Visualization of Community Structures in Scientific Networks

Yuan Huang
Department of Physics
University of Massachusetts, Amherst
Amherst, MA, 01002
yuanh@physics.umass.edu

1 Introduction

Detecting community structure is an important task in the field of complex networks. Community structure refers to that the vertices in networks tend to cluster into different groups with a large number of within-group edges and a small number of edges between groups. In this paper, we focus on the community structure in the citation networks of academic articles. The analysis of the underlying community structure in 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; identifying gaps in research for researchers and students; improving 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.

As I start this project, I wish to use various network visualization to answer the following questions: 1) How do different subfields of physics evolve in time? Which subfields are fast growing in recent years? 2) Who are the physicists with most impact in a specified field of physics? How are the top physicists interacting with each other (through citation and collaboration)? 3) What are the top institutions for a specified field? How are the institutions collaborating with each other? So far, the project includes the first two questions regarding the subfield analysis and top physicists in a specified field. Hopefully the intitutions analysis can be realized in the future.

2 Related Works

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. The study of the community structure has been attracting more and more interest in the field of scientific network study. In recent years, a growing number of clustering algorithms for categorical data have been proposed to study the community structure of the scientific networks [5, 6, 7, 8, 9, 10].

In 2008, Leicht and Newman [5] proposed a general divisive hierarchical clustering algorithm for network. The method try to maximize the modularity of the graph using the eigenvalues and eigenvectors of the modularity matrix. By iteratively divide each group into subgroups according to the leading eigenvectors of the modularity matrix, the algorithm can achieve the clustering that maximize the global modularity.

At the same year, Blondel et. al. [6] proposed a highly efficient agglomerative clustering algorithm. Here each network node is initially assigned to a unique community. Then, the algorithm iteratively groups the pair of nodes that maximizes the increase of the global modularity into the same community, until a maximal modularity is achieved. This algorithm is computationally more efficient than the previous divisive approach but has the disadvantage of requiring more memory.

In 2010, Chen and Redner [1] investigated the community structure of physics subfields in the citation network of well-cited Physical Review Publications (with more than 100 citations). In their paper, they applied the divisive hierarchical algorithm based on modularity maximization [5] to identify community within networks and performed analysis on the major communities. They also examined communities decade by decade and found a small number of significant links between communities that are widely separated in time.

In 2011, Zhao and Zhang [7] proposed a new clustering method for detecting community structures. In this method, individuals and their relationships are denoted by weighted graphs, then the graph density gives a better quantity depict of whole correlation among individuals in a community, so that a reasonable clustering output can be presented. By constructing a non-binary hierarchical tree of clusters, this method has two important features: one is that the hierarchical tree is much smaller that clearly highlights meaningful clusters; the other one is the clusters can overlap with each other which does reflect the complexity of our real world.

3 Data Sets

The data set has been requested from the American Physics Society website and contains over 450,000 articles from Physical Review Letters, Physical Review, and Reviews of Modern Physics dating from 1893 to 2015. The data sets have two parts: (1) citing article pairs, which consists of pairs of APS articles that cite each other in format ”10.1103/PhysRevSeriesI.11.215, 10.1103/PhysRevSeriesI.1.1”(first string is the doi of the citing article, while the second string is the doi of the cited article); (2) article metadata which consists of the basic metadata of all APS journal articles including doi, title, journal, date, volume, page, author, affiliations, and so on (each article is stored in individual json files).

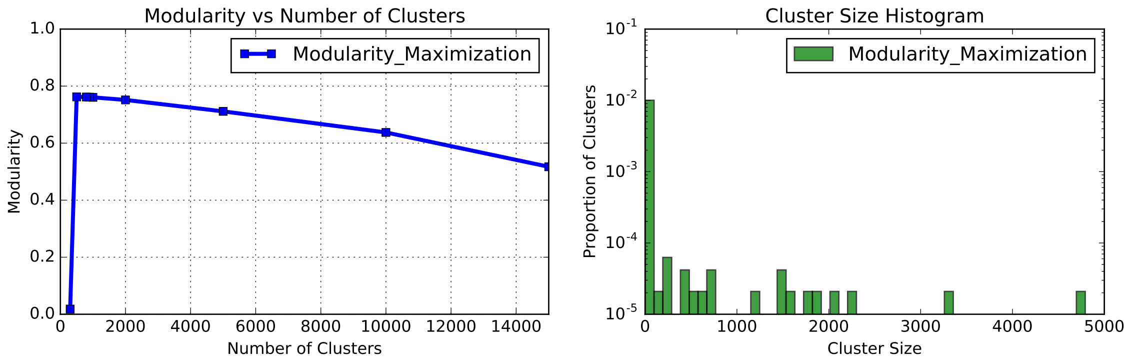
In this project, we will focus on all the articles published in Physical Review Letter, Review of Modern Physics, and Physical Review X from 2007 to 2015, and perform clustering algorithm on the citation network of these articles. To get this subset of citation network, we load all the article metadata and citation data into a mongodb database, and use the journal and date features in the metadata to filter the citation data. This results in a network of 27695 articles and 91346 citation links between them.

