Bibliometrics for Librarians Session 3

Presented by Phillip Doehle and Clarke Iakovakis
Oklahoma State University

Course Home: https://pdoehle.github.io/bibliometrics-for-librarians/
Details: http://www.ala.org/lita/bibliometrics-librarians

Course Outline

- 7/9: Introduction to bibliometrics & citation analysis
- 7/15: Overview of citation metrics & data sources
- 7/22: Network analysis
- 7/30: Altmetrics, summing up, and next steps

Bibliometrics Lesson 1

If you cloned the Azure Notebook early on, delete **library(pubmedR)** from this code chunk:

In [ ]:

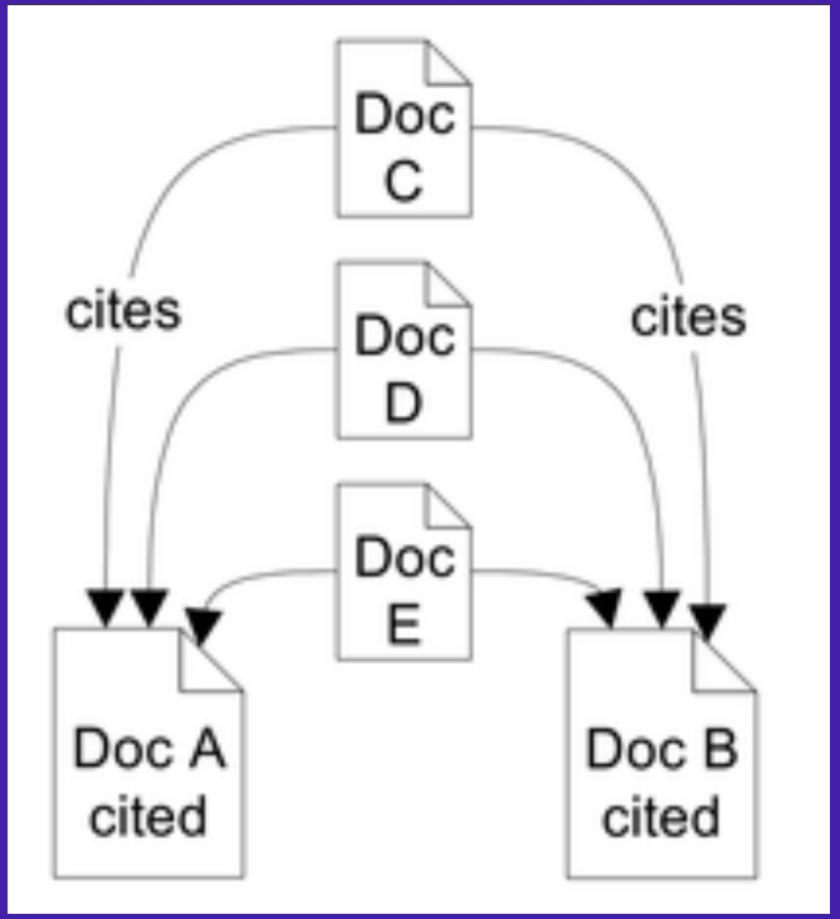
```
# load packages
library(rcrossref)
library(citecorp)
library(dplyr)
library(purrr)
library(stringr)
library(tidyr)
library(readr)
library(ggplot2)

# increase number of columns and rows displayed when we print a table
options(repr.matrix.max.cols=100, repr.matrix.max.rows=20)
```

Outline for today

- Bibliographic Network Analysis. Concepts
  - Co-citation networks
  - Bibliographic coupling
  - Direct citations
  - Co-authorship
  - Term co-occurrence
- Bibliographic Network Analysis Tools
- Network Analysis
  - Networks/Graphs
  - Centrality
  - Connectedness
  - Clustering

Co-citation networks

- Two publications are co-cited if there is a third publication that cites both publications
- The larger the number of publications by which two publications are co-cited, the stronger the co-citation relation between the two publications
- This has been the dominant metric over the last few decades, until recently
- Co-citation networks can also be constructed for **journals** and **authors**

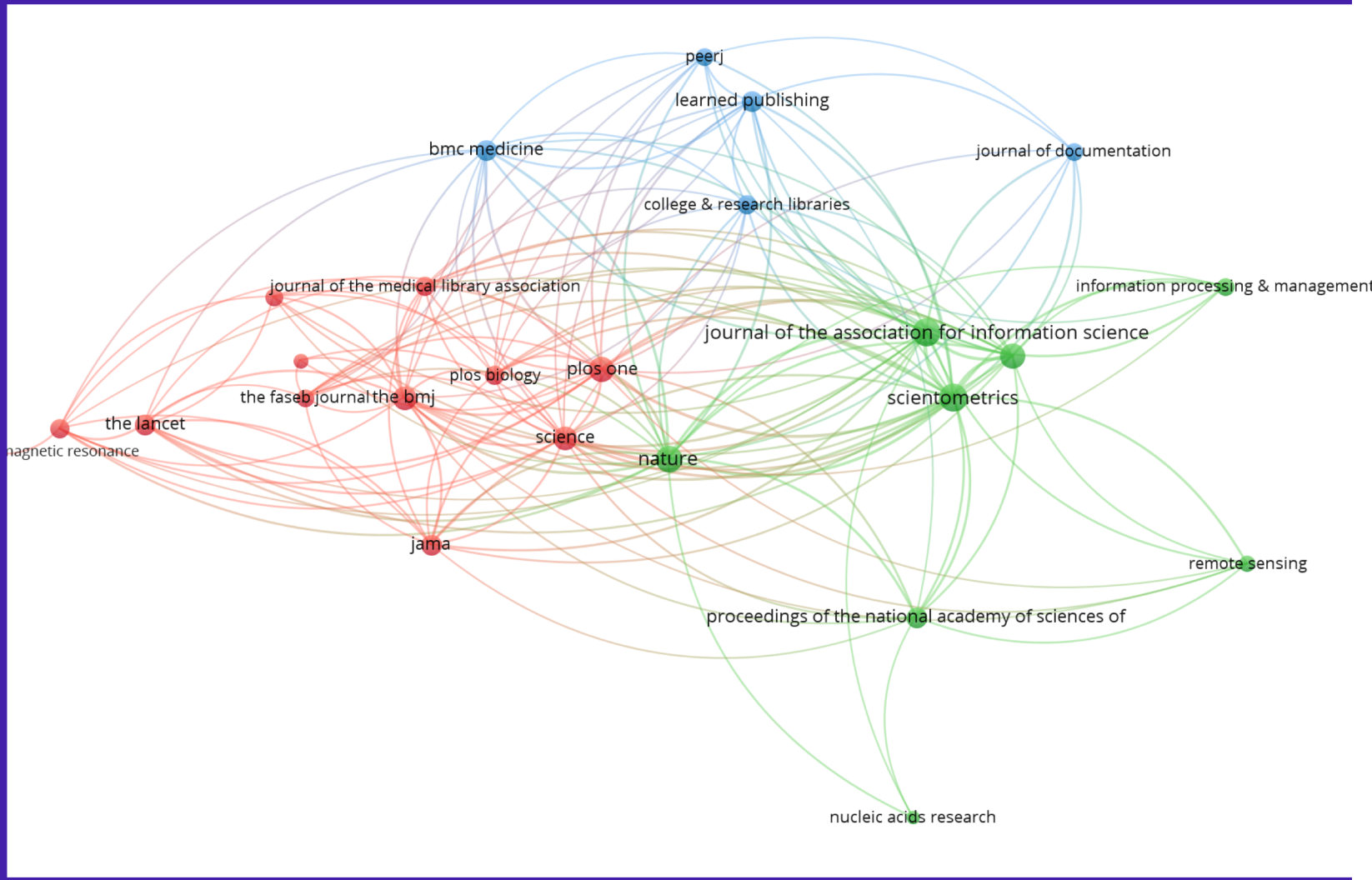


"Visualization_of_co-citation_analysis_methods" by Bela Gipp and Joeran Beel on Wikipedia at https://en.wikipedia.org/wiki/File:Visualization_of_co-citation_analysis_methods.png.

Nees Jan van Eck and Ludo Waltman, "Visualizing Bibliometric Networks," in *Measuring Scholarly Impact: Methods and Practice*, ed. Ying Ding, Ronald Rousseau, and Dietmar Wolfram (Cham: Springer International Publishing, 2014), 285–320, http://vosviewer.com/download/f-x2.pdf.
White, H.D., & McCain, K.W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. Journal of the American Society for Information Science, 49(4), 327–355.

