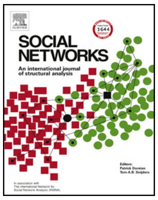
Social Networks 43 (2015) 73–85

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet



Weaving the fabric of science: Dynamic network models of science's unfolding structure

Feng Shi[a], Jacob G. Foster[b], James A. Evans[a,c,*]

[a] Computation Institute, University of Chicago, 5735 S Ellis Ave, Chicago, IL 60637, USA
[b] Department of Sociology, University of California Los Angeles, 375 Portola Plaza, Los Angeles, CA 90095, USA
[c] Department of Sociology, University of Chicago, 1126 East 59th Street, Chicago, IL 60637, USA

ARTICLE INFO

Keywords:
Link prediction
Hypergraphs
Random walks
Multi-mode networks
Science studies
Metaknowledge

ABSTRACT

Science is a complex system. Building on Latour's actor network theory, we model published science as a dynamic hypergraph and explore how this fabric provides a substrate for future scientific discovery. Using millions of abstracts from MEDLINE, we show that the network distance between biomedical things (i.e., people, methods, diseases, chemicals) is surprisingly small. We then show how science moves from questions answered in one year to problems investigated in the next through a weighted random walk model. Our analysis reveals intriguing modal dispositions in the way biomedical science evolves: methods play a bridging role and things of one type connect through things of another. This has the methodological implication that adding more node types to network models of science and other creative domains will likely lead to a superlinear increase in prediction and understanding.

1. Introduction

Science can be viewed as a complex system (Foote, 2007; Evans and Foster, 2011). It is built up from strong interactions between diverse, differentiated components and manifests emergent, often unexpected collective behavior at all scales: periods of incremental effort punctuated by bursts of controversy or transformation. This recent characterization of science is strikingly similar to the one proposed by Bruno Latour, Michel Callon and others in work going back to the 1980s. In their conception, science is a complex, dynamic network in which scientists, institutions, concepts, physical entities and forces "knit, weave and knot" together (Latour, 1987, p. 94) into an overarching scientific fabric (Latour, 1987, 1999; Latour and Woolgar, 1986; Callon, 1986). In Latour's view, components of the network can stabilize over time into social or natural *things*[1]—nodes (or groups of nodes) that become more "fact-like" as they become more tightly coupled to other nodes at the center of a techno-scientific network.

Latour's work focuses on the politics of things and thing-making, but in doing so clarifies the fundamentally multi-mode character of scientific networks. After Latour, any single-mode view, focused only on co-authorship networks between scientists (Newman, 2001, 2004; Martin et al., 2013), or co-occurrence networks between concepts (Foster et al., 2013), must be understood as partial and provisional. In this paper, we argue and then empirically demonstrate that the networks described by Latour do more than trace the past politics of science; they act as a substrate for future scientific discovery. This perspective immediately enriches and extends a classic network-oriented perspective on human problem solving. Newell and Simon (Newell and Simon, 1972) describe problems as situated in a "network of possible wanderings," through which a problem solver may seek a solution (p. 51). By wandering across conceptual links in the network, the solver can collect, imagine, or assemble parts of a solution—or the ingredients of a scientific hypothesis. Consider the many paths available once the network of science is enriched along Latourian lines: A scientist could conjecture that two proteins interact within a human cell because she has seen them in the same or adjacent research

* Corresponding author at: Sociology Department, University of Chicago, 1411 E. 54th Place, Chicago, IL 60637, USA. Tel.: +1 773 834 3612.
E-mail address: jevans@uchicago.edu (J.A. Evans).

[1] In *Making Things Public*, Latour points out that the old word "Thing" originally designated a type of archaic assembly, as the Icelandic Althing: "Thus, long before designating an object thrown out of the political sphere and standing there objectively and independently, the Ding or Thing has for many centuries meant the issue that brings people together because it divides them" (p13) (Weibel and Latour, 2005). Although Latour typically calls nodes in the network actors or "actants" (nonhuman things), we use the term "thing" to generically reference them all.

http://dx.doi.org/10.1016/j.socnet.2015.02.006

articles; because they have been studied by the same scientist; because they react with the same small molecule; because they are implicated in the same disease; or because they can be isolated or analyzed with the same method. In this way, the complex network of science provides a rich substrate on which scientists "think".

Here we apply this perspective to the multi-mode network of biomedicine. We first map the complex web of scientists, chemicals, diseases, and methods, and provide a descriptive account of the ways in which things combine in published biomedical research. Then we ask how network structure determines how the field of biomedical science evolves. More concretely, we investigate whether the linkages between biomedical "things" inscribed by scientific articles can predict the formation of new ties in the network. This is no small task: there are many reasons for two things to be connected! To give one example, two scientists who have never coauthored a paper and who study disparate topics with disjoint methods may nevertheless write a paper together because one joins the other's institute. Links of this kind are hard to predict without the relevant information; indeed, in this paper we exclude institutions from our analysis. Moreover, as we show below, the majority of new links actually occur between things that are "near neighbors" in the network of scientists, chemicals, diseases, and methods. This raises an important question: are there particular paths in the network of possible wanderings—particular forms of proximity—that make the formation of new ties more likely? In other words, are there dispositions that channel scientists' exploration of this complex network?

Before turning to our analysis, we note one further complication with immediate consequences for our representation strategy. James March, a colleague and coauthor of Herbert Simon, championed a distinct theory of problem solving—the "garbage-can model" (Cohen et al., 1972)—in which problems and solutions are mixed randomly (i.e., in the garbage can). Solutions that happen to "stick to" nearby problems are deemed successful. The garbage-can model suggests the need to go beyond the standard network representation, in which things are connected dyadically to other, related, things. According to this alternative view, science is not just a network of dyadic ties; it is also collection of garbage cans (i.e., research projects leading to research articles). Research articles draw together *groups* of things that have stuck—authors, methods, chemicals, diseases (and occasionally garbage). The outcome of this assembly process cannot be accurately represented by projecting the group gathered by an article onto a unipartite network of things, i.e., connecting two things if they appear together in the same article. This representation loses precious information about the context of their co-appearance, the gathering that brought them together. The trace of such a complex assembly process is better formalized as a hypergraph, in which things are combined in (possibly overlapping) sets. Our approach here follows this intuition and models science as a dynamic hypergraph, in which articles are hyperedges and contain nodes of several distinct types. Using the formalism of hypergraphs to model heterogeneous assemblies hews more closely to Latour's picture than a dyadic, unipartite network, as Latour consistently advocates greater concreteness in our descriptions of groups and the processes that bring them together (Latour, 2005).[2] The hypergraph framework developed by Taramasco et al. (2010) is close to ours in spirit; however, they focus on formal measures of paper composition such as the fraction of repeated associations, while we focus on the dynamics that drive new associations.

We proceed in the following steps. In Section 2, we define our terms and the hypergraph representation. In Section 3, we perform a detailed descriptive analysis of the evolving hypergraph documented in MEDLINE. Here we find that the distance between things in the hypergraph of biomedical science is surprisingly small, once things of many types (e.g., methods, diseases, chemicals) are included; two steps is the modal shortest path between disconnected things. This result implies that the hypergraph is dominated by local structures. In Section 3, we examine the local structure of this network by considering the immediate network neighborhoods of different kinds of nodes. We then introduce a local random walk model to approximate "possible wanderings" through this network. In Section 4, we use the transition probabilities from the random walk model to define the proximity of different things, and use these proximities as features to predict the local evolution of the network in a logistic regression framework. This proximity-based classifier has excellent performance (AUC $\geq$0.9),[3] which we verify in a 10-fold cross validation (Fawcett, 2006). We interpret our logistic regression as a simple model of the practices that collectively weave the network of science. The logistic weights reflect modal dispositions of the scientific imagination; some forms of proximity make a new connection more conceivable and likely to be followed than others. We find that biomedical science tends to "link" across rather than within types of things, which underlines the importance of incorporating increased complexity—multiple types of things—in any study of scientific reasoning or discovery.

