Comparison of methods for clustering citation networks

Lovro Šubelj
University of Ljubljana,
Faculty of Computer and
Information Science
Ljubljana, Slovenia
lovro.subelj@fri.uni-lj.si

Nees Jan van Eck
Leiden University,
Centre for Science and
Technology Studies
Leiden, Netherlands
ecknjpvan@cwts.leidenuniv.nl

Ludo Waltman
Leiden University,
Centre for Science and
Technology Studies
Leiden, Netherlands
waltmanlr@cwts.leidenuniv.nl

There is an extensive literature on graph partitioning and community detection in networks [2]. This literature studies methods for partitioning the nodes in a network into a number of groups, often referred to as communities or clusters, where nodes belonging to the same cluster should be relatively strongly connected to each other, while nodes belonging to different clusters should be only weakly connected [3].

Which methods for graph partitioning and community detection perform best in practice? The literature does not provide a clear answer to this question, and if the question can be answered at all, then most likely the answer will be dependent on the type of network that is being studied and on the type of partitioning that one is interested in.

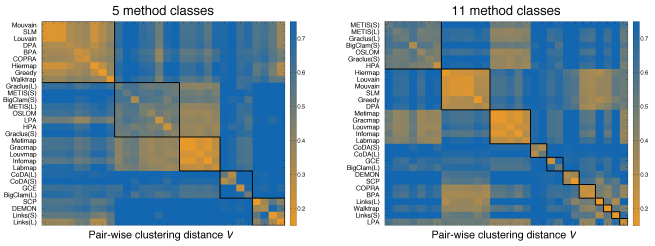
We address the above question in one specific context [1]. We are interested in grouping scientific publications into clusters based on their direct citation relations and we expect each cluster to represent a set of publications that are topically related to each other. We therefore compare the performance of different clustering methods when applied to citation networks that have been collected from the Web of Science bibliographic database.

We consider a large number of methods including spectral and statistical methods, modularity optimization, matrix factorization, map equation algorithms, link clustering, label propagation, random walks and methods based on cliques, to name a few. We first conduct a direct pair-wise comparison of the clusterings obtained using different methods. Despite a large number of methods considered, these can be divided into only a handful of truly different classes, whereas the differences between the classes can be rather substantial (Fig. 1).

We next compare standard statistical properties of the clusterings including cluster size distributions, robustness to random perturbations [4], network modularity [6] and other. We also focus on a number of properties that are of special relevance in the context of citation networks of publications. These include fraction of citations covered, effective range of the clusterings, orders of magnitude covered by cluster sizes, and method uncertainty and complexity.

However, to obtain a deep understanding of the differences between clustering methods, we believe that analyzing the statistical properties of clusterings is not sufficient. Understanding the differences between methods also requires an expert-based assessment of different clusterings. This is a challenging task that involves a number of practical difficulties, but we nevertheless make an attempt to perform such an expert-based assessment for publications in the field of Library & information science [7].

Since none of the considered methods performs indeed satisfactory according to all desired criteria, we discuss strengths and weaknesses of different methods.



Pair-wise distances between the clusterings obtained by different methods. Panels show the heatmaps of the clustering distances for Library & information science citation network, where the methods are grouped into 5 and 11 classes (left and right, respectively). The clustering distance is measured using variation of information [5], where lower values (orange) correspond to greater similarity between the clusterings.

[1] L. Šubelj, N. J. Van Eck and L. Waltman. Clustering of scientific publications based on citation relations: A comparison of different methods. *PLoS ONE*, in submission, 2015.
[2] S. Fortunato. Community detection in graphs. *Phys. Rep.*, 486(3-5):75–174, 2010.
[3] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *P. Natl. Acad. Sci. USA*, 99(12):7821–7826, 2002.
[4] B. Karrer, E. Levina, and M. E. J. Newman. Robustness of community structure in networks. *Phys. Rev. E*, 77(4):046119, 2008.
[5] M. Meila. Comparing clusterings: An information based distance. *J. Multivariate Anal.*, 98(5):873–895, 2007.
[6] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, 2004.
[7] N. J. Van Eck and L. Waltman. CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *J. Infometr.*, 8(4):802–823, 2014.

This work has been supported in part by the Slovenian Research Agency Program No. P2-0359, by the Slovenian Ministry of Education, Science and Sport Grant No. 430-168/2013/91, and by the European Union, European Social Fund.