Journal Co-citation network

- Each circle is a journal
- Circle size = number of citations the journal has received
- Journals located close to each other indicate stronger relationship (based on co-citations) than journals far away from each other

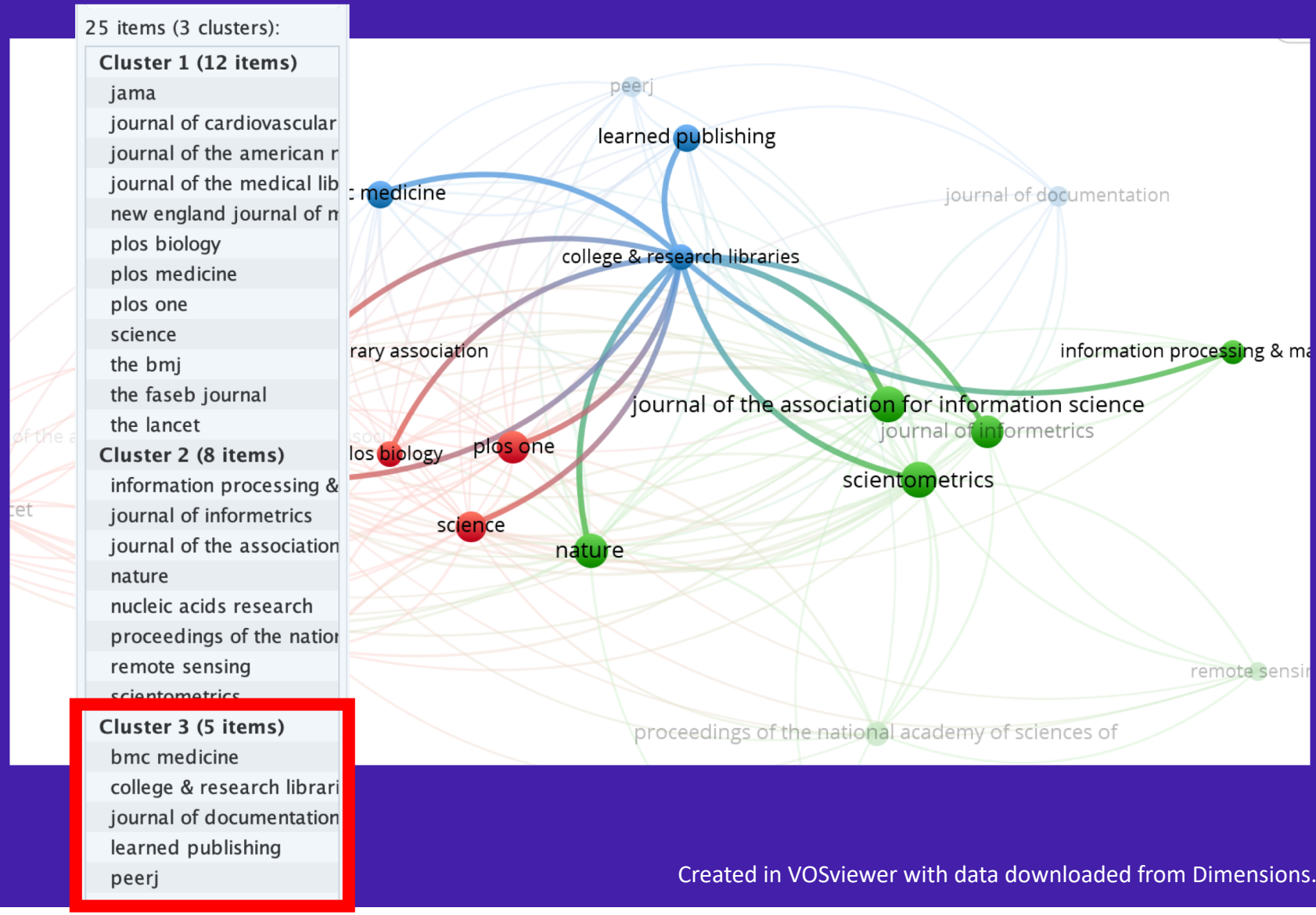


Nees Jan van Eck and Ludo Waltman, "Visualizing Bibliometric Networks," in *Measuring Scholarly Impact: Methods and Practice*, ed. Ying Ding, Ronald Rousseau, and Dietmar Wolfram (Cham: Springer International Publishing, 2014), 285–320, http://vosviewer.com/download/f-x2.pdf.

Created in VOSviewer with data downloaded from Dimensions. Includes Journals with at least 30 references to the journal

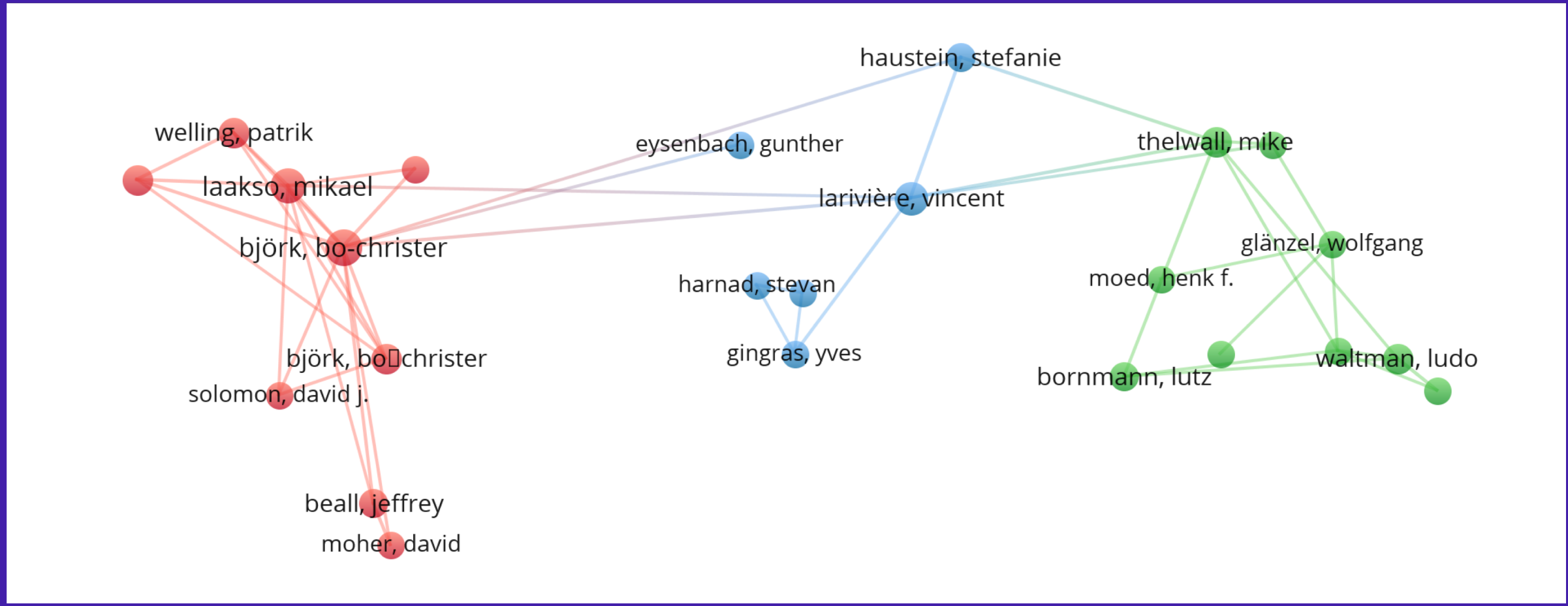
Journal Co-citation network

- Thus in this case, *C&RL* has been co-cited with 12 other journals
- It has also been **clustered** with four other journals based on connectedness



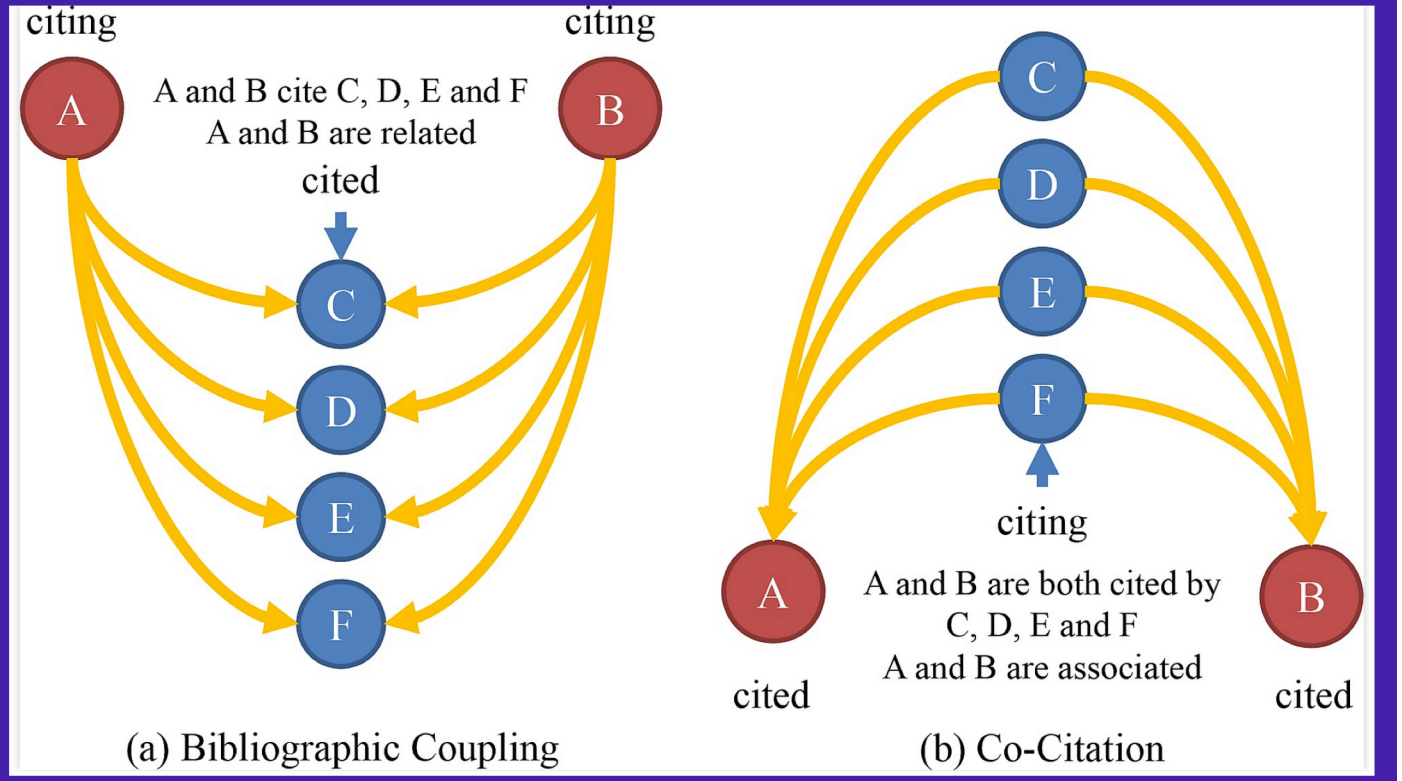
Created in VOSviewer with data downloaded from Dimensions.

