Network Analysis for the Digital Humanities: Principles, Problems, Extensions

Deryc T. Painter, *Arizona State University*
Bryan C. Daniels, *Arizona State University*
Jürgen Jost, *Max Planck Institute for Mathematics in the Sciences*

Abstract: Traditional historical scholarship struggles to keep up with the rapid pace of modern scientific publication trends. Even focusing on a particular scientific field, the rate of new publications far outpaces even the most studious historian's research capacity. This essay summarizes an approach to this problem that uses computational techniques of network analysis. As a complement to close analysis of particular documents, network analysis can give a large-scale perspective on the history of science, identifying relational patterns across a vast number of documents that might otherwise require an entire career to digest. To demonstrate the power of the approach, the essay applies network theory to a corpus of publications in evolutionary medicine. Four distinct networks, including those focused on authors, keywords, and citations, quickly unearth a range of relevant historical information. The essay illustrates how interpretable historical conclusions are drawn from a variety of quantitative metrics. The aim is to provide an overview of network techniques for historians looking to add robust network analysis to their research repertoire.

BACKGROUND MATERIALS

Making things digital means representing them as bit strings. By itself, this is not yet very helpful. It can serve as a basis to extract structure by computational methods. The questions then are what structure to extract, what structure to expect to emerge from the data, and how to make that structure interpretable for humans. We are in the humanities, after all.

Deryc T. Painter is a postdoctoral researcher at Arizona State University, working in the Laubichler Lab for Computational HPS. A computational historian, he studies innovation, interdisciplinarity, and modeling at varying scales. ASU-SFI Center for Biosocial Complex Systems, Arizona State University, Tempe, Arizona 85281, USA; deryc.painter@asu.edu.

Bryan Daniels is Assistant Research Professor in the ASU-SFI Center for Biosocial Complex Systems at Arizona State University. He investigates the logic of collective behavior in living systems, integrating empirical data with theoretical concepts from statistical physics, model selection, and information theory. He is coediting a special issue of *Theory in Biosciences* entitled "Quantifying Collectivity." ASU-SFI Center for Biosocial Complex Systems, Arizona State University, Tempe, Arizona 85281, USA; bryan.daniels.1@asu.edu.

Jürgen Jost has been a director of the Max Planck Institute for Mathematics in the Sciences in Leipzig since 1996. He is also an external faculty member of the Santa Fe Institute in New Mexico. He works in geometry, analysis, discrete structures, and information theory and on a wide range of complex systems. Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany; jjost@mis.mpg.de.

Isis, volume 110, number 3. 0021-1753/2019/0110-0008$10.00.

To achieve this, we need to start with some structural hypotheses. That is, we need to agree on some class of structures and then see what specific features and properties the particular structure possesses.

A general such class is that of networks. The basic idea is that the system to be analyzed consists of elements that stand in pairwise relations. The relations may simply be qualitative (present vs. absent); they may also be directed (from *A* to *B*, but not necessarily also from *B* to A); and they may also be quantitative—that is, they may possess strengths or weights.

An easy example: data about the email exchanges within a certain group of people. These people constitute the elements or agents—or, in the formal terms of network analysis, the vertices or nodes—of the network. There is a relation—formally, a link or an edge in the network—between *A* and *B* if an email message between them has been recorded. We can just leave it at that and formally analyze the resulting network, or we can decide to represent more details. For instance, we can distinguish whether there has been a message from *A* to *B*, one from *B* to *A*, or both. That would yield a directed network. We can also count how many emails have passed in either direction; assigning that number to the corresponding edge would produce a weighted network.

We can then analyze the network and draw conclusions about the underlying social structure. People that receive many emails may be important, and those that send many emails might be influential. We can look at whether important people preferentially exchange emails with each other or whether they instead communicate with less well connected people. In the first case, the network, or by implication the underlying social structure, would be called assortative; in the second case it is disassortative.

This example is drawn from the social sciences, not by accident or chance, but because social scientists were the first to represent relations in terms of networks systematically and analyze the large-scale structural features of those networks. This started with work of Jacob L. Moreno and Helen H. Jennings in the 1930s. Stanley Milgram pointed out the important "small world" property of social networks. Social networks, while not necessarily very regular, do not follow a simple random connectivity pattern but share a particular structural property. This property is the fact that even though the network may be quite large, when one picks two arbitrary nodes one needs only a relatively small number of edges to get from one to the other. This property was formalized by Duncan J. Watts and Steven H. Strogatz and identified as a nearly universal property of empirical networks, not only social ones. Mark S. Granovetter recognized the importance of weak ties for the cohesion of social networks.[1] A social network may be modularized—that is, it may consist of a couple of subcommunities that are internally well connected but have only relatively few links with other communities. When an agent wants to establish an indirect connection—that is, a connection via intermediaries to another community—he or she typically needs to go through certain other agents to which only weak ties exist. To establish distant connections, then, one should not necessarily approach one's own closest friends but should instead seek certain people that one is only weakly acquainted with. Of course, in hindsight, much of this does not look so surprising; the important point is that these things can not only be detected as universal features of social systems but also quantified via formal network analysis.

CREATING OUR CORPUS

Network analysis in the digital humanities typically begins by defining a specific corpus of texts. This definition may be chosen, for example, to represent a specific scientific field or group of

[1] J. L. Moreno and H. H. Jennings, "Statistics of Social Configurations," *Sociometry*, 1938, *1*(3–4):342–374; Jeffrey Travers and Stanley Milgram, "An Experimental Study of the Small World Problem," *ibid.*, 1969, 32:425–443; Duncan J. Watts and Steven H. Strogatz, "Collective Dynamics of 'Small-World' Networks," *Nature*, 1998, 393:440–442; and Mark S. Granovetter, "The Strength of Weak Ties," *American Journal of Sociology*, 1973, 78:1360–1380.

people or set of ideas. A careful definition of the corpus is nontrivial and is key to creating meaningful conclusions.[2]

Here, we use as an example a corpus of publications from individuals who self-identify as interested in evolutionary medicine. These individuals were identified through the International Society for Evolution, Medicine, and Public Health global directory (EvMed Network) and the editors and contributing authors for two major evolutionary medicine textbooks. We used the Thomson Reuters Web of Science (WoS) to identify and download all available PDFs, resulting in a corpus of 6,456 publications that appeared from 1971 through 2017.[3]

In this essay, we examine a single time period within the overall corpus: the year 2007. We use metadata for each publication gathered from WoS, including author names, institutional affiliations, publication journal, and date. Even using this corpus for a single year, we can construct a variety of useful networks to explore various relations among authors, their publications, the content, and citations. Though we will not focus here on changes over time, the dynamics of a particular corpus can also be studied by concatenating networks in a time series.[4]

As emphasized in other contributions to this Focus section, most corpora require a substantial amount of human attention to reach a state sufficiently clean for analysis.[5] For instance, a surprisingly thorny issue is that of ambiguous mappings from names in the database to actual people. A single scientist may publish using slight variants of the same name, and, conversely, two distinct individuals may carry and publish under the same name. Automated tools can help sort this out, but human effort is currently still necessary. In our corpus, we disambiguated names manually. We should note, though, that it was previously found with a disambiguated corpus that such errors will amount to only a few percent overall.[6]

ORGANIZING DATA AS NETWORKS

After the texts are gathered and disambiguated, it is time to begin organizing the data to create networks.

