Dormán

28 May, 2020

Abstract

This paper examines how Wikipedia articles in different fields of science are related to each other conceptually, and what kinds of network phenomena can be attributed to them. First, a ranking of fields based multiple measures of cross-field contribution to scientific knowledge are introduced. Next, the analysis extends to a more detailed, field-to-field distribution of hyperlinks. To characterize fields by their network behaviour, logistic regressions are fitted to field membership variables, revealing significant differences between fields in network properties such as neighbor connectivity and clustering. By refining the process of data acquisition, introducing new measures of scientific contribution and adding complexity to the predictive model, this framework may provide the basis for analysing how knowledge is transferred within and between fields of science.

Contents

List of Figures

List of Tables

1 Introduction

Scientific discoveries emerge via the - often incidental - interaction of concepts and empirical observations. In order to accurately describe specific phenomena and not get lost in their infinite complexity, the domain and terminology of the analysis needs to be restricted. One has to take certain statements and concepts established by other means as given, and upon them build a new framework suitable to examine a different segment reality. This division of intellectual labor is what makes interdisciplinary interactions so important, and it is what directed my attention to the current subject matter.

The aim of this paper is to describe in detail - using the tools of network science - how the different academic fields build on each other's accumulated knowledge and terminology, and to uncover field-specific network phenomena. For the analysis, I use the network of Wikipedia articles and hyperlinks, which provides a general, time-independent documentation of the interrelated structure of scientific concepts. In the following section, I introduce the data acquisition procedure, along with the resulting network and its properties. In **section 3**, I rank individual fields based on their relative contribution to collective knowledge, by applying previously defined measures to the Wikipedia hyperlink dataset. Furthermore, I examine tendencies in field-to-field interaction. In **section 4**, I use predictive analysis to show that individual fields can be distinguished based on their network behaviour. The **last section** concludes.

2 Methods

2.1 Data acquisition

For this study I used the ***wiki-topcats*** dataset made publicly available by Stanford University (see Leskovec and Krevl (2014) for reference). This network was obtained in 2011 by taking the largest strongly connected component of the entire network of Wikipedia hyperlinks, filtering it for the top 100 categories, and extracting the largest strongly connected component once again. The nodes are Wikipedia articles along with their respective ID, title and category labels, and the edges are directed hyperlinks. An edge points from article A to article B, if the former contains a reference to the latter. The network is unweighted and has a few number of self-loops, but no parallel edges. An article may belong to multiple categories.

As I was exclusively interested in scientific articles, I relied on the classification of academic fields provided by Wikipedia itself (Wikipedia, nd). This structure is illustrated in **Table I**. I excluded the broad field of Professions and Applied Sciences, as I wanted to focus on the theoretical connec-

broad field	field
Formal sciences	computer science
	logic
	mathematics
	systems science
Natural sciences	biology
	chemistry
	earth science
	physics
	space science
Humanities and social science	anthropology
	linguistics
	philosophy
	economics
	political science
	sociology
	psychology
	history
	religion
	arts
	geography
	interdisciplinary studies

Table 1: Classification of academic fields.
Fields that have been omitted or merged with other fields are written in gray. (Source: self-made)

tions between fields. To obtain articles from a given category I used string matching methods, with mainly manual instructions based on some finer categories that are not included in this table, but listed on the Wikipedia page. Due to the lack of available data (and slightly ironically) I also had to exclude systems science (the dataset was collected earlier and many articles in this relatively new field have been written only recently). Furthermore, in order to keep the attempted narrow focus of the analysis, I did not include history, religion, arts and interdisciplinary studies. Also, I merged geography and earth sciences for the sake of simplicity.

2.2 Building the network

Based on these previously determined fields, a total number of **293** sub-categories were selected, containing **42 878** articles altogether. Next, I extracted all hyperlinks belonging to these articles and built the graph, which I then filtered for the largest weekly connected component[1]. The resulting network contains **39 710** nodes and **287 725** edges, with an average degree of **7.25**. I calculated the in- and out-degrees, the neighbor connectivity (the sum of in- and out-degrees averaged over all predecessors and successors of the node) and the clustering coefficient of each node within this network. **Figure 1.** shows that both in- and out-degrees roughly follow a power law distribution.

