Project Proposal

Revisit The Structure of Scientific Collaboration Networks Using PubMed Data

Min Chen
Indiana University

6/15/2017

Abstract

In this project, the author plans to implement the methods and techniques established by Newman(2000, 2011) and replicate the results of the paper using self-downloaded PubMed data. Further, the author plans to extend Newman's paper by evaluate the evolution of the collaboration network as well as consider the potential differences of sub-networks formed by authors of articles related to magraine or some other diseases. I expect to find that these networks establish small world property, high centrality, with a power-law distribution [1] of the degree.

[1] Maybe with a cut-off as discussed by Newman(2000)

Contents

1 Introduction

The scientific collaboration networks has a long history starting from the concept of Erdos number. Inspired by a series of paper published in early 2000 by M.E.J Newman (2000, 2001), who talks extensively about the scientific collaboration networks, the author of this proposal decides to apply the method and techniques established by Newman to a new set of data from PubMed to get a better understanding of this network.

In Newman's papers, he included three sets of data: MEDLINE (biomedical research), the Los Alamos e-Print Archive (physics) and NCSTRL (computer science), (from 1995 to 1999) and discussed several key concepts, namely small world property, degree distribution, clustering and centrality. Some of his results in his 2000 paper are listed below:

1. All 3 scientific collaboration networks establish a small-world property and the degree of separation is about five or six.
2. The network is highly clustered however, the MEDLINE data has a much lower value than the other two "hard science" sources
3. The degree distributions follows a power-law form with an exponential cutoff.

This project is a combination of data-driven project and replication project in nature. The author obtained a subset of data from PubMed containing roughly 25,000 paper regarding the disease migraine and plan to apply the methods and techniques provided in Newman's paper to calculate the correpsonding statistics and metrics in this particular collaboration network and evaluate the results.

There are several aspect of the data that make this project interesting. First of all, PubMed data is in nature similar to MEDLINE data, therefore, the author could have a chance to verify if Newman's result on MEDLINE data can be generalized to other biomedical database. Secondly, the PubMed data has a time span from 1946 to 2016, compared to Newman's 5 year window. The author will first obtain the results at the entire time span and then cut it into 5-year windows to study the evolution of the network. Thirdly, the author is also scraping other area such as cancer on PubMed. Instead of looking at different discipline like physics, computer science, I am exploring within the domain of biomedical but in different sub areas. If enough additional data could be obtained, the author plan to evaluate the network structure in and between these sub areas.

Finally, Newman's paper covers almost all the key concepts in the network science course and replication of this paper using a self generated dataset could be a good practice to help understanding these concepts. Some extensions mentioned in the previous paragraph will provide an opportunity to learn network science in a more efficient and self-motivated way.

2 Literature review

There is a substantial literature on collaboration network and Newman's papers set the foundation in this area. Besides the key results listed in the previous paragraphs, he also pointed out a measure of strength of collaborative ties which takes account of the number of papers a given pair of scientist have written together as well as the number of other coauthors with whom they wrote these papers. This idea turns the network edges into a weighted one. (Newman, 2001). Some recent researches focus on the evolution of collaboration networks. For example, Prosperi et al. (2016) identified relevant country-specific kinship trends over time and found that authors who are part of a kin tend to occupy central positions in their collaborative networks. Cocciaa and Wang (2016) focus on the convergence of international collaboration patterns between the applied and basic sciences, which they claimed to be one of contributing factors that supports the evolution of modern scientific fields. Petersen (2015) quantified the impact of weak, strong, and super ties in the scientific collaboration network and found that super ties contribute to above-average productivity and a 17% citation increase per publication, thus being a major factor in science career development.There are other researchers focus on the collaboration network in specific fields. Seltzer and Hamermesh (2017) focus on the collaboration in economic history compared to economics in general and concluded co-authorships in economic history are more likely to be formed of individuals of different seniority as compared to economics generally. Budner and Grahl (2016) explored the large-scale social structure of the music industry and Andrikopoulos and Kostaris (2016) analyzed the small-world property in collaboration networks in accounting research.

3 Methods and Plan

3.1 Data

The current data set that has been scraped from PubMed is the all paper regarding the disease migraine. There are 24853 records (papers) with a total number of 55764

authors. This dataset has all migraine related research paper on PubMed from 12/1/1946 to 8/23/2016.

The author is also trying to scrap some other potential datasets from PubMed. There are two potential directions:

1. Get the data related to some other diseases, like cancer and headache, etc
2. Get the data from PubMed in a limited time span, for example 2011/01/01 to 2016/01/01 regardless of the topics.

Using the first potential datasets, the different sub-network of the whole network can be evaluated. We may get an idea if the structures of the sub networks of collaborations are different from one and each other.

Using the second potential dataset, we could see if the structures of the migraine literature authors is substantially different from the network of collaboration on PubMed as a whole.

A final point is that the author plans to extend Newman's paper by looking at the evolution of the network if the data has a long time span. The migraine data has this property, so will the first sets of potential data. For the whole topic dataset, due to the size, it is hard to get data within a extended time period.

3.2 Methods

The author plans to apply the methods presented by Newman(2000) to calculate the main statistics and structure of the collaboration network established by the datasets discussed in previous section. These key statistics and structure concepts include: average degree, degree distribution[2], average shortest path length, centrality etc.

Besides, the author plans to extend Newman's paper in the following ways: Exploring the evolution of the network across time; checking if the structures are different for different sub-fields including the whole network regardless the topic. Finally, according to Newman(2011), the author will try to implement the network in a weighted undirected graph and explore the properties.

[2]whether it follows a power law with cut-off

References

[1] Andreas Andrikopoulos and Konstantinos Kostaris, *Collaboration networks in accounting research*. Journal of International Accounting, Auditing and Taxation, Volume 28, 2017, Pages 1-9, ISSN 1061-9518, https://doi.org/10.1016/j.intaccaudtax.2016.12.001.

[2] Budner, Pascal and Jrn Grahl, *Collaboration Networks in the Music Industry*. CoRR abs/1611.00377 (2016): n. pag.

[3] Mario Coccia and Lili Wang, *Evolution and convergence of the patterns of international scientific collaboration*. PNAS 2016 113 (8) 2057-2061; published ahead of print February 1, 2016, doi:10.1073/pnas.1510820113

[4] M. E. J. Newman, *The structure of scientific collaboration networks*. PNAS 2000 98 (2) 404-409; doi:10.1073/pnas.98.2.404

[5] M. E. J. Newman, *Scientific collaboration networks. I. Network construction and fundamental results.*. Phys. Rev. E 64 , no. 1 (2001): 016131.

[6] M. E. J. Newman, *Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality.*. Phys. Rev. E 64 , no. 1 (2001): 016132.

[7] Mattia Prosperi, Iain Buchan, Iuri Fanti, Sandro Meloni, Pietro Palladino, and Vetle I. Torvik, *Kin of coauthorship in five decades of health science literature*. PNAS 2016 113 (32) 8957-8962; published ahead of print July 25, 2016, doi:10.1073/pnas.1517745113

[8] Alexander Michael Petersen, *Quantifying the impact of weak, strong, and super ties in scientific careers*. PNAS 2015 112 (34) E4671-E4680; published ahead of print August 10, 2015, doi:10.1073/pnas.1501444112

[9] Seltzer, Andrew and Hamermesh, Daniel S., *Co-Authorship in Economic History and Economics: Are We Any Different?*. (May 2017). NBER Working Paper No. w23404. Available at SSRN: https://ssrn.com/abstract=2968242