2. Hypergraph representations

We begin by representing the scientific system as a bipartite network with two kinds of elements: *things* and *articles*. In our case, scientific articles record the outcome of assembly processes in which different types of thing (scientists, methods, and topics) are combined. A bipartite graph between things and articles is equivalent to a hypergraph over several node types: hyperedges correspond to articles and nodes correspond to things (Faust, 1997; Borgatti and Everett , 1997). One common approach to the analysis of natively bipartite or hypergraph-like networks is to project the whole network onto a certain node type. For example, in a co-authorship network, two scientists become linked when they coauthor a paper together (Newman, 2001, 2004; Martin et al., 2013). Other work has studied chemical networks, linking two chemicals if they appear in the same article (Foster et al., 2013). Such projections, however, leave out important information from the original multi-mode hypergraph. They fail to distinguish the simultaneous co-presence of several elements (authors, chemicals, etc.) and the serial appearance of subsets of those elements. They also omit any relational information connecting elements of different types (e.g., authors and chemicals). To appropriately describe the heterogeneity in types of things and the article-thing structure, we propose the following multi-mode hypergraph representation.

Formally, let $\mathcal{G} = (V, E)$ be a hypergraph. V is the set of nodes (things) and $V = \bigcup_{\alpha \in I} V^{(\alpha)}$ where $V^{(\alpha)}$ corresponds to nodes of a certain type, indexed by $\alpha \in I$, which can be authors, objects of

[2] Hypergraphs are mathematically equivalent to bipartite graphs in which articles (hyperedges) are represented as a distinct type of node that connects other *things* together. We detail this similarity below, but retain the hypergraph language because hyperedges (or node sets) corresponds intuitively to the image of an article containing scientific "things".

[3] Area Under the ROC Curve (AUC) is a popular scalar measure summarizing classifier performance. A random classifier achieves an AUC of 0.5, and higher AUCs correspond to better performance. If we choose, at random, a pair of disconnected nodes that *will* be connected in the future and a pair that will not, a classifier with AUC = 0.9 will assign a higher score to the first pair 90% of the time.

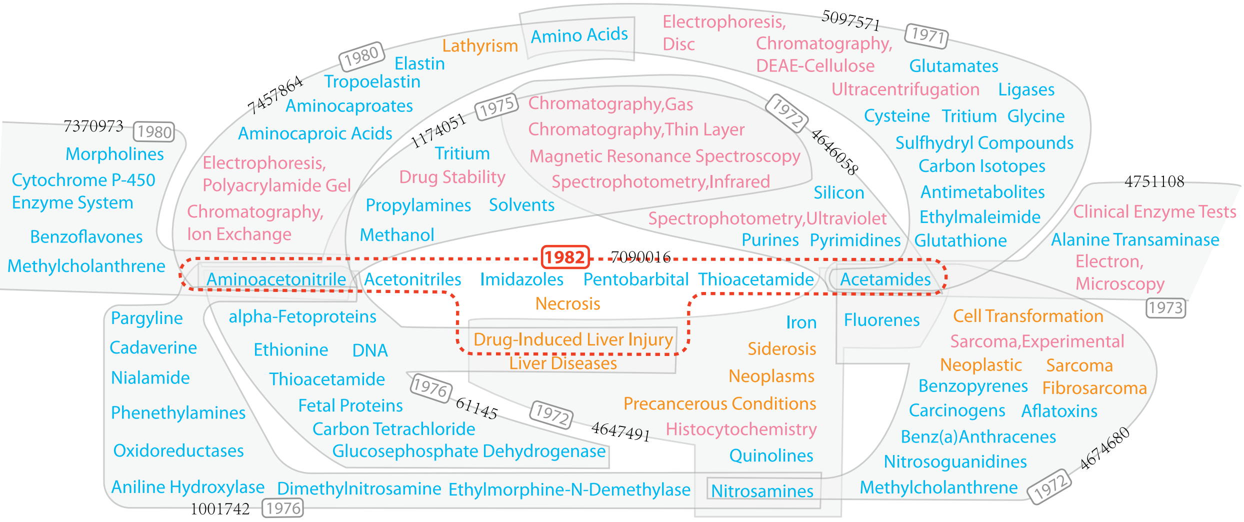


Fig. 1. A random sample from the hypergraph of MEDLINE through 1982, featuring the common neighborhoods of two simple organic compounds, acetamides (including thioacetamide) and aminoacetonitrile. Authors are not included; chemicals, diseases and methods, which are colored in blue, orange and pink, respectively, are enclosed by hyperedges corresponding to actual papers in MEDLINE with PMIDs shown along the edges. The red dotted hyperedge, a hepatotoxicology article from *Toxicology Letters*, links the two chemicals in a 1982 study of how aminoacetonitrile prevents liver injury induced by thioacetamide in rats. Other linking articles were published in the related subfields of toxicology, cancer biology, pharmacology and synthetic bio- and organic chemistry. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

study, methods of investigation, etc. We will use Greek letters such as α to denote node types and Roman letters such as u or v to denote the nodes themselves. E is the set of hyperedges and $E = \{e : e \subset V,\ |e| \geq 2\}$, a collection of subsets of nodes whose cardinality is no less than 2. This is a reasonable constraint in scientific publication because an article must contain a scientist and something—anything—studied. We will use the term "link" exclusively for edges of size 2, the usual meaning of a link in traditional graphs. To avoid any confusion, we define several hypergraph statistics that generalize comparable measures used in the study of simple, unimodal, dyadic graphs.

Node Degree The degree $d(u)$ of a node u is the number of hyperedges that contain u, i.e., $d(u) = |\{e \in E : u \in e\}|$.

Edge Size The size $d(e)$ of a hyperedge e is the number of nodes contained in e, i.e., $d(e) = |e|$. The major distinction between a hypergraph and a graph is that a hyperedge can combine (or contain) more than 2 nodes. Note that $d(e)$ is just the degree of e in the bipartite representation (in which hyperedges are distinct node types and each hyperedge e is connected to all of the nodes $u \in e$).

Neighborhood The neighborhood $\Gamma(u)$ of a node u is defined as $\Gamma(u) = \{v \in V : \exists e, \text{s.t.}\{u, v\} \subset e\}$, that is, the set of all nodes v such that there is at least one hyperedge containing both u and v. Nodes in $\Gamma(u)$ are "neighbors" of u. Note that $|\Gamma(u)|$ is not necessarily equal to $\sum_{e \ni u} d(e)$ unless one allows $\Gamma(u)$ to be a multiset, because a pair of nodes u and v can be linked by multiple hyperedges (cf. a simple graph).

Graph Distance The graph distance (distance in short) between nodes u and v is the number of hyperedges along the shortest path that connects u and v. A path from u to v is a sequence of nodes $u_1, u_2, \ldots, u_n$ such that $u_1 = u$, $u_n = v$, and $\{u_i, u_{i+1}\} \in e$ for some $e \in E$.

3. MEDLINE as a hypergraph

We apply this representational scheme to the National Library of Medicine's MEDLINE dataset (see Fig. 1 for an illustration). MEDLINE contains the metadata and abstracts of 19,916,562 articles in the biomedical literature from 1865 to 2010. We analyze four types of nodes central to the biomedical field: authors, chemicals, diseases, and methods (Leydesdorff et al., 2012). Disambiguated author names for each paper are obtained from Smalheiser and Torvik's Author-ity tool (Torvik and Smalheiser, 2009). Each paper is annotated with MeSH (Medical Subject Heading) terms, from which we extract the chemicals and/or diseases studied in a paper, as well as the methods used. The papers indexed by PMIDs[4] provide the hyperedges; all things (authors, chemicals, diseases, and methods) in a paper are combined together by that hyperedge. In total, the dataset involves 9,300,182 authors, 9159 chemicals, 4390 diseases, and 2370 methods. Note that about 30 MeSH terms belong to multiple categories; each such term is split into multiple nodes belonging to their corresponding categories. These nodes have no measurable influence on our findings. To make the correspondence concrete, consider the PMID 7457864, which identifies a paper published in 1980. That paper brings together 6 chemicals (elastin, tropoelastin, aminocaproates, aminocaproic acids, amino acids, aminoacetonitrile); two methods (polyacrylamide gel electrophoresis, ion exchange chromatography); one disease (lathyrism); and five authors (J.A. Foster, C.B. Rich, M.D. DeSa, A.S. Jackson, and S. Fletcher) Foster et al. (1980), see Fig. 1.

