SU17: NETWORK SCIENCE: 13951

Revisit The Structure of Scientific Collaboration Networks Using PubMed Data

Min Chen

Indiana University

June 2017

Overview

Introduction

Abstract

- Plans to implement the methods and techniques established by Newman(2000, 2011)
- Replicate the results of the paper using self-downloaded PubMed data
- Extension1: Evaluate the evolution of the collaboration network
- Extension2: Consider the potential differences of sub-networks formed by authors of articles related to magraine or some other diseases.

Foundation Papers: Newman(2000, 2001)

- Datasets
  1. MEDLINE (biomedical research)
  2. The Los Alamos e-Print Archive (physics)
  3. NCSTRL (computer science)
- Dataset span: from 1995 to 1999
- Key concepts discussed: small world property, degree distribution, clustering and centrality

Key results of Newman(2000, 2001)

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

Proposed Project

- Nature: combination of data-driven project and replication project
- PubMed Data obtained:
  - 25,000 paper regarding the disease migraine (more are coming)
  - similar to MEDLINE data in nature, part of Newman's results can be tested
  - time span from 1946 to 2016, possible to study the evolution of the network
  - trying to download data from other diseases, such as cancer, will compare network in different sub areas of biomedical field

Proposed Project (cont'd)

- Goal: 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 will provide an opportunity to learn network science in a more efficient and self-motivated way.

Literature review

Related Studies and Main Results

- Measure of strength of collaborative ties, which turns the network into a weighted one (Newman, 2001)
- 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 (Prosperi et al. 2016)
- Convergence of international collaboration patterns between the applied and basic sciences (Cocciaa and Wang 2016)
- Super ties contribute to above-average productivity and a 17% citation increase per publication, thus being a major factor in science career development. (Petersen 2015)

Collaboration Network in Specific Fields

- Co-authorships in economic history are more likely to be formed of individuals of different seniority as compared to economics generally (Seltzer and Hamermesh 2017)
- Large-scale social structure of the music industry (Budner and Grahl 2016)
- Small-world property in collaboration networks in accounting research (Andrikopoulos and Kostaris 2016)

Methods and Plan

Data: Obtained

1. PubMed data regarding disease migraine.
   - 24853 records (papers)
   - 55764 authors
   - Scope: all migraine related research paper on PubMed from 12/1/1946 to 8/23/2016
2. PubMed data regarding cancer (Ongoing downloading)
   - 150756 records (papers) so far
   - Scope: all cancer related research paper on PubMed from 9/1/2016 to 06/2017
   - Progress: download month by month, hope to get to all data back to 1946

Data: Potential

- Get the data related to some other diseases, like diabetes, hypertension, headache, etc
- Get the data from PubMed in a limited time span, for example 2011/01/01 to 2016/01/01 regardless of the topics. (This may require a lot more time, perhaps cannot be done before the summer term ends)

Methods: Replication Part

- Apply the methods presented by Newman(2000) to calculate the main statistics and structure of the collaboration network
- Key statistics and structure concepts include: average degree, degree distribution[1], average shortest path length, centrality etc.

[1] Whether it follows a power law with cut-off

Methods: Extension Part

- Explore the evolution of the network across time
- Check if the structures are different for different sub-fields including the whole network regardless the topic.
- Implement the network in a weighted undirected graph and explore the properties. [2]

[2] According to Newman(2011)

References


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.


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


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


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


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


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


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


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


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