Organizing data in a network is obviously a process of abstraction. The network represents only a formal structure of relations; it dismisses all their content. Analyzing such data as networks allows researchers to study massive amounts of data and glean novel insights otherwise obfuscated when using traditional techniques, which cannot easily handle the same number of

---

[2] Randi Reppen, "Building a Corpus: What Are the Key Considerations?" in *The Routledge Handbook of Corpus Linguistics*, ed. Anne O'Keeffe and Michael McCarthy (New York: Routledge, 2010), pp. 59–65; Sue Atkins, Jeremy Clear, and Nicholas Ostler, "Corpus Design Criteria," *Literary and Linguistic Computing*, 1992, 7:1–16; and William Crawford and Eniko Csomay, *Doing Corpus Linguistics* (New York: Routledge, 2015).

[3] Randolph M. Nesse, *EvMed Network*, 2018, https://isemph.org/evmednetwork; Wenda R. Trevathan, Elucid O. Smith, and James J. McKenna, *Evolutionary Medicine*, 2nd ed. (Oxford: Oxford Univ. Press, 1999); Peter D. Gluckman, Alan A. Beedle, and Mark A. Hanson, *Principles of Evolutionary Medicine*, 1st ed. (Oxford: Oxford Univ. Press, 2009); and Kai Li, Jason Rollins, and Erjia Yan, "Web of Science Use in Published Research and Review Papers, 1997–2017: A Selective, Dynamic, Cross-Domain, Content-Based Analysis," *Scientometrics*, 2018, *115*:1–20.

[4] Norbert Marwan *et al.*, "Complex Network Approach for Recurrence Analysis of Time Series," *Physics Letters A*, 2009, 373:4246–4254; Reik V. Donner *et al.*, "Recurrence-Based Time Series Analysis by Means of Complex Network Methods," *International Journal of Bifurcation and Chaos*, 2011, *21*:1019–1046; and Yue Yang and Huijie Yang, "Complex Network-Based Time Series Analysis," *Physica A: Statistical Mechanics and Its Applications*, 2008, 387(5–6):1381–1386.

[5] Kenneth D. Aiello and Michael Simeone, "Triangulation of History Using Textual Data"; Julia Damerow and Dirk Wintergrün, "Hitchhiker's Guide to Data in the History of Science"; and Abraham Gibson and Cindy Ermus, "The History of Science and the Science of History: Computational Methods, Algorithms, and the Future of the Field": all in this Focus section.

[6] M. E. J. Newman, "The Structure of Scientific Collaboration Networks," *Proceedings of the National Academy of Sciences*, 2001, 98:404–409; and Albert-Laszlo Barabási *et al.*, "Evolution of the Social Network of Scientific Collaborations," *Physica A*, 2002, *311*:590–614.

people and texts. In social networks, the elements or agents are usually people, but the type of relation or interaction may vary. Moreover, the data usually provide only proxies for the social relations one is really interested in, like friendship. For instance, one may have the number of phone or email exchanges between pairs of people, but not their content. Or one may have coauthorship between scientists as an indication of collaborations or citations of papers as indicators of scientific communities.[7] In many cases such proxies work remarkably well, because statistical fluctuations average out, and there are basic correlations between the underlying relations and their traces in the data.

Usually, there is still some freedom in constructing the network. When we want to represent the semantic structure of a text, we may look at word co-occurrences in sentences, paragraphs, or pages.[8] The words will be the elements of the network, but we still need to decide whether to consider different grammatical forms of the same root as different words. We may set a threshold and link two words in the network if they co-occur at least five times on a page of our text. When we want to have a more detailed representation, we can also consider a weighted network, with the weight of a link given by the number of co-occurrences.

We then come to a fundamental question. When we construct such networks—say, word co-occurrence networks for different texts or different languages—the resulting networks will not be identical. The question then is whether a class of networks from a particular domain, like the coauthorship networks, still typically have some formal properties in common that distinguish them from networks from other domains, like email contacts between students. And, conversely, are there systematic relations between different networks formed by the same class of agents, like citation relations and joint attendance at conferences? Or, for instance, are networks from Indo-European languages more similar to each other than to language networks from other families?

These questions are qualitative in nature, but network analysis can make them quantitative and often visual.

VISUALIZATION

Network analysis still depends on human guidance and interpretation. The network structure and the results of the network analysis have to be made accessible to a human observer. In principle, there is an easy scheme: draw the network as a graph. One chooses positions for the nodes and connects two nodes that are linked in the network by a line.

While in some cases it may be obvious how to place the nodes in such a drawing, in particular when the network is as regular as Network A in Figure 1, this has some pitfalls. Just changing the position of a single node, as in Network B, without changing the connectivity of the network, may make the network look very different. More generally, important structural features of a network, like the community structure, can be blurred or highlighted, depending on how the network is depicted. Therefore, network visualizations should be interpreted with

---

[7] Nini Gu, "Metadata: Reidentification Using Telephone Data Is Easier Than You Think," *Chicago Policy Review (Online)*, 2016; Sarika Sharma *et al.*, "Using an Ethnography of Email to Understand Distributed Scientific Collaborations," *iConference*, 2015; Sameer Kumar, "Co-Authorship Networks: A Review of the Literature," *Aslib Journal of Information Management*, 2015, *67*:55–73; Marjan Cugmas, Anuška Ferligoj, and Luka Kronegger, "Scientific Co-Authorship Networks," arXiv:1711.00770, 2017; Ted A. Skolarus *et al.*, "Assessing Citation Networks for Dissemination and Implementation Research Frameworks," *Implementation Science*, 2017, *12*:97; and Jason Portenoy, Jessica Hullman, and Jevin D. West, "Leveraging Citation Networks to Visualize Scholarly Influence over Time," *Frontiers in Research Metrics and Analytics*, 2017, *2*:8.

[8] Mehri Sedighi, "Application of Word Co-Occurrence Analysis Method in Mapping of the Scientific Fields (Case Study: The Field of Informetrics)," *Library Review*, 2016, *65*(1/2):52–64; and Anne Veling and Peter Van Der Weerd, "Conceptual Grouping in Word Co-Occurrence Networks," *IJCAI*, 1999, *99*:694–701.

