ANALYSIS OF CITATION NETWORKS

Anita Valmarska [1,2], Janez Demšar [1]

[1] Faculty of Computer and Information Science, Ljubljana, Slovenia
[2] Jožef Stefan Institute, Ljubljana, Slovenia
anita.valmarska@ijs.si, janez.demsar@fri.uni-lj.si

1 Introduction

Citation networks are directed networks in which one paper cites another. Reasons for citations are various. In most cases the authors cite older publications in order to identify the related body of work, to substantiate claims or establish precedence, or to legitimate their own statements or assumptions. In the scientific world citations are used also to critically analyze or correct earlier work. Intuitively, it can be expected that papers would more often cite other papers from the same research subfield. To confirm this hypothesis, we wanted to find out whether we could detect the research subfields within a single research field, i.e., psychology, using only a citation network of papers published in the given research field. To this end we applied one of the state-of-the-art algorithms for community detection in the hope that we would be able to differentiate among different topics addressed in psychology research.

2 Data collection and network construction

To the best of our knowledge, there is no central repository with publications from psychology. Consequently, we decided to crawl the pages connected with psychology in Wikipedia. From each of the visited pages we collected the references identified by their DOIs in the reference section. This resulted in a collection of 63,826 unique papers.

Next, we queried the Microsoft Academic Research data (MAS) and collected information about the scientific papers citing the initial set of collected papers. This allowed us to construct a citation network whose core contained papers published in the field of psychology. The resulting network consists of 948,791 vertices and 1,539,563 edges.

Due to the nature of our data collection process, we had to perform an initial data pre-processing in order to extract the papers that had a significant impact on the field of psychology. This resulted with a new network of 3,918 vertices connected by 5,732 edges.

3 Community detection and naming the communities

The process of identification of research subfields in the citation network was translated into the problem of community detection. For the purpose of our research, we applied the Louvain method [1]. It is a simple, efficient, and easy to implement method for identifying communities in large networks.

The method is divided into two phases that are repeated iteratively. At the beginning, each of the n vertices of the graph $G(V, E)$ is assigned to a different community. In this initial partition there are as many communities as there are vertices. Then, for each vertex v we consider the neighbors u of v and we evaluate the gain of modularity that would result by removing v from its community and placing it in the community of u. The vertex is placed in the community for which this gain is maximal, and only if the gain is positive. This process is applied repeatedly and sequentially for all vertices until no further improvement can be achieved. The second phase of the method consists of building a new network whose vertices are now the communities found during the previous phase. Once the second phase is completed, it is possible to iteratively re-apply the two phases on the obtained networks.

Part of the evaluation of the detected communities was to name them and examine their connections. Due to the vast quantity of available data and unfamiliarity with the field of psychology, we named the communities based on the cosine similarity between our initially collected psychological papers and relevant texts for each of the APA (*American Psychological Association*) divisions of psychology.

4 Results

The community detection algorithm implemented in Pajek [4] detected 52 communities. The division into communities can be observed in *Figure 1*. The smallest cluster included 7 papers, while the largest cluster was constructed of 230 psychological publications. In *Figure 1* we can observe the extracted and named communities.

5 Conclusion

Results obtained by the network analysis and community detection are encouraging. The visual representation of the communities reveals sensible relationships between psychology subfields. However, the nature of data collection and the influence of our subjective judgment on community naming offer opportunities for further improvement. This involves improved data collection, developing new and improved methods for community detection, and employing better measures for text similarity. In further work, we would also like to explore the methodology proposed by Grčar et al. [2].

References

1. V. D. Blondel, J. L. Guillaume, R. Lambiotte, E. Lefebvre. ”Fast unfolding of communities in large networks”, *Journal of Statistical Mechanics-Theory and Experiment*, vol. 10, no. 10, 2008.
2. M. Grčar, N. Trdin, N. Lavrač. ”A Methodology for Mining Document-Enriched Heterogeneous Information Networks” , *The Computer Journal*, bxs058, 2012.
3. T. Kamada, S. Kawai. ”An algorithm for drawing general undirected graphs”, *Information Processing Letters*, vol. 31, no. 1, pp. 7–15, 1989.
4. W. de Nooy, A. Mrvar, V. Batagelj. *Exploratory Social Network Analysis with Pajek*, Cambridge University Press, New York, 2011.

