The Structure of Scientific Disciplines

A Comparative Analysis of Bibliometric & Semantic Networks

Vladimir Borel

A thesis presented for the degree of
Doctor of Philosophy

Sociology
University of California, Riverside
United States
April 24, 2023

Contents

List of Figures

List of Tables

Abstract

This paper examines the structure and dynamics of science across 16 fields by analyzing both bibliographic and semantic networks. The study finds that there are few structural differences between the fields, indicating that the underlying structure of scientific knowledge is relatively uniform. The analysis reveals that there are common patterns of co-citation and term co-occurrence, highlighting the interconnected nature of scientific research.

1 Introduction

The debates on the structure and dynamics of science have often been based on abstract ideas rather than empirical evidence. However, the recent trend towards digitalization of peer-reviewed scientific articles has made it possible to test these theories using empirical data. This provides a valuable opportunity to advance our understanding of the structure and dynamics of science.

These theories ask a range of questions about the nature of science and its relationship to other fields of knowledge. Some of the key questions they address include:

- Are there regular patterns in the way that the structure of science changes?
- Is science a distinct field of knowledge or is it embedded in larger domains, is the knowledge generated by scientific methods fundamentally different from other types of knowledge?
- Are there clear distinctions between science and non-science or between different types of science (e.g. "hard" and "soft" sciences)?

This paper aims to make three types of contributions. The first set of contributions is theoretical and involves synthesizing and unifying different conceptions of scientific development in graph-theoretic terms. The second set of contributions is empirical and involves deriving and testing hypotheses from the literature on scientific development. The third set of contributions is practical and focuses on developing a standard analysis pipeline applicable to a wide variety of fields. This is necessary due to the increasing difficulty of coordinating research efforts and bridging "academic silos" in the face of the exponentially increasing output of technical knowledge and the fact that researchers only have local knowledge about their immediate neighbors.

When discussing the structure of a body of knowledge, I am referring to two main aspects: the citation structure and the semantic structure. The citation structure refers to the way in which ideas and information are connected through references and citations, while the semantic structure refers to the meaning and relationships between the concepts and ideas within the body of knowledge. Both of these aspects of structure are important for understanding the organization and development of a particular field of knowledge.

2 Theory

2.1 Revolutionary Science

Kuhn (2012) posits that scientific knowledge advances through periods of incremental progress punctuated by abrupt, revolutionary changes in understanding. According to Kuhn, during normal science, scientists work within an accepted paradigm, which is a set of assumptions, concepts, and methods that define a scientific discipline at a given time. Scientists in a paradigm share a common language, a set of beliefs, and a set of practices that enable them to conduct research and generate new knowledge within the paradigm.

However, as anomalies and inconsistencies between what the models predict and what is observed accumulate within the paradigm, scientists may begin to question its validity, and a crisis may emerge. During this crisis, scientists may propose new theories and methods that challenge the existing paradigm. If a new theory is accepted, it may eventually replace the old paradigm, leading to a scientific revolution.

2.2 Fractal Science

Abbott's (2001) fractal division of disciplines is a model that suggests that any discipline can be divided into smaller sub-disciplines, which in turn can be further divided into smaller sub-sub-disciplines, and so on, creating a fractal-like structure.

Figure 1 (p. 14)

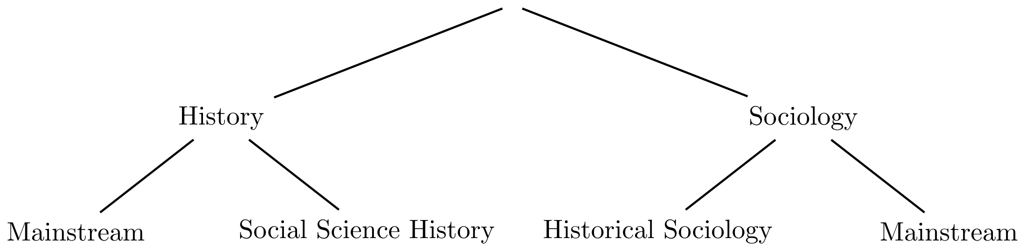

Figure 1: Fractal Distinctions

2.3 High Consensus, Rapid Discovery Science

Collins (1994) posits that achieving high levels of consensus and rapid scientific discovery in science research relies on the use of research technologies and techniques that generate consistent, observable, and reliable data, resulting in a high degree of reproducibility. As a result, conflict and disagreement are limited to the frontiers of research.

The natural sciences have experienced high levels of consensus and rapid scientific discovery since the post-scientific revolution of the 1600s, but this has not been solely due to empiricism, mathematization, measurement precision, or the experimental method that existed before this era. Instead, the distinguishing feature of this time period has been the creation of research technologies that facilitate the production of new data and observations.

The social sciences face obstacles not solely due to a lack of common ideas, as Kuhn suggests, but also from a dearth of technologies that produce consistent data and results: "The most glaring lack within social science is a self-generating lineage of research technologies" (p. 165).