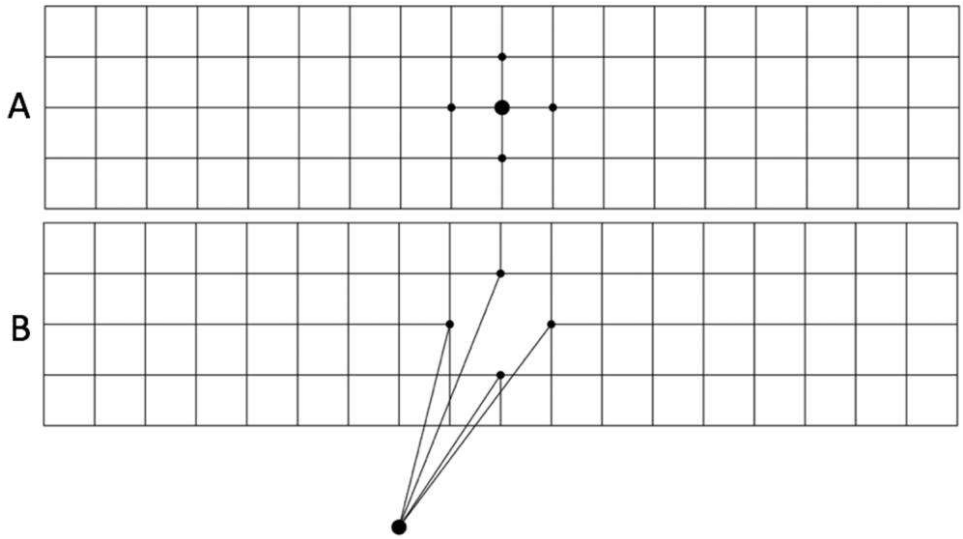

Figure 1. Two grid networks. Network A is a regular grid, with one node and its neighbors emphasized. Network B is the same regular grid with the position of one node changed, preserving all neighborhood relations.

caution.[9] Other representations of a network—for instance, through its curvature or eigenvalue distribution—may be more indirect and abstract but can be readily standardized and therefore allow for a better comparison between different networks.

CONCEPTS AND TOOLS OF NETWORK ANALYSIS

Let us assume, then, that we have organized our data in a network and that we want to extract some qualitative formal properties—or, better, that we have networks for different datasets from the same domain and that we want to extract some common properties of those networks that distinguish them from networks from other domains.

Two networks constructed from different datasets from the same domain are not expected to be identical, but at best similar, and we want to quantify the degree of similarity. The methods typically depend on the comparison of large-scale statistical properties. These statistical features allow for a quick coarse comparison and for the identification of particular features that distinguish networks from a particular domain.

In particular, such questions can be approached by measuring appropriate *statistics*: we will make use of mathematical functions that transform a detailed structure (in this case, a network) into single informative numbers. These can be network-level statistics (e.g., average number of connections per individual), statistics describing the local structure (e.g., the number of connections emanating from a particular individual or connection), or the contribution of a single piece of the network to the global structure (e.g., the centrality of a particular connection).

We consider first a simple coauthorship network, which can be represented by an unweighted and undirected graph. That is, for two nodes A and B, there is either an edge between them, in which case we write $A \sim B$ and call A and B neighbors (indicating that the two authors have published a paper together), or they are not connected (indicating that they have not). The *degree* of a node is defined as the number of its neighbors.

[9] Gibson and Ermus, "History of Science and the Science of History" (cit. n. 5).

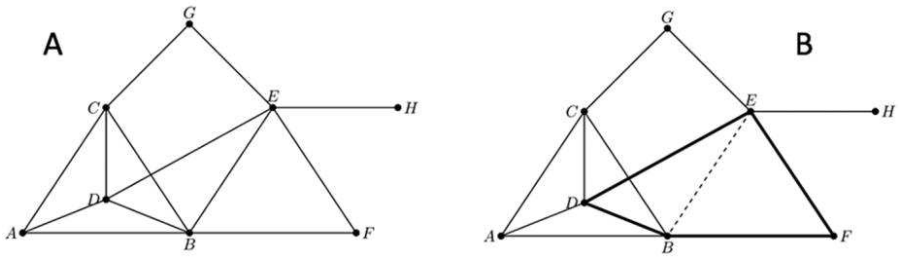

Figure 2. Two similar networks. Network A is a graph with nodes *A*, *B*, *C*, *D*, *E*, *F*, *G*, and *H*. Network B is the same network with the edge (*B*,*E*) cut, emphasizing triangles.

In Network A of Figure 2, for instance, $A \sim B$, but *A* is not connected to *E*, *F*, or *G*. *A* has three neighbors, *B*, *C*, and *D*, and therefore its degree is 3. Nodes *B* and *E* are best connected, having degree 5, whereas *H*, possessing only a single neighbor (*E*), has degree 1. One can then calculate statistics for how many nodes there are of a given degree; in our example, there are two nodes (*F* and *G*) of degree 2 and two nodes (*C* and *D*) of degree 4. For a general network, we denote by $d(n)$ the number of nodes of degree n. It turns out that for empirical networks, this degree distribution typically behaves like a power of n (such that $d[n]$ is proportional to n^{α}), with exponent α between −3 and −2.[10] Thus we have examples of both a node-level property (the degree) and a corresponding network-level property (the degree distribution).

Beyond the degree, which provides information about local connectivity, there is a zoo of network statistics that capture properties both of individual nodes and edges and of larger connectivity patterns. Particularly useful in the context of digital humanities are measures of clustering, assortativity, and centrality. In social and conceptual networks, the neighbors of a particular node are often also neighbors of each other (quantified by the clustering coefficient); nodes with large degree are often more likely to connect to one another (quantified by assortativity measures such as the rich club coefficient); and we are often interested in the nodes and edges that are central in that they connect otherwise disconnected areas of the network (quantified by centrality measures such as betweenness centrality).

Many network statistics are computed with respect to nodes. Yet edges are what make a network a network, and it therefore seems more appropriate to look at a statistic for edge rather than node properties. The simplest such quantity is the so-called Ricci curvature of a graph. Originally introduced by R. Robin Forman in 2003, it has been forged into an efficient tool for network analysis in work by R. P. Sreejith and colleagues in 2007 and by Emil Saucan and colleagues in 2018.[11] The curvature of an edge between two nodes *X* and *Y* is defined as

$$R(X, Y) = 4 - \deg X - \deg Y. \quad (1)$$

The minus signs and the 4 come from the origin of this concept in Riemannian geometry (see, e.g., Jürgen Jost's *Riemannian Geometry and Geometric Analysis*), and for our present

[10] Réka Albert and Albert-László Barabási, "Statistical Mechanics of Complex Networks," *Reviews of Modern Physics*, 2002, 74:47–97.

[11] R. Robin Forman, "Bochner's Method for Cell Complexes and Combinatorial Ricci Curvature," *Discrete and Computational Geometry*, 2003, 29:323–374; R. P. Sreejith *et al.*, "Systematic Evaluation of a New Combinatorial Curvature for Complex Networks," *Chaos Solitons and Fractals*, 2017, *101*:50–67; and Emil Saucan *et al.*, "Discrete Curvatures and Network Analysis," *MATCH Communications in Mathematical and in Computer Chemistry*, 2018, 80:605–622.

purposes we can simply consider them as historical conventions.[12] For Network A in Figure 2, for instance,

$$R(A, B) = -4, R(E, H) = -2, R(D, E) = -5, R(B, E) = -6, \quad (2)$$

and in fact the edge (B, E) has the most negative value of the curvature. We may suspect that this is the most important edge in the network, as it connects two highly connected nodes. This is a general heuristic principle of network analysis.[13] Another empirical finding is that the curvature distribution of a network typically has more than one hump and that this corresponds to a superposition of a structural and a functional network.[14] This feature seems to hold quite generally, from biological to social and linguistic networks, and this merits closer examination.