Author co-citation network



Bibliographic coupling

- Two publications are bibliographically coupled if there is a third publication that is **cited by** both publications
- The opposite of co-citation:
  - "Bibliographic coupling is retrospective whereas co-citation is essentially a forward-looking perspective."



Nees Jan van Eck and Ludo Waltman, "Visualizing Bibliometric Networks," in *Measuring Scholarly Impact: Methods and Practice*, ed. Ying Ding, Ronald Rousseau, and Dietmar Wolfram (Cham: Springer International Publishing, 2014), 285–320, http://vosviewer.com/download/f-x2.pdf.

Eugene Garfield, "From Bibliographic Coupling to Co-Citation Analysis via Algorithmic Historio-Bibliography: A Citationist's Tribute to Belver C. Griffith," 2001. https://garfield.library.upenn.edu/papers/drexelbelvergriffith92001.pdf

Jarneving, J. (2007). Bibliographic coupling and its application to research-front and other core documents. Journal of Informetrics, 1(4), 287–307.

Figure: Xing, Lizhi (2017): Bibliographic coupling and co-citation.. PLOS ONE. Figure. https://doi.org/10.1371/journal.pone.0184055.g001

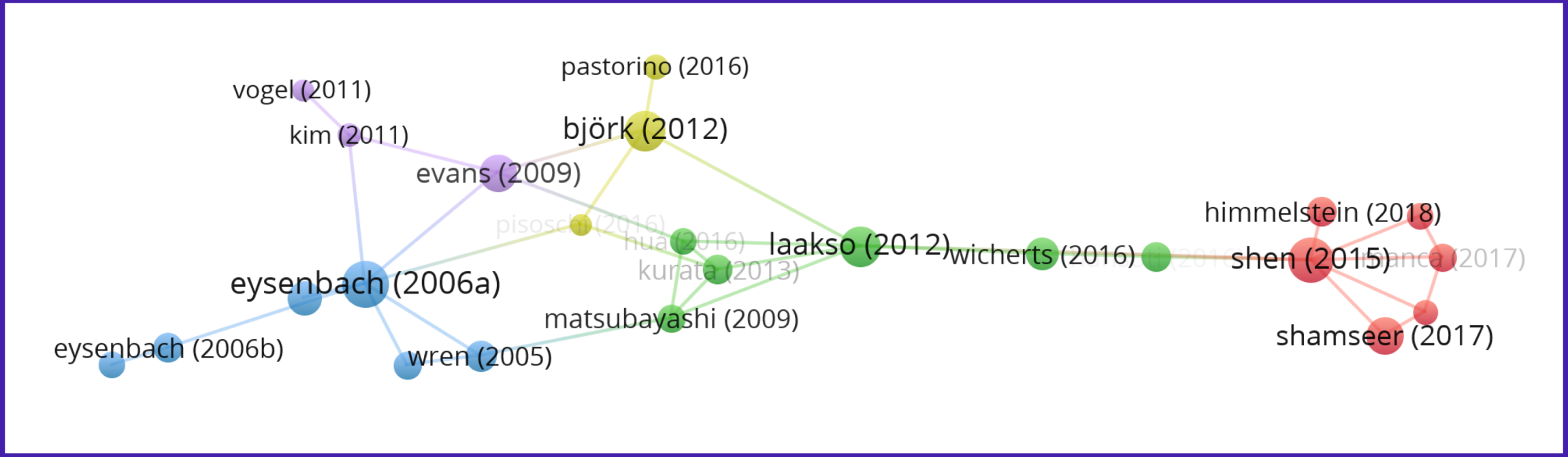
Direct citations (also called cross citations)

- Relationship is based on the number of times the publications within the set cite each other
- Citation links are treated as **undirected** by VOSviewer. Hence, no distinction is made between a citation from item A to item B and a citation in the opposite direction.
- Can also be computed for journals, authors, organizations, or countries

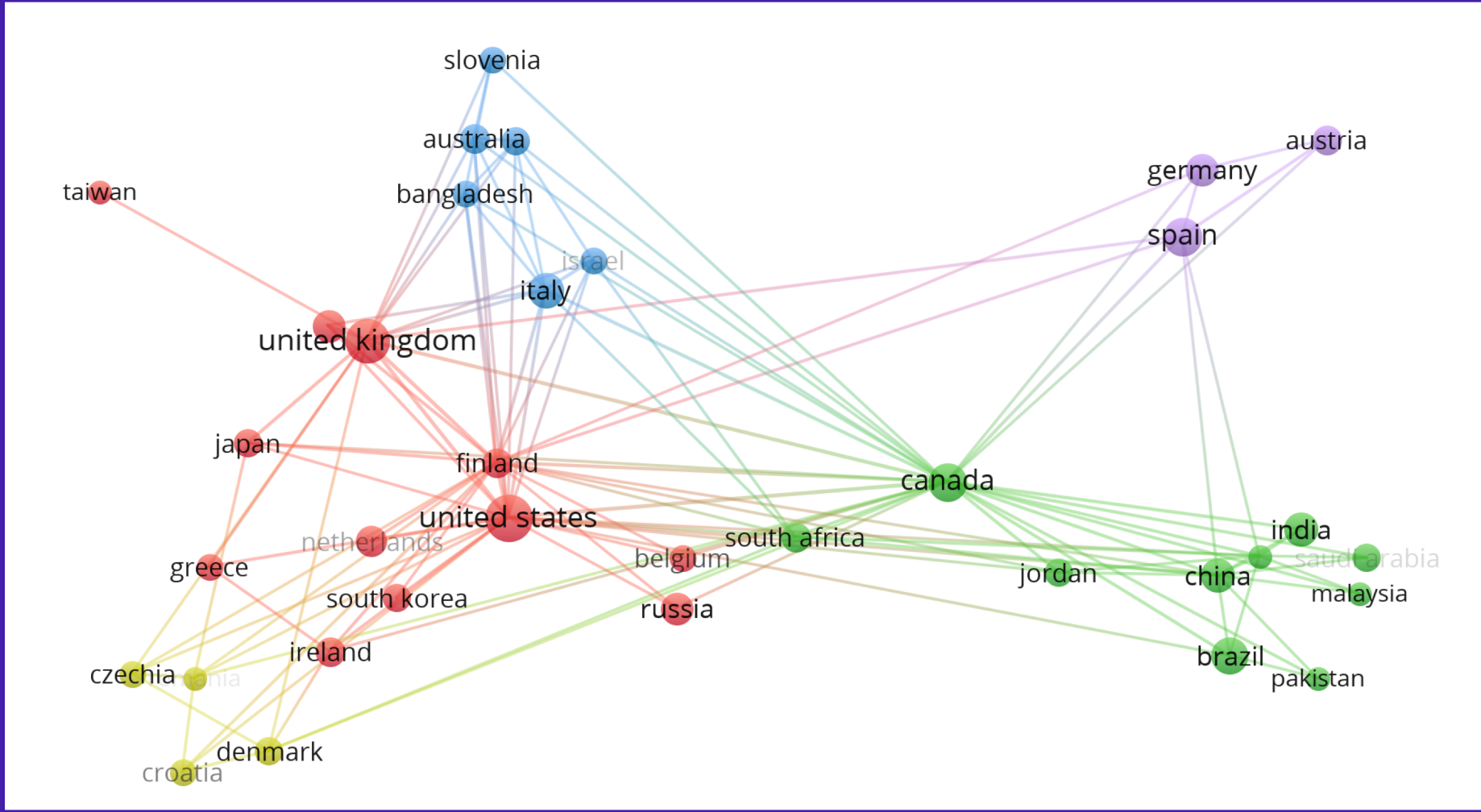
Nees Jan van Eck and Ludo Waltman, “Visualizing Bibliometric Networks,” in *Measuring Scholarly Impact: Methods and Practice*, ed. Ying Ding, Ronald Rousseau, and Dietmar Wolfram (Cham: Springer International Publishing, 2014), 285–320, http://vosviewer.com/download/f-x2.pdf.