2.4 Attention in Science

The "Law of Small Numbers" is a concept first introduced by Collins (1998), which is based on the idea of a "limited attention space" which sets an upper bound on the possible number of subfields in a field. This concept has been further developed by other researchers, with Price (1986) suggesting that the number of items that can be held in attention is around five, while Collins himself revised this number to seven in 1994. In other words, the "Law of Small Numbers" is analogous to a carrying capacity.

A binomial distribution can be used to describe the number of groups competing for attention in situations where there are two possible outcomes for each group (i.e. success or failure in acquiring the resource). The binomial distribution can be used to calculate the probability of a given number of groups successfully acquiring the resources. For example, if there are 10 groups competing for a finite number of resources, the binomial distribution can be used to calculate the probability of each possible number of groups (from 0 to 10) successfully acquiring the resources. I expect to observe a binomial distribution with a mode of anywhere between 1 and 7 for the number of subfields in a field, but most likely between 5 and 7.

2.5 Phases of Science

Mullins (1973) proposed a model for the development of scientific fields that incorporates elements from both the "invisible college" (De Solla Price and Beaver, 1966; Crane, 1969, 1972; Paisley, 1972) and "scientific revolutions" (Kuhn, 2012). According to Mullins, the development of scientific fields is determined by the interactions and competition between different "theory groups", which are networks of scientists who share similar ideas and interests.

Mullin's model consists of four stages, which are as follows:

1. Normal: In the normal stage of theory group development, the communication network among researchers is loosely coupled, meaning that there is only a weak connection between the different researchers within the group. This stage is characterized by low clustering and density, and individual researchers are largely focused on solving specific problems within the field. This stage is similar to what Kuhn (1962) referred to as the "puzzle-solving" or "paradigmatic" phase of scientific development, in which researchers are largely focused on applying established theories and methods to the study of specific phenomena. In this phase, there is little attention to or interest in alternative paradigms or theories, and the focus is on making incremental advances within the existing framework.

2. Network: In the network stage of theory group development, the patterns of communication among researchers begin to change due to the emergence of one or more influential ideas or theories that attract the attention of multiple researchers. This leads to a shift in the focus of the group and the development of a consensus among its members, which allows for the creation of an alternative paradigm or approach to the problem they are studying. As the group becomes more focused and cohesive, communication within the group increases, while communication with external researchers or other groups declines. In this phase, the group's success in advancing their ideas and theories is crucial for maintaining its growth and attracting new members.

3. Cluster: The third stage of theory group development, known as the cluster phase, is characterized by the emergence of a power-law distribution of communication ties among researchers. In this phase, clusters of researchers typically form around the most productive or influential members of the group, who can often be found in one or a few institutions. This creates a positive feedback loop, where the ideas and theories of the most productive researchers are reinforced and strengthened by the input and support of the other members of the cluster. This can lead to the institutionalization of the group and its ideas, as the cluster becomes increasingly recognized and respected within the broader field. As the group's ideas and theories evolve, they may also begin to diverge from the established concepts and theories of the parent discipline, reflecting the group's growing focus and specialization. According to Mullin (1973), the degree of divergence from the parent field is a function of the group's size and isolation, with larger and more isolated groups being more likely to develop their own distinct paradigms and theories.

   At this point, the group can take one of two possible trajectories, depending on the parent discipline's reaction to the group's ideas. If the parent discipline rejects or ignores the group's ideas, the group may become isolated and continue to develop their ideas independently, remaining in the cluster stage. This is often referred to as a "revolutionary" trajectory, as the group is effectively challenging the existing paradigms and theories of the parent discipline. On the other hand, if the parent discipline adopts or diffuses the group's ideas, the group may become integrated into the broader field and recognized as an "elite" group. In this case, the group's ideas may become mainstream and widely accepted within the parent discipline.

4. Speciality: The final stage of theory group development, known as the specialty phase, is only reached by "elite" groups whose ideas are adopted and diffused by the parent discipline. In this phase, the group's members may begin to scatter and move to different institutions or locations, weakening the connections within the group and enabling the diffusion of their ideas to a wider audience. This process of scattering and diffusion can lead to the routinization and institutionalization of the group's ideas, as they become integrated into the broader field of knowledge and accepted as a new paradigm or approach. Over time, the dense network of communication within the group may begin to loosen again, returning to the more loosely-coupled, normal pattern of communication seen in the earlier stages of theory group development.

The transition between the different phases of theory group development is not guaranteed, and groups may fail or die off before reaching the later stages. The success of a group and its ability to advance through the different phases is contingent on a number of factors, including its ability to grow and attract new members at a faster rate than alternative groups. This in turn is dependent on factors such as the group's apparent success in advancing their ideas, the presence of strong intellectual leaders within the group, the support and recognition of the broader community, the availability of research centers and materials to support the group's work, and the output of textbooks and other materials that can help to promote and disseminate the group's ideas.

3 Methods

3.1 Data

Having established the possibility of linking citation networks and concept networks as clusters at various levels of analysis, I will now delve into a discussion of the data used to identify the structure and dynamics of scientific communities. The analysis of this data will enable me to evaluate the accuracy of the various theories of scientific development discussed earlier and to provide a realistic account of the fields analyzed, given that the data is reliable and sufficient.