[1] The disconnected nodes mainly feature members of institutions (such as the London School of Economics) and less known American scholars. The largest connected component within the disconnected part of the network is a small collection of geography-related lists.

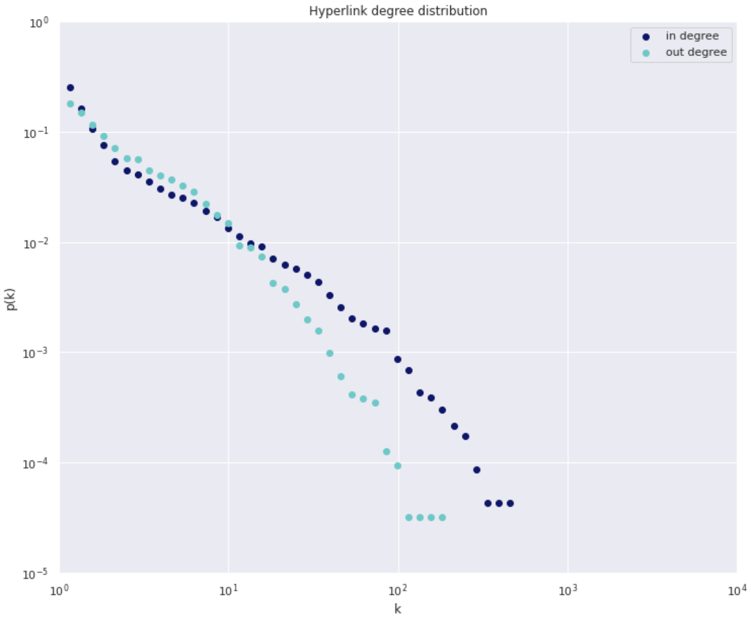

Figure 1: Distribution of in- and out-degrees in the Wikipedia hyperlink network . (Source: self-made)

As a base of comparison, I performed degree-preserving randomization on the network, using the ***joint degree sequence*** method of the Networkx Python package (see Hagberg et al. (2008) for reference). This is a procedure that generates a random directed graph based on a previously determined joint degree sequence, which assigns to each out-in degree pair i, j the number of links pointing from nodes with out-degree i to nodes with in-degree j. I assigned each node its own in- and out-degree in the actual hyperlink network. This way, while the degree distribution of the network is completely preserved, field-specific attributes are erased - apart from those that necessarily emerge from the degree-distribution. The randomized network is thus a useful tool for isolating true field-specific phenomena from other structural properties.

2.3 Overlapping fields

Articles are distributed between the various fields as shown in **Figure 2**. Taking overlaps into account, I divided the number of articles that are shared by several fields equally between these fields (an article that belongs to n fields counts as $\frac{1}{n}$ in each field). This overlapping structure also affects edges: there's no way to tell, which field a particular hyperlink is related to, if the two nodes it connects belong the multiple fields. As an example, let's suppose there's a link between Albert Einstein, who belongs to both *physics* and *philosophy*, and Kurt Gödel with field labels *mathematics*

