Mapping the Academic Landscape of Numerical Cognition through Interactive Citation Networks

Nibu Jacob, Kia Ekbia, and Samuel Bentum

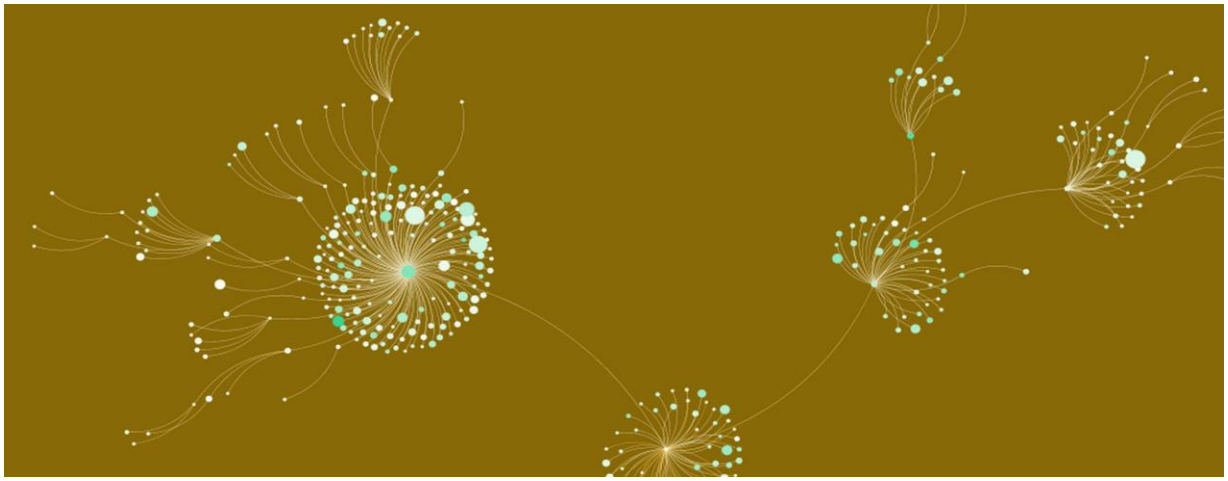

Figure 1. Communities among research papers detected by Blondel algorithm

Abstract—The evolution of research in an academic discipline can be represented through visualization techniques. Citation data has proven valuable in detecting communities among research papers and powering the visualization of the academic landscape. Attempts have been made in fine-tuning citation data and creating algorithms for community detection as well as ranking. Ranking algorithms are applied to direct citation networks whereas community detection algorithms are applied to co-citation networks. In this paper, we attempt to combine the results from both approaches and produce a visualization that shows both structure and ranking. We further make this visualization interactive so that researchers can customize the network by adding their domain knowledge. We hope that this approach will be a valuable research aid in exploring sub-disciplines and finding focus areas.

Index Terms—Citation network analysis, Network visualization, Co-citation

INTRODUCTION

The purpose of this project is to visualize the knowledge diffusion process in the academic sub-discipline of "numerical cognition" using an interactive citation network visualization that shows the clusters and ranking within the network. David Braithwaite (the client of this project) needs a visualization that extracts the main bodies in the extensive literature related to numerical cognition. Such a visualization would help him identify the core research papers and linkages between them that are most relevant to the development of the sub-discipline over the years.

Conventional method of searching and filtering a large database like Google Scholar or Web of Science leaves you with a list of articles that does not convey any insights on its own. A visualization of the same list of articles is significantly more informative. It can show the relationships between the articles, their relative importance, and clustering within the network of articles. An interactive citation network visualization goes one step further by letting the researcher input his or her own domain knowledge in the visualization, thereby customizing the visualization to the specific research goal.

Citation network is a scale-free network similar to WWW, social networks, semantic web, protein network, and airline network. This means that some articles in the network have relatively very high number of citations received and they are called the “hubs”. Analytic approaches used in citation network analysis can be potentially applied to many real-world networks and vice versa.

In addition to providing the client with an interactive visualization tool and workflow that may be reused by other researchers, this project aims to find the prominent research articles, authors, journals, and schools of thought (communities or clusters) in the field of numerical cognition.