3.1.1 Sample

There are two main ways of bounding citation networks. The first involves sampling based on journals (Beam et al., 2014; Moody and Light, 2006), while the second involves sampling based on a topical search (Borrett et al., 2014; Callon et al., 2005; Chavalarias and Cointet, 2013). In this study, I used the former method, selecting articles that were published in one of the 5 peer-reviewed journals with the highest impact factor.

The main advantage of sampling based on journals is that it will allow us to focus on articles published in high-quality, reputable journals. This can help ensure that the data used in the study is of a high quality and is representative of the broader field of research. Additionally, journal-based sampling can make it easier to compare the results of different studies, as it provides a common frame of reference.

Bradford's law (Bradford, 1985; Hjørland and Nicolaisen, 2005) is a principle that describes the relationship between the number of scientific journals in a particular field and the number of citations that articles in those journals receive. It states that a small number of highly-cited journals tend to publish a disproportionate number of highly-cited articles, while a large number of low-cited journals tend to publish a disproportionate number of low-cited articles.

This law can be used to support the choice of only selecting a small number of journals in each scientific field under study. Focusing on a select group of highly-cited journals, can increase the chances of finding and reading the most influential and highly-cited articles in that field. It's worth noting that Bradford's law is not a hard and fast rule, and there may be cases where it is necessary or beneficial to review a larger number of journals in a particular field. However, as a general principle, Bradford's law can be a useful guide when selecting journals for review.

The data used in this study consists of peer-reviewed scientific articles collected by Clarivate Web of Science. These data include each document's citation information, as well as its title, abstract, key words, and publication date.

3.1.2 Fields

The fields being analyzed include Artificial Intelligence, Astronomy & Astrophysics, Economics, Ethnic & Cultural Studies, Gender Studies, Genetics & Genomics, Geometry, Geophysics, Human Resources & Organizations, Immunology, International Business, Language & Linguistics, Law, Material Engineering, Neurology, Political Science, Probability & Statistics, Sociology, Biochemistry. A complete list of the journals for each of the fields can be found in the Appendix.

These fields were selected for analysis because they represent a diverse range of disciplines, including both hard sciences, as well as soft sciences. This diversity allows for a comprehensive examination of how scientific communities function across different areas of study. Additionally, these fields are dynamic and rapidly advancing, providing a rich source of data for analysis and making them well suited for studying the structure and dynamics of scientific communities.

Furthermore, these fields were selected because they enable not only the analysis of individual fields but also the use of a comparative framework. Comparing and contrasting the structures and dynamics of scientific communities across different fields allows for a deeper understanding of the similarities and differences among them, and the opportunity to identify common as well as idiosyncratic patterns and trends that may be relevant to theories of scientific development.

3.1.3 Tools

Python was used for the analysis (van Rossum and Drake, 2010), and several libraries were particularly noteworthy, including Networkx (Hagberg et al., 2008), PyMC (Wiecki et al., 2023) and Polars (Vink et al., 2023), among many others.

3.2 Networks

3.2.1 Citation

A citation network represented as a directed unweighted graph, with nodes representing individual documents and edges indicating a citation from one document to another.

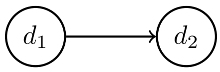

Figure 2: Citation

Figure 2 depicts a citation diad in which document d_1 cites document d_2.

3.2.2 Co-Citation

A co-citation network represented as an undirected weighted graph, with nodes corresponding to individual documents and edge weights representing the frequency with which two documents are co-cited.

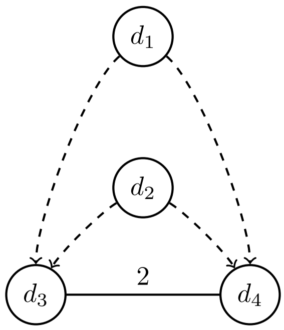

Figure 3: Co-Citation

As shown in Figure 3, the edge weight of 2 between documents d_3 and d_4 reflects the fact that they are co-cited by two documents, d_1 and d_2.

3.2.3 Co-Occurrence

A co-occurrence network represented as an undirected weighted graph, with nodes representing individual terms and edge weights indicating the frequency with which two terms co-occur within a document.

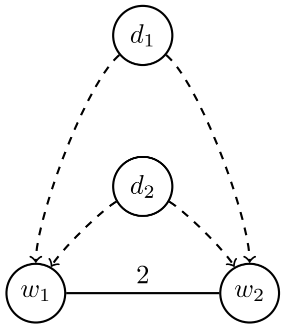

Figure 4: Co-Occurrence

As shown in Figure 4, the edge weight of 2 between terms w_3 and w_4 reflects the fact that they co-occur in both documents, d_1 and d_2.

3.3 ERGMs

Exponential Random Graph Models (ERGMs) are commonly employed to model and test various hypotheses about the structural characteristics of social networks. This approach allows for the investigation of both local and global properties of a network and enables the identification of significant factors that contribute to its formation. By examining network properties such as the presence of triangles or the average shortest path between nodes, we can determine the extent to which these features played a role in shaping the network's structure.