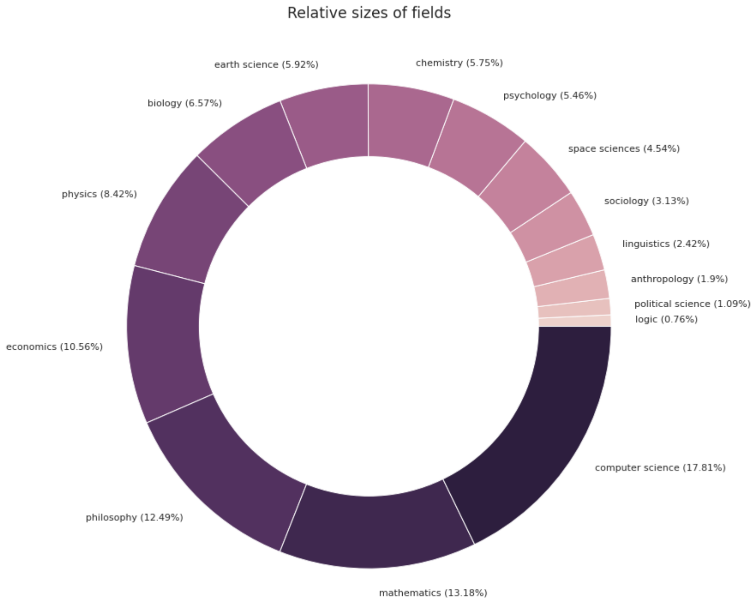

Figure 2: Relative sizes of fields
. (Source: self-made)

and *philosophy*. What can we say about this connection? Is this a within- or a cross-field link? Is is between two philosophers, or a physicist and mathematician, or a physicist and a philosopher? Resolving this problem requires a detailed analysis of context, which is unfeasible due to the large amount of data and beyond the scope of this analysis. Therefore, I decided to give equal weights to all of these possibilities, by splitting the edges between combinations. In the example above, this means that a link from Eistein to Gödel is counted as four distinct edges (from *philosophy* to *philosophy*, from *philosophy* to *mathematics*, from *physics* to *philosophy* and from *physics* to *mathematics*), each with a **0.25** weight. This way, I could calculate the share of *within-field-out-degrees* and *within-field-in-degrees* for each node, which express how often an article refers to, or is referred to by another article from the same field, as opposed to articles from different fields.

3 Transfer of knowledge across disciplines

3.1 Measures of scientific contribution

Rinia et al. (2002) examined knowledge transfer between fields of science via bibliometric methods. Their database contained all papers from the year 1999, published in journals included in the Science Citation Index, and the references between them. Disciplines were categorized based on the ISI classification of journals. Relying on previous studies, the authors developed three different metrics to measure the relative contribution of individual fields to collective scientific progress. According to their assumption, this contribution can be captured in the distribution of cross-field references, which reveals how much researchers in one field utilize results obtained in other fields. Although they acknowledge the weaknesses of this methodology, they conclude that it can provide useful indications about the value of research and potentially inform funding decisions.

They defined the *relative external use* of publications within field j as $\frac{(1-\alpha_j)}{\alpha_j} \cdot \frac{\sum_{i \neq j} R_{i,j}}{\sum_{i \neq j} R_i}$, where α_j is the relative size of j (calculated from the number of publications in each field), and the second fraction is the share of references to field j of all references by other fields (based on Rinia et al., 2002, pp. 351, 365). This measure thus favors small fields, as well as fields that received a large share of all references. As an alternative that takes into account the absolute number of articles and references, the *external citation average* is defined as $\frac{\sum_{i \neq j} R_{i,j}}{P_j}$, where the counter is the number of all references received by j, and the denominator is the number of articles in j (based on Rinia et al., 2002, pp. 351, 367). A third option is the *import/export ratio*: $\frac{\sum_{i \neq j} R_{i,j}}{\sum_{i \neq j} R_{j,i}}$, which expresses the number of references to j from other fields relative to the number of references from j to other fields (based on Rinia et al., 2002, pp. 351, 358). The authors found that based on the first two measures, *Multidisciplinary Sciences* rank the highest, followed by *Basic Life Sciences*, *Pharmacology* and *Environmental Sciences*. Four fields - *Multidisciplinary Sciences*, *Basic Life Sciences*, *Physics* and *Geo Sciences* - were associated with an import/export ratio higher than one, meaning that publications in external fields rely more heavily on them than they rely on external knowledge. (Rinia et al., 2002)

3.2 Ranking fields based on Wikipedia hyperlinks

My approach differs from the one described above in two aspects: (1) I used Wikipedia articles instead of publications as a measure of cross-disciplinary transfer of knowledge and (2) I restricted the analysis to theoretical fields and excluded applied sciences such as *Engineering* and *Medicine*. Relying on Wikipedia articles that store accumulated knowledge may overlook the short-term, gradual development of fields, which is captured in the separate publication of discoveries. However, it provides a more general description of how the different terminologies are interrelated, irrespective of their temporal evolution.