Direct citation network

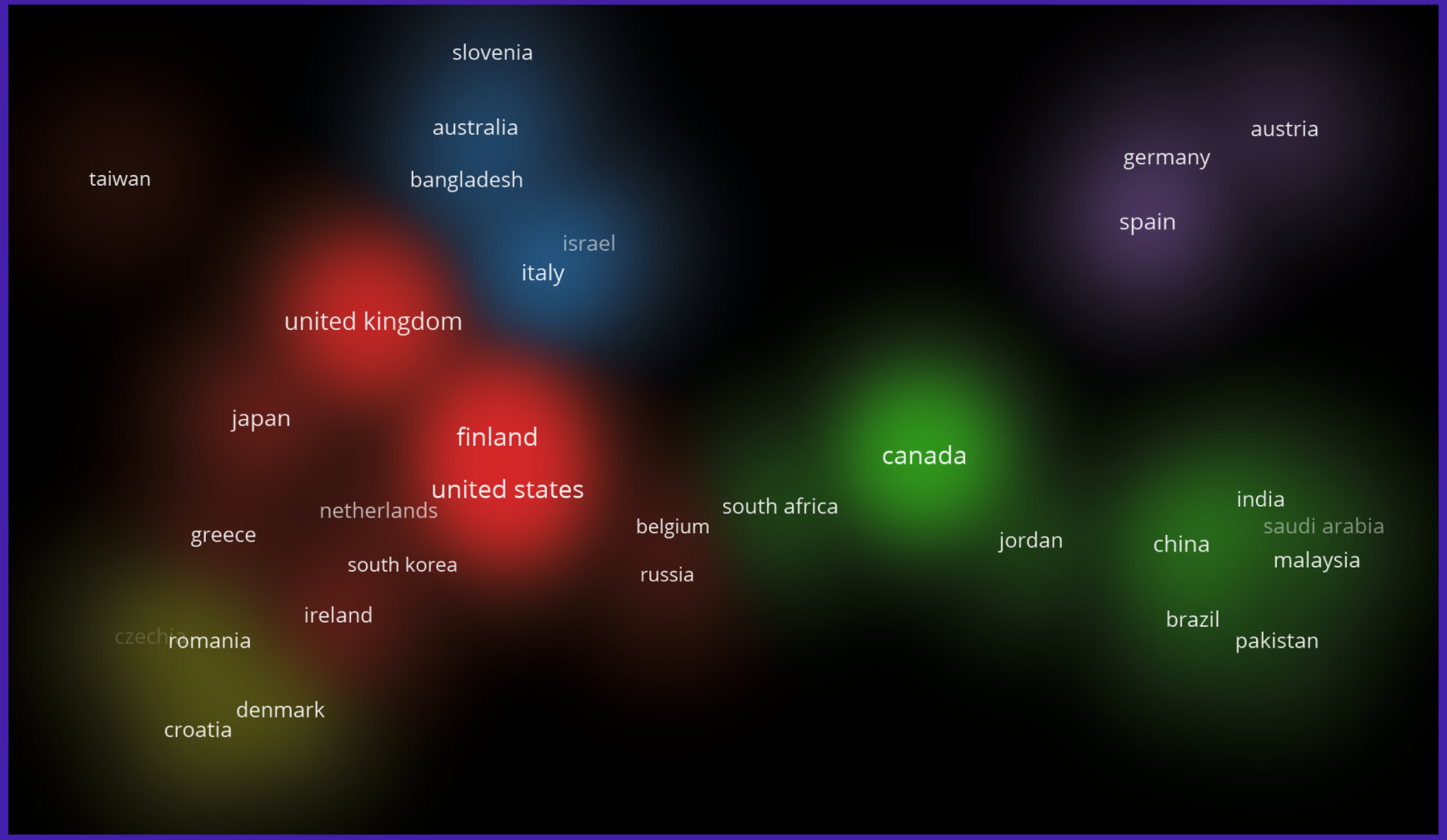
- Circle size = number of citations
- Proximity = stronger relationship (based on citation links)



Direct citation network – nations



Direct citation network (density visualization)

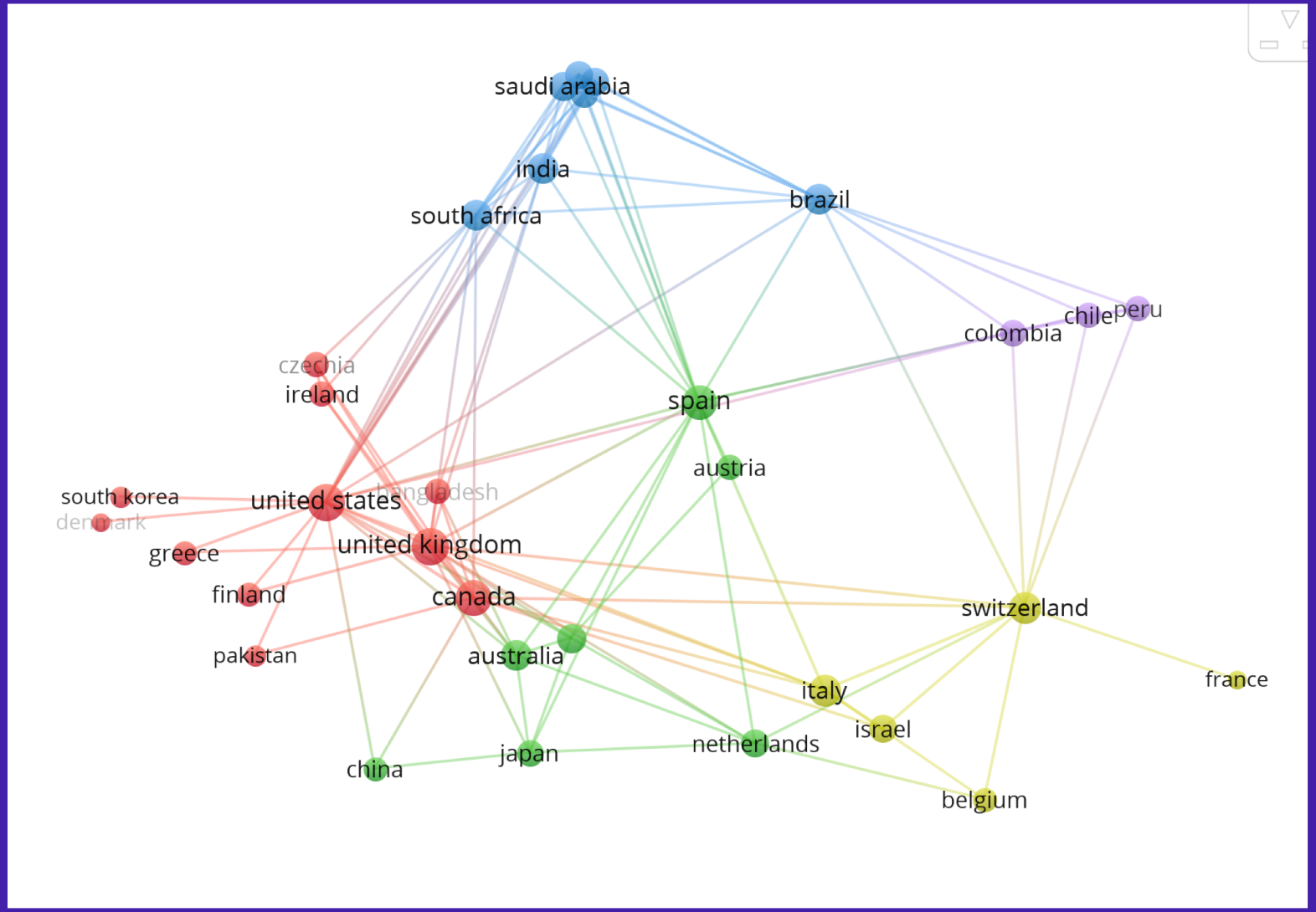


Co-authorship network

- Researchers, research institutions, or countries are linked to each other based on the number of publications they have authored jointly