- *Nibu Jacob is a graduate student at Indiana University Bloomington. E-mail: jacobnibu@gmail.com*
- *Kia Ekbia is a graduate student at Indiana University Bloomington. E-mail: kiekbia@indiana.edu*
- *Samuel Bentum is a doctoral student at Indiana University Bloomington. E-mail:sbentum@indiana.edu*

1 RELATED WORKS

Ever since Garfield (1964) showed that the network of citations strongly coincide with the network of historical events [1], citation analysis has been used as an effective strategy to study the evolution of academic landscapes. Earlier researches in this field focused on ranking the publications based on centrality measures calculated from local direct citations [2]. More recently, the importance of indirect citation links and clustering within the network has been shown [3]. Various PageRank algorithms that consider many factors in addition

to local ties are widely used in finding prominent articles in a citation network [4].

Unlike a co-citation network, a direct citation network is better suited to study the history and development of an academic discipline [5]. Direct citation network analysis can detect large and young emerging clusters earlier [6]. CitNetExplorer is a software tool that was developed to visualize a direct citation network over time, however, it lacks interactivity features for customization by a researcher for a specific topic of interest [5].

Waltman, L. and Yan, E. describe a method of extracting a citation network using Sci2 software tool (with the database plugin added) [7] [8]. This method has the advantage that, by using a custom property file, a researcher can include any variables of interest in the extracted network. Sci2 also does a lot of data preparation automatically on loading the data.

Analysis of transient articles (the more recent articles with peak citations in a short time) can reveal emerging trends and topic bursts in a citation network [9]. CiteSpace is a software tool that has been successfully used in this regard [10].

Ying Ding was able to show a high correlation between the "prestige rank" of a paper and the weighted PageRank, which can be calculated using different algorithms [11]. Jon Kleinberg developed the HITS algorithm to find authoritative sources of information in the WWW [12]. When applied to a citation network, this algorithm calculates an authority score (ranging from 0 to 1) for each article in the network, higher scores indicating more authoritative articles.

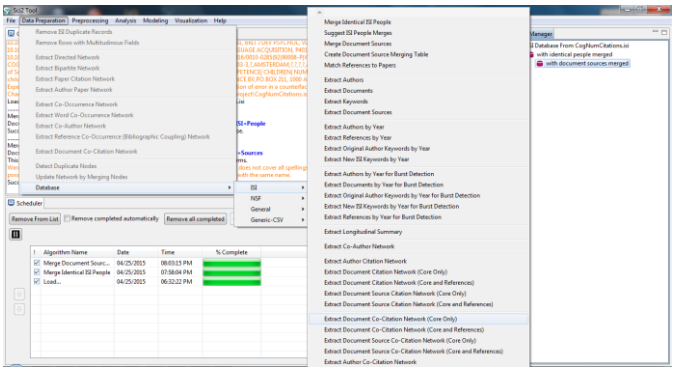

Figure 2. Screenshot of Sci2 for data preparation

Scale-free networks are known to exhibit community structures and many algorithms have been developed to detect these communities. The Blondel algorithm (also called the Louvain method) is a modularity-maximization algorithm that can detect communities in a fast and efficient way [13]. Comparative studies of community-detection algorithms have demonstrated that Blondel algorithm has excellent performance [14] [15].

Large networks having many interconnections can give messy visualizations that are difficult to examine and interpret. Pathfinder algorithms can prune such networks to a smaller subset that is easier to visualize. MST-Pathfinder algorithm is an efficient variant of the Pathfinder algorithm that can be used in a co-citation network to reduce the number of articles to a representative smaller sample [16].

2 Data Preparation

Using a list of keywords provided by the client, searched and downloaded a list of articles in ISI format from Thomson ISI's Web of Science. This dataset had a total of 6,379 records published from 1919 to 2015.

Among the few software tools available for loading files in ISI format, the Java-based tool, Sci2 (version 0.5.2 with the ISI database plugin), was selected for further processing of this dataset. This particular tool offered the best flexibility and automated cleaning options for working with citation data (Figure 2).

A search result in Web of Science may fetch duplicate records if two records of the same article have a different first author or the source or author name is written differently. To remove such duplicate records, Sci2 offers merge options based on identical people and sources.