We show some descriptive statistics for active things (those that appeared in at least one paper) in each year from 1950 to 2008 in Fig. 2. There are few records prior to 1950 and records in 2009 and afterwards are not complete. In the first panel, we plot the total number of each type of thing active in a given year. Then we show the average degree of a thing of each type, i.e., the number of articles in which an author, chemical, disease, or method typically appears. Finally, in the third panel we show the average number of things of each type in an article. There is an apparent surge of scientific activity in the 1960s (Fig. 2(a)), which is revealed by a rapid increase in the number of methods and chemicals. We note that the sudden increase of things is partially attributable to the

[4] A PMID (PubMed identifier) is a unique number assigned to each paper in the database.

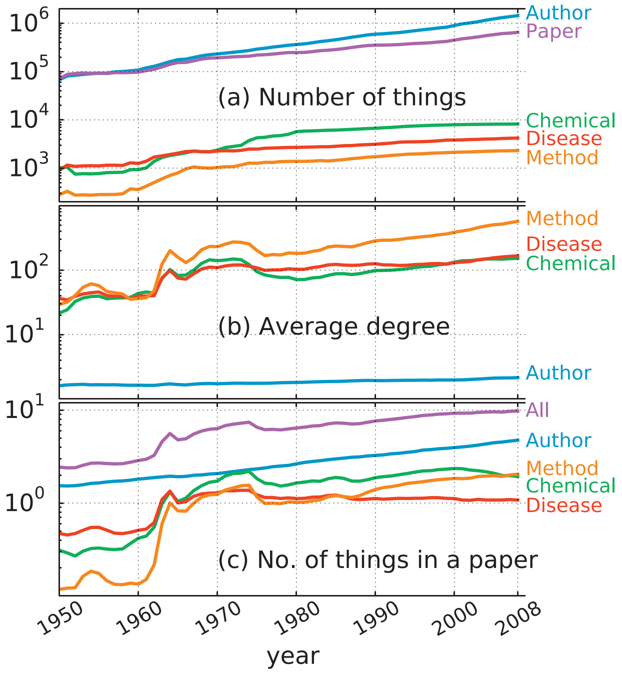


Fig. 2. (a) Numbers of active things (i.e., those that appear in at least one article) each year, and number of papers produced each year. (b) Average degree of things (i.e., average number of papers attached to a thing) each year. (c) Average number of things in a paper, including all things taken together.

introduction of a revised MeSH system at that time. There is also a notable increase in the number of papers in the 60s, however, which rules out the possibility that the jump in numbers of things is purely an annotation artifact. Furthermore, we recalculated all the descriptive network statistics using only articles with at least one chemical, disease or method annotation; we find that changes are negligible (Appendix A).

During the past three decades, the number of papers produced per year is significantly less than the number of active authors (compare blue and magenta lines from 1980 to the present). This observation seems at first blush to conflict with Fig. 2(b) in which the average degree of authors (i.e., the average of the number of papers a person writes) is approximately two for the past two decades. This latter fact might suggest that the number of papers should be comparable to the number of authors. In hypergraphs, however, a hyperedge can be shared by more than two nodes; thus the average degree of authors (number of papers per author) can stay relatively flat while the total number of authors outpaces the total number of papers. Indeed, we find that scientific collaborations have grown larger (Fig. 2(c)). Taken together, these findings suggest that as collaborations have grown larger and more frequent, productivity (number of papers divided by number of authors) has decreased (Martin et al., 2013).

We also note that ingredients of the "modal" paper have remained relatively consistent since the early 1970s: *one* disease studied in combination with *two* chemicals by an increasing but relatively small number of people (two to four). While the number of diseases studied has remained stable, the number of *methods* in a paper has increased since the early 1980s, and "multi-method" papers are now standard, with two methods the average. This rise tracks the log-linear growth in the number of scientific authors per paper, and suggests that researchers, as repositories of methodological skills, may be partially responsible for the increase in the number of methods per paper, just as demand for more methods in a paper may be partially responsible for the increase in the number of authors. The 5 most popular and least popular methods since 1980 and the number of papers annotated with those methods are listed in the table below.[5]

Most popular methods	Papers
Tomography, X-ray computed	184,429
Immunohistochemistry	183,453
Magnetic resonance imaging	177,659
Polymerase chain reaction	164,671
Cloning, molecular	140,043
Least popular methods	**Papers**
Radiesthesia	4
Cell migration assays, macrophage	4
Nerve expansion	2
Speleotherapy	1
Interpleural analgesia	1

4. Dynamic structure of the MEDLINE hypergraph

4.1. Characterization of new links

The longitudinal nature of the data allows us to trace and characterize the links formed between things. For a pair of things that appears in an article, we investigate the relative distance between them in the previous five years. We find that most links are formed between pairs of nodes two steps away—friends of a friend. This lays down the foundation for our link prediction model. Fig. 3 shows the fraction of links formed each year that are repeated (distance 1), involve newcomer nodes, connect nodes two steps away (distance 2), or link nodes at greater distances in the previous five years. We compare these fractions with an estimate of the number of link "opportunities" at each time, ascertained by calculating the fraction of pairs of things at a given distance when pairs are selected at random, i.e., regardless of whether they do or do not become linked by an article.

Consider the links formed between author–author pairs in Fig. 3(a1). From ~20% (in 1950) to ~40% (in 2008) of the author–author links formed in a given year represent repeat collaborations, while ~80% (in 1950) to ~60% (in 2008) are new links, unpublished in the previous five years. We break these new links into three distinct categories. The first includes new links that involve a debuting author. In 1950, approximately 60% of all links were both new and contributed by authors making their MEDLINE debut. This number decreased to about 20% in 2008. The decrease is a necessary consequence of recent trends in collaboration (Guimerà et al., 2005): as collaborations grow, the fraction of links contributed by each author shrinks. From 20% (1950) to 40% (2008) of the links are new and formed between existing authors. A large fraction of these new links are formed between authors previously two steps from one another, and this fraction becomes larger in recent decades. This observation implies that when people acquire new collaborators, they tend to select those who share collaborators, methods, chemicals, or diseases—an intuition we confirm later (Fig. 5). Chemical–chemical links shown in Fig. 3(c1) are dominated by repeat links (Foster et al., 2013). Few new chemicals are added to the annotation system each year and their contributions are barely visible. As with authors, the majority of new links are formed between pairs of chemicals at distance two from one another. This pattern also holds for other links formed between non-human things (Appendix B). The composition of human-nonhuman links follows a slightly different pattern (see

[5] The first eight most popular methods are "general methods" such as Treatment Outcome, Risk Factors, and Follow-Up Studies; therefore, we do not include them in the table below, though all methods are used in our analysis. Removing these "general methods" does not change our pattern of results.