ERGMs rely on the fundamental theoretical assumption of dependence, which posits that the presence of certain relationships between documents can impact the formation, persistence, or prevention of other relationships. To illustrate, when an article, let's say article A, cites both articles B and C, it is more probable that articles B and C will also cite each other.

In their general form, ERGMs are written as (Hunter et al., 2008):

$$P(Y = y) = \frac{exp(\theta' g(y))}{k(\theta)}$$

The model specifies that the probability of a network (on the left-hand side) is a function of terms on the right-hand side, which represent hypothesized network features that deviate from chance expectations, where:

- Y is the random variable for the state of the network (with realization y)
- $g(y)$ is a vector of model statistics (ERGM terms) for network y
- θ is the vector of coefficients for those statistics
- $k(\theta)$ represents the quantity in the numerator summed over all possible networks

The numerator can be written as:

$$log(exp(\theta' g(y))) = \theta_1 g_1(y) + \theta_2 g_2(y) + \cdots + \theta_p g_p(y)$$

The denominator of the probability distribution is a normalizing constant that ensures that the distribution sums up to one. Computing this constant requires summing over the space of all possible networks on n nodes, but the number of possible configurations grows exponentially with the number of nodes, specifically $2^{(n(n-1)/2)}$ for undirected graphs and $2^{(n(n-1))}$ for directed graphs. Therefore, exact computation of this sum is infeasible. To address this, Markov Chain Monte Carlo (MCMC) methods are commonly used to generate samples and estimate the values of the parameters that maximize the likelihood of the observed network. These estimates capture the strength and direction of the effects of different network statistics on the likelihood of observing the network.

The above formula can be rewritten in terms of the covariate vector θ:

$$logit(Y_{ij} = 1|y_{ij}^c) = \theta' \delta(y_{ij})$$

Where:

- y_{ij}^c is the complement of y_{ij}, i.e. all dyads in the network other than y_{ij}
- y_{ij}^+ as the same network as y except that $y_{ij} = 1$
- y_{ij}^- as the same network as y except that $y_{ij} = 0$
- $\delta(y_{ij})$ is given by $g(y_{ij}^+) - g(y_{ij}^-)$ which measures how the sufficient statistic $g(y)$ changes if the (i, j)th edge is "toggled" on or off.

To summarize, for each of the sufficient statistics, a matrix $g_p(y)$ of size $n \times n$ is created. Each entry g_{ij} represents the effect that the presence of an edge between nodes i and j has on the corresponding network statistic, while holding the rest of the network constant.

To help us interpret the coefficents, let us take an example. If only edges and triangles are included in the model, the conditional log-odds of two documents being co-cited or two terms co-occurring, keeping the rest of the network constant, is:

$$\theta_{edges} \times \text{change in the number of ties} + \theta_{triangles} \times \text{change in number of triangles}$$

The total number of ties in the network always increases by 1 with the addition of any new tie, therefore, for an edge that creates no triangles, the conditional log-odds is simply θ_{edges}. If the addition of the edge creates one triangle the conditional log-odds becomes $\theta_{edges} + 1 \times \theta_{triangles}$. If the addition of the edge creates one triangle the conditional log-odds becomes $\theta_{edges} + 2 \times \theta_{triangles}$. Etc. Note that the edge covariate is always included and is treated as an intercept in the model.

To obtain the corresponding probability, the expit (or inverse logit) of θ is taken:

$$P(e) = \exp(\theta)/(1 + exp(\theta))$$

3.4 Measurements

Following Jiao et al. (2017) I define the network measurments and the expected direction of their effect

3.4.1 Density effect (Edges)

Density refers to the proportion of connections in a social network relative to the total possible connections.

This is interpreted as the intercept of the model. The entry of the 'density' matrix M_{ij} indicates how the presence of that edge changes the number of edges in the graph, holding the rest of the network constant. It a matrix of ones - with the upper right triangle masked in the case of undirected networks.

3.4.2 Small-World effect (Triangles)

Social networks tend to exhibit a much higher number of triangles and transitive triads than what would be predicted by random graphs with comparable density.

Triangles, which represent a set of three documents or concepts who are mutually connected, are a common feature of social networks and are thought to play an important role in social cohesion and network structure (Kossinets and Watts, 2006; Newman, 2018).

The entry of the triangle matrix M_{ij} indicates how the presence of that edge changes the number of triangles in the graph, holding the rest of the network constant.

Overall tendency of networks to be composed of local densities (closure). Are the ties randomly distributed of do they form local structures.A graph with a low triangle effect would look homogeneous distribution of substructures is what we would expect from a random graph

Lumps appear (local structures)

3.4.3 Closure effect (Transitivity)

A triad is a group of three nodes in a graph. A triad can either be open or closed. An open triad is a group of three nodes that are connected by two edges (Figure 5a), while a closed triad is a group of three nodes that are connected by three edges (Figure 5b).

(a) Open Triad (b) Closed Triad

Figure 5: Triads

Transitivity is defined as the ratio of the number of closed triads in the graph to the number of open triads in the graph.

$$T = \frac{3t_c}{t_o}$$

Where t_c is the number of closed triads and t_o is the number of open triads.