Returning to the claim that (B, E) is the most important edge for the network: that claim is somewhat mitigated by the observation that even if we cut that edge, we could still easily get from B to E by going through either D or F as a single intermediate step. For instance, if B and E are authors, in the new network they would no longer be direct coauthors but would still share the two coauthors D and F. This is indicated in Network B of Figure 2.

As a simple example of larger connectivity patterns, consider Figure 3. Cutting the edge (A,B) would disconnect the left graph into two pieces with three nodes each, whereas cutting the edge (G,H) in the right graph would have no such effect. One might say that the left graph consists of two modules or communities, (A,C,D) and (B,E,F), that are connected only by the single edge (A,B).

This suggests a general principle of community detection. Try to disconnect the network into large subnetworks by cutting as few edges as possible. The rationale is that a community or module should be well connected within but only sparsely connected to other modules. For larger networks, identifying those edges may become computationally difficult, and there exist many heuristics for that purpose.[15]

2007 EVOLUTIONARY MEDICINE COAUTHORSHIP NETWORK

Let us now return to our evolutionary medicine corpus and examine how statistics at both the local and the network level, as well as clustering techniques for community detection, can help investigators create a large-scale, robust understanding. First, we examine the overall coauthorship network. To visualize the network we use VOSviewer, and we calculate network statistics using Cytoscape and ORA.[16]

The coauthorship network, visualized in Figure 4, contains 397 nodes in 49 separate connected components. The average individual actively worked with approximately 8 other coauthors, a statistic that describes the research practices from fields contributing to evolutionary medicine.[17]

---

[12] Jürgen Jost, *Riemannian Geometry and Geometric Analysis* (Cham: Springer, 2017).

[13] See Sreejith *et al.*, "Systematic Evaluation of a New Combinatorial Curvature for Complex Networks" (cit. n. 11).

[14] Saucan *et al.*, "Discrete Curvatures and Network Analysis" (cit. n. 11).

[15] Santo Fortunato, "Community Detection in Graphs," *Physics Reports*, 2010, *486*(3–5):75–174.

[16] Nees Jan Van Eck and Ludo Waltman, *VOSviewer Manual* (Leiden: Univ. Leiden, 2009); Paul Shannon *et al.*, "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks," *Genome Research*, 2003, *13*:2498–2504; and Kathleen M. Carley, "ORA: A Toolkit for Dynamic Network Analysis and Visualization," in *Encyclopedia of Social Network Analysis and Mining*, ed. Reda Alhajj and Jon Rokne (New York: Springer, 2018), pp. 1–10.

[17] Barabási *et al.*, "Evolution of the Social Network of Scientific Collaborations" (cit. n. 6); Newman, "Structure of Scientific Collaboration Networks" (cit. n. 6); and M. E. J. Newman, "Coauthorship Networks and Patterns of Scientific Collaboration," *Proc. Nat. Acad. Sci.*, 2004, *101*(suppl. 1):5200–5205.

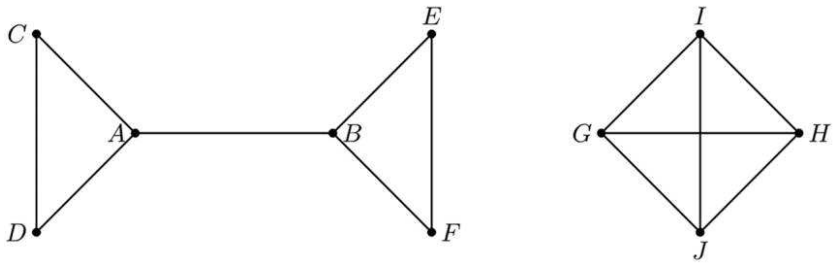

Figure 3. Connectivity patterns. Two networks with structurally different edges and the same *R*.

An immediate benefit of a network representation is the number of basic metrics quickly available. For instance, ranking authors by degree identifies Dan J. Stein as the author who collaborated with the most individuals, 94. Combining the coauthorship network with data about the number of publications and citations for each author (indicated by the thickness of edges and size of nodes in Figure 4) begins to reveal other key authors and publications. Keith D. Lindor and Stein both published more than any other author from our corpus, with 21 separate publications, followed by Jacobus J. Boomsma, with 14 publications.

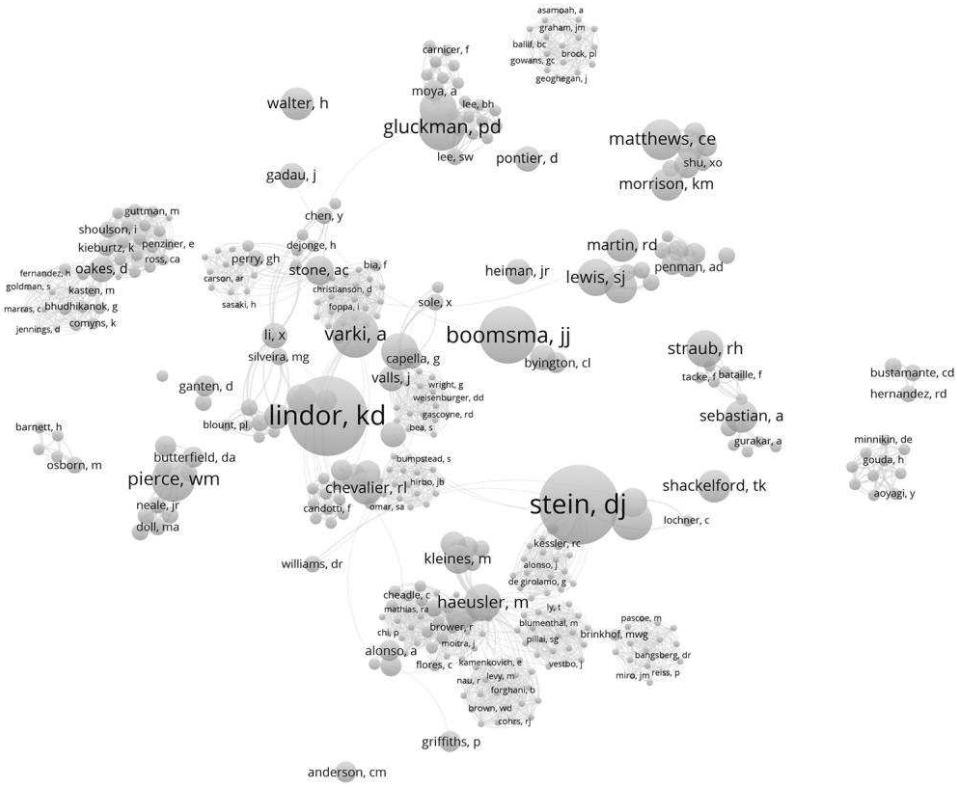

Figure 4. The 2007 evolutionary medicine coauthorship network. The size of the labels and nodes is scaled to the number of documents for the individual. The network image excludes authors with fewer than 5 publications in 2007, and nodes are selectively labeled to increase visibility. A thicker edge represents multiple papers coauthored by the two individuals.