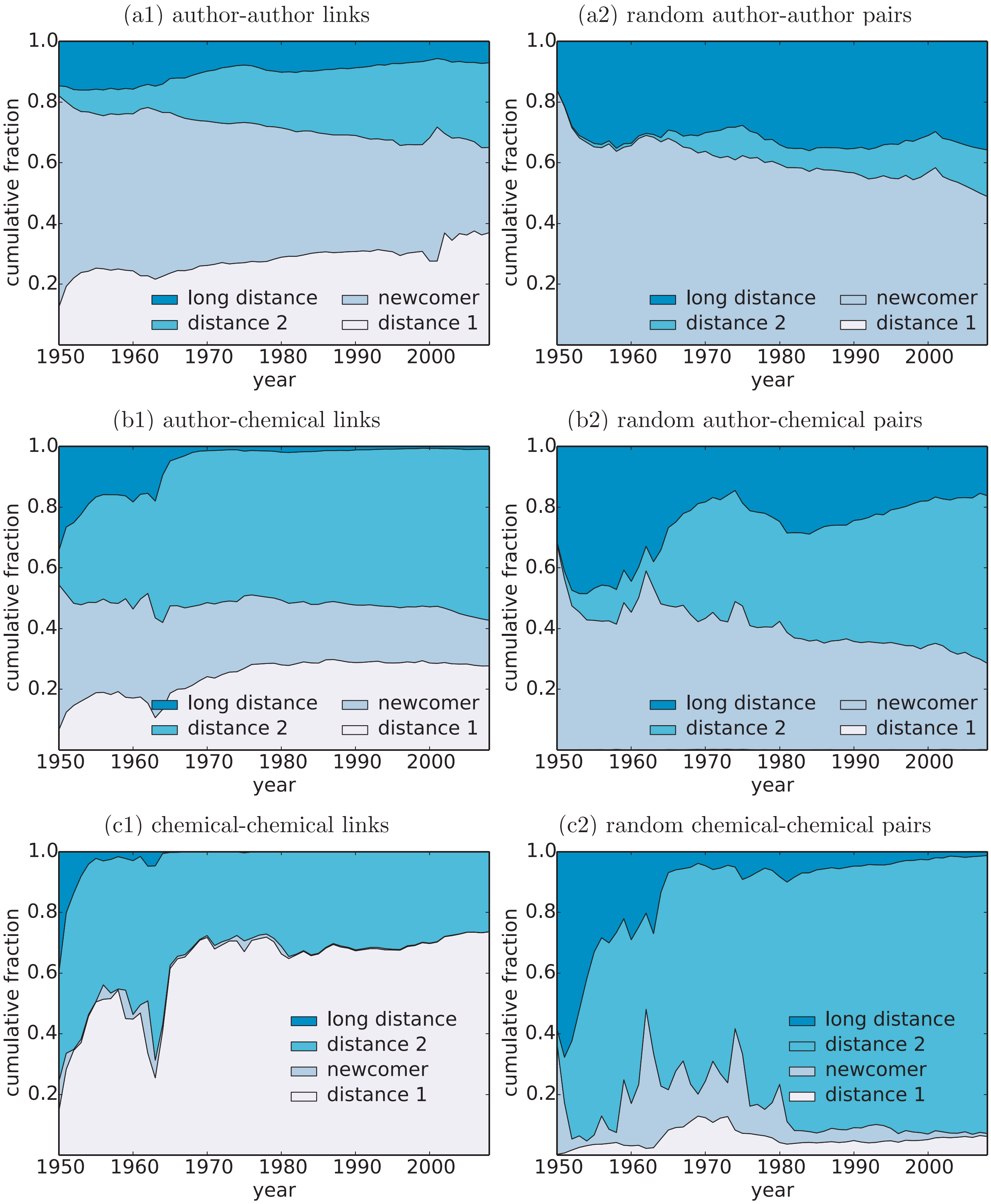


Fig. 3. Left column: Fraction of links formed each year that are repeated, involve newcomer nodes, connect nodes two steps away, or link those at greater distances in the previous five years, between (a1) authors and authors, (b1) authors and chemicals, (c1) chemicals and chemicals. Right column: Distribution of relative distance between random pairs of things that may or may not form links. Note that a large fraction of pairs are only two steps away.

Fig. 3(b1) for author–chemical links and Appendix B for others). Between 20 and 30% of these links are repeated each year, with new ones dominated by pairs formerly of distance 2. Taken together, these findings suggest that when a scientist chooses a new topic to study or adopts a new method for her investigation, she is highly likely to choose something directly related to her current expertise (or something used by a collaborator). In all cases, statistics of observed links (those that actually appear in the hypergraph) significantly differ from what would be expected if pairs of nodes were chosen at random; nevertheless, we note that the fraction of things that are more than 2 steps away from one another is still very small (except for pairs involving authors), i.e., *the vast majority of existing nodes that have not yet been connected are at distance two*. This last point highlights the topological distinction between a complex

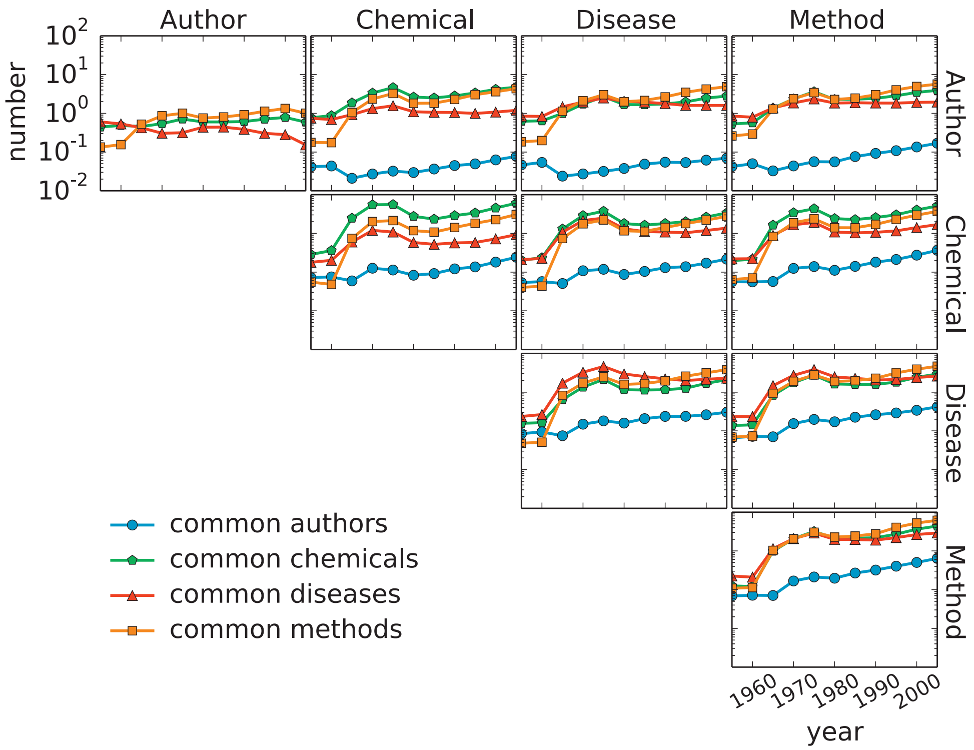


Fig. 4. Average number of common neighbors between each of the four types in each year. Types of the two end nodes are given by row and column labels. The number of common author neighbors in the first panel is less than 0.01 over the entire period and hence not shown in the plot. Most things at distance two have more things of same type in between, e.g., there are almost twice as many chemicals as other things between chemicals and chemicals (first panel in the second row); note the semi-log scale.

hypergraph of science and a simple network composed of single-type nodes (e.g., chemicals). Because two things can be connected along a path going through *anything* else—authors, chemicals, diseases and methods—and because diseases, chemicals, and methods especially have high average degree (appear in a large number of papers), the hypergraph is dominated by local structure.

While a large proportion of the links formed each year are repeated or contributed by newcomers, we are interested in scientific discovery and transformation that builds on the current substrate by weaving new connections; in other words, the formation and establishment of new links between existing things. Given the large size of the system (about 20 million papers and 9 million authors) and the relatively small number of opportunities for long range connection (Fig. 3), we focus our study on the local structure of the network—on pairs of things at distance two. We call such a pair “end nodes;” we will show later that their dynamics—whether they link or not—are largely determined by the local network neighborhood.

4.2. *Common neighborhoods*

To probe the local structure of the network, we first follow the approach of Newman and collaborators (Martin et al., 2013) and count the number of common neighbors between end nodes of various types. We split our data (1950–2008) into five-year chunks. For the hypergraph aggregated over every time window $[t-5, t]$, the average number of common neighbors between two nodes at distance two is shown in Fig. 4. Each panel in the figure evaluates the common neighbors for a particular combination of end node types. For example, the first panel shows the average number of common neighbors between two authors at distance two. Common neighbors are likewise divided into common author neighbors (circles), common chemical neighbors (pentagons), common disease neighbors (triangles), and common method neighbors (squares).