Publication	Articles
Cognition	210
Journal of experimental child psychology	154
Neuropsychologia	138
Neuroimage	130
Developmental science	104
Journal of cognitive neuroscience	91
Cortex	80
Developmental psychology	77
Child development	76
Behavioral and brain sciences	72

Table 1. Top 10 sources in "numerical cognition" research

After matching the references to the articles in the dataset (to find links between articles and their references), a citation network as well as a co-citation network of the core articles (articles in the dataset excluding their references) were extracted. These networks had articles in the range of 3000 to 6000 and significantly more number of interconnections between them.

We used the igraph package in R to prune the citation network based on number of keywords (that the client is most interested in) and other parameters like authority score, page-rank, year of publication, and the number of citations received. For the co-citation network, we removed isolated articles and used MST-Pathfinder algorithm (in Sci2) to prune the network.

The citation network has ranking information (authority score and page-rank) whereas the co-citation network can show the community structure with a community-detection algorithm. To display both ranking and structure in the same visualization, we mapped the ranking information in citation network to the co-citation network using R. The source code for this can be found on GitHub [17].

3 Data Visualization

Figure 6 shows the top 20 authors based on number of articles authored in the dataset as well as the number of citations they received globally. A list of top 10 publication sources based on the number of articles in the dataset can be seen in Table 1.

A line chart (Figure 3) shows the evolution of research activity in the field, in terms of number of articles and references published, distinct authors, sources, and ISI keywords.

Two visualizations (Figure 4 and 5) showing top 100 research papers based on authority score and page-rank respectively show the performance of the two indicators in a direct citation network. The two networks are laid out using DAG Layout plugin in Gephi.

Figure 7 shows the screenshot of interactive visualization of the co-citation network of research papers showing their communities (based on semantic similarity) and relative ranking (based on authority score and page-rank).

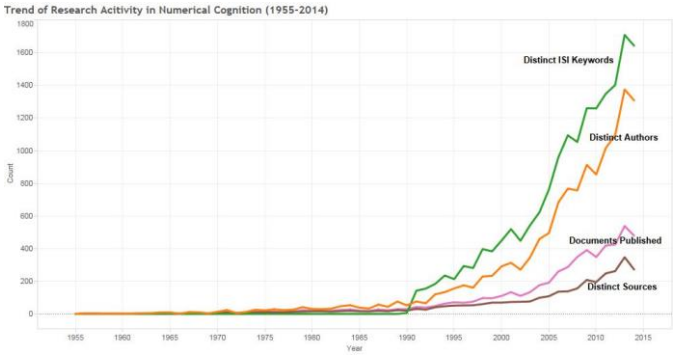

Figure 3. Trend of research activity in "numerical cognition"