Collaborative groups are shown by clusters in a coauthorship network.[18] Betweenness centrality can bring out this structure, measuring the extent to which an individual acts as a bridge between otherwise isolated clusters.[19] Quantifying a node's betweenness centrality involves the idea of a shortest path: given any two nodes s and t, there is at least one shortest path that connects them by moving along edges of the networks. The betweenness centrality of a node v is then defined as the fraction of all of the network's shortest paths that pass through node v:

$$b_v = \sum \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{3}$$

where σ_{st} represents the total number of shortest paths between nodes s and t, $\sigma_{st}(v)$ represents the number of those paths that pass through v, and the sum runs over all possible other nodes s and t.

The complete coauthorship network indicates that Kathleen Carole Barnes connected two well-established collaborative group clusters. The network also indicated that Sir Peter D. Gluckman and Dyanne Wilson were central in connecting several working-group clusters with their work on the effect of in vitro fertilization on childhood growth.

The Forman-Ricci curvature can similarly be used to identify important edges by finding those that, if removed, would greatly lessen the flow of information. Calculating the Forman-Ricci curvature of the edges in Figure 4, we find that the edge connecting Rasmus Nielsen and Melissa J. Hubrisz, a pair that published together 7 times in 2007, has the most negative curvature, −1165. Forman-Ricci curvature is significantly negatively correlated to edge betweenness centrality.[20] Thus we infer that he cooperation between Nielsen and Hubrisz connects disparate communities in the network. Also, while computing betweenness centrality is expensive, as all paths need to be evaluated, Forman-Ricci curvature as a local quantity is very easy to compute, and it is often a good proxy for betweenness centrality.

PARSING COAUTHOR RELATIONSHIPS USING METADATA

Identifying important individuals and publications in the whole of evolutionary medicine provides a macroscopic field-wide view. We can also gain insight by limiting our scope and parsing the corpus into smaller pieces. In particular, the interdisciplinary nature of evolutionary medicine lends itself to analysis of partial coauthorship networks. This allows for comparison between networks as well as the discovery of specific types of collaborations.

Evolutionary medicine is the combination of two distinct scientific disciplines, evolutionary biology and human health and disease. Evolutionary medicine began with the publications "The Dawn of Darwinian Medicine" and *Why We Get Sick: The New Science of Darwinian Medicine*, defining a new field that brought the core tenets of evolution into human health and disease.[21] In order to evaluate this endeavor, we categorized authors and the journals containing their publications. Each individual from the EvMed Network was assigned to one of three categories, "evolution" or "medicine" or "other," depending on his or her professional interests.

---

[18] Wolfgang Glänzel and Andras Shubert, *Measuring and Evaluating Science–Technology Connections and Interactions* (Dordrecht: Elsevier, 2005), pp. 695–716; and Eldon Y. Li, Chien Hsiang Liao, and Hsiuju Rebecca Yen, "Co-Authorship Networks and Research Impact: A Social Capital Perspective," *Research Policy*, 2013, *42*:1515–1530.

[19] M. Barthélemy, "Betweenness Centrality in Large Complex Networks," *European Physical Journal B*, 2004, *38*:163–168; Stephen P. Borgatti, "Centrality and Network Flow," *Social Networks*, 2005, *27*:55–71; and M. E. J. Newman, "A Measure of Betweenness Centrality Based on Random Walks," *ibid.*, pp. 39–54.

[20] Sreejith *et al.*, "Systematic Evaluation of a New Combinatorial Curvature for Complex Networks" (cit. n. 11).

[21] George C. Williams and Randolph M. Nesse, "The Dawn of Darwinian Medicine," *Quarterly Review of Biology*, 1991, *66*:1–22; and Nesse and Williams, *Why We Get Sick: The New Science of Darwinian Medicine* (New York: Vintage, 2012).

Individuals with a clinical career focus were assigned to "medicine," and researchers with a primary focus on evolutionary biology were assigned to "evolution." Remaining scientists and members of the general public were categorized as "other." Similarly, journals were assigned a category ("evolution," "medicine," "general interest," or "other") on the basis of their assumed reader base. Then each publication could be categorized, for instance, as appearing in an evolution journal while its authors specialized in medicine.

In 2007, individuals with an evolution background published 219 times in evolution journals and 238 times in medical journals. By isolating individuals from the EvMed Network with an evolution background who published in medical journals, we identified individuals who were bringing the principles of evolution into the medical domain. This illustrates how multiple questions are addressed using one corpus.

Dan J. Stein published more than any other author, with 18 publications in medical journals. He was followed by Sir Peter D. Gluckman, Suzanna Lewis, Charles E. Matthews, and William M. Pierce, Jr., with 8 publications each. Stein also accrued the most citations, with 624, followed by Lewis with 459 citations. Additionally, Stein collaborated with the most researchers, 72. He also collaborated with Sing Lee, a professor at the University of Hong Kong, which produced the smallest Forman-Ricci curvature, –96.

KNOWLEDGE MAPS FOR EVOLUTIONARY MEDICINE

Looking deeper into our dataset, we can move beyond the authors to explore the ideas presented in each publication.

Keyword analysis offers a simple way to capture words and short phrases that appear most frequently. Connecting keywords into a network of those that appear in the same publication gives insight into the landscape of knowledge.

The keyword co-occurrence network (KCN) in Figure 5 was created using publications from our partial "evolution to medicine" coauthorship network (discussed in the previous section) and identifying keywords in each document. Keywords are defined as words that appear significantly more frequently than in a relevant reference corpus, here the *Baker-Brown Corpus for General American English*. For this study, we used WordSmith Tools.[22] For an in-depth discussion of WordSmith Tools, please refer to the section "Computational Responsibility" in "Triangulation of History Using Textual Analysis," in this Focus section.

In a KCN, the nodes represent unique keywords and the edges between represent a shared document. KCNs identify large-scale ideas and, if analyzed in a time series, trends in those ideas.

By definition, keywords from the same document form a completely connected subnetwork, or clique. Sangyoon Yi and Jinho Choi showed in 2012 that keyword networks might not exhibit a small-world network structure but instead tend toward locally clustered, scale-free networks.[23] This is different from coauthorship networks, whch tend toward small-world, scale-free networks.[24] When two or more documents share a keyword, these cliques become loosely connected, thus creating several subgraphs connected through a shared language.

---

[22] Paul Baker, *Baker-Brown Corpus for General American English* (Edinburgh: Edinburgh Univ. Press, 2006); Mike Scott, *WordSmith Tools Version 5* (2008), p. 122; and Aiello and Simeone, "Triangulation of History Using Textual Data" (cit. n. 5).

[23] Sangyoon Yi and Jinho Choi, "The Organization of Scientific Knowledge: The Structural Characteristics of Keyword Networks," *Scientometrics*, 2012, 90:1015–1026.

[24] Watts and Strogatz, "Collective Dynamics of 'Small-World' Networks" (cit. n. 1); L. A. N. Amaral *et al.*, "Classes of Small-World Networks," *Proc. Nat. Acad. Sci.*, 2000, 97:11149–11152; and Albert-László Barabási and Eric Bonabeau, "Scale-Free Networks," *Scientific American*, 2003, 288(5):60–69.