3.4.4 Clique effect (Cliques)

Tendecy for there to be parts of the graph where multiple nodes all co-occur more than by random change

If I discuss A all other concepts in the clique it belongs to. highly related to each other distinct topical or functional units within the network

This build on triangles (which are a clique) but captures more information

It asks whether triangles overlap to form larger structurees rather than just isolated (random) triangles

Densly connected always connected

3.4.5 Silo effect (Components)

silos refer to isolated areas of research or knowledge that are not well connected to other areas or fields limit the flow of information and influence within the network Hyper specialization of field of knowledge Talk to one another but never talk to nodes outside their primary group

3.4.6 Popularity effect (k-star)

The popularity effects in most classes were not obvious, which meant the degree (the total number of actors selecting an individual) of individuals in the class networks had little difference.

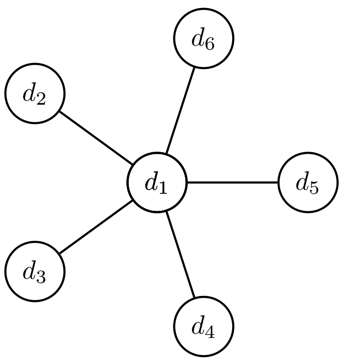

Figure 6: Star

What we mean by a strong k star effect Control for 9 stars in order to determine if 10 star is significant 10 star is a bunch of 9 stars

3.4.7 Mediating effect (betweenness centrality)

Nodes with high betweenness centrality in this type of network may represent important bridge terms or pivot terms that connect different topics or themes within the corpus of documents. The betweenness centrality c_b of node n is given by:

$$c_b(i) = \sum_{j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$$

Where σ_{ij} is the total number of shortest paths from node i to node j and $\sigma_{ij}(n)$ is the number of those paths that pass through node n. In other words, it is the proportion of all shortest paths between nodes i and j that pass through node n. The average betweenness centrality over all nodes of the network is taken as another network statistic.

Indicates good integration but maybe structural fragility (if central nodes are removed, speperate components)

3.4.8 Community effect (Louvain)

The Louvain community detection method consists of two main steps. Initially, each node is assigned to a separate community. Then, for each node, the algorithm attempts to optimize the modularity of the network by evaluating the potential gain in modularity achieved by moving the node to each of its neighboring communities. If no gain is achieved, the node remains in its original community.

In the second step, a new network is constructed where each node represents a community from the previous step. The edges between the new nodes are weighted by the sum of the weights of the edges between the nodes in the corresponding communities in the original network. The Louvain method is then applied to this new network, and the process is repeated until no further improvement in modularity can be achieved.

Modularity is a measure of the degree of segregation of a network into communities. It is calculated as the difference between the fraction of edges within a given group and the expected fraction of edges if they were randomly distributed in the network. The change in modularity ΔQ achieved by moving the node to each of its neighboring communities is measured as

$$Q = \frac{1}{2m} \sum_{i,j} [A_{ij} - \frac{k_i k_j}{2m}] \delta(c_i, c_j)$$

where:

- Q is the modularity index
- m is the total number of edges in the network
- A_{ij} is the weight of the edge between nodes i and j
- k_i and k_j are the degrees of nodes i and j, respectively
- c_i and c_j are the community assignments of nodes i and j
- $\delta(c_i, c_j)$ is the Kronecker delta, which is equal to 1 if nodes i and j are in the same community and 0 otherwise.

4 Results

4.1 Desciptive Statistics

Bibliometric networks exhibit the following characteristics (Holme and Kim, 2002):

- They are large and sparse graphs, meaning that only a small fraction of all possible edges actually exist (Watts and Strogatz, 1998; Newman, 2003; Jackson, 2010)
- The average geodesic (shortest path) length grows logarithmically (Leskovec et al., 2007)
- The degree distribution approximately follows a power law decay (Barabási and Albert, 1999)
- They exhivit a high degree of clustering (Newman, 2003; Kretschmer, 2004)

4.2 Co-Citation

[Table 1 about here]

4.2.1 Density Effect (Edges)

Both co-citation and co-occurrence graphs are found to be sparse for every discipline, which is indicated by the negative and significant coefficients. A sparse network is characterized by a relatively low proportion of connections between documents in comparison to the total number of potential connections that can exist within the network.

4.2.2 (Triangles)

As suggested by prior research, the social networks under investigation demonstrate a higher prevalence of triangles than what would be expected by random chance. The observed excess of triangles in the current study is consistent with previous findings on the presence of triadic closure in social networks, where nodes tend to form connections with the neighbors of their neighbors (Burt, 2000).

All coefficients that are statistically significant for triangles exhibit a positive sign, which suggests that triangles are present in the social network under investigation at a rate higher than that which would be expected by chance alone.

4.2.3 Small-World effect (Transitivity)

The *transitivity* term is positive and significant, this indicates that the presence of a tie between nodes i and j, and between nodes j and k, increases the likelihood of a tie between nodes i and k, all other things being equal. This can be interpreted as evidence of a "triadic closure" effect, where nodes that are connected to the same neighbor are more likely to form a direct tie between themselves.