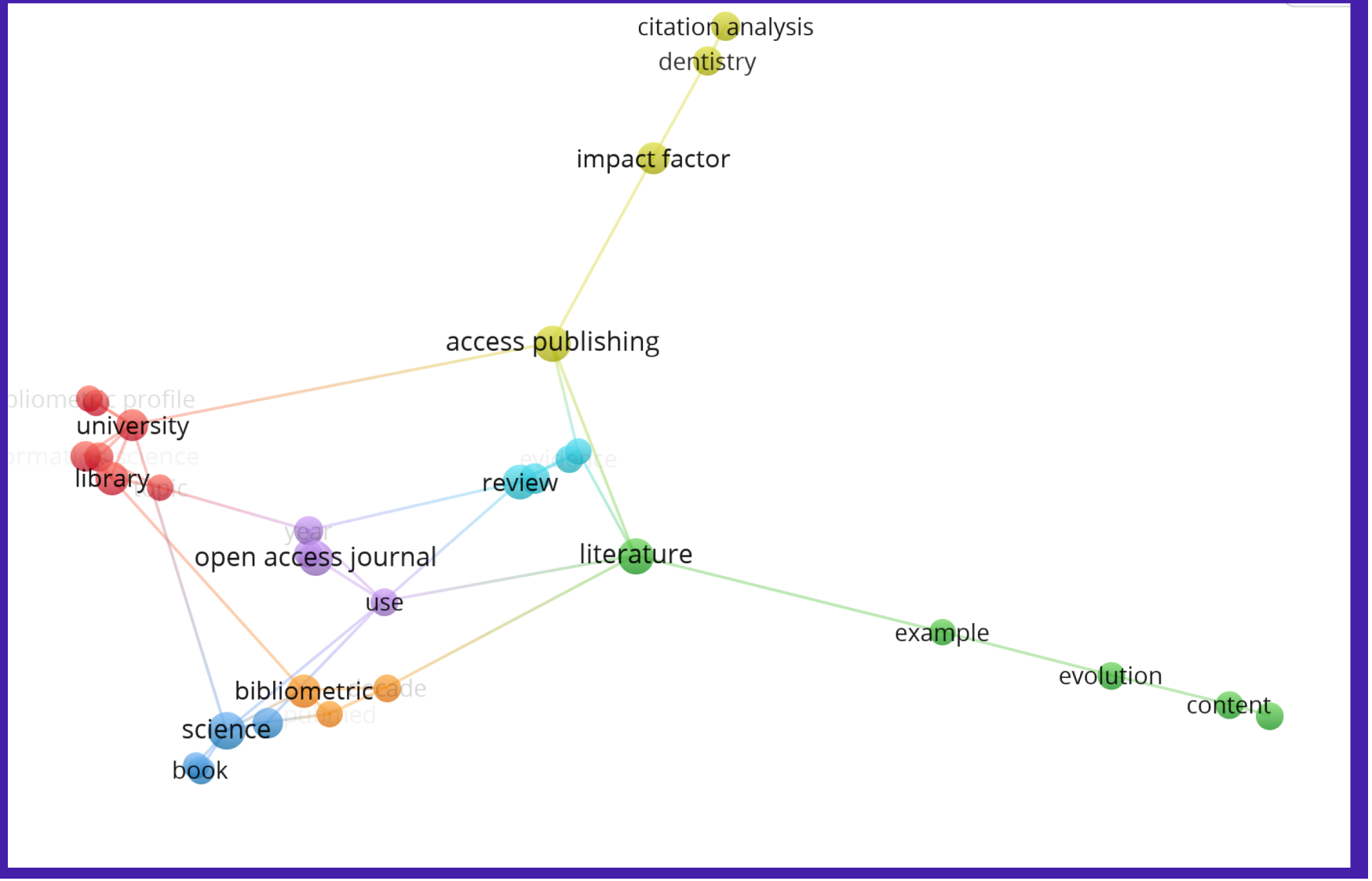
Nees Jan van Eck and Ludo Waltman, “Visualizing Bibliometric Networks,” in *Measuring Scholarly Impact: Methods and Practice*, ed. Ying Ding, Ronald Rousseau, and Dietmar Wolfram (Cham: Springer International Publishing, 2014), 285–320, http://vosviewer.com/download/f-x2.pdf.

Co-authorship network – nations



Term co-occurence

- Can be extracted from title, abstract, subject headings



VOSviewer

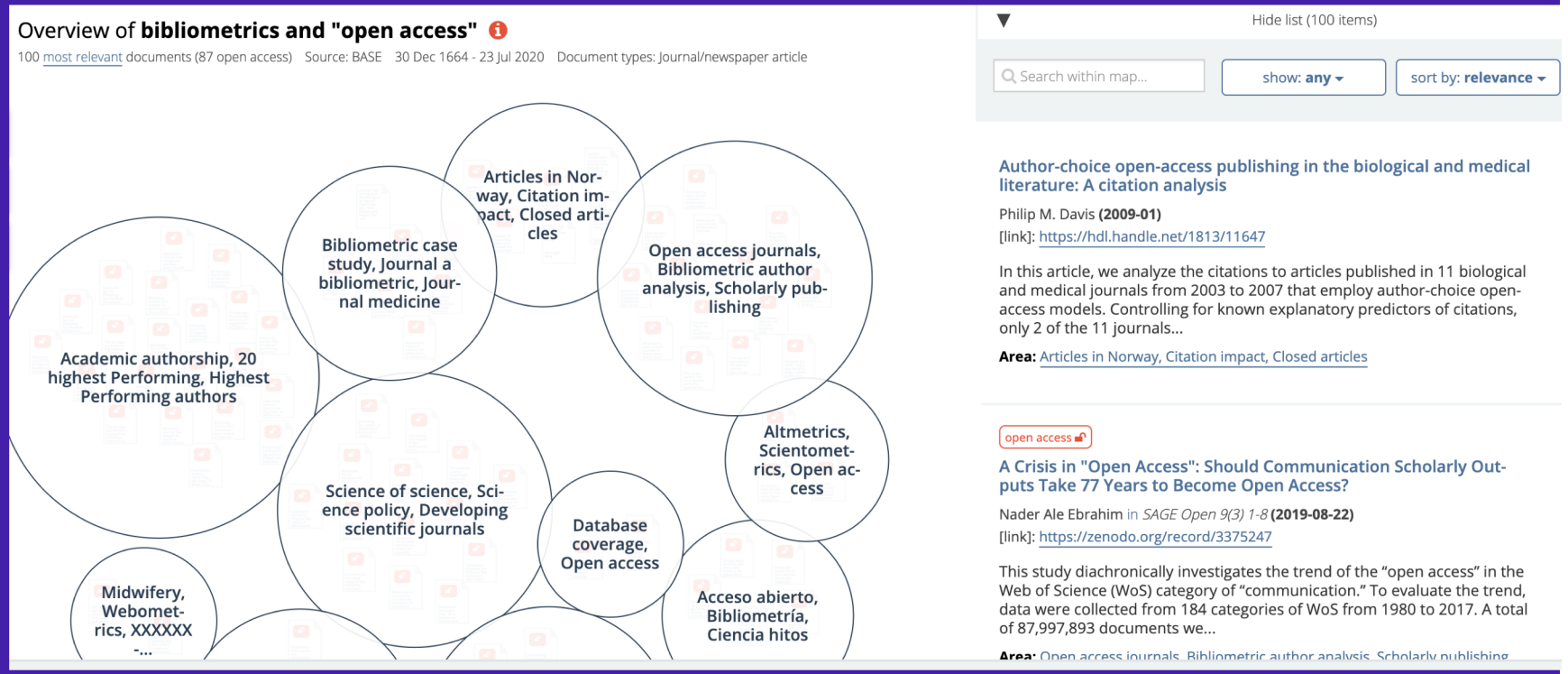

Nees Jan van Eck


Ludo Waltman

- https://www.vosviewer.com/
- a software tool for constructing and visualizing bibliometric networks
- Developed by Nees Jan van Eck and Ludo Waltman at Leiden University's Centre for Science and Technology Studies (CWTS).
- Fast, intuitive, usable, and dynamic, with excellent documentation
  - Video tutorials
    - Introduction: https://www.youtube.com/watch?v=9dTWkNRxUtw
    - Visualization of Bibliographic network using VosViewer https://www.youtube.com/watch?v=nLBEU2Tul9Q
- Brett Williams, "Dimensions & VOSViewer Bibliometrics in the Reference Interview," *code4lib,* https://journal.code4lib.org/articles/14964

Open Knowledge Maps

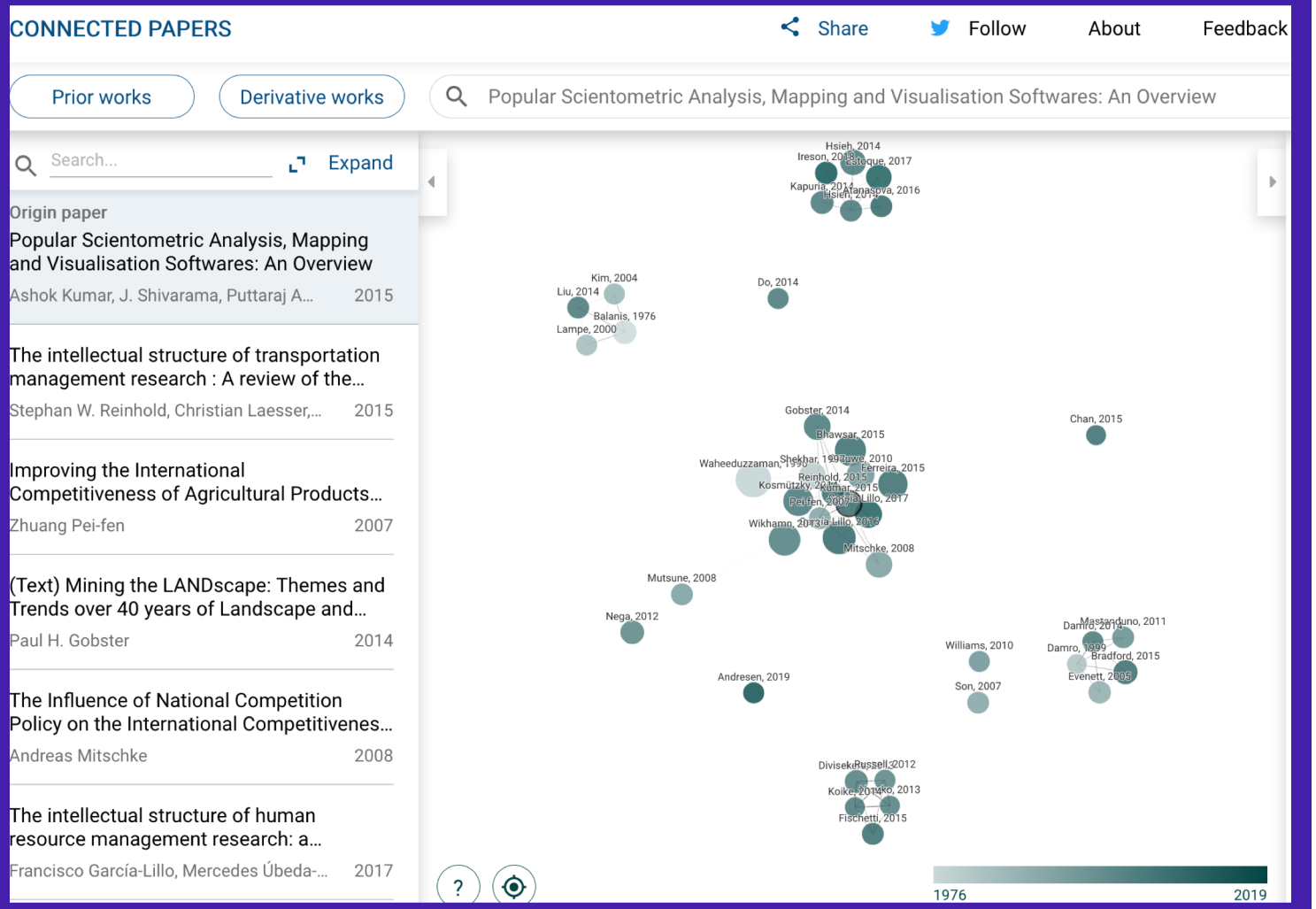
- https://openknowledgemaps.org/
- Uses PubMed and BASE (Bielefeld Academic Search Engine) relevancy rankings
- Creates knowledge maps based on article metadata