Note how the composition of the common neighborhood changes with end node types. Most of the time things at distance two have more things of same type in between, e.g., there are almost twice as many chemicals as other things between chemicals and chemicals (first panel in the second row in Fig. 4). The structure of common neighborhoods influences the formation of new links. In Burt’s calculus of structural constraint, authors at distance two within the social network are structurally identical relative to common neighbors (Burt, 1992). More generally, nodes with one or more common neighbors are more likely to connect because of triadic closure (Rapoport, 1953). Martin et al. found that triadic closure operates in collaboration networks: the more common collaborators two scientists share, the more likely they will form a collaboration in the future (2013); see Kossinets and Watts (2006), Backstrom et al. (2006), and Crandall et al. (2008) for similar studies of closure in online contexts. We observe the same phenomenon here (Fig. 5). The triadic closure phenomenon also obtains for all other types of end nodes—methods, diseases and chemicals. Nevertheless, we describe another metric of proximity, assessed via random walks, that is much more predictive of new ties in the hypergraph of science.

4.3. *Local random walks*

To obtain a dynamic view of the local structure, we define a random walk on the hypergraph as in (Cooper et al., 2011). On a simple graph, the random walk is a stochastic process $X(t)$ whose state space is the set of nodes. In one step, the random walker moves from node u to any of u’s neighbors with a transition probability $1/d(u)$ (where $d(u)$ is the degree of u). In the hypergraph setting, a hyperedge usually combines multiple nodes. There are thus multiple destinations available through that hyperedge. Accordingly, each step of a random walk on a hypergraph has two stages: at a given node, it first picks at random a hyperedge attached to the

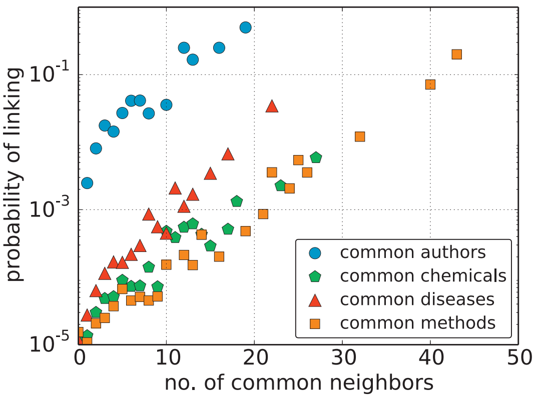


Fig. 5. Conditional probability of forming a link in 2000 between authors given the number of common author neighbors, common chemical neighbors, common disease neighbors and common method neighbors, respectively, during (1995, 2000). The conditional probability is calculated as the fraction of author–author pairs linked in 2000 among all author–author pairs with a given number of common neighbors.

node; then it moves to a random node within that hyperedge. We then define the transition probability $P^{(\alpha)}(u, v)$ between u and v through a given type α to be the probability that a random walker, departing randomly from u or v, will arrive at the other in two steps *through* a node of type */alpha*. The formal definition of the random walk transition probabilities is given in Appendix C.

Roughly speaking, this transition probability can be thought of as an Adamic/Adar similarity score, extended to hypergraphs. The original Adamic/Adar score was designed to measure the similarity between two nodes through their common features, putting more weight on rarer features (Adamic and Adar, 2003). It was applied by Liben-Nowell et al. with common neighbors as features in the form $\sum_{w\in\Gamma(u)\cap\Gamma(v)} 1/\log d(w)$, and works surprisingly well for predicting links in many social networks (Liben-Nowell and Kleinberg, 2007). Our random walk measure weights common neighbors by their degrees, but also weights hyperedges by their sizes to account for the fact that ties within larger collaborations will likely be proportionately weaker Newman (2001). In addition, the transition probability defined here takes into account the number of 2-step paths between two things through a given node type, as well as the degree of the node along each path. As a result, we argue that transition probabilities provide a *much* more sensitive description of the network structure between nodes at distance two than just the raw number of common neighbors—and a more sensitive assessment of their "proximity." We calculate the transition probability through each of the four types and present the average transition probabilities across all pairs in Fig. 6 to compare with the average number of common neighbors. Details of the calculations and the distribution of transition probabilities are given in Appendix C.

5. Weaving the fabric

The random walk model not only serves as a refined assessment of structural connectivity; it also provides a simplified but plausible model for the generative scientific practices through which biomedical science evolves. Scientists do not wander randomly through articles, of course. Each has her own preferences and strategies, and each is subject to unmeasured external forces that shape how she selects a topic for study or finds someone with whom to collaborate. Here we aim only to investigate the extent to which new discoveries *can* be explained by a local random walk; specifically, the extent to which the formation of new links between things is constrained by the local structure of the scientific network. We note that there is a rich literature on link prediction of complex networks (Liben-Nowell and Kleinberg, 2007; Menon and Elkan, 2011; Hasan and Zaki, 2011). Most of the work in that field employs either the idea of homophily or of triadic closure, i.e., nodes "similar" or "close by" are more likely to form links. As shown above, a large number of common neighbors between two authors suggests that the two are similar (e.g., they study the same chemicals, or work with the same people) and hence are more likely to form a new link. Here we do not aim to compete with the state of the art in link prediction (although our model performs well in predicting new links), and the size of our network makes many of the most recent exponential random graph models impractical (Snijders, 2002; Snijders et al., 2006; Robins et al., 2007; Goodreau, 2007). Rather, we focus on quantifying the underlying dynamics that systematically shape the formation of new links between things in science.

5.1. The random walk model

We propose the following socio-cognitive model for exploring the dynamics of link formation. For a pair of things at distance two, assume that the probability of forming a link between them is a function of random walk transition probabilities and corresponding weights. Transition probabilities reflect the proximity between two things under mental wandering; those with high proximity may be perceived as "similar" or "relevant" to one another as a result. The estimated weights reflect modal dispositions favoring proximity through different intermediate types of things. For a given time window $[t-\Delta t, t)$, during which end nodes u and v are two steps away from each other, let Y_{uv} be an indicator such that $Y_{uv}=1$ if u and v form a link (appear in the same paper) at year t and 0 otherwise. The probability of $Y_{uv}=1$ conditioned on the random walk proximities between u and v is defined as

$$P(Y_{uv}=1|\mathbf{c},\mathbf{p}(u,v)) = f(c_0+\sum_{\alpha\in I} c_\alpha p^{(\alpha)}(u,v)). \tag{1}$$

where $\mathbf{c}=(c_\alpha)$, $\alpha\in I=\{author, chemical, disease, method\}$, and c_0 corresponds to a "background" similarity independent of proximity (equivalently, similarity) under mental wandering. $\mathbf{p}(u, v) = (p^{(\alpha)}(u, v))$, $\alpha \in I$, is calculated as in equation C.4 on the hypergraph aggregated over $[t-\Delta t, t)$. The ideal window length Δt is not obvious. We use $\Delta t=5$ in our analysis, but our results are robust to the choice of window size as long as it is large enough to cover normal cycles of production (see Appendix D for a detailed discussion on the effect of window size). To retain a consistent network and avoid noise introduced by newcomers, we restrict our analysis to things that are active (i.e., that appear in at least one article) in both the $[t-\Delta t, t)$ period and year t.

Note that c_α could, in principle, vary across individual scientists, representing idiosyncratic dispositions, but here we focus on the average disposition across the system, and hence we treat c_α as constant over all pairs in a given time-slice. As such, the coefficient c_α measures the weight given to each type of proximity under our random walk model.