Figure 3. displays the ranking of fields I obtained based on the total number of external citations and the three measures introduced by Rinia et al.. In absolute terms, physics has the highest number of external citations (over **16 000**). The leading articles are about entire subfields (*Physics*, *Astronomy*, *Quantum Mechanics*), as well as core concepts (*Electron*, *Sun* and *Atom*) and pioneer physicists (*Einstein* and *Newton*). Physics is tightly followed by philosophy (with around **15 000** external citations), with articles about the greatest philosophers (*Aristotle*, *Plato*, *Marx*, *Kant* and surprisingly, *Einstein* again), the field of *Ethics* and the concept of *Capitalism.* The external citation average and the relative external use yield almost identical rankings, with *logic* and *political science* in leading positions. This means that these two fields are referred to by articles from other fields more than what would be proportional to their relatively small size. *Mathematics* and *computer science* rank last, being the largest and most introverted fields. The three most prominent natural sciences, *physics*, *chemistry* and *biology* have a higher than one import/export ratio, which is no surprise considering the large amount of knowledge accumulated in these fields. The three smallest fields, *logic*, *anthropology* and *political science* also belong in this category, which is probably because due to their size, they cannot reciprocate the amount of references given to them by larger fields. In the case of *logic*, it may also be part of the explanation that as a highly abstract field, it uses less external knowledge compared to how often it is used in other fields. One thing I find surprising is that *mathematics* is not in this league, despite its similarly abstract nature.

3.3 Preferential attachment across fields

To get a more detailed picture of cross-field referencing tendencies, I calculated the matrices shown in **Figure 4**. On the left, the joint distribution of hyperlinks is shown for the real network. A value in row i and column j expresses the share of references received by field j of all references given by field i (values in a row sum up to one). Only proportions higher than **0.05** are annotated. There are high values in the diagonal, as within-field references are the most frequent. One outlier is *logic* with a larger share of references to both *mathematics* and *philosophy* individually then itself. Another such example is *political science* that references articles from *economics* equally often, and articles from *philosophy* more often then articles within the field. *Anthropology* and *sociology* show similar, but weaker such tendencies. On the other end of the scale we find computer science with more than 80% of its references pointing within the field, which makes sense if we consider its autonomous nature. Yet, the lack of references to *mathematics*, *logic* and even *physics* is counter-intuitive. On the receiver side, *philosophy* stands out by having a significant share of references from many fields, including *anthropology*, *linguistics*, *logic*, *political science*, *psychology* and *sociology. Physics* is often cited by *chemistry* and *space sciences*, while *political science* and *sociology* heavily rely on *economics.* These observations are more or less in line with common intuitions. On the right side, the same matrix obtained from the randomized network is shown. The exact values are not presented as they are irrelevant to the point. The difference between the two matrices is apparent: no within- or cross-field preferences are present in the randomized network. References from each field are distributed in the same way, based on field size. This confirms that the phenomena observed in the real network are not structural necessities, but are related to the nature of individual fields.