Knowledge Map of bibliometrics AND "open access"
https://openknowledgemaps.org/map/8b1804cf25ceab46937af06aa9481a20

Connected Papers

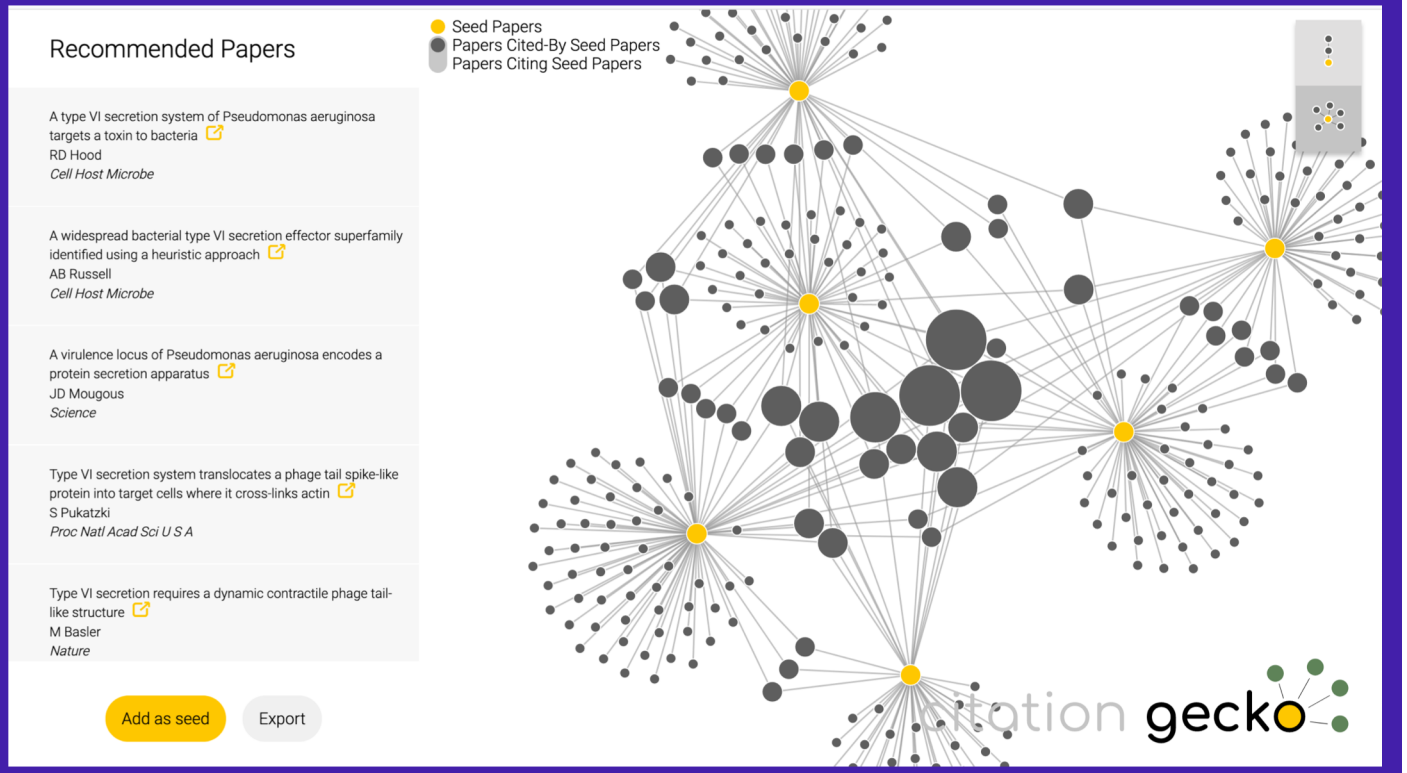
- https://www.connectedpapers.com/
- Similarity based on co-citation and bibliographic coupling (not direct citation tree)
- Connected to Semantic Scholar https://www.semanticscholar.org
  - a free, AI-powered search and discovery tool
  - Open Research Corpus: http://api.semanticscholar.org/corpus/



https://www.connectedpapers.com/main/51ef073316e2eb29fa8ee79078264f1aa3bd05de/Popular-Scientometric-Analysis-Mapping-and-Visualisation-Softwares-An-Overview/graph

Citation Gecko

- https://citationgecko.azurewebsites.net/
- Add “seed papers” from BibTeX, Papers, Zotero, Mendeley
- View a network of papers cited by and papers citing the seed papers



Other bibliometric mapping software

- CitNetExplorer https://www.citnetexplorer.nl/
- Gephi (not specific to citations) : https://gephi.org/
- CiteSpace: http://cluster.cis.drexel.edu/~cchen/citespace/
- CoCites (PubMed) http://www.cocites.com/
- Sci2 (site currently has expired certificate) https://sci2.cns.iu.edu/user/index.php

Next steps

- See https://musingsaboutlibrarianship.blogspot.com/2020/06/more-researchliterature-mapping-tools_16.html
- Ashok Kumar, J. Shivarama, and Puttaraj A. Choukimath, "Popular Scientometric Analysis, Mapping and Visualisation Softwares: An Overview," 2015, https://www.semanticscholar.org/paper/Popular-Scientometric-Analysis%2C-Mapping-and-An-Kumar-Shivarama/51ef073316e2eb29fa8ee79078264f1aa3bd05de?p2df
  - See the Connected Papers network for this paper: https://www.connectedpapers.com/main/51ef073316e2eb29fa8ee79078264f1aa3bd05de/Popular-Scientometric-Analysis-Mapping-and-Visualisation-Softwares-An-Overview/graph

Next steps

NIH Bibliometrics Training Series is a comprehensive introduction to network analysis, with examples in Sci2, Gephi, and Inkscape

https://www.nihlibrary.nih.gov/services/bibliometrics/bibliometrics-training-series

Next steps

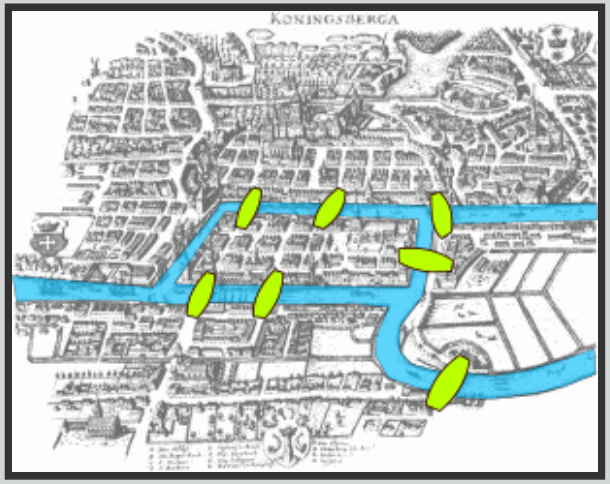
- Access the webpage and follow the instructions under Exercises and Videos for session 3 at https://pdoehle.github.io/bibliometrics-for-librarians/
  - Clone the new Azure notebook at https://notebooks.azure.com/clarke-iakovakis/projects/intro-to-bibliometrics-network
  - Download VOSviewer and complete the steps on the website
- Contact us if you need help
  - clarke.iakovakis@okstate.edu
  - doehle@okstate.edu

Network Analysis

Centrality, Connectedness, and Clustering

Networks/Graphs

- Graph theory provides a powerful tool for bibliometricians
- It allows an investigator to abstract the relationship between multiple objects with multiple relationships
- Network analysis began with Euler's solution to the Konigsberg (Kaliningrad, Russia) Bridge Problem



By Bogdan Giuşcă - Public domain (PD),based on the image, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=112920

Networks/Graphs

Formal Definition: A network or graph, G is a pair of two related sets: $G = (V, E)$ consisting of a set V of verticies (or nodes) and a set E of edges, links, or arcs.