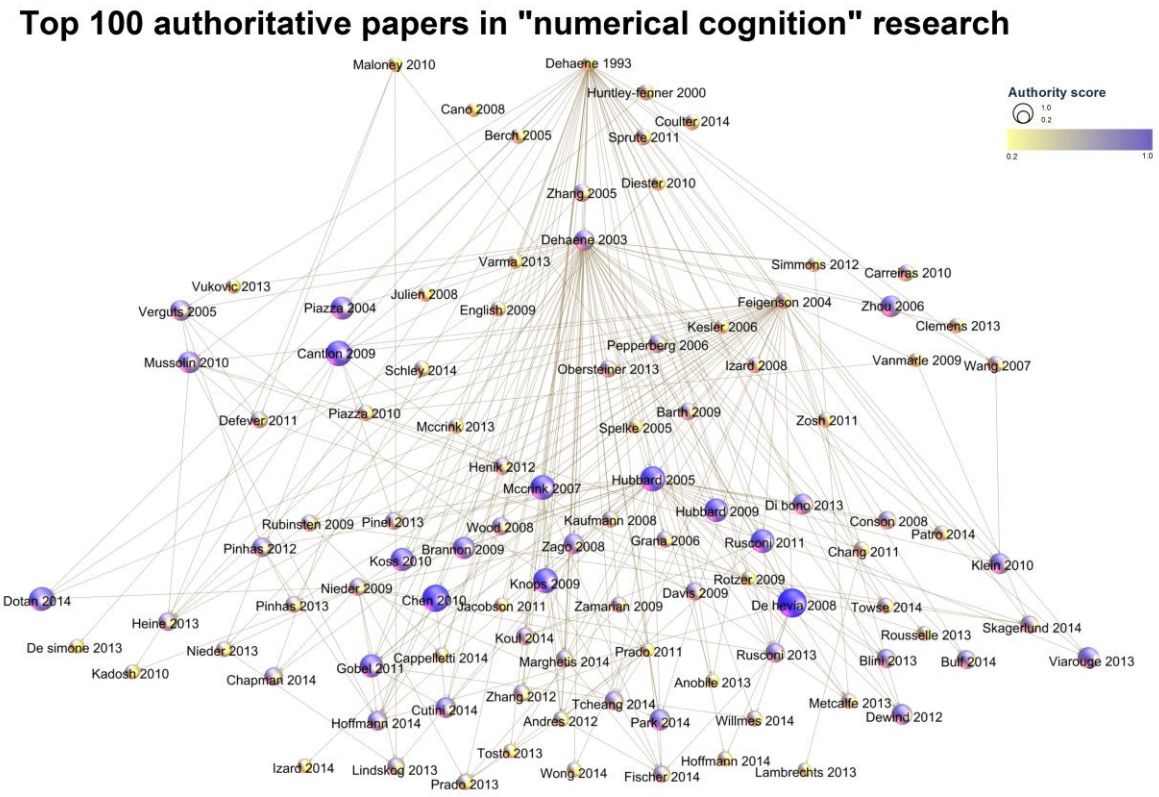

Figure 4. Top 100 papers based on authority score

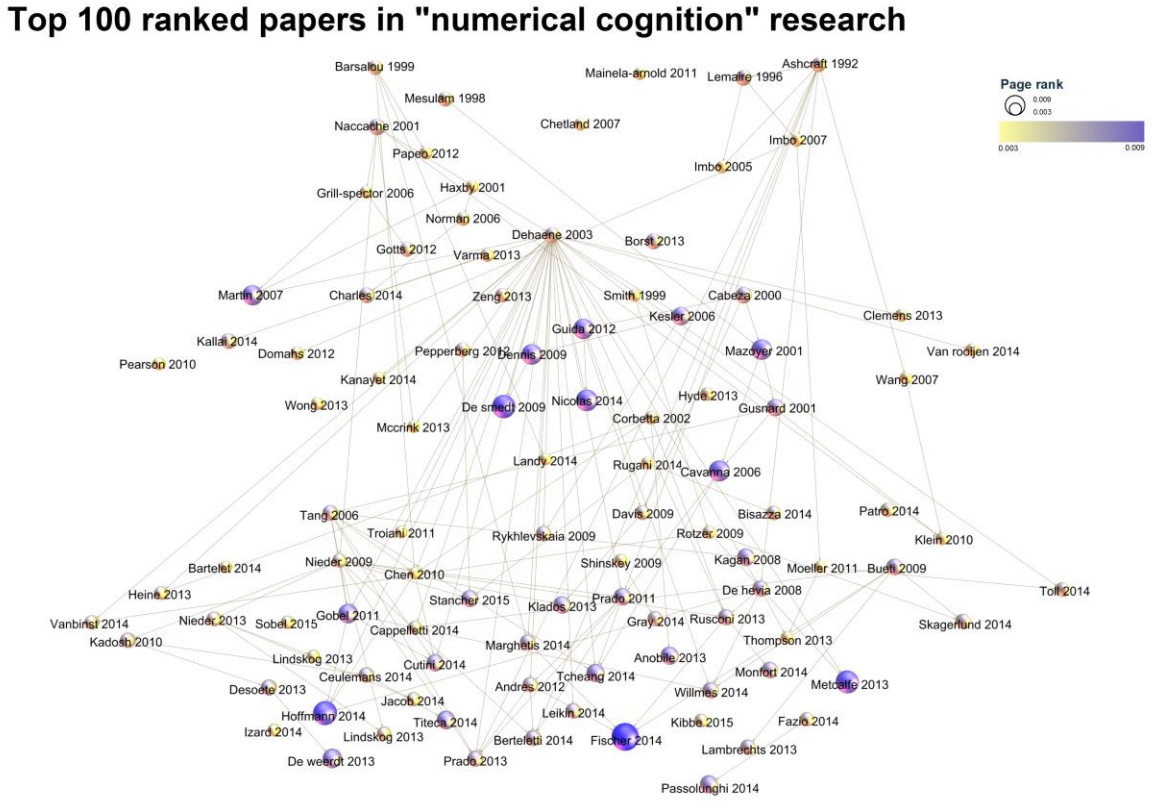

Figure 5. Top 100 papers based on page-rank