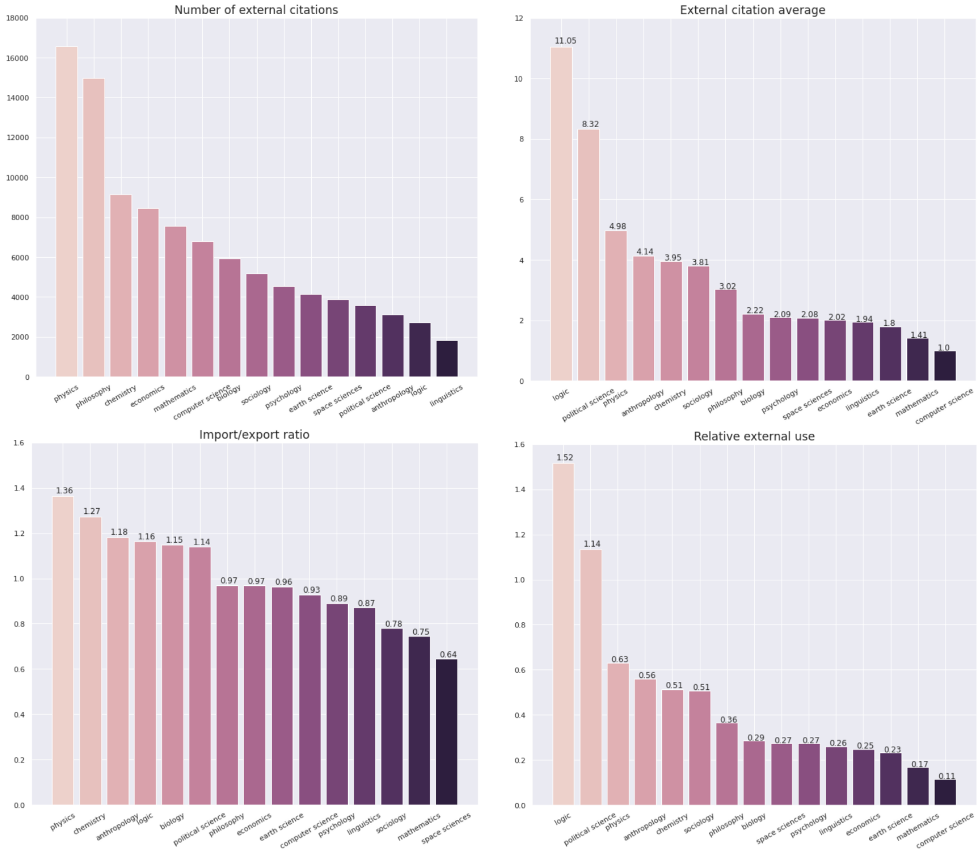

Figure 3: Ranking of scientific fields on Wikipedia, based on the absolute and average number of external citations (upper left and right), import/export ratio (lower left) and relative external use (lower right).

(Source: self-made)

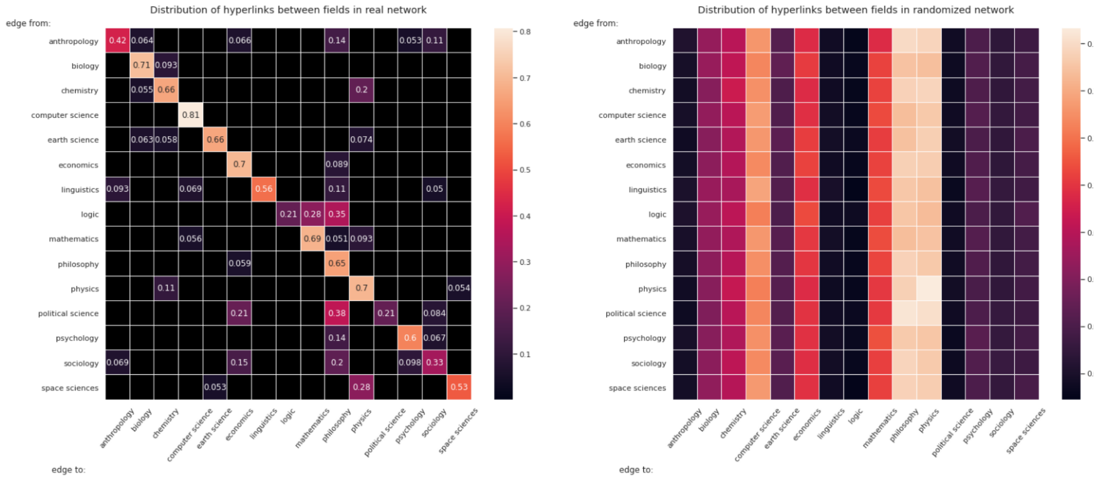

Figure 4: Cross-field hyperlinks in the real (left) and randomized (right) networks . (Source: self-made)