To fit this model to the data, we adopt the logistic function for f. Fitting the model is thus equivalent to a logistic regression, although alternative formulations with probit or identity link functions yield similar results. To allow for the possibility that coefficients change in relative importance over slow timescales, we first split time into 5-year windows: [1950,1955], [1955,1960], . . ., [2000,2005]. During each window $[t-5, t]$, $t=1955, 1960, \ldots,$ 2005, for each possible combination of end node types (e.g., authors and methods, chemicals and diseases, etc.) we predict the probability that two end nodes will appear in published combinations as a function of their random walk proximities using the logistic

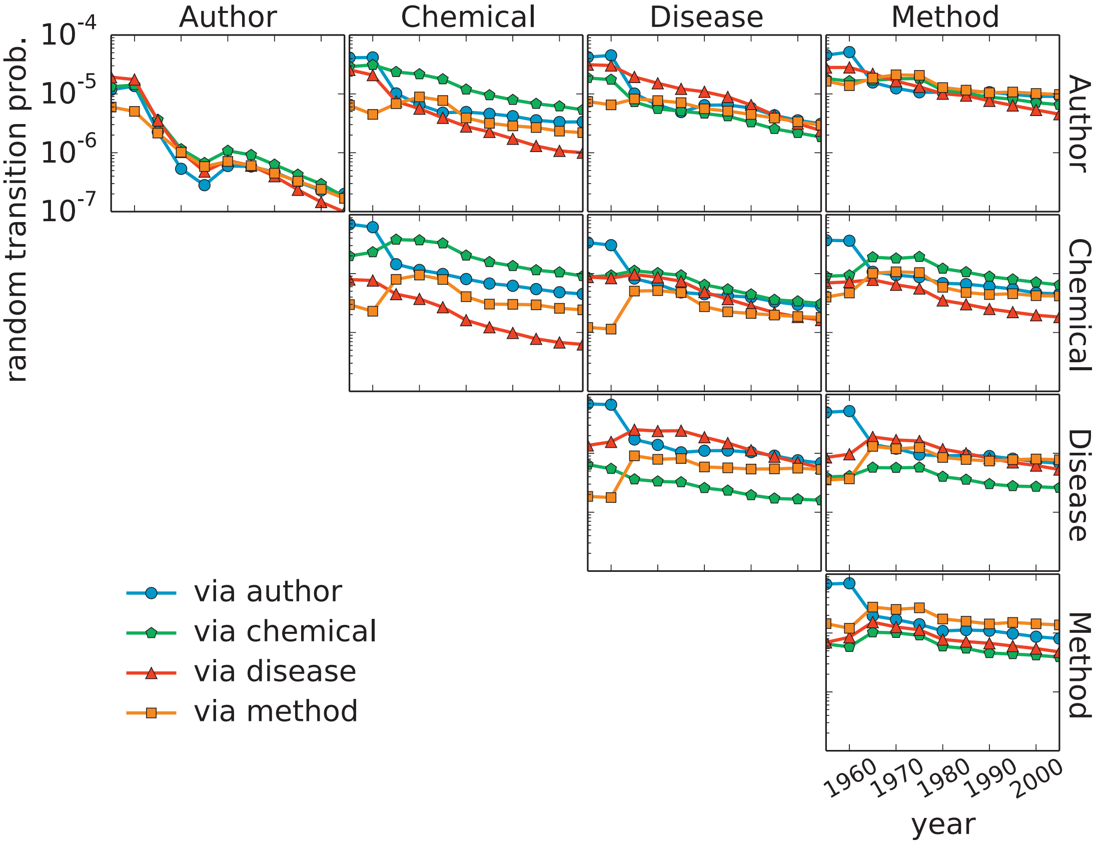


Fig. 6. Average transition probability via each of the four types at each year between two nodes of types corresponding to row and column labels. See Appendix C for the distribution of transition probabilities across the population and detailed discussions.

Table 1
Estimated coefficients of the logistic model (Eq. (1)) for chemical–chemical pairs at each time slice during 1950–2005.

	c_{author}	$c_{chemical}$	$c_{disease}$	c_{method}
1950–1955	366 ± 61	2771 ± 347	2567 ± 519	3432 ± 784
1955–1960	427 ± 52	1884 ± 228	2313 ± 484	782 ± 450
1960–1965	528 ± 56	4319 ± 79	5728 ± 414	6959 ± 309
1965–1970	289 ± 47	3572 ± 66	7454 ± 389	4500 ± 218
1970–1975	520 ± 30	2652 ± 36	6288 ± 254	4897 ± 150
1975–1980	672 ± 35	3369 ± 42	8341 ± 309	9481 ± 230
1980–1985	1249 ± 48	3665 ± 43	14557 ± 356	13831 ± 259
1985–1990	1400 ± 48	4175 ± 47	15505 ± 405	10443 ± 240
1990–1995	1399 ± 49	5084 ± 50	19282 ± 394	8140 ± 201
1995–2000	1442 ± 54	5894 ± 54	20186 ± 428	8931 ± 230
2000–2005	1516 ± 69	6557 ± 67	20439 ± 453	11799 ± 245

p-values for all coefficients are less than.001.

model (Eq. (1)).[6] In total 110 logistic models are fitted (11 time windows and 10 possible combinations of types). Fitting diagnostics show that the c_α's are significant (Table 1), confirming that proximities under local random walk are associated with link formation. Due to space limitations, only results for chemical–chemical pairs are included in Table 1; regression results for all the 110 logistic models are included in Appendix F, as well as a sensitivity analysis that tests alternative time windows. We evaluate and interpret our random-walk model in the following section. There we also compare that model to one using number of common neighbors of various types as predictors. The model with random walk proximity has much greater predictive power, exhibiting high levels of discrimination, while the model with common neighbors is little better than chance.

5.2. Model evaluation

The simple logistic model based on random walk proximity has excellent predictive power. The area under the ROC curve (AUC) obtained from a 10-fold cross validation for each model is shown in Fig. 7. As a comparison, the AUC's for a logistic model with random walk proximities replaced as predictors by the number of common neighbors are shown in the figure as well. With the random walk proximities as predictors, the model achieves a much higher AUC (and hence predictive power) than it does with the number of common neighbors as predictors. The model with random walk proximities supports an AUC of greater than 90% (generally considered excellent) while common neighbors have an AUC of roughly

[6] In fitting each logistic model, all possible pairs of things of the desired types are used except for author–author pairs because of the large number of authors (at least 10^{10} author–author pairs for each window). Hence for each window a random sample of 3×10^7 author–author pairs is used.

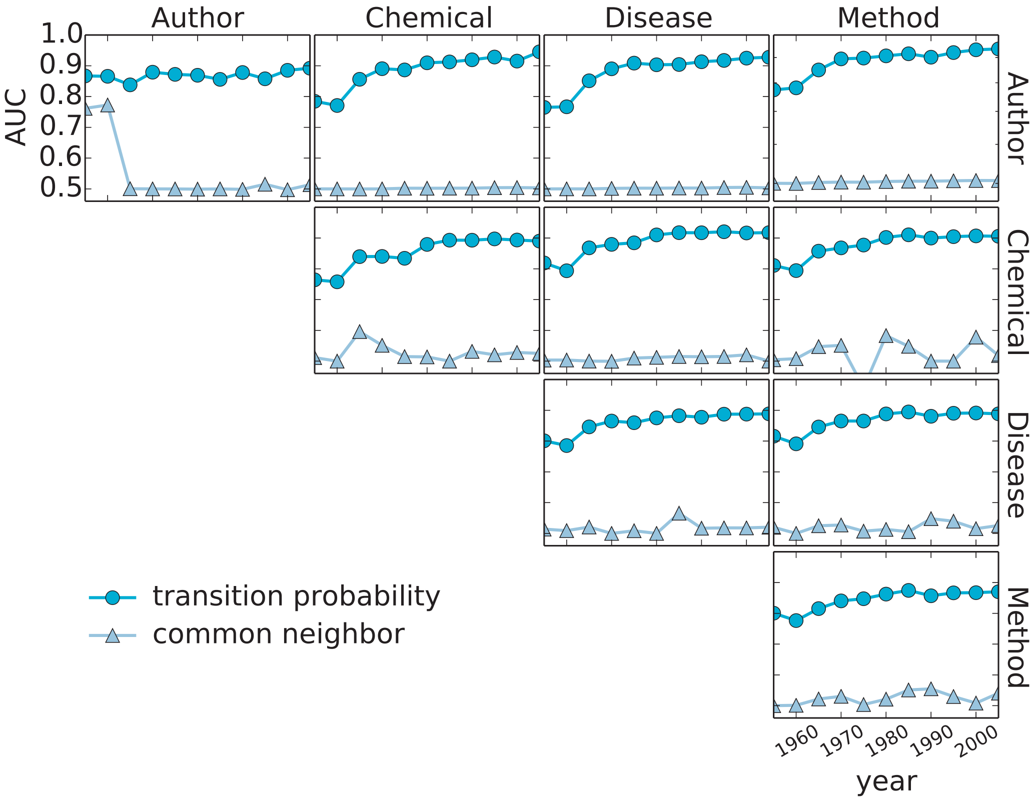


Fig. 7. Area under the ROC curve (AUC) obtained from a 10-fold cross validation for each model, measuring the predictive power of logistic models. Circles correspond to models using the random walk proximities as predictors and triangles to models using the number of common neighbors as predictors.