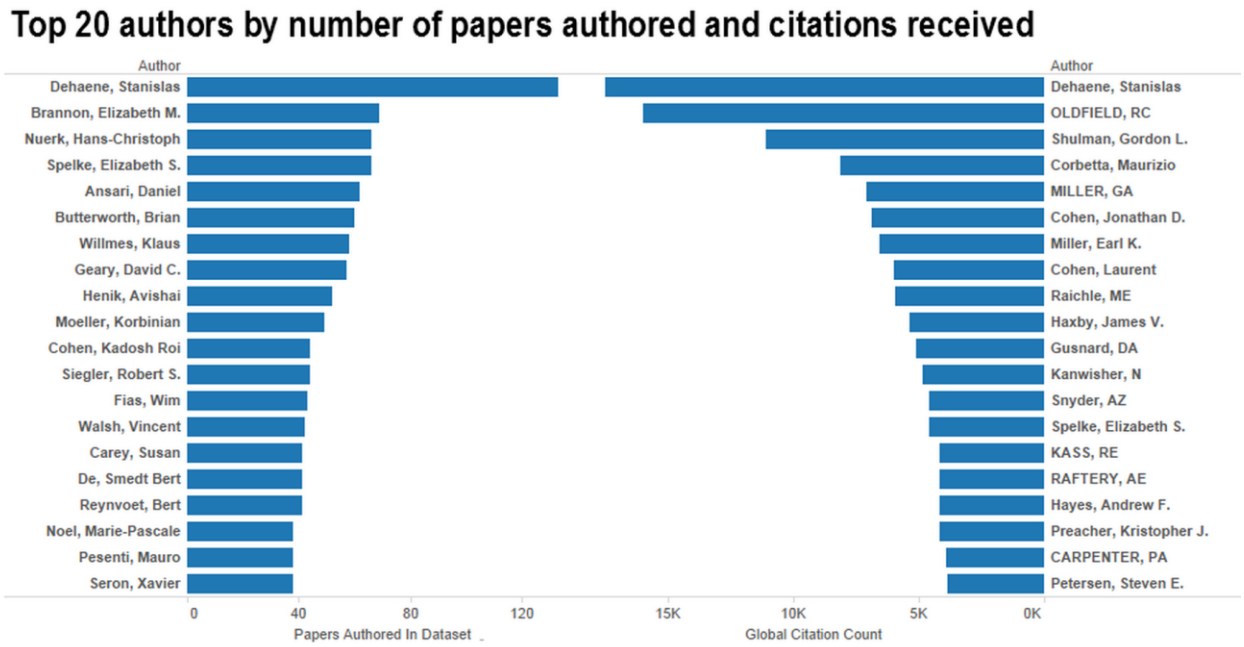

Figure 6. Top 20 authors based on number of papers authored in the dataset and citations received globally