4 Predicting field membership with network attributes

Apart from measuring the contribution of the different fields to collective knowledge and scientific terminology, the goal of this paper was to uncover field-specific network phenomena. I decided to use logistic regression as a tool to characterize fields by their network properties. For each field, I created a dummy variable: an $\{0;1\}^N$ array (where N is the total number of articles), which indicates, whether a particular node belongs to the field or not. As there were no heavy correlations between field dummies[2], I fitted a logistic regression separately to all of them. As independent variables, I used the following attributes: out-degree (*out_degree*), neighbor connectivity (*neighbor_conn*), clustering (*clustering*) and the share of within-field references of all references to and from the article (*p_within_field_in* and *p_within_field_out*, respectively). In-degree was omitted due to its high correlation with out-degree. Because of the overlaps between fields, the latter two variables were averaged for each article. To illustrate what this means, let's assume an article belongs to both *biology* and *chemistry*. The value of the *p_within_field_out* variable for this article is then calculated by first taking the share of its references to articles in biology and chemistry individually, and then averaging the two. I chose this method for reason described in **section 2.3**. A different approach - for instance considering the union of references to these two fields as within-field references - would have been equally acceptable. The results are presented in **Figure 5**. On the left, estimated parameters for the real network are shown, while the figure on the right features the same estimations for the

[2] The lowest correlation, **-0.181** was found between *philosophy* and *computer science*, and the highest, **0.087** between *mathematics* and *logic*.

randomized network. Coefficients are annotated in the rectangles of the heatmap, and insignificant parameters are masked. The ROC curves for both networks are shown in **Figure 6.** along with accuracy measures, indicating that all regressions perform above chance, although the performance is higher for most fields in the real network.

As is apparent from **Figure 5.**, neighbor connectivity and clustering are insignificant for almost all dummies in the case of the randomized network. This means that any connections involving these two variables observed in the real network are actual field-specific properties, independent from its degree-distribution. In the case of neighbor connectivity, *computer science* stands out with by far the lowest parameter value, and is followed by *psychology* and *earth science* in the negative range. On the other hand, *physics* and *space sciences* are associated with the highest neighbor connectivity. Although the clustering coefficient is not significant for all fields, it is significantly lower in the case of *mathematics* and *economics*, while higher for *linguistics*, *logic* and *space sciences*. This implies that *mathematics* and *economics* tend to refer to articles in a broader scope, which are less related to each other. On the contrary, *linguistics*, *logic* and *space sciences* use densely connected concepts.

The *p_within_field_in* and *p_within_field_out* coefficients mostly reflect what has already been discussed in the previous sections. Articles in *logic*, *political science*, *anthropology* and *sociology* are less likely to cite and be cited by other articles within their own field, than other disciplines. The *p_within_field_out* coefficients highlight two additional fields in this category: *linguistics* and *space sciences*. All of these are smaller fields, which at least partly explains why their coefficients are lower. Also, while *logic* and *space sciences* are tightly connected to one or two fields (*mathematics* and *philosophy*, and *physics*, respectively), the other four fields have a more spread out, interdisciplinary character. Articles in *mathematics* and *philosophy* are the most likely to receive citations from their own field, while *computer science*, despite its strong tendency to produce within-field references, has a small negative *p_within_field_in* coefficient. To break down what this means, we need to consider the absolute number of references. **Figure 4.** reveals that computer science receives 5% of references from *mathematics*, a field producing over **30 000** citations. Meanwhile, *mathematics* receives many references from *logic*, but that is the smallest field in the data. *Philosophy* receives many links from several disciplines, but it has almost **30 000** internal references. These finer details account for the observed differences between the otherwise similarly abstract and introverted fields. The out-degree parameter estimates suggest that articles in *physics* cite other articles most frequently, followed by *logic* and *philosophy*. Articles in *computer science* are the least likely to cite other articles, which also contributes to the explanation of their low *p_within_field_in* value, and low neighbor connectivity.