50% (equivalent to chance) (Fawcett, 2006). We further test the robustness of our results by carrying out an event history analysis with the transition probabilities as predictors as in (Kossinets and Watts, 2006). The event history analysis yields the same pattern of results as the logistic models. The relative sizes of coefficients for the transition probabilities are the same in both models, confirming the robustness of the conclusions we describe below (Appendix E).

This logistic model admits an interpretation consistent with our claim that the structure of the network, coupled with scientific dispositions, largely determines how science evolves. Logistic regression is equivalent to a single-layer perceptron, a neural network model (Hinton, 1992; Bishop, 1995; Ripley, 2007). In essence, the perceptron takes as input the random walk proximity $p^{(\alpha)}(u, v)$ of the nodes in question, under different intermediate node types α; weights the input from each random walk proximity according to c_α; sums up the weighted input; and "fires" with a probability given by the logistic function:

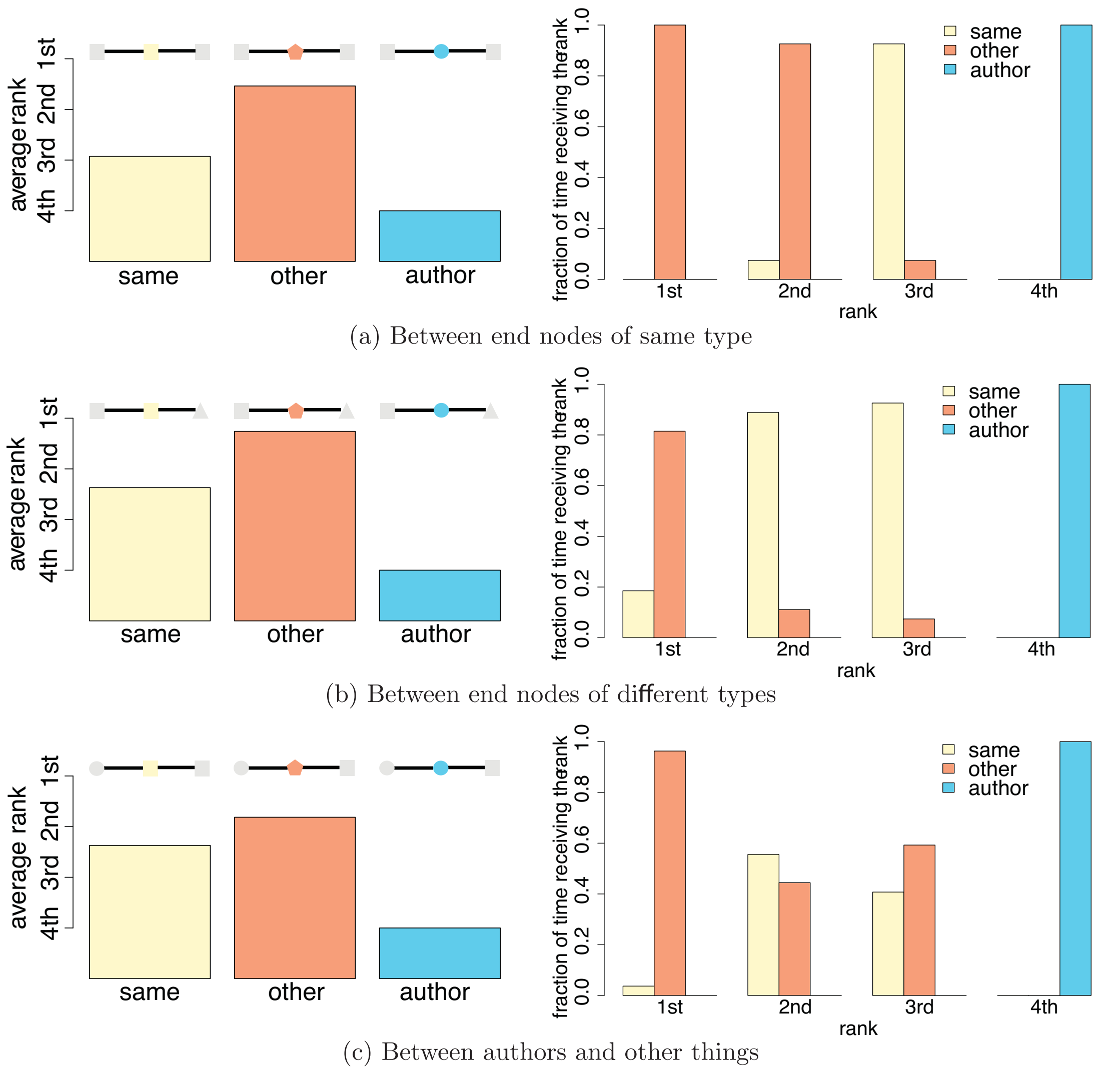
$$P(fire|\mathbf{c}, \mathbf{p}(u, v)) = \frac{1}{1 + e^{(c_0 + \sum_{\alpha \in I} c_\alpha p^{(\alpha)}(u,v))}}. \qquad (2)$$

where "firing" corresponds to making a connection between u and v. While simple neural networks like this one have well-known technical limitations (Hinton, 1992; Bishop, 1995; Ripley, 2007), the predictive power of our model suggests that the random walk transition probability does a much better job of capturing the cognitive proximity between two things—or, even better, the perceived viability of their combination—than number of shared neighbors. It also suggests that we might fruitfully interpret the actual values of the perceptron weights as something quite close to a disposition to respond differentially to changes in different forms of proximity (Bourdieu, 2004). The dispositions (coefficients) of all logistic models are shown in Appendix F.

5.3. *Patterned dispositions*

We now take a closer look at the weights as dispositions. To compare dispositions over all time periods and types of end nodes (i.e., u and v in Eq. (2)), we first rank the dispositions (coefficients) of each logistic model. We assign these ranks to three bins: (1) dispositions to establish new links through non-human intermediaries (chemicals, methods, diseases) that are of the *same* type as either end node; (2) dispositions to establish new links through non-human intermediaries of *different* type than the end nodes; and (3) the disposition to combine through authors. The average rank of dispositions in each bin when forming links between non-human end nodes of the same type (i.e., chemicals and chemicals, diseases and diseases, methods and methods) is exhibited in the top row in Fig. 8, along with the fraction of the time that dispositions in each bin receive the corresponding rank. Similarly, the average rank of dispositions when forming links between distinct non-human end nodes (i.e., chemicals and diseases, chemicals and methods, diseases and methods) is shown in the second row in Fig. 8; the average rank of dispositions when forming links between authors and any other node type, excluding authors, is shown in the bottom row in Fig. 8. Despite folk theories of free association, which would suggest that scientists reason analogically within type (in which case they would be most sensitive to changes in proximity induced by nodes of the same type), we find that most of the time scientists are more sensitive to (i.e., disposed toward) proximity through intermediaries that are of *different* type than the end nodes.

This surprising finding has two interpretations and a strong methodological implication. The first interpretation emphasizes the practical task of assembling the ingredients for a publishable research project. Because the typical paper in this space will involve at least one chemical, one method, and one disease—and because



(a) Between end nodes of same type

(b) Between end nodes of different types

(c) Between authors and other things

Fig. 8. Left Column: Average rank of disposition on intermediate types that are the *same* as end node types, that are *different* from end nodes types, and that are *authors*. Here we consider (a) non-human end nodes of same type, (b) non-human end nodes of different types, and (c) authors and other non-human things. Right Column: Fraction of cases in which disposition toward a certain intermediate type receives the corresponding rank.

the intermediate nodes suggesting the plausibility of a new link are themselves reasonable candidates for inclusion in the paper—it makes sense that scientists would be especially sensitive to proximity induced by nodes of a different type. This search procedure would lead more quickly to a paper with the necessary building blocks.