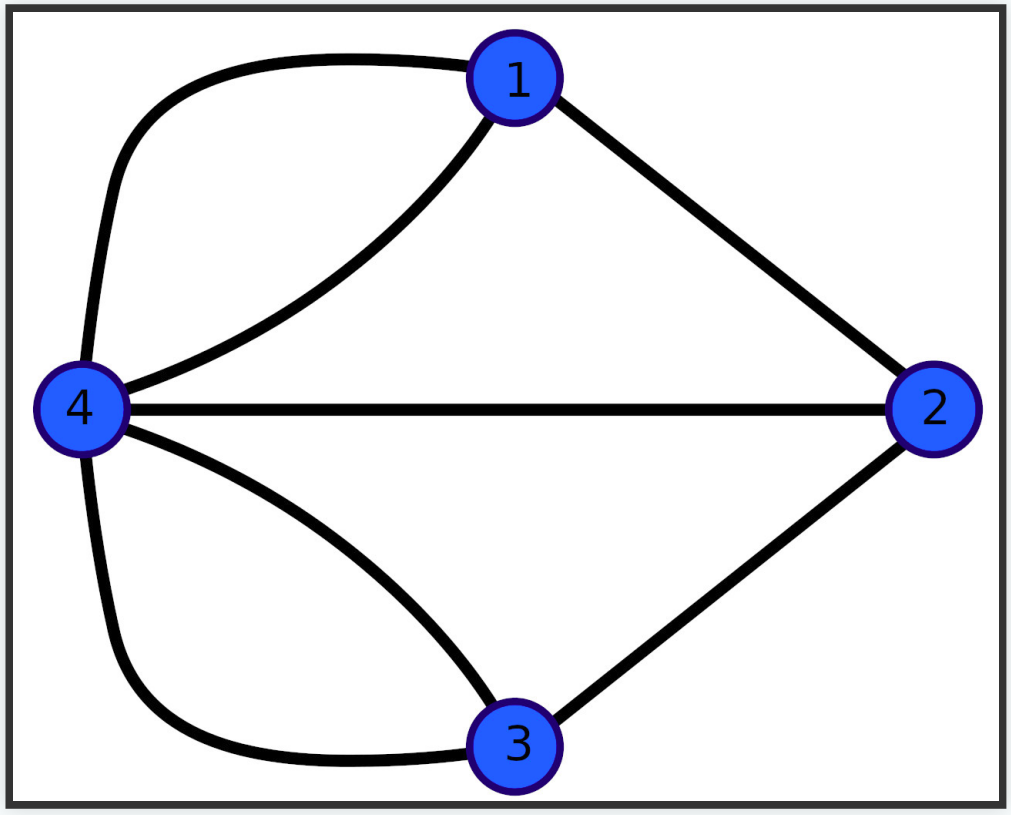
- In sociological research, nodes are sometimes referred to as *actors*.

Representing Graphs

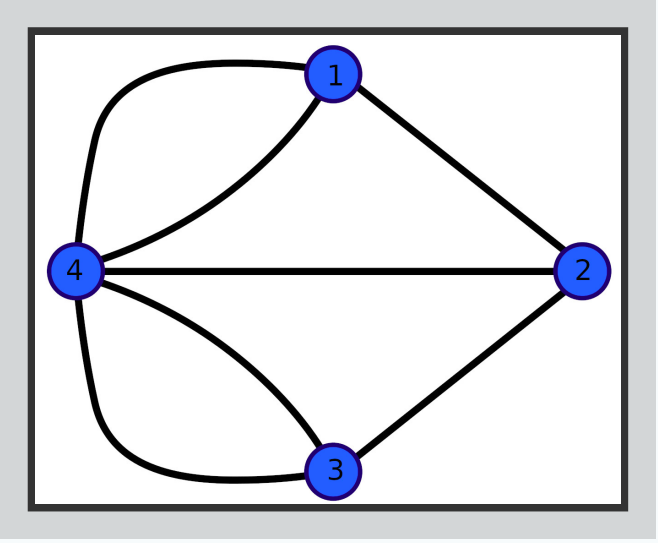
There are three standard ways of representing graphs:

1. Two-dimensional graphs
2. Adjacency matrices
3. Adjacency lists

Two-Dimensional Graphs

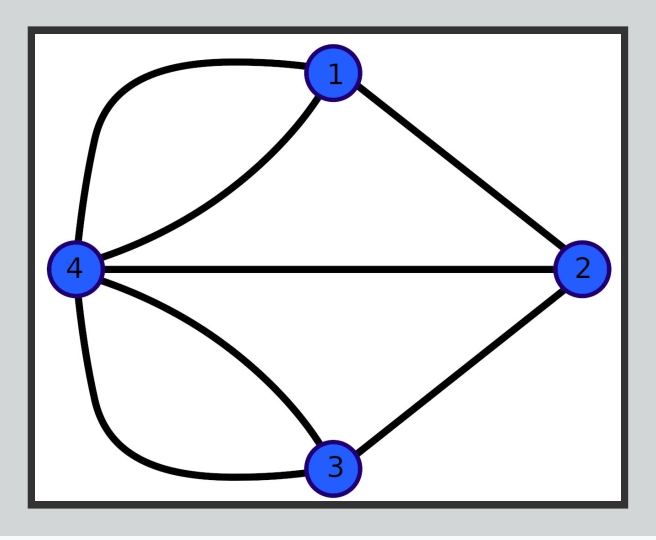


Adjacency Matrices



$$A = \begin{bmatrix} 0 & 1 & 0 & 2 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 2 \\ 2 & 1 & 2 & 0 \end{bmatrix}$$

Adjacency Lists



$$A = \{(1,2),(1,4),(1,4),(2,3),(2,4),(3,4),(3,4)\}$$

Connectedness

Density

Definition: A *complete graph* is a graph where every node is connected to every other node.

Definition: Density, $D = \frac{\text{Number of edges}}{\text{Number of edges in a complete graph with the same number of nodes}}$.

Density

Question: How many ways can I choose 2 items from the set $\{A, B, C, D, E, \ldots\}$ containing n letters.

Answer:

$$\frac{n!}{2!(n-2)!} = \binom{n}{2}$$

Number of Edges in a *Complete* Graph (Undirected)

- A graph $G = (V, E)$ consists of two sets V (verticies) and E (edges).
- In a *complete* graph with N nodes, every node connects to every other node.
- We can think of an edge from E as choosing two nodes from V where order does not matter.
- The maximum number of edges (the number in a *complete* graph) is:

$$\binom{\text{Number of nodes}}{2} = \binom{N}{2}$$

Calculating Density for an Undirected Graph with N Nodes and E Edges

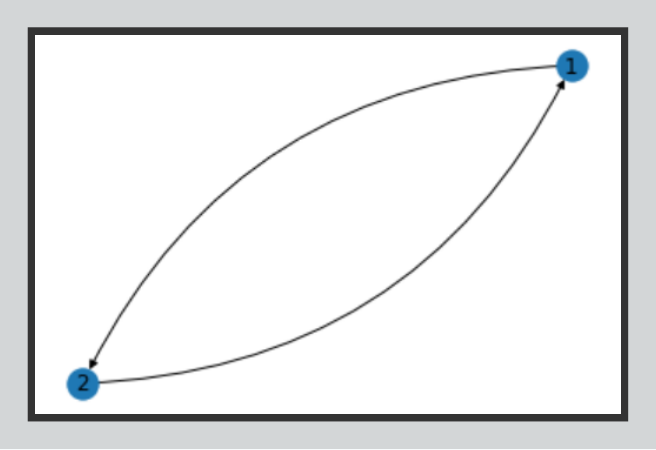
$$
\begin{aligned}
D &= \frac{\text{Number of edges}}{\text{Number of edges in matching complete graph}} \\
&= \frac{E}{\binom{N}{2}} \\
&= E \cdot \frac{2!(N-2)!}{N!} \\
&= \frac{2E}{N(N-1)}
\end{aligned}
$$

Calculating Density for an Undirected Graph with N Nodes and E Edges

- To calculate the density of an *undirected* graph with N nodes and E edges, use the following formula:

$$D = \frac{2E}{N(N-1)}$$

Calculating Density for a Directed Graph with N Nodes and E Edges



- A complete, *directed* graph with N nodes has twice as many edges as a complete *undirected* graph with N edges.
- The denominator of my ratio will be twice as large (i.e., divide by 2):

$$D = \frac{E}{N(N-1)}$$

Density

- Density gives us a measure of how connected the nodes in a graph are to each other.
- Density is given as a ratio.
- A value of 1 indicates every nodes is connected to every other node.
- A value of 0 indicates no connections between nodes.

Centrality

- Each measure examines a different aspect of leadership.
- We will discuss four different centrality measures:
    - Degree
    - Closeness
    - Betweenness
    - Eigenvector

Degree Centrality

- The number of neighbors a node is connected to:

$$\deg(i) = \sum_j m_{ij},$$

  where $m_{ij} = 1$ if i connects to j and $m_{ij} = 0$ otherwise.
- In a graph with N nodes, this measure can be standardized by dividing the degree centrality of a node by $(N - 1)$.
- Be sure to differentiate in and out degrees if working with a directed graph.