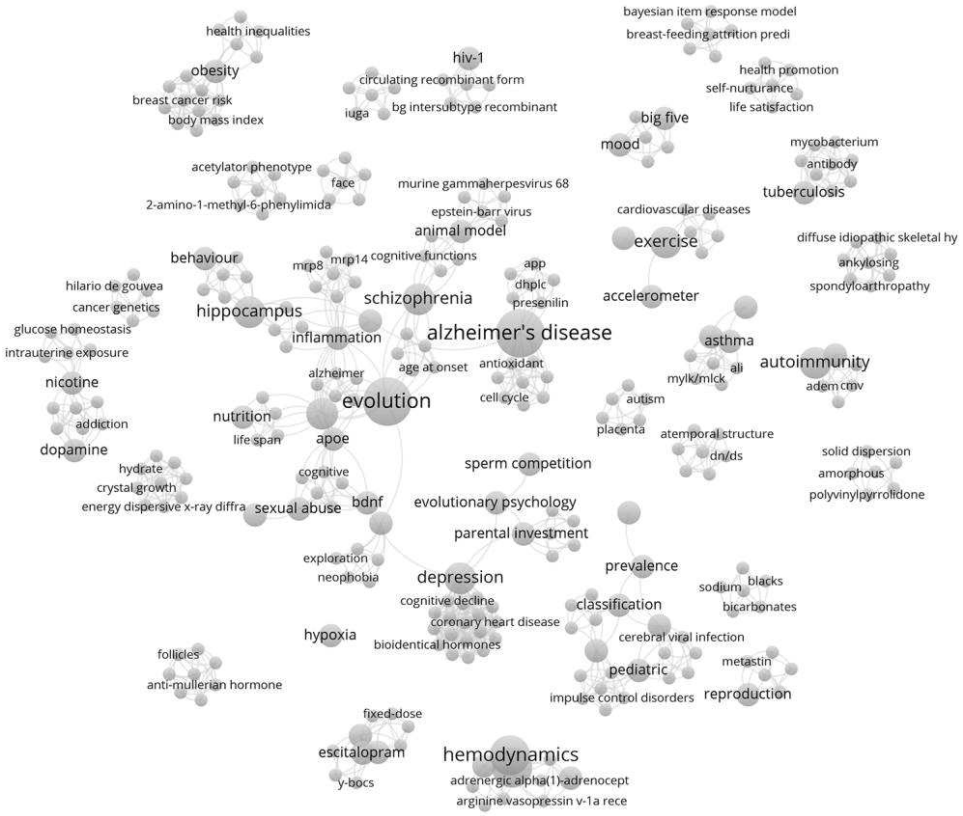

Figure 5. The 2007 keyword co-occurrence network for evolutionary biologists publishing in medical journals. The size of the labels and nodes is scaled to the number of occurrences of the keyword. The network image is selectively labeled to increase visibility. The thickness of the edges represents the number of shared documents. This network contains keywords from publications in a medical journal by authors who focus primarily on evolution, the same individuals mentioned in the section "Parsing Coauthor Relationships Using Metadata."

In the "evolution to medicine" KCN shown in Figure 5, "Alzheimer's disease" and "evolution" occur the most frequently, with 5 occurrences, and they are followed by "hemodynamics" and "rats" (not labeled), both with 4 occurrences. This indicates that a large number of individuals interested in evolution are publishing about Alzheimer's disease in medical journals. "Antibody," "biosynthesis," "HIV," "macrophage," "mycobacterium," "regulatory," "serodiagnosis," and "t cell" had the highest citation count, with 198 each. "Alzheimer's disease" occurs the most and averages 46.8 citations per occurrence within the corpus. The number of occurrences illustrates the level of attention from individuals in the EvMed Network, and the number of citations quantifies importance within the broader scientific community.

"Exercise" and "Alzheimer's disease" have the highest betweenness centrality, with 0.708 and 0.655, respectively. A high betweenness centrality in a KCN indicates that a node creates a bridge between two separate parts of the network. Bridge keywords connect different knowledge areas and create conceptual connections between major research themes using common language. By examining the Forman-Ricci curvature of edges, we identify the "evolution"– "expression" and "evolution"–"behavior" edges as having the most negative curvature values, –778

and –682, respectively. This confirms the assumption that evolution is important to the cohesion of evolutionary medicine as a field.

BIBLIOGRAPHIC COUPLING NETWORK

Up to this point, our study has focused on how individuals and keywords relate to one another. In the final two sections, we focus on how publications relate to one another, using two different kinds of networks. Bibliographic coupling and co-citation networks are similar because they both use citation as a means to measure similarity. Bibliographic coupling networks represent publications as nodes and a shared citation in their bibliographies as an edge. Co-citation networks represent the references from those publications as nodes, with edges representing publications that cite both references. That is, in the bibliographic network two papers are linked if they cite the same paper, whereas in the co-citation network two papers are linked when they are cited by the same paper. As we show, the differences in how these networks use citations create different descriptions of evolutionary medicine.

First, we examine the bibliographic coupling network, a type of knowledge map.[25] A cluster of publications all tightly linked together indicates a level of shared knowledge. This could be shared citations of an experimental technique, bedrock citations in a particular field, or different insights built on a similar foundation of citations. Core areas of a bibliographic coupling network represent the status quo. The denser regions of the network imply a level of acceptance or standards between the nodes. A forthcoming study shows that innovative publications are more likely to appear in the loosely connected areas of a bibliographic coupling network.[26] The loosely connected publications (bottom left in Figure 6) contain more unique bibliographic citations than the core (right side in Figure 6), therefore increasing the possibility of innovation by building on unique work. By incorporating fewer ubiquitous citations and including unique citations, publications are able to incorporate new knowledge areas into existing schools of thought.

In Figure 6, "Gluckman (2007b)" is connected to "Godfrey (2007)." The two publications share 22 citations in common. In contrast, "Gluckman (2007b)" and "Ellison (2007)" share only 3 citations. On the sole basis of how many citations their bibliographies have in common, the Gluckman and Godfrey publications appear to be more similar to each other than the Gluckman and Ellison publications. When the two pairings are examined more closely, it is revealed that both the Gluckman and the Godfrey papers involve early life development and its relationship to disease origin. Meanwhile, the Gluckman and Ellison articles share only the commonality of disease in a general sense. Thus bibliographic coupling is a relatively quick and accurate measure of similarity between two documents.[27] Clearly, the measure is not perfect: two documents might still be similar despite citing different literature.

The Ricci curvature helps to identify connections bridging tightly connected regions. We discovered that "Gluckman (2007)" and "Boomsma (2007)" exhibited a Forman-Ricci curvature

---

[25] M. M. Kessler, "Bibliographic Coupling between Scientific Papers," *American Documentation*, 1963, *14*:10–25; and Dangzhi Zhao and Andreas Strotmann, "The Knowledge Base and Research Front of Information Science, 2006–2010: An Author Co-Citation and Bibliographic Coupling Analysis," *Journal of the Association of Information Science and Technology*, 2018, 85:348–357.

[26] Deryc T. Painter, Bryan C. Daniels, and Manfred D. Laubichler, "Innovations Are Disproportionately Likely in the Periphery of a Scientific Network," *Theory in Biosciences* (forthcoming).