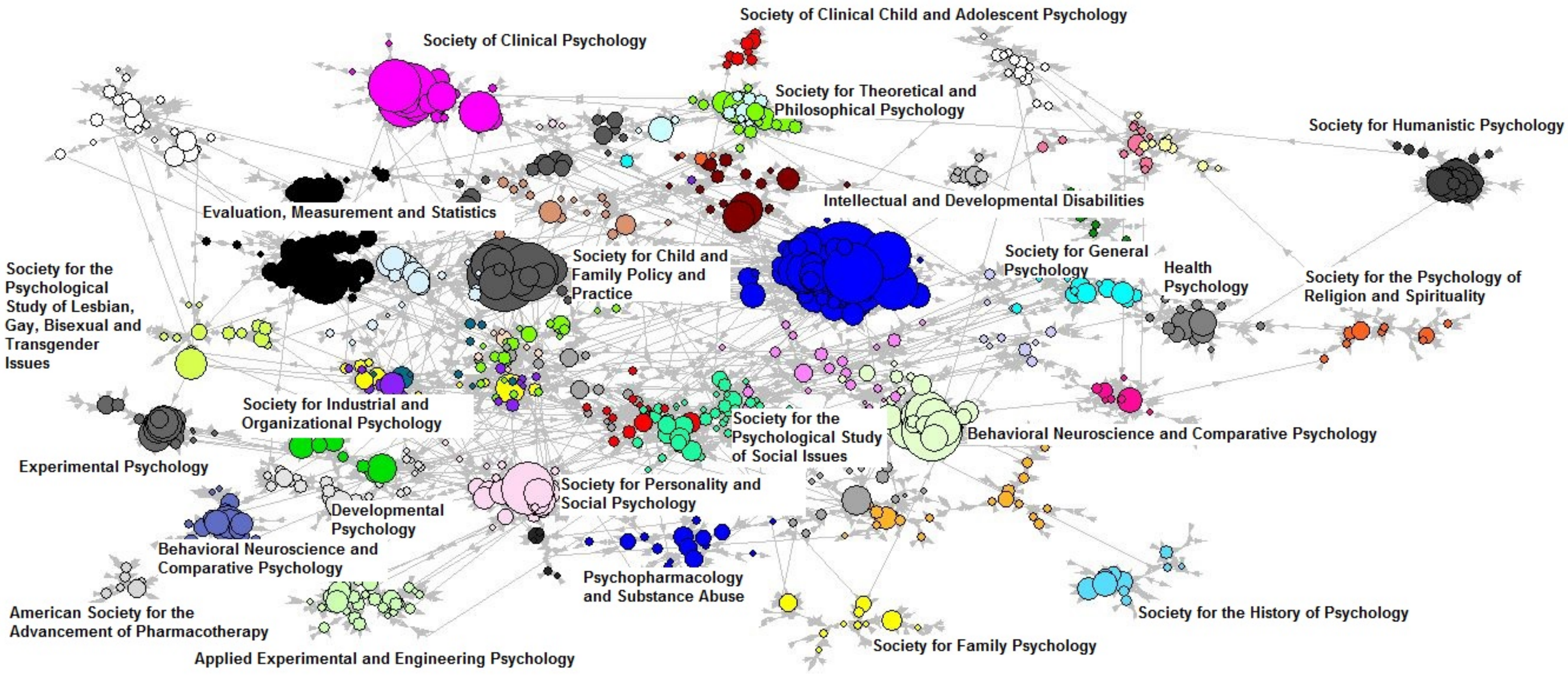
Society of Clinical Child and Adolescent Psychology
Society of Clinical Psychology
Society for Theoretical and Philosophical Psychology
Society for Humanistic Psychology
Intellectual and Developmental Disabilities
Evaluation, Measurement and Statistics
Society for the Psychological Study of Lesbian, Gay, Bisexual and Transgender Issues
Society for Child and Family Policy and Practice
Society for General Psychology
Health Psychology
Society for the Psychology of Religion and Spirituality
Society for Industrial and Organizational Psychology
Society for the Psychological Study of Social Issues
Behavioral Neuroscience and Comparative Psychology
Experimental Psychology
Society for Personality and Social Psychology
Developmental Psychology
Behavioral Neuroscience and Comparative Psychology
Psychopharmacology and Substance Abuse
American Society for the Advancement of Pharmacotherapy
Society for the History of Psychology
Society for Family Psychology
Applied Experimental and Engineering Psychology

Fig. 1. *Community detection of psychological papers. Louvain method returned a total of 52 communities. The smallest cluster included 7 papers, while the largest cluster was constructed of 230 psychological publications. Vertex size represents the betweenness value of the vertex. Network is visualized with the Kamada-Kawai [3] visualization algorithm. Algorithm was applied twice: firstly for separation of the communities, and then for optimization of the position of vertices within their community. Communities were named based on the measures for cosine similarity between our initial psychological papers and the relevant texts for each of the APA divisions.*