The Community Structure of Scientific Collaboration in Physics — A Citation Network Analysis of the Physics Review Publications

Yuan Huang *yuanh@physics.umass.edu*

Problem Statement. In this project, we are going to study the community structure of citation networks within the Physical Review family of journals using clustering algorithms. The analysis of the underlying community structure in the citation networks provides numerous benefits for the development of the research field, including building a foundation for future research through the acknowledgement of past research activities; identify gaps in research for researchers and students; improve the integration between theory and practice, and so on [1, 2, 3, 4]. On the other hand, finding clusters in network-structured data is a fundamental problem that has many applications in social networks and other related fields.

Methodology. In recent years, a growing number of clustering algorithms for categorical data have been proposed based on various centrality measures. Ref [5] gives a good review about these algorithms. In this project, we are going to use a clustering method proposed in Ref [6]. In this method, each individual nodes and their relationships are denoted by weighted graphs, and the measure of graph density is used to give a quantity depict of whole correlation among individuals in a community. The clustering algorithm is a process that detects all dense subgraphs and construct a hierarchically nested system to illustrate their inclusion relation. A heuristic process is applied here for finding all quasi-cliques with density in various levels. The code for the algorithm will be implemented in Python language.

Data Sets. The data set has been requested from the American Physics Society website (http://journals.aps.org/datasets) and contains over 450,000 articles from Physical Review Letters, Physical Review, and Reviews of Modern Physics dating back to 1893. The data sets have two parts:

- Citing article pairs: This data set consists of pairs of APS articles that cite each other.
- Article metadata: This data set consists of the basic metadata of all APS journal articles.

Experiments. In order to evaluate the clustering method, we are going to compare its results with that of preceded methods, such as K-Means method and modularity maximization method proposed in [7]. Since we don't know the ground-truth clustering labels for the data, we will use some internal metrics such as Silhouette Coefficient or Calinski-Harabaz Index to evaluate the performance of the clustering method. Another possibility is to extract the clustering label from the PACS code (PACS is a hierarchical partitioning of the whole spectrum of subject matter in physics, astronomy, and related sciences to identify fields and sub-fields of physics.) in the metadata of the articles. With the sub-fileld information, we can use external metrics such as F-score, or normalized mutual-information to evaluate the clusters.

Related Work and Novelty. The analysis of citation and co-authorship network has been performed for many research fields, such as organic LED [2], learning analysis [3], sustainable science [4], and so on. Many clustering algorithms have been proposed to study the community structure of the academic networks [5, 7, 6]. For the Physics Review dataset, the citaion network analysis has also been performed in Ref [1], where they used a modularity maximization clustering method. In this project, we will apply a new clustering method based on the graph density metrics which has been tested for other well-known network datasets but not in a large-scale citation network before. In this project, we also want to study the time-evolving features of different sub-fields in physics to exhibit the future trend in physics.

References

[1] P. Chen, S. Redner. *Community structure of the physical review citation network*, Journal of Informetrics, **4** (2010) 278-290.

[2] Y. Kajikawa, Y. Takeda. *Citation network analysis of organic LEDs*, Technological Forecasting and Social Change, **76** (2009) 1115-1123.

[3] S. Dawson, D. Gasevic, G. Siemens, S. Joksimovic. *Current State and Future Trends: A Citation Network Analysis of the Learning Analytics Field*, In Proceedings of the Fourth International Conference on Learning Analytics And Knowledge (LAK 2014). ACM, New York, NY, USA, 231-240.

[4] Y. Kajikawa, J. Ohno, Y. Takeda, K. Matsushima, H. Komiyama. *Creating an academic landscape of sustainability science: an analysis of the citation network*, Sustain Science, (2007) **2**:221-231.

[5] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas. *Comparing community structure identification*, J. Stat. Mech. (2005) P09008.

[6] P. Zhao, and C. Zhang. *A new clustering method and its application in social networks*, Pattern Recognition Letters **32** (2011) 2109-2118.

[7] M. E. J. Newman. *Fast algorithm for detecting community structure in networks*, Phys. Rev. E **69**, 066133 (2004).

[8] Y. Ding. *Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks*, Journal of Informetrics **5** (2011) 187-203.

[9] M. E. J. Newman. *Coauthorship networks and patterns of scientific collaboration*, PNAS, vol. 101, suppl. 1, 5200-5205 (2004).

[10] M. E. J. Newman, and M. Girvan. *Finding and evaluating community structure in networks*, Phys. Rev. E **69**, 026113 (2004).