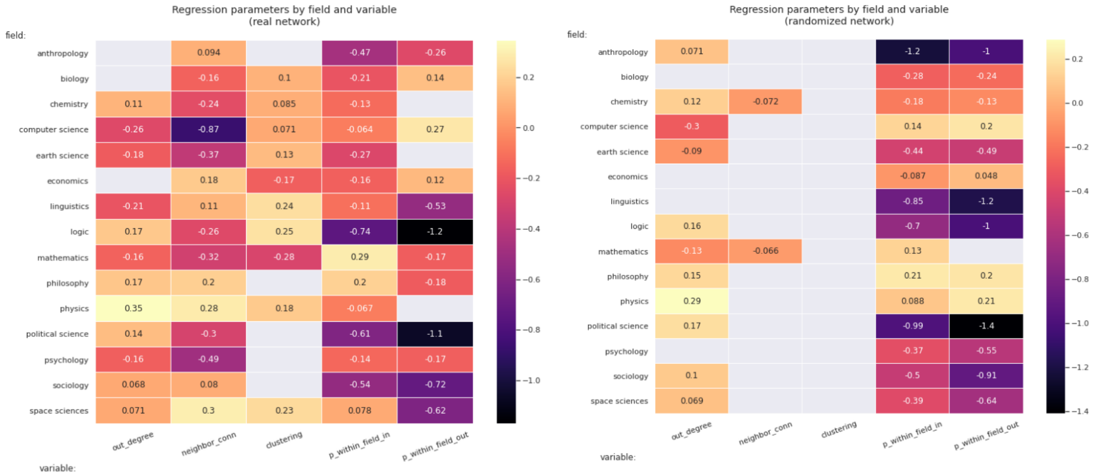

Figure 5: Regression coefficients for the real (left) and randomized (right) networks. Insignificant parameters are masked. (Source: self-made)

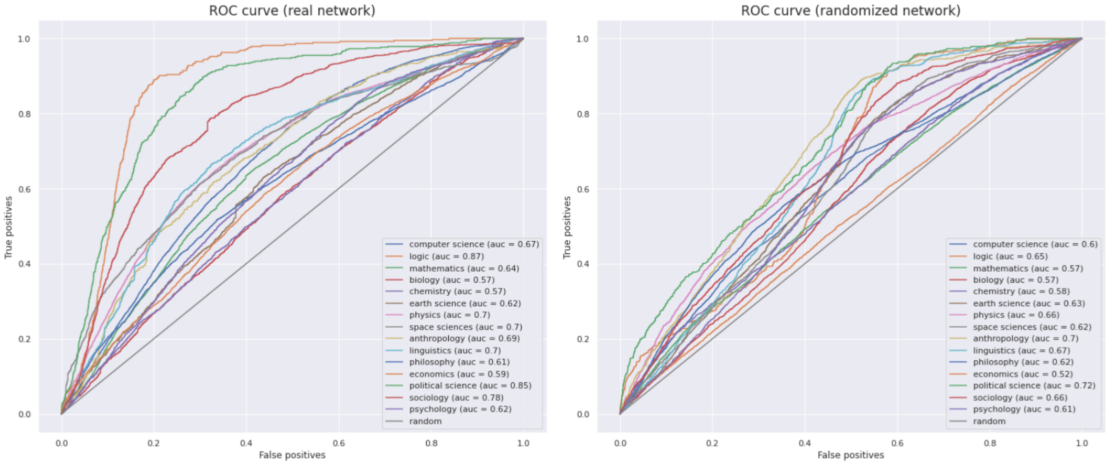

Figure 6: ROC curves and accuracy measures of regressions performed on the real (left) and randomized (right) networks.

(Source: self-made)

5 Conclusion

In this paper, I introduced the network of Wikipedia articles and hyperlinks, in an attempt to characterize scientific fields based on how they share information and terminology, and their behaviour within the network of knowledge. Relying on measures defined by Rinia et al. (2002), I ranked each field by its relative external use, external citation average and import/export ratio. This ranking indicates that, while in absolute terms the most robust natural science, *physics* has the greatest external relevance, fields such as *logic* and *political science* are also very important relative to their small sizes. The analysis also revealed that fields show quite specific tendencies in the ways they refer to, and are referred to by articles in other fields, even beyond intuitive semantic relations between them. By the tools of predictive statistics it has been shown that network properties, such as neighbor connectivity and clustering, vary between fields independent from their other attributes.