Degree Centrality

- Within a cocitation network, the degree centrality of an author is the number of authors within the network that have coauthored papers with the auther in question.
- Within a ciatation network, degree centrality corresponds to the flat citation count discussed in week one.

Closeness Centrality

- Compute the average distance from a given node to other nodes:

$$\text{close}(i) = \sum_{j \in V, i \neq j} \frac{\gamma(i,j)}{(N-1)},$$

  where $\gamma(i,j)$ is the minimum distance between i and j.
- A low value indicates that on average, the node is a short distance from other nodes.

Closeness Centrality

- Low values indicate the node is close to others.
- High values indicate the node is far from others.
- Within a citation network, a publication with a low value distributes is knowledge efficiently across the network.

Betweenness Centrality

- Given any path through a graph, how many times will that path go through the node in question?
- An illustration from history: Kashgar, Xinjiang and the oasis kingdoms of the Tarim basin.
- For node v, follow these steps to calculate betweenness:
    1. Identify all node pairs in the rest of the network.
    2. For each pair, calculate the following ratio: $\frac{\text{Number of shortest paths through } v}{\text{Total shortest paths}}$.
    3. Add all the ratios together.

$$\text{betweenness}(v) = \sum_{s \neq t \neq v} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}},$$

where $\sigma_{s,t}$ is the number of shortest paths between s and t (s and t are nodes that are not v) and $\sigma_{s,t}(v)$ is the number of those shortest paths that go through v.

Betweenness Centrality

- The maximum value of of betweenness for a given node v in a graph with N nodes is $\frac{N^2-3N+2}{2}$.
- Betweenness can be normalized as follows:

$$\frac{2 \cdot \mathrm{betweenness(v)}}{N^2 - 3N + 2}$$

Betweenness Centrality

- Within a citation network, a publication with high betweenness centrality acts as a gatekeeper.
- These papers are important because they transfer information between publications.
- They serve as a bridge.
- Betweennes has also been used to examine interdisciplinarity of publications.

Eigenvector Centrality

- Eigenvector centrality attempts to measure the influence of a node on a network by ranking how *key* it is relative to the other nodes.
- The more connected a node is to well-connected neighbors, the higher its rank.
- The node and its immediate neighbors both affect a node's ranking.

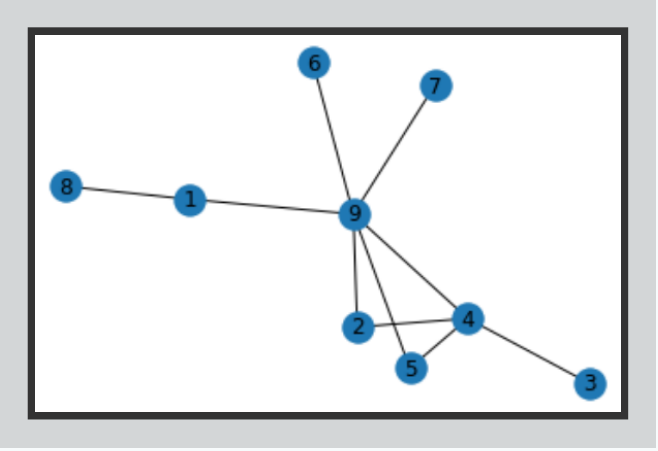
Applications of Eigenvector Centrality

- Google's Page Rank algorithm is a variant of Eigenvector Centrality.
- It has been applied to predicting academic positions and examining the influence of editors on the quality of articles within Wikipedia. (Diallo 1008)

Computing the Eigenvector Centrality

1. Solve the eigenvector equation $\mathbf{Ax} = \lambda\mathbf{x}$ for the largest eigenvalue.
2. Normalize the eigenvector so all the values either add to 1.
3. The nth element in the eigenvector corresponds to the nth node, and its magnitude corresponds to that node's ranking.

Eigenvector Centrality–Why it Works



- Taking a random walk, my journey will pass through certain nodes more often.
- If I run this experiment long enough and keep a count for each node, the node rank order will converge to a set ordering.
- This ordering corresponds to the eigenvector rank ordering.
- You will get a chance to test this in your homework.

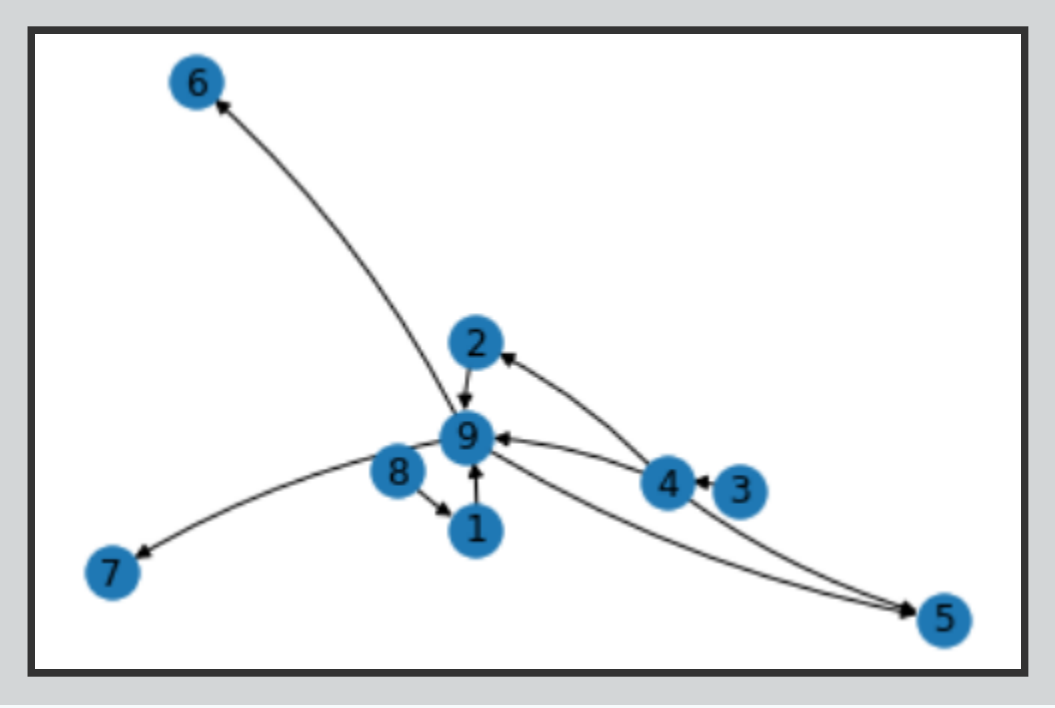
Eigenvector Centrality–The Math

- Changing entries in the adjacency matrix to probabilities, creates a transition matrix, $\mathbf{P}$, for a time-homogeneous Markov process.
- If the Markov chain is irreducible (positive probability of getting from any node to any other node) and not periodic, then there is a stationary distribution π (i.e., $\pi\mathbf{P} = \pi$) and

$$\lim_{k\to\infty} \mathbf{P}^k = \mathbf{1} \cdot \pi$$

- Note: Eigenvector Centrality is usually computed with right eignvectors. In this case, we would start with the transpose of the adjacency matrix before computing the eigenvector.

Eigenvector Centrality and Google Page Rank



Eigenvector Centrality–Things to Remember

- This metric can provide insight into the probability of a paper appearing within a citation chain.
- We can use Eigenvector centrality to understand an actor's role on the flow of information within a network.
- If I cannot get to any node from any other node, this technique does not work (dangerous for directed networks).
- A node's ranking is affected by its neighbors.

Clustering

- A clustering coefficient measures how often a friend of a friend is also a friend.
- The clustering coefficient, C, for a graph is

$$C = \frac{3 \cdot \text{Number of triangles in a network}}{\text{Number of connected triples}}$$

- C will always be between 0 and 1, i.e., $0 \leq C \leq 1$.
- This ratio can measure a social network to get a sense of how interconnected a community is.

References

- Anthonisse, J. M. (1971). The rush in a graph. Amsterdam: Mathematisch Centrum.
- Bavelas, A. (1948). A mathematical model for small group structures. Human Organization, 7(3), 1630.
- Borgatti, S. P. (1995). Centrality and AIDS. Connections, 18(1), 112115.
- Diallo SY, Lynch CJ, Gore R, Padilla JJ. Identifying key papers within a journal via network centrality measures. Scientometrics. 2016 Jun;107(3):1005–20.
- Freeman, L. C. (1977). A set of measures of centrality based on betweenness. Sociometry,
  40(1), 3541.
- Leydesdorff, L. (2007). Betweenness centrality as an indicator of interdisciplinarity of scientific journals. Journal of the American Society for Information Science & Technology, 58(9), 13031319.
- Newman, M. E. J. (2003). The structure and function of complex networks. SIAM Review, 45(2), 167256.
- Rousseau R, Egghe L, Guns R. Becoming Metric-Wise: A Bibliometric Guide for Researchers.
- Sabidussi, G. (1966). The centrality index of a graph. Psychometrika, 31(4), 581603.
- Serfozo R. Basics of Applied Stochastic Processes [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009 [cited 2020 Jul 23]. (Gani J, Heyde C, Jagers P, Kurtz TG, editors. Probability and Its Applications). Available from: http://link.springer.com/10.1007/978-3-540-89332-5