The clustering analysis and visualizations of the citation network and community structures can help the users to learn about the research history in physics and identify interesting fields. After identifying a subfield which is potentially interesting, the visualizations of the community networks can help the users to find out the top physicists that are working on the selected field, papers with the most impacts in the field, and see the trend of each field in physics.

4 Exploratory Analysis

In this project, we apply a hierarchical clustering algorithm based on modularity maximization to cluster the network of 27695 articles and 91346 citation links. The first step is to use pymongo to load the mongo database in python. Then we apply the clustering algorithm for the citation network data.

Modularity maximization is a bottom-up hierarchical clustering algorithm which repeatedly merges two clusters into one in order to optimize the modularity of the clustering. This algorithm works as follows: Starting from a state where each node is in its unique community, we repeatly join pairs of communities into one, until we reach the maximal modularity. I implement this algorithm from scratch using python and apply this clustering algorithm on the citation network. The modularity function peaks at $Q = 0.762$, when the number of cluster is 500, shown in Fig 1a. This value is rather high in real world network data, indicating a strong community structure in our citation network. As we perform analysis on the clustering result, one observation is that the size of clusters features a power-law distribution, where the largest cluster has nearly 5000 articles, while the majority of clusters are rather small with less than 1000 articles, shown in Fig 1b. In this project, we decide to focus on the large clusters with more than 1000 articles (there are 13 large clusters in total) since they are more representative as different fields in physics.



(a) Modularity function.

(b) Cluster size distribution.

Figure 1: The modularity function, and the cluster size distribution function in the simulation with the modularity maximization algorithm on the citation network.

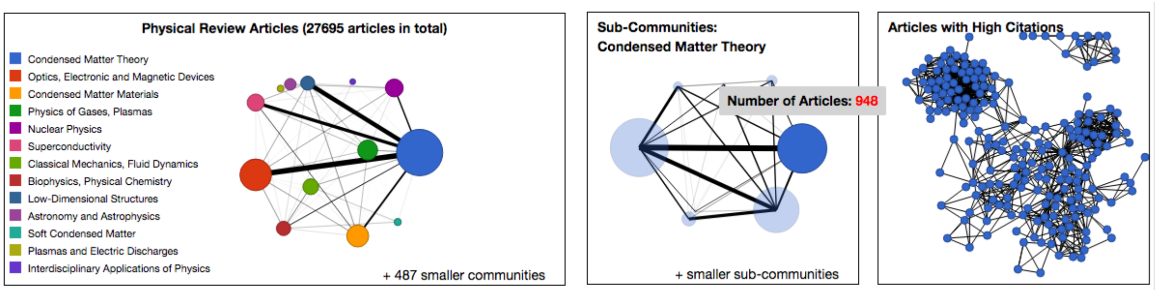
With the obtained clustering result, we can perform different analysis on the dataset. First, we query the database using pymongo library to obtain the papers with high citation numbers in each cluster. As we look at those papers with high impact, we summarize the topic in each cluster and use it as the cluster's name. Then, we extract the number of articles published over time for each cluster to study the trend in physics research. Another useful analysis is the authors with the most publications in each field of physics. From this analysis, we can identify the physicists with high impact in different fields. All the results from these analysis are dumped into json files for visualizations.

5 Implementations of Visualizations

The implementation of the visualizations are based on D3, NVD3, bootstrap, jqurey libraries. The landing page and layout design are based on start bootstrap templates. The visualizations on the webpage include two parts: the visualizations for articles, and physicists.

5.1 Articles

The network of articles is visualized using D3 force-layout bubble charts(Fig 2). The size of bubbles represents the number of articles in each community, while the width of links represents the number of citations between two communities. Clicking on each bubble in the first and second charts will trigger the chart to its right to show the corresponding structures inside the clicked cluster. In the right most chart with network of each individual articles, clicking on each node will open a new tab in your browser and go to the APS webpage of the selected article.