4.3 Co-Occurrence

[Table 2 about here]

References

Abbott, Andrew. 2001. *Chaos of Disciplines*. Chicago, IL: University of Chicago Press.

Barabási, Albert-László and Réka Albert. 1999. "Emergence of Scaling in Random Networks." *Science* 286:509–512.

Beam, Elizabeth, L. Gregory Appelbaum, Jordynn Jack, James Moody, and Scott A. Huettel. 2014. "Mapping the Semantic Structure of Cognitive Neuroscience." *Journal of Cognitive Neuroscience* 26:1949–1965.

Borrett, Stuart R., James Moody, and Achim Edelmann. 2014. "The Rise of Network Ecology: Maps of the Topic Diversity and Scientific Collaboration." *Ecological Modelling* 293:111–127.

Bradford, Samuel. 1985. "Sources of Information on Specific Subjects." *Journal of Information Science* 10:176–180.

Burt, Ronald S. 2000. "The Network Structure Of Social Capital." *Research in Organizational Behavior* 22:345–423.

Callon, M., J. Courtial, and F. Laville. 2005. "Co-Word Analysis as a Tool for Describing the Network of Interactions between Basic and Technological Research: The Case of Polymer Chemsitry." *Scientometrics* 22:155–205.

Chavalarias, David and Jean-Philippe Cointet. 2013. "Phylomemetic Patterns in Science Evolution—The Rise and Fall of Scientific Fields." *PLOS ONE* 8:e54847.

Collins, Randall. 1994. "Why the Social Sciences Won't Become High-Consensus, Rapid-Discovery Science." *Sociological Forum* 9:155–177.

Collins, Randall. 1998. "The Sociological Eye and Its Blinders." *Contemporary Sociology* 27:2–7.

Crane, Diana. 1969. "Social Structure in a Group of Scientists: A Test of the "Invisible College" Hypothesis." *American Sociological Review* 34:335–352.

Crane, Diana. 1972. *Invisible Colleges: Diffusion of Knowledge in Scientific Communities*. Chicago: University Of Chicago Press, n edition edition.

De Solla Price, Derek J. and Donald Beaver. 1966. "Collaboration in an Invisible College." *American Psychologist* 21:1011–1018.

Hagberg, Aric A, Daniel A Schult, and Pieter J Swart. 2008. "Exploring Network Structure, Dynamics, and Function Using NetworkX." .

Hjørland, Birger and Jeppe Nicolaisen. 2005. "Bradford's Law of Scattering: Ambiguities in the Concept of "Subject"." In *Context: Nature, Impact, and Role*, edited by Fabio Crestani and Ian Ruthven, Lecture Notes in Computer Science, pp. 96–106, Berlin, Heidelberg. Springer.

Holme, Petter and Beom Jun Kim. 2002. "Growing Scale-Free Networks with Tunable Clustering." *Physical Review E* 65:026107.

Hunter, David R., Mark S. Handcock, Carter T. Butts, Steven M. Goodreau, and Martina Morris. 2008. "Ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks." *Journal of Statistical Software* 24.

Jackson, Matthew. 2010. *Social and Economic Networks*. Princeton University Press.

Jiao, Can, Ting Wang, Jianxin Liu, Huanjie Wu, Fang Cui, and Xiaozhe Peng. 2017. "Using Exponential Random Graph Models to Analyze the Character of Peer Relationship Networks and Their Effects on the Subjective Well-being of Adolescents." *Frontiers in Psychology* 8.

Kossinets, Gueorgi and Duncan J. Watts. 2006. "Empirical Analysis of an Evolving Social Network." *Science* 311:88–90.

Kretschmer, Hildrun. 2004. "Author Productivity and Geodesic Distance in Bibliographic Co-Authorship Networks, and Visibility on the Web." *Scientometrics* 60:409–420.

Kuhn, Thomas S. 2012. *The Structure of Scientific Revolutions*. University of Chicago Press.

Leskovec, Jure, Jon Kleinberg, and Christos Faloutsos. 2007. "Graph Evolution: Densification and Shrinking Diameters." *ACM Transactions on Knowledge Discovery from Data* 1:2–es.

Moody, James and Ryan Light. 2006. "A View from above: The Evolving Sociological Landscape." *The American Sociologist* 37:67–86.

Mullins, Nicholas C. 1973. "The Development of Specialties in Social Science: The Case of Ethnomethodology." *Science Studies* 3:245–273.

Newman, Mark. 2018. *Networks*. Oxford, New York: Oxford University Press, second edition, new to this edition:, second edition, new to this edition: edition.

Newman, M. E. J. 2003. "The Structure and Function of Complex Networks." *SIAM Review* 45:167–256.

Paisley, William. 1972. "The Role of Invisible Colleges in Scientific Information Transfer." *Educational Researcher* 1:5–19.

van Rossum, Guido and Fred L. Drake. 2010. *The Python Language Reference*. Number Pt. 2 in Python Documentation Manual / Guido van Rossum; Fred L. Drake [Ed.]. Hampton, NH: Python Software Foundation, release 3.0.1 [repr.] edition.