[27] Rey Long Liu, "A New Bibliographic Coupling Measure with Descriptive Capability," *Scientometrics*, 2017, *110*:915–935; Kevin W. Boyack and Richard Klavans, "Co-Citation Analysis, Bibliographic Coupling, and Direct Citation: Which Citation Approach Represents the Research Front Most Accurately?" *Journal of the American Society for Information Science and Technology*, 2010, *61*:2389–2404; and Thierson Couto *et al.*, "A Comparative Study of Citations and Links in Document Classification," in *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries* (2006), pp. 75–84.

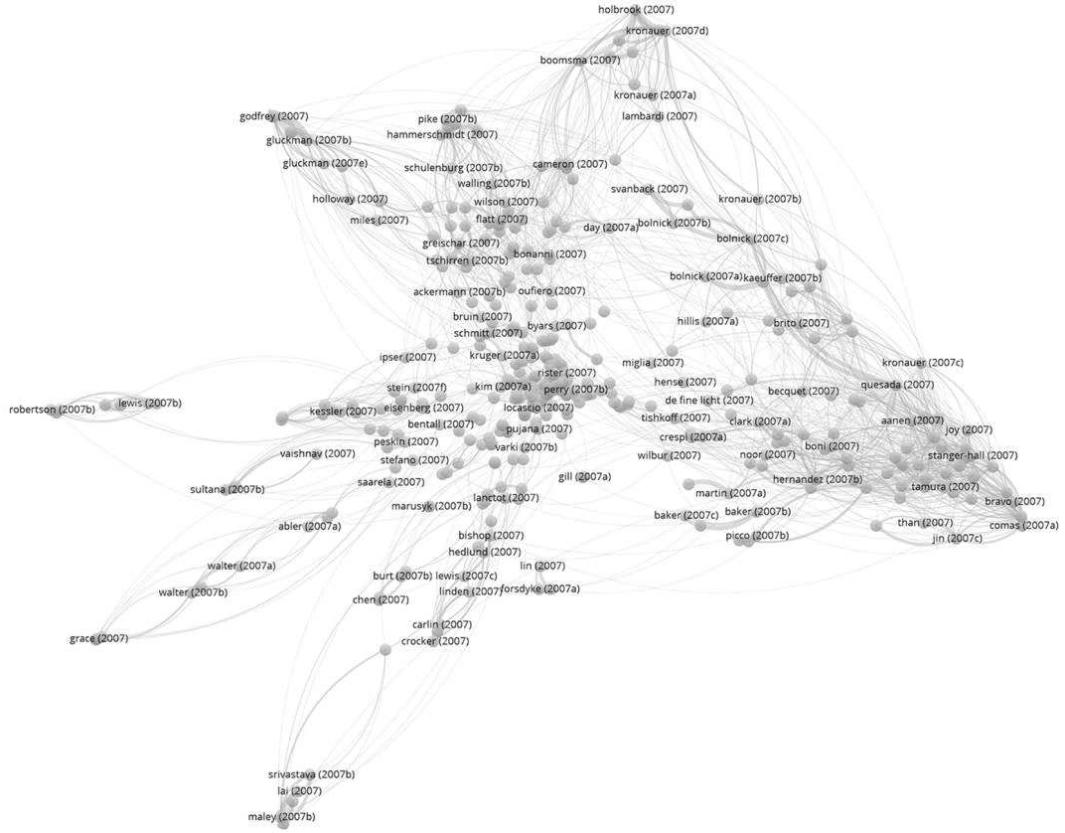

Figure 6. Bibliographic coupling network for the 2007 evolutionary medicine corpus. The size of the nodes and labels is scaled to the number of citations received by each publication. The network is selectively labeled to increase visibility. Thicker edges indicate more shared citations in the bibliographies. The right side of the network shows an example of a highly connected core (or rich club), and the bottom left is the periphery.

of –62. "Comas (2007)" and "Coscolla (2007)" also had –62. Ricci curvature has not been systematically studied in relation to innovation. This is an area that requires further attention. The negative curvature might indicate that these two edges connect two otherwise loosely connected subgraphs, making them important for the cohesive knowledge structure of evolutionary medicine.

Bibliographic coupling networks offer insight into the thinking of the authors as expressed in their choice to cite particular papers. The negative Forman-Ricci curvatures show us the edges of these schools of thought. The networks also illuminate properties of the community as a whole, differentiating works into core versus peripheral areas.

CO-CITATION NETWORK

In a co-citation network, each node represents a publication cited within the corpus, and an edge between nodes indicates that both are cited in the same corpus publication. In this way, co-citation edges represent the evolutionary medicine publications connecting two pieces of foundational knowledge.

These cited publications are the foundation of knowledge on which the ideas in the corpus were based. With metrics like betweenness centrality and clustering coefficient, we use our co-citation network to identify foundational knowledge areas.[28]

The clustering coefficient estimates the likelihood that a particular node will have shared neighbors. Nodes with a large clustering coefficient are locally part of a highly connected clique, while those with a small clustering coefficient connect neighbors that are not themselves connected. Specifically, the clustering coefficient for node v is defined as

$$C(v) = \frac{\text{number of closed triplets involving } v}{\text{total number of triplets involving } v}, \tag{4}$$

where a "triplet" is any set of three nodes and a "closed triplet" forms a triangle in which all three edges between the three nodes are present. This measure can be used in a co-citation network to identify areas of similar content.

Figure 7 displays the co-citation network for the 2007 EvMed corpus. In the displayed image, we constrain the network to publications cited a minimum of 50 times in the corpus. The image contains 404 nodes with 7,093 connections, while the unthresholded, complete co-citation network boasts 22,231 nodes. This is substantially larger than the 618 documents from the original corpus.

The complete co-citation network reveals that "Thompson et al. (1994)" was cited by the most publications in the corpus, 17, followed by "Posada and Crandall (1998)" and "Thompson et al. (1997)," with 16 and 11, respectively. These three publications deal primarily with bioinformatics tools and techniques. The Forman-Ricci curvature allows us to identify edges that connect different groups. The edge between "Posada (1998)" and "Thompson (1994)" had the most negative Forman-Ricci curvature, –270. "Posada (1998)" and "Tajima (1989)" followed, with the edge at –265. Here, then, we see "Posada (1998)" with multiple important edges. It is difficult to credit any one publication from the original corpus for these edges, as multiple articles are responsible for both edges. "Schwander et al. (2005)" has the highest betweenness centrality. While betweenness centrality can identify important foundational bridge

[28] Matteo Pontecorvi and Vijaya Ramachandran, "A Faster Algorithm for Fully Dynamic Betweenness Centrality," *Journal of Mathematical Sociology*, 2011, 25:163–177; and Watts and Strogatz, "Collective Dynamics of 'Small-World' Networks" (cit. n. 1).