I would like to acknowledge the limitations of this analysis and suggest ways to enhance the framework. The quality of the data can be improved by using more sophisticated extraction methods. Aspert et al. (2019) describe Wikipedia as a network of multiple types of nodes and edges. An edge may connect two articles with a hyperlink between them, or an article and a category. This allows for a hierarchical structure, and one can find more and more articles, proceeding along the depth of classification. It is also possible to define other measures of cross-field knowledge transfer, as the ones described in this paper do not account for some finer details that have only been discussed in a qualitative manner. Furthermore, it may be fruitful to utilize the correlations between field dummies in predictive analysis, for a more precise characterization of fields. Adding interactions of variables and additional network properties (such as betweenness centrality) could also improve on the model. As for the base of comparison, with more available computational capacity one may perform the degree-preserving-randomization several times and average the results for greater robustness.

Finally, I would like suggest a potential direction of research that I could not include in this paper due to its limited scope. Biglan (1973) defined a scheme for classifying academic fields along three dimensions: *hard/soft*, *pure/applied* and *life/non-life*. Although the validity of this framework is debated, Simpson (2017) could extract two of the dimensions from the mere distribution of course offerings across institutions, using dimension reduction methods. The idea is to use the same dimension reduction methods on the network-related properties of fields. According to my hypothesis, differences in network behavior should seclude regions in the feature space that correspond to some meaningful classification. It would be equally interesting to observe that the emerging dimensions are in line with Biglan's model or produce some entirely different scheme.

The source code of the entire project is available on **GitHub**.

References

Aspert, N., Miz, V., Ricaud, B., and Vandergheynst, P. (2019). A graph-structured dataset for wikipedia research. *Companion Proceedings of The 2019 World Wide Web Conference*, pages 1188–1193. Downloaded: `https://dl.acm.org/doi/pdf/10.1145/3308560.3316757` (25 May, 2020).

Biglan, A. (1973). The characteristics of subject matter in different academic areas. *Journal of Applied Psychology*, 57(3):195–203. Downloaded: `https://pdfs.semanticscholar.org/6663/512cd3ef5e8a7f7e5509e12790cfe5516a29.pdf` 28 May, 2020.

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using networkx. In Varoquaux, G., Vaught, T., and Millman, J., editors, *Proceedings of the 7th Python in Science Conference*, pages 11–15. Downloaded: `https://networkx.github.io/documentation/latest/` (12 May, 2020).

Leskovec, J. and Krevl, A. (2014). Wikipedia network of top categories. In *SNAP Datasets: Stanford Large Network Dataset Collection*. Downloaded: `https://snap.stanford.edu/data/wiki-topcats.html` (2 April, 2020).

Rinia, E., van Leeuwen, T., Bruins, E., van Vuren, H., and van Raan, A. (2002). Measuring knowledge transfer between fields of science. *Scientometrics*, 54(3):347–362. Downloaded: `https://link.springer.com/content/pdf/10.1023/A:1016078331752.pdf` (23 May, 2020).

Simpson, A. (2017). The surprising persistence of biglan's classification scheme. *Studies in Higher Education*, 42(8):1520–1531. Downloaded: `https://www.tandfonline.com/doi/full/10.1080/03075079.2015.1111323?scroll=top&needAccess=true` (23 May, 2020).

Wikipedia (n.d.). List of academic fields. Downloaded: `https://en.wikipedia.org/wiki/List_of_academic_fields` (25 May, 2020).

Source of the cover image: n.a. (2019): Structure of brain networks is not fixed. *Neuroscience News*. Downloaded: `https://neurosciencenews.com/brain-network-structure-14435/` (21 May, 2020)