Vink, Ritchie, Stijn De Gooijer, Alexander Beedie, J Van Zundert, Gert Hulselmans, Cory Grinstead, Marco Edward Gorelli, Matteo Santamaria, Daniël Heres, IbENPC, Jorge Leitao, Marc Van Heerden, Colin Jermain, Ryan Russell, Chris Pryer, Adrián Gallego Castellanos, Jeremy Goh, Moritz Wilksch, Illumination-K, Joshua Peek, Max Conradt, Liam Brannigan, Yu Ri Tan, Elbaro, Nicolas Stalder, Søren Havelund Welling, Adam Gregory, Jakob Keller, and Paq. 2023. "Pola-Rs/Polars: Python Polars 0.17.2." Zenodo.

Watts, Duncan J. and Steven H. Strogatz. 1998. "Collective Dynamics of 'Small-World' Networks." *Nature* 393:440–442.

Wiecki, Thomas, John Salvatier, Ricardo Vieira, Maxim Kochurov, Anand Patil, Michael Osthege, Brandon T. Willard, Bill Engels, Colin Carroll, Osvaldo A Martin, Adrian Seyboldt, Austin Rochford, Luciano Paz, Rpgoldman, Kyle Meyer, Peadar Coyle, Marco Edward Gorelli, Oriol Abril-Pla, Ravin Kumar, Junpeng Lao, Virgile Andreani, Taku Yoshioka, George Ho, Thomas Kluyver, Kyle Beauchamp, Alexandre Andorra, Demetri Pananos, Eelke Spaak, Benjamin Edwards, and Eric Ma. 2023. "Pymc-Devs/Pymc: V5.3.0." Zenodo.

Table 1: Co-Citation ERGM Model

	Artificial Intelligence	Economics	Ethnic & Cultural Studies	Gender Studies	Genetics & Genomics	Geometry	Geophysics	Human Resources & Organizations	Immunology	International Business	Language & Linguistics	Material Engineering	Neurology	Political Science	Probability & Statistics	Sociology
Density	-3.736***	-4.816***	-1.544*	-2.905***	-4.387***	-2.231***	-4.055***	-5.133***	-3.275***	-4.305***	-2.883***	-4.027***	-4.059***	-3.741***	-4.8***	-3.287***
	(0.532)	(0.57)	(0.673)	(0.682)	(0.463)	(0.604)	(0.447)	(0.462)	(0.425)	(0.515)	(0.508)	(0.468)	(0.47)	(0.46)	(0.596)	(0.523)
Triangles	1.947***	2.308***	1.138**	3.009***	1.987***	1.972***	2.864***	2.197***	1.261**	2.867***	2.093***	3.362***	2.099***	0.895***	2.422**	3.605***
	(0.497)	(0.673)	(0.393)	(0.711)	(0.532)	(0.378)	(0.387)	(0.463)	(0.419)	(0.552)	(0.295)	(0.609)	(0.35)	(0.168)	(0.799)	(0.625)
Stars	-0.042*	0.015*	0.018*	0.056	0.046***	-0.085**	0.008*	0.01***	0.025	0.011***	-0.029*	0.039***	0.015*	0.008***	0.131***	-0.03
	(0.023)	(0.008)	(0.008)	(0.048)	(0.005)	(0.033)	(0.004)	(0.002)	(0.02)	(0.001)	(0.013)	(0.011)	(0.008)	(0.002)	(0.025)	(0.02)
Betweenness	-0.02	-0.009	-0.033	0.002	-0.009	0.001	-0.082	-0.048	0.005	-0.014	-0.006	-0.011	-0.055	-0.144	-0.005	0.014
	(1.022)	(1.005)	(0.984)	(1.024)	(1.002)	(0.988)	(1.003)	(0.984)	(0.982)	(1.003)	(0.991)	(0.993)	(1.019)	(0.996)	(1.01)	(1.019)
Cliques	0.147	-0.685	-0.114	-0.951*	-0.313*	-0.009*	-0.486***	-0.15	-0.091	-0.865*	-0.136**	-1.095***	-0.321**	-0.009	-0.399	-1.051***
	(0.232)	(0.44)	(0.106)	(0.557)	(0.19)	(0.005)	(0.151)	(0.172)	(0.116)	(0.397)	(0.05)	(0.336)	(0.115)	(0.014)	(0.434)	(0.323)
Components	1.698***	2.132***	3.494***	3.449***	3.188***	1.918***	-0.946**	-0.683*	3.236***	1.732***	2.135***	2.116***	1.309**	0.187	3.392***	2.954***
	(0.437)	(0.48)	(0.455)	(0.498)	(0.563)	(0.536)	(0.364)	(0.389)	(0.522)	(0.442)	(0.443)	(0.487)	(0.49)	(0.359)	(0.594)	(0.501)
Louvain	-0.716*	-5.229***	-5.908***	-3.807***	-4.732***	-3.3***	-3.465***	-1.501***	-2.835***	-3.52***	-4.392***	-4.915***	-4.087***	-0.619*	-3.709***	-3.927***
	(0.341)	(0.37)	(0.378)	(0.37)	(0.434)	(0.521)	(0.353)	(0.228)	(0.423)	(0.421)	(0.426)	(0.407)	(0.452)	(0.312)	(0.416)	(0.366)
Date	0.348	0.305	-1.052	0.448	0.518	-0.252	-0.826*	-0.102	0.35	0.282	-0.228	0.924**	-0.092	-0.039	1.357***	0.733*
	(0.411)	(0.426)	(0.655)	(0.378)	(0.36)	(0.562)	(0.448)	(0.382)	(0.305)	(0.397)	(0.468)	(0.365)	(0.411)	(0.423)	(0.389)	(0.442)

