HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI
HELSINKI UNIVERSITY LIBRARY

ASSESSING CITATION NETWORK CLUSTERING AS INDICATOR NORMALIZATION TOOL

Riku Hakulinen
Helsinki University Library
Helsinki, Finland

Eva Isaksson
Helsinki University Library
Helsinki, Finland

INTRODUCTION

Helsinki University Library participates in planning the publication metrics of the research assessments for University of Helsinki (UH) [1]. This creates urgency in understanding the most recent developments in methodology used by organizations providing large scale analyses of scientific impact. We tested the network clustering tool developed by CWTS, Leiden [2] on a ~67k set of social sciences oriented publications.

CHARACTERISTICS OF THE NETWORK

An undirected citation network (cnw) was gathered using bibliographic data from two databases [3]. Search query was 'Social AND Science in Title or Abstract', which should create a multidisciplinary collection of publications that share topics. In order to show that our network is a realistic enough cnw, the degree distribution, which is a central feature of any network, is plotted in Fig1.

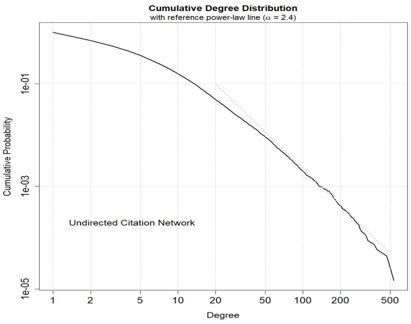

Fig1. Cumulative degree distribution of the network as a log-log plot. It indicates a typical form for a citation network, including an approximate power-law tail.

The mean degree of the network is 6.0, additional characterizing numbers are presented in Table1. Transitivity is a network size-dependent measure of the number of triangles in the network, compare to Table2.

Nodes	Links	Triangles	C_ Global	C_ Local
67605	201350	112453	0.06	0.12

Table1. Basic numerical features of the network. C_Global is the Ratio of the Means version of transitivity, and C_Local then the Mean of the Ratios version [4].

CLUSTERING FOR NORMALIZATION

Leiden and Louvain algorithms for clustering were used and mean normalized citation scores (MNCS) for UH publications in the set were calculated. As a measure of robustness, standard error was created by varying resolution, see Fig2., using mainly default values for other parameters. The MNCS robustness does not in any obvious way correlate with quality function value, but an average tendency for MNCS values can be seen.

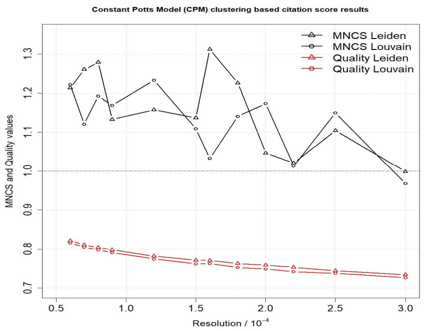

Fig2. Variation of MNCS for 232 UH publications as a function of resolution parameter value. Also shown CPM quality function values that indicate similar performance of Leiden and Louvain algorithms for this network. Average MNCSs with errors are 1.16±0.10 (Le) and 1.13±0.08 (Lo).

Comparing, Louvain had a small advantage in that its results fluctuated somewhat less on average for these resolutions. The produced number of clusters seemed high (N>2000) and the size variation was large. These are probably caused by sparsity of the network, see text below Table2. Also, each checked clustering by Leiden algorithm with a high quality function value Q (e.g., >0.7) contained only internally connected clusters, as expected [2].

QUALITATIVE RESULTS

An observation was that changing modestly the resolution, clustering method could report two very close Qs, but produce dissimilar clustering results – like the one depicted in Fig3. with Leiden/CPM and Q~0.84. In this partial cnw, there are two UH publications and as a qualitative classification, relating author keywords are listed below:

(1) Facebook; Well-being; Social network; Social support; Personality; Extraversion (2) CRIME; EXPERIENCES; SOCIETY; DESIGN; AGE.

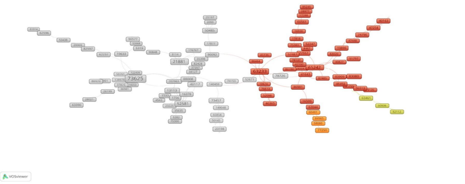

Fig3. A partial cnw representing two clustering results. This network is either one cluster or divided into four (colors). In the latter, the two UH publs are assigned to two clusters (red,gray).

In order to further study the number of clusters produced by Leiden algorithm, links were added to the cnw randomly so that small detached parts of the networks were given links, Table 2.

Nodes	Links	Triangles	C_ Global	C_ Local
67605	203849	112453	0.06	0.11

Table2. Characteristics of the randomly modified version of the cnw. The new network had the same mean degree 6.0 as the unmodified cnw.

This modification reduced the number of clusters N and made their size distribution somewhat more even. It is concluded that detached small networks are here a plausible partial cause for large N.

CONCLUSIONS

Calculated MNCSs for 232 UH publications were found robust for specific resolution intervals, but some arbitrariness in clustering for weakly connected areas of cnw was observed. Also, the results with almost perfect Q had typically one giant cluster and the rest were very small ones, not a useful classification. The contents of observed Leiden algorithm clusters with quality 0.7≤Q≤0.9 seemed reasonable for this network.

REFERENCES

[1] 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

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

[3] Part of data sourced from Web of Science (https://clarivate.com/products/web-of-science/) and part from Dimensions, an interlinked research information system provided by Digital Science (https://www.dimensions.ai)

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