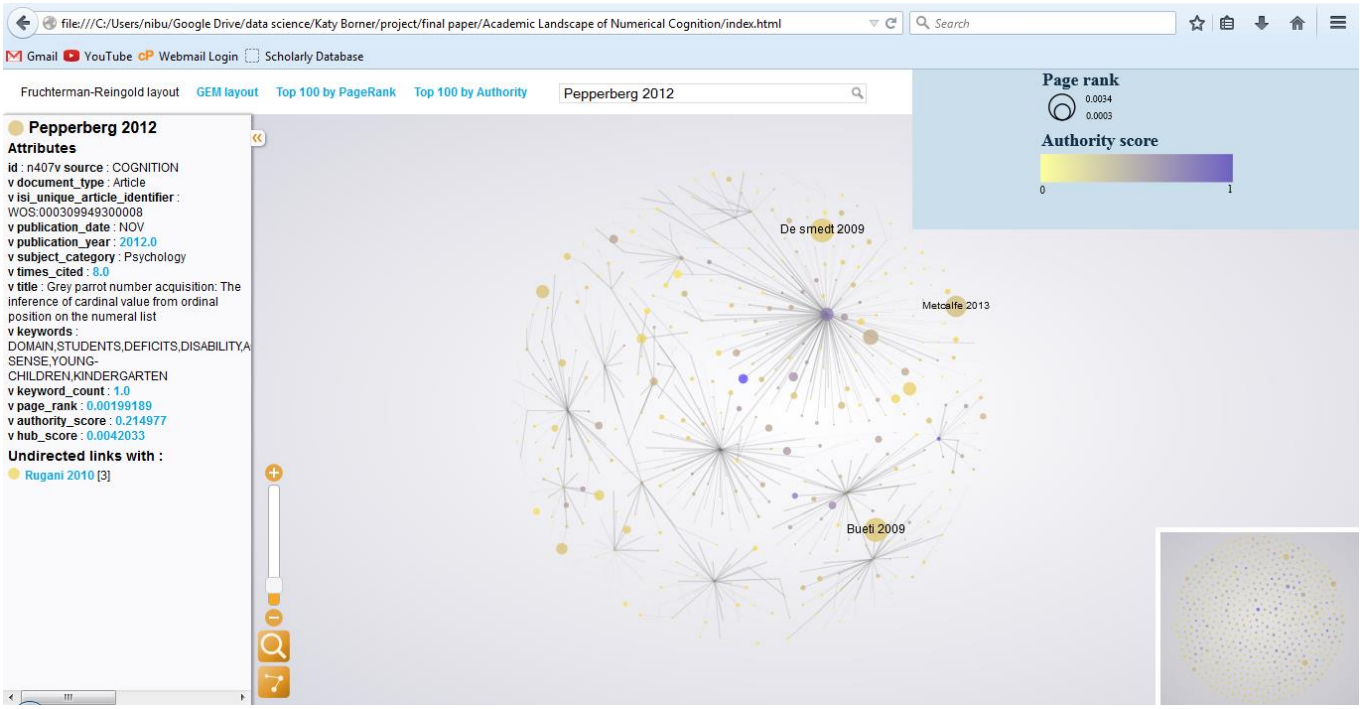

Figure 7. Screenshot of the interface of interactive visualization

4 Discussion of results

We learned from our initial workflow iterations that a citation network can be too huge for many visualization tools out there to handle. Even if the tool can process big networks, the resulting visualization is too cluttered for human inspection and interpretation. Networks of size less than 500 nodes produced the best results for analysis.

The community-detection algorithms tend to produce layouts that lead the attention to "hub" articles. In a citation network, the hubs tend to be review articles that cite many articles in a specific research topic. Even though this is highly indicative of semantic similarity, a researcher could be more interested in finding the authoritative articles in the network.

Blondel's community detection algorithm detects communities but finding out what topic or keyword has constituted a community is left to the researcher to find out. Sub-setting the articles in a cluster for topical analysis of their titles, keywords, or abstracts to find a common theme or keyword could be helpful in giving a label to the communities. The accuracy of this approach is, however, questionable.

The lack of a single software tool that can search the databases for citation data and make relevant visualizations is a concern. In our research, CiteSpace is the visualization tool that came close to achieving that goal [9]. However, CiteSpace has a learning curve that can prevent its widespread adoption.

Citation network being a directed acyclic graph, DAG layout was very suitable and created an insightful visualization with regards to the historical perspective of the research field [18].

The very broad nature of this dataset with regards to the keywords used in searching for the articles could potentially create a mirage of what may appear to be a significant contributor to a network. Thus care must be taken in selecting which keywords to extract scientific journals from.

The initial dataset from the client showed over 5,000 nodes and 12, 000 edges. An initial screening using some popular tools like GUESS (in Sci2) and Cytoscape yielded some challenges with overall clarity of the network compared to Gephi tool, which could easily handle them. It was evident we had to scale down the network to accommodate these types of tools. However, even with a scaled down dataset, it is always important to understand how the articles are related based on their hub activities/scores to get a better picture of the co-citation network.