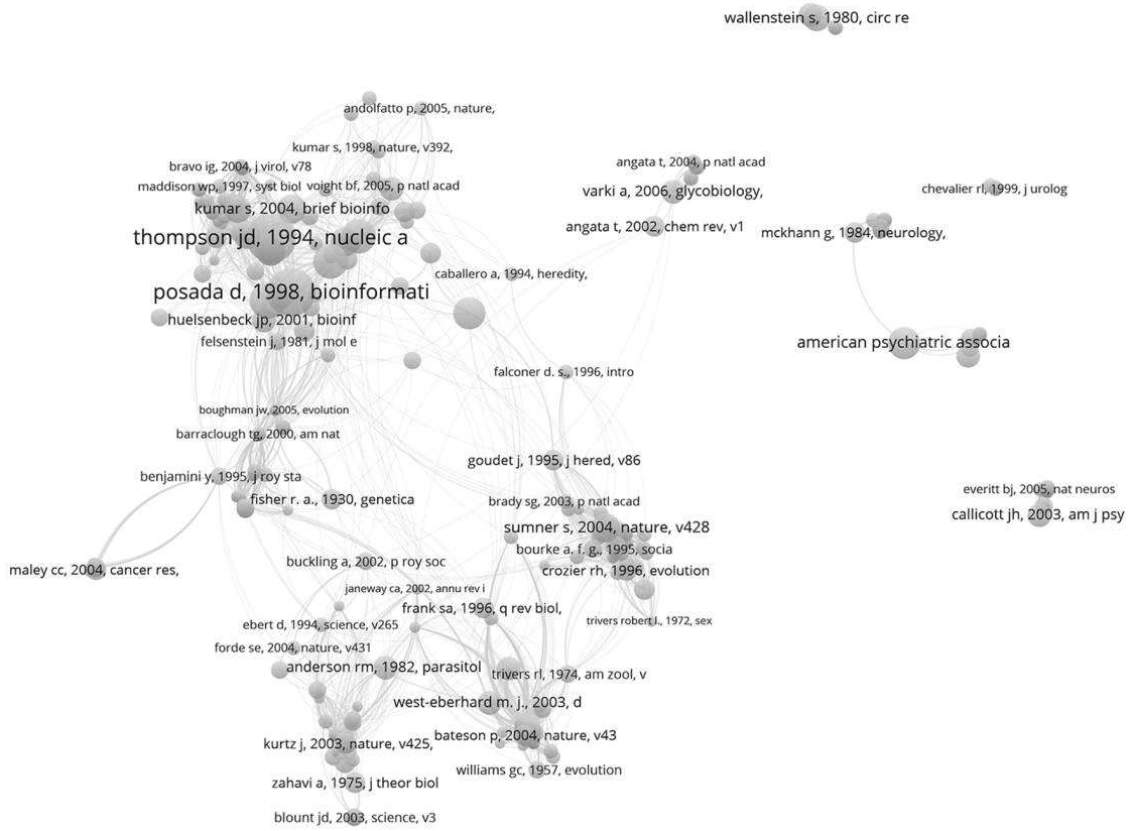

Figure 7. Co-citation network for the 2007 evolutionary medicine corpus. The size of the nodes and labels is scaled to the citation count in the 2007 evolutionary medicine corpus. The network is selectively labeled to increase visibility. The thickness of the edges represents the number of times each pair was cited together. The dense areas are examples of foundational knowledge areas within evolutionary medicine.

nodes, the Forman-Ricci curvature not only identifies bridge edges but can also be used to identify foundational knowledge communities.

CONCLUSION

Many data can be represented as networks. These structures can then be formally analyzed to extract important qualitative features, like cohesion versus modularity, distribution of triangles and other motives, symmetries, node duplications, and so on. It then remains to interpret those features. One will typically find that networks from different domains can be readily distinguished through particular formal properties, like the distribution of curvatures, degrees, or eigenvalues.[29]

Our intention is to illustrate the diversity of insight and utility we can achieve when we examine a well-constructed corpus using a network-based approach. Using a carefully cleaned corpus of publications in evolutionary medicine, we highlighted four distinct networks that unearth a range of historical information. The overall coauthorship network (Figure 4) identifies key individuals in evolutionary medicine from 2007, including the most prolific publishers and collaborators as well as some of the most-cited individuals. The section "Parsing Coauthor Relationships Using Metadata" explains that by sorting the corpus through metadata, we are able to isolate interdisciplinary ties that are of particular interest in bringing together the two halves of evolutionary medicine. This focuses on individuals with an evolutionary biology background who publish in medical journals, illuminating individuals whose research and publishing practices closely align with those within evolutionary medicine.

Networks can also be used as a starting point for understanding textual content. Figure 5 tracks significant keywords from the publications mentioned in "Parsing Coauthor Relationships Using Metadata," providing a conceptual overview of how evolutionary biology was entering medicine in 2007. A keyword co-occurrence network connects these keywords on the basis of shared publications and produces clusters of keywords around larger-scale ideas central to the field. Keywords that bridge the larger clusters can be identified using betweenness centrality and Ricci curvature.

Finally, bibliographic data can be used to identify specific content areas. A bibliographic coupling network (Figure 6) illustrates the conceptual similarity between publications on the basis of the number of citations their bibliographies share. This network representation groups and sorts the publications by similarity without the time-intensive process of reading every publication. It also identifies more mainstream publications as those that share a larger number of citations within a tightly connected core group. A co-citation network (Figure 7) links publications on the basis of their being cited together in the same publication. This groups foundational documents into the knowledge map that gave rise to the corpus.

These are by no means the only kinds of graphs available to historians and philosophers of science. There are organizational networks that link individuals within institutions and social networks that link conference attendees.[30] It seems that such large-scale comparisons have not yet been systematically analyzed from a formal perspective.

---

[29] Anirban Banerjee and Jürgen Jost, "On the Spectrum of the Normalized Graph Laplacian," *Linear Algebra and Its Applications*, 2008, *428*(11–12):3015–3022; and Banerjee and Jost, "Graph Spectra as a Systematic Tool in Computational Biology," *Discrete Applied Mathematics*, 2009, *157*:2425–2431.

[30] Gautam Ahuja, Giuseppe Soda, and Akbar Zaheer, "The Genesis and Dynamics of Organizational Networks," *Organization Science*, 2012, 23:434–448; and Alvin Chin *et al.*, "Using Proximity and Homophily to Connect Conference Attendees in a Mobile Social Network," in *Proceedings of the 32nd International Conference on Distributed Computing Systems Workshops* (Macau, China: IEEE, 2012), pp. 79–87.

And, of course, one also wishes to analyze how particular networks change over time. For instance, with yearly resolved data, we could study how the two fields of evolutionary biology and medicine come together by creating links in the various networks, check whether the two communities then form their own intrinsic clusters at the expense of cross-community connections, determine how new actors enter the field and how new topics emerge, and so on. Such work can also benefit from tools from the theory of dynamical systems, as described in Jürgen Jost's *Dynamical Systems*.[31]

The purpose of this study is to provide examples of how a network approach to the history of science can enrich our understanding. Computational history allows historians to add quantitative analysis to an already rich historical tradition. It is our sincerest hope that after reading this essay, more historians will want to include network analysis in their research toolkit as a primary or supplementary means of investigation.

[31] Jürgen Jost, *Dynamical Systems: Examples of Complex Behavior* (Berlin: Springer, 2005).