Social Network Analysis and Big Scholarly Data

Yogendra Singh

University Librarian

Swami Rama Himalayan University

Email - yogi5240@gmail.com

After this talk

- You should have a basic knowledge of Social Network Analysis
- You have an basic understanding of Big Scholarly Data
- You have an understanding of how Social Network Analysis can be applied to Big Scholarly Data

Social Network

- Society is made of individuals (like wife and husband). They are known as nodes/actors/vertices in Social Network parlance
- Individuals have relations
- Information flows between these relations
- These relations are known as ties/links/images in Social Networking
- Social network is a network (group) of individuals which have certain type of relations for a particular information flow

Look at this picture

Now the big question is

- Who matters among this crowd? There could be different answers depending upon different point of views
- The analysis of this type of social network using graph theory is called Social Network Analysis
- Since Scholarly networks are the network of people (co-authors), it can well be applied to large scholarly data or Big Scholarly Data

Big Scholarly Data (BSD)

- BSD refers to millions of scholarly records available today due to tremendous changes in scholarly communication cycle

- BSD may include

  - E-books, articles, reports, standards, patents etc., published by major commercial and not for profit organizations - sciencedirect.com, tandfonline.com, doaj.org etc.
  - Abstracting and Indexing databases- Scopus, Web of Science, EBSCO, Google Scholar
  - Academic social networks- Academia, ResearchGate, Mendeley etc.
  - Many other type of scholarly data

Three major scholarly data providers

Sl. No.	Brand-name	Publisher	Coverage	No. of Records
1.	Google Scholar	Google	Full Universe of Knowledge/ All Formats	350+ million
2.	Web of Science	Clarivate Analytics	Bibliographic Information including citations and other details including abstract	90+ million
3.	Scopus	Elsevier Science	Bibliographic Information including citations and other details including abstract	75+ million

Big data analysis methods

- Statistical analysis
  - Suitable for smaller datasets
- Scholarly text mining
  - Can be used with big data
- Scholarly Network Analysis (or Social Network Analysis)

Scholarly Network Analysis/ Social Network Analysis: Important measures include Centralities

- Average path length
- Clustering coefficient
- Centralities

Average Path Length

- **Average path length**: Average distance of any two nodes in a network is known as Average path length

Clustering Coefficient

- **Average path length**: Average distance of any two nodes in a network is known as Average path length

- **Clustering coefficient:** is a measure of the degree to which nodes in a graph tend to **cluster** together

Centrality Measures

- **Average path length:** Average distance of any two nodes in a network is known as Average path length
- **Clustering coefficient:** is a measure of the degree to which nodes in a graph tend to **cluster** together
- **Centrality Measures:** They measure how central (important) a node is in a network

Network Centrality

- Which individuals (nodes) are important (Central)
- Measurement of importance is called Centrality in SNA
- Centrality may mean differently for different people and in different context

Why are Centrality and Centralization Important?

- Access to information and ideas
- Interaction among members of the network
- Control the flow of information, resources, and other network content
- Visibility
- Ability to act together collectively

Multiple Ways to Calculate Centrality

- Degree
- Closeness
- Betweenness
- Eigenvector

Calculating Centrality

- **Degree** – Proportional to the number of other nodes to which a node is links – Number of links divided by (n-1).

Calculating Centrality

- **Degree** – Proportional to the number of other nodes to which a node is links – Number of links divided by (n-1).

- **Closeness** – The sum of geodesic distances (shortest paths) to all other points in the graph. Divide by (n-1), then invert.

Calculating Centrality

- **Degree** – Proportional to the number of other nodes to which a node is linked – Number of links divided by (n-1).

- **Closeness** – The sum of geodesic distances (shortest paths) to all other points in the graph. Divide by (n-1), then invert.

- **Betweenness** – The extent to which a particular point lies ‘between’ other points in the graph; how many shortest paths (geodesics) is it on? A measure of brokerage or gatekeeping.

Calculating Centrality

- **Degree** – Proportional to the number of other nodes to which a node is links – Number of links divided by (n-1).

- **Closeness** – The sum of geodesic distances (shortest paths) to all other points in the graph. Divide by (n-1), then invert.

- **Betweenness** – The extent to which a particular point lies ‘between’ other points in the graph; how many shortest paths (geodesics) is it on? A measure of brokerage or gatekeeping.

- **Eigenvector**– A weighted measure of centrality that takes into account the centrality of other nodes to which a node is connected. That is, being connect with other central nodes increases centrality. E.g., secretary of powerful person. Google’s page rank algorithm is based on a variation of this approach.

Network Analysis Tools Applied to BSD

Software/ Access	Platform/ Language	Description
CiteSpace/ Free	Windows, IOS/ Java	Visualizing and analyzing trends and patterns in scientific literature; knowledge domain visualization, best for WoS datasets
Gephi/ Free	Windows/Linux/IOS Java	Exploratory Data Analysis; Social Network Analysis; Link Analysis
iGraph/ Free	Windows/IOS C/R/Python/Perl	A collection of network analysis tools with the emphasis on efficiency, portability and ease of use
NetworkX / Free	Windows/IOS Python	Creation, manipulation, and investigation of the structures, dynamics, and functions of complex networks
Pajek/ Free	Windows/IOS C/R	Analysis and visualization of large networks having some thousands or even millions of vertices

Types of Scholarly Networks could be Generated by Applying SNA to BSD

- Co-Author Network
  - Personal Network
  - Organizational Network
  - Geographic Network
- Co-Word Network
- Co-Citation Network

BSD Analysis Applications

- Scientific Impact Evaluation
  - Article Impact
  - Author Impact
  - Journal Impact
  - Institutional Impact

BSD Analysis Application - Academic Recommendations

- Literature Recommendations
- Expert Recommendations
- Collaboration Recommendations
- Priority Recommendations

Scholarly Data Analysis: Steps

- Data Collection
  - Download desired dataset from appropriate source
- Data cleaning
  - Most difficult task as same name, institute, department is represented in different ways even by the same individual
- Create graph using the data
- Use graph for further processing

Source and Software Needed

- An appropriate data source to download desired datasets
  - I used Scopus to download research data of IIT Roorkee
- A software tool to clean data
  - I used OpenRefine an open source software to clean the data, however, quite a bit was done manually
- A software tool to create the graph
  - An online tool Table2Net was used
- Process the graph for further obtaining necessary measures
  - Gephi was used for this purpose

Steps in Analysing BSD through SNA

- Download a dataset
- Clean the data by some cleaning software such as OpenRefine and Manually
- Create Graph File through some scientific network creating online tool such as Table2Net or Scopus2Net
- Analyse that Graph file in Network Analysis software such as Gephi. You can calculate all SNA measures using Gephi

Conclusion

- Application of Social Networking Tools to Big Scholarly Data is going to be big area of interest to scientometricians as very large BSD is generated daily.
- These measures can be used to evaluate the authors, institutions, subject areas or countries objectively.
- Special areas of interests, possible collaboration opportunities can be easily identified.
- As the impact of the publications can be easily identified, it will have great impact in policy making.
- Librarians can also use SNA for analyzing in-house generated data such as circulation, reference data, even footfall data.

THANK YOU