As discussed earlier, performing a Hyperlink-Induced Topic Search (HITS) algorithm to determine the hub and authority score is one unique way of categorizing articles by their level of influence to other articles (or authors). In our analysis, it also became apparent to use the PageRank scores to establish the relative importance of the articles. We could see a trend between the authority scores of the articles and the PageRank scores of the corresponding documents, i.e., a proportional index. These two ways of ranking helps to identify which key authors' works have major prominence across the scientific communities.

Key findings can further be expounded using the keywords within these articles to identify which citation links share that similar concept. This then creates the basis of communities of knowledge as described above in the community detections of the network. With this approach, the client (and/or researchers) can review the communities of interests and potentially modify them with an addition or subtraction of nodes in an interactive fashion.

5 Conclusion

Numerical cognition is an area of active research with near exponential growth in the number of research papers being published in the recent years. Network visualization has shown clearly differentiating areas of focus and communities in the field. Authority score and page-rank measures provided additional insights to detect individual papers of importance in the field. Citation analysis has proven to be useful in delineating the development of the academic discipline, however, there is a lack of easy-to-use software tools that researchers can employ in their literature review. Reducing the large body of research papers to a small but representative subset asks for a certain level of software skills from the part of the researcher. This is an area for potential future work in citation analysis.

Acknowledgments

The authors wish to thank David Braithwaite (Carnegie Mellon University) and Katy Borner (Indiana University) for their guidelines and support.

References

[1] S. Jung, and A. Segev, Analyzing future communities in growing citation networks. *Knowledge-Based Systems*, 69:34–44J., 2014.

[2] Y. Jiang, Locating active actors in the scientific collaboration communities based on interaction topology analyses. *Scientometrics* 74(3):471–82E, 2008.

[3] M. Marra et al., The value of indirect ties in citation networks: SNA analysis with OWA operator weights, Inform, 2015.

[4] M. Nykl, et al., PageRank variants in the evaluation of citation networks. *Journal of Informetrics* 8:683–692J.S., 2014.

[5] N.J. Van Eck, and L. Waltman, CitNetExplorer: A new software tool for analyzing and visualizing citation networks, *Journal of Informetrics* 8:802–823W.-K., 2014.

[6] K. Fujita, Detecting research fronts using different types of weighted citation networks. *J. Eng. Technol. Manage*. 32:129–146H., 2014.

[7] L. Waltman, and E. Yan, PageRank-Related Methods for Analyzing Citation Networks. Measuring Scholarly Impact: Methods and Practice 94-95K., 2014.

[8] Sci2 Team. (2009). Science of Science (Sci2) Tool. Indiana University and SciTech Strategies, https://sci2.cns.iu.edu

[9] C. Chen, CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *J. Am. Soc. Inf. Sci.*, 57: 359–377R., 2006.

[10] C. Chen, The CiteSpace Manual., 2014, available at http://cluster.ischool.drexel.edu/~cchen/citespace/CiteSpaceManual.pdf

[11] Y. Ding, "Applying weighted PageRank to author citation networks." *Journal of the American Society for Information Science and Technology* 62.2: 236-245S.P., 2011.

[12] J.M. Kleinberg, "Authoritative sources in a hyperlinked environment." *Journal of the ACM* 46, no. 5:604-632., 1999.

[13] V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E.L.J.S. Mech, "Fast unfolding of communities in large networks." *J. Stat. Mech*:P10008, 2008.

[14] S. Fortunato, and A. Lancichinetti, "Community detection algorithms: a comparative analysis: invited presentation, extended abstract," paper presented at the meeting of the VALUETOOLS, ACM, pp. 27, 2009.

[15] T.N. Dinh, and M.T. Thai, "Community Detection in Scale-Free Networks: Approximation Algorithms for Maximizing Modularity," *IEEE Journal on Selected Areas in Communications* 31 (6), 997-1006, 2013.

[16] A. Quirin, O. Cordón, V.P.G. Bote, B. Vargas-Quesada, and F. de Moya Anegón, "A quick MST-based algorithm to obtain Pathfinder networks (∞, n - 1)," *JASIST*, 59, 1912-1924, 2008.

[17] Available at https://github.com/jacobnibu/Network-Reduction

[18] Paul Paulson, DAG Layout plugin, 2012, available at https://marketplace.gephi.org/plugin/dag-layout/