Table 2: Co-Occurrence ERGM Model

	Artificial Intelligence	Economics	Ethnic & Cultural Studies	Gender Studies	Genetics & Genomics	Geometry	Geophysics	Human Resources & Organizations	Immunology	International Business	Language & Linguistics	Material Engineering	Neurology	Political Science	Probability & Statistics	Sociology
Density	-5.458***	-5.11***	-4.663***	-4.763***	-5.0***	-5.329***	-4.948***	-4.973***	-4.761***	-4.089***	-5.486***	-5.252***	-4.971***	-5.23***	-5.43***	-7.008***
	(0.305)	(0.341)	(0.289)	(0.284)	(0.238)	(0.35)	(0.251)	(0.326)	(0.295)	(0.241)	(0.315)	(0.357)	(0.274)	(0.351)	(0.28)	(0.495)
Triangles	2.377***	1.409***	1.347***	1.96***	0.952**	-0.476	1.411***	0.65**	0.15	2.84***	2.46***	0.993***	1.459***	0.555	2.393***	0.78
	(0.472)	(0.245)	(0.285)	(0.508)	(0.385)	(0.514)	(0.413)	(0.238)	(0.507)	(0.378)	(0.602)	(0.244)	(0.376)	(0.34)	(0.288)	(0.805)
Stars	0.013***	0.008***	0.008***	0.007***	0.004***	0.003***	0.005***	0.011***	0.006***	0.013***	0.003***	0.007***	0.006***	0.005***	0.011***	0.002***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.0)	(0.0)	(0.001)	(0.001)	(0.0)	(0.003)	(0.0)	(0.001)	(0.001)	(0.001)	(0.002)	(0.0)
Betweenness	-0.441	-0.047	-0.015	-0.014	0.015	-0.005	-0.022	-0.012	-0.002	-0.022	0.001	0.006	0.008	0.0	-0.037	0.029
	(0.985)	(0.981)	(0.992)	(1.002)	(0.999)	(0.981)	(1.006)	(1.001)	(0.998)	(0.99)	(1.0)	(1.004)	(1.021)	(1.007)	(0.987)	(1.002)
Clustering	-0.14	0.128	0.221	0.17	-0.282	-0.289	-0.117	-0.051	0.016	-0.025	-0.062	-0.293	-0.213	-0.126	-0.004	-0.27
	(1.005)	(0.988)	(0.986)	(0.991)	(0.993)	(1.006)	(0.993)	(0.992)	(0.998)	(1.009)	(1.007)	(1.012)	(0.994)	(1.0)	(1.002)	(0.985)
Transitivity	-0.002	0.012	0.003	-0.01	0.009	-0.007	-0.006	-0.007	-0.009	0.016	0.008	0.003	0.0	-0.02	0.012	0.007
	(0.99)	(1.002)	(0.976)	(0.992)	(0.993)	(1.014)	(0.995)	(1.005)	(1.013)	(1.005)	(1.026)	(0.997)	(0.989)	(1.019)	(1.016)	(0.991)
Cliques	-0.462*	-0.072*	-0.086	-0.513*	0.073	0.494**	-0.018	0.042	0.228	-0.965***	-0.399	0.014	-0.214	0.13*	-0.277***	0.424
	(0.212)	(0.042)	(0.075)	(0.232)	(0.151)	(0.185)	(0.166)	(0.043)	(0.225)	(0.18)	(0.296)	(0.037)	(0.142)	(0.064)	(0.058)	(0.394)
Components	-2.081***	-1.655***	-1.405***	-2.189***	-5.229***	-4.051***	-2.206***	-2.063***	-3.224***	-1.618***	-1.03*	-2.271**	-2.407***	-1.613***	-2.966***	-0.02
	(0.366)	(0.441)	(0.419)	(0.447)	(0.534)	(0.668)	(0.409)	(0.42)	(0.734)	(0.45)	(0.485)	(0.856)	(0.753)	(0.511)	(0.337)	(0.803)
Louvain	-0.663***	-0.252*	-0.693***	-0.285	-0.095	-0.108	-0.781***	-0.369*	0.256	-3.844***	-5.334***	-0.301	-3.64***	1.195***	-0.37**	-3.759***
	(0.168)	(0.152)	(0.199)	(0.201)	(0.195)	(0.255)	(0.252)	(0.19)	(0.24)	(0.48)	(0.542)	(0.352)	(0.652)	(0.29)	(0.14)	(0.678)