Physical Review Articles (27695 articles in total)
Condensed Matter Theory
Optics, Electronic and Magnetic Devices
Condensed Matter Materials
Physics of Gases, Plasmas
Nuclear Physics
Superconductivity
Classical Mechanics, Fluid Dynamics
Biophysics, Physical Chemistry
Low-Dimensional Structures
Astronomy and Astrophysics
Soft Condensed Matter
Plasmas and Electric Discharges
Interdisciplinary Applications of Physics
+ 487 smaller communities
Sub-Communities:
Condensed Matter Theory
Number of Articles: 948
+ smaller sub-communities
Articles with High Citations

(a)

(b)

Figure 2: Visualizations of the network of articles.

The evolution of communities over time is visualized using stacked area charts implemented with NVD3 library(Fig 3). The visualizations support selections, probing, and changing layout. Select one area in the left chart will trigger the right chart to show the corresponding sub-fields in the select field. Selecting in either the left most cluster bubble chart or the left stacked area chart will simultaneously trigger all the rest charts to respond.

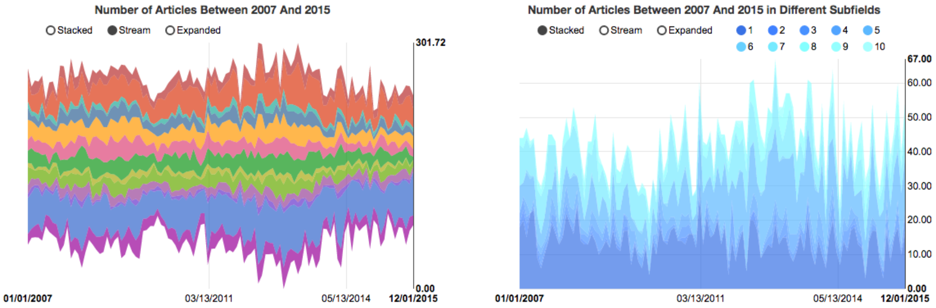


Figure 3: Visualizations of the evolution of different fields in time.

5.2 Physicists

The physicists in different fields is visualized using D3 force-layout bubble charts(Fig 4). The left chart is the same with the first chart in the network of articles. The right one shows the physicists from different fields, represented by bubbles with color denoting the field, and size denoting the number of publications. The typical number of publications in different fields vary dramatically, therefore I normalize the size according to the largest node in each cluster. One interesting observation is that the authors in nuclear physics and astrophysics tend to publish much more papers than other fields. Clicking on a bubble in the left chart will highlight the selected field in the right chart. Clicking on each individual node in the right bubble chart will open a new tab in your browser and start a Google search for the selected physicist.

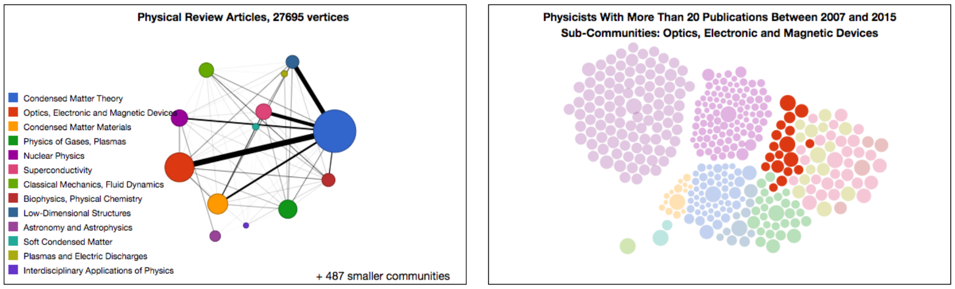


Figure 4: Visualizations of the physicists in different fields.

6 Links

The link to the webpage is `http://www-edlab.cs.umass.edu/~yuanh/590v/project/pages/index.html`.

The whole dataset can be requested at `https://journals.aps.org/datasets`.

7 Acknowledgements

I want to thank the American Physical Society for offering the dataset for this project. Image credits to Martin Driver, deviantart.net., and society6.com. The webpage template is adapted from Start Bootstrap.

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] E. A. Leicht, and M. E. J. Newman. *Community structure in directed networks*, Physical Review Letters, (2008) **100**, 118703.

[6] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, E. (2008). *Fast unfolding of communites in large networks*, Journal of Statistical Mechanics: Theory and Experiment, P1008, 17425468.

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

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

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

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