Assessing citation network clustering as indicator normalization tool

Riku Hakulinen[1] and Eva Isaksson[2]

[1] *Riku.Hakulinen@helsinki.fi*, [2] *Eva.Isaksson@helsinki.fi*
University of Helsinki, Helsinki University Library, Research Services, PO Box 53,
00014 Helsinki (Finland)

Introduction

Attempts to use bibliometrics in assessing research performance requires a normalization procedure to cover publications from different fields. The classification behind this normalization has traditionally been broad categories based on scientific fields of journals. During recent years, due to the need for a more specific method, citation network clustering has been utilized to partition research output into (micro) research fields. Also, machine learning approaches with the aim of classifying publications based on their full text have emerged, but it seems that the time for these as a reliable basis for classification may not yet have come. In this work we study the effect of choosing a particular network clustering method on the bibliometric impact of University of Helsinki (UH).

The purpose of this effort is to test whether these clustering classifications that are network science wise perhaps coherent [Traag et al., 2019], produce also meaningful classifications, and to study how varying clustering methods and their parameters affect the UH results. We use publication data from several fields with different publication and citation practices and focus on testing Leiden and Louvain algorithms [Traag et al., 2019]. This results in statistics of classification robustness and, for example, a comparison of normalized citation scores (NCSs) for chosen sets of UH publications in varying clustering classifications. The NCSs here are produced by considering our dataset as the research publication output of the world.

Dataset

As a starting data we collected for this paper a set of 93.6 thousand publications from Web of Science, obtained through basic search 'Topic = Social science', limited to publication years 1990-2019. Citation counts from WoS were used in this work. A more comprehensive set of bibliographic data is collected for the poster from various sources including Scopus and Dimensions.

Method

Citation links are identified from bibliographic data and an undirected citation network is compiled from that as a list of publication index pairs. The network is used as input in the clustering tool *RunNetworkClustering*, provided by CWTS through GitHub: CWTSLeiden/networkanalysis. Further work on the data and clustering results is done using the statistical software R and the networks are visualized with VOSviewer.

Preliminary Tests

Algorithms

At least three clustering methods are tested on our data: Leiden algorithm with Modularity and Louvain and Leiden algorithms with Constant Potts Model (CPM) as quality functions. First observation was that with the still modest network of about 42k nodes and 100k links from the 93.6k publications, it was not trivial to find a suitable value for the resolution parameter. We used 0.00015 in CPM and 0.7 in Modularity.

Comparing NCSs

We calculated MNCSs for the UH set of 182 publications in 33 (Le/Mod), 51 (Le/CPM) and 58 (Lo/CPM) clusters. We also tabulated a comparison of three sets of NCS values for an exemplifying set of six publications from 31 UH publications sharing a cluster in each clustering result.

Themes within Clusters

Characterization of a network cluster content can have terms from a broader category or field, like Web of Science categories, which we used in the first test to label (in all three clustering results) the cluster containing the mentioned 31 publications.

Web of Science Categories as Label-sets

All three clusters produced by the three algorithms (between 5k and 7k publications in each), containing the 31 UH publications had the following five categories as the most numerous covering about half of all category designations:

[1] "Ecology"
[2] "Environmental Sciences"
[3] "Environmental Studies"
[4] "Geography"
[5] "Science & Technology"

When based on counts in the cluster, the order of these terms varied between types of clustering. More tests and possibly more specific terms are required to allow conclusions about the contents of these and other clusters.

Numerical Results

The results in Table 1 can be interpreted so that the normalizations based on classifications from all three clustering methods produce here similar, but different impact results.

Table 1. Example NCS values for arbitrary six of the 31 (see text) UH publications following normalization based on classification from three clustering methods and age of publication.

NCS	Publ_1	Publ_2	Publ_3
Leiden, Modularity	1.73	4.47	0.92
Leiden, CPM	1.69	4.35	0.90
Louvain, CPM	1.71	4.42	0.91
NCS	**Publ_4**	**Publ_5**	**Publ_6**
Leiden, Modularity	2.12	1.01	1.04
Leiden, CPM	2.06	0.98	1.01
Louvain, CPM	2.10	1.00	1.03

E.g., Publ_5 only gets above "world average" with Leiden/Mod –version of clustering, and MNCSs for the whole UH set were 1.28, 0.98 and 1.01 (Le/Mod to Lo/CPM). To have a better understanding from the perspective of an organization that purchases services based on these methods, this will be tested with more data and varying algorithm parameter values like the resolution.

Figure 1. Visualization of Leiden/CPM cluster network. Includes about 42k publications in 426 linked clusters, compare to Figure 2. Cluster 38 contains 5181 publications of which 31 have an author with UH affiliation.

The large difference in amount of clusters in Figures 1 and 2 follows from that using CPM as the quality function produced about seven times more linked clusters with similar quality function values (~0.81) and with only 1.2 times larger total amount of clusters. The largest few clusters were of similar size in all three, but for CPMs the cluster sizes were more evenly distributed.

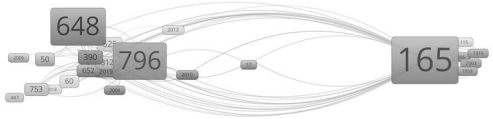

Figure 2. Visualization of Leiden/Modularity cluster network. Includes 64 linked clusters and is constructed from the same citation network as Figure 1. Cluster 165 contains 5715 publications of which 31 with UH affiliation.

Statistics and Conclusions

In addition to calculating average values and error bars using results from cluster analyses of the collected data, the aim is to connect some network properties like the number of vertices or even transitivity [Newman, 2002] with an indicator of quality of the clustering as classification. This will help to clarify results, e.g., the shown tentatively observed MNCS difference, in the context of UH publications.

Acknowledgments

The authors thank Ed Noyons from CWTS for discussing the application of Leiden algorithm for the University of Helsinki 2012-2016 publication dataset [Noyons & Mälkki, 2019].

References

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45, 167-256.

Noyons, E. & Mälkki, A. (2019). *Research performance analysis for the University of Helsinki 2012-2016/17.* Leiden: CWTS. Retrieved April 5, 2019 from: http://hdl.handle.net/10138/298733

Traag, V. A., Waltman, L. & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9.