There is a more intriguing interpretation, based on a detailed consideration of the cognitive underpinnings of the creative task. We note this cognitive interpretation as a possible contributing factor to the patterned dispositions we estimate. The scientist needs to call to mind a focal entity. She then starts to wander out from that entity across the collectively woven fabric of science, searching for plausible new connections. The very task of calling an entity of a particular type to mind, however, makes it more difficult to retrieve other entities of the same type, due to a well-attested phenomenon called “retrieval-induced forgetting” (Anderson et al., 2000).[7] When trying to recall an entity of a particular type (say, a disease), initial activation of the larger category in the retrieval task necessarily summons other members of the category. In order to retrieve a *particular* member of the category, the others must be, in effect, inhibited. When retrieval of a particular entity is repeated and becomes practiced, inhibition can last for a non-trivial period of time (in experiments, up to 20 min) (Norman et al., 2007).[8] In this context, it is natural that associations through things of distinct type should be preferred. Once a scientist is thinking about the focal disease, for example, neighboring diseases may be effectively suppressed from memory as perceived proximity through diseases attenuates. Note also that this cognitive interpretation of the pattern of dispositions does not discount the essentially social context in which dispositions are deployed: an individual scientist is thinking over the fabric of science she and her colleagues have collectively woven, a fabric (following Latour) that contains other scientists, chemicals, methods, diseases, etc.

The methodological implication of this finding is striking. Much research on networks considers particular types of things (e.g.,

[7] While this process could occur in some scientists some of the time, we do not argue that this necessarily scales to the macro-behavioral phenomena observed across MEDLINE in this study. Here we consider its suggestive alignment with our observed phenomena.

[8] Indeed, neural network models suggest that memory representations of entities other than the one repeatedly retrieved are actually weakened, making these other entities harder to retrieve (Norman et al., 2007).

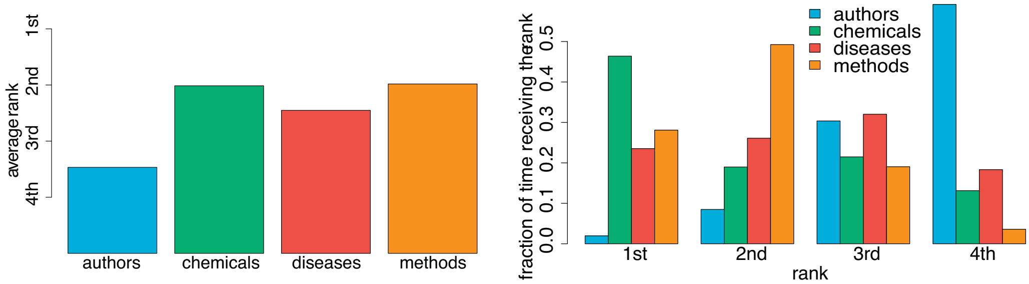


Fig. 9. Left: Average rank of the contribution each type of thing made to the links formed between all pairs of things over all time periods. Right: Fraction of time the contribution from a certain intermediate type receives the corresponding rank over all links and all time periods.

authors, creators, chemicals) in isolation (Foster et al., 2013; Uzzi and Spiro, 2005; Newman, 2004; Martin et al., 2013). If things tend to connect disproportionately through *different* types of things, however, adding more types of nodes will capture disproportionately more of the search and discovery process, and will perform disproportionately better in predicting new links. Our hypergraph framework suggests a refinement of the classic formulation of Latour's non-human agency, in which many different types of things combine, but in patterned and predictable ways sensitive to the types of thing in question.

5.4. Heuristic test of predicted search paths

We have so far studied the local connectivity between things and the modal disposition toward certain kinds of proximity. Random walk proximity, appropriately weighted, yields an excellent predictive model of new link formation. But what evidence do we have that scientists actually follow the paths suggested by their dispositions, and not others? To probe this question, we treat the term $c_\alpha p^{(\alpha)}(u, v)$ in the logistic model as a proxy for the influence of type-α things on the formation of new links. It effectively represents how much proximity induced by nodes of type α contributes to the decision to form a link. (More precisely but less evocatively, each $c_\alpha p^{(\alpha)}$ contributes linearly to the log odds of link formation.) It is hard to assess whether this proxy truly characterizes the cognitive paths followed without tracing the precise, historical exposure that biomedical researchers had to each scientific thing (e.g., diseases, chemicals, methods), as well as their efforts to weave these exposures and mental wanderings into hypotheses and publishable experiments. Nevertheless, a trace of the cognitive path followed may remain in the published article. For example, things along an "important" path may appear in the resulting article that asserts the new link. Furthermore, if things of a certain type are associated with the most consequential path between two things u and v, then we would expect more things of that type to accumulate in the article that connects u and v.

To make the discussion more concrete, consider a pair of methods, "perfusion" and "molecular cloning," which were connected for the first time in 1990 in Bullock et al. (1990) (PMID 1690810). Their weighted random walk proximities suggest that chemicals provided the most consequential path—the one that most increased the log odds of their ultimate connection. These methods were linked in prior articles to several of the same chemicals, and when the methods co-appear for the first time, the article includes three of those shared chemicals (Ion Channels, Immune Sera, and Lipid Bilayers) out of only four chemicals annotated in the paper. The presence of these shared chemicals in the article provides an empirical trace of tie formation, strongly suggesting that the methods connected "through" these chemicals, as predicted by the weighted proximities derived from the logistic model. We formalize this intuition as a "path accumulation trace" and assess its consistency with our estimated path influence. We note that this "path accumulation trace" is an intuitive and heuristic test for our claim that scientists and scientific institutions tend to search along paths ranked by the product of random walk proximities and estimated dispositions. It cannot confirm that science searches out new possibilities in the way we describe. As we show below, however, the high correlation between path accumulation traces and predicted search paths is suggestive and points to a common collective pattern in the evolution of science.

For a pair of nodes, u and v, which appear together in article e at year t and were at distance two in the interval $[t-5, t)$, we define the path accumulation trace for things of type-α to be the fraction of type-α things in the paper e that were common neighbors of u and v during $[t-5, t)$. Formally, letting $\Gamma(u)$ be the neighborhood of u during $[t-5, t)$, the path accumulation trace for type-α things is computed as

$$\frac{|\{w : w \in e, w \in V^{(\alpha)}, w \in \Gamma(u) \cap \Gamma(v)\}|}{|\{w : w \in e, w \in V^{(\alpha)}\}|}. \qquad (3)$$

For each pair of nodes linked in a given year, we compute the path accumulation trace for the four types of things as in (3) and compare the rank of these traces with the rank of $\{c_\alpha p^{(\alpha)}\}_{\alpha \in I}$ obtained from the logistic model as in (1). The median of the Spearman correlation between the two ranks is 0.8 (see Appendix G for details), indicating that the two ranks are highly correlated for at least half of the pairs. We do not expect results from the logistic model to align perfectly with the empirical path traces, because they are two different assessments. In fact, for certain pairs, the two ranks are negatively correlated, which results in a population mean of the Spearman correlations at 0.6. Most of those negatively correlated pairs have low odds of linking under the logistic model, and hence the underlying process of accumulation may not be well-captured by the random walk model. The high overall correlation between the two measurements, however, suggests that scientists and scientific things often connect via the paths that we estimate as most consequential in the logistic model.

Following this argument, we turn back to the logistic model and use the term $c_\alpha p^{(\alpha)}(u, v)$ as a measure of the contribution from type-α things to the link formed between u and v. Noting that the AUC's of the logistic models (Fig. 7) are relatively stable across all models since 1970, we restrict the analysis to slices since 1970 (i.e., [1970, 1975], . . ., [2000, 2005]). For each time window $[t-5, t]$, we rank the four terms $c_\alpha p^{(\alpha)}$, $\alpha \in \{author, chemical, disease, method\}$ for each link formed at t. In this way, contributions are comparable across all links and all years. The ranks of the contributions from the four types of things, averaged over all links and time windows since 1970, are shown in Fig. 9(a). Methods and chemicals dominate as the first and second most important paths through which links are formed. A close look at the distribution of ranks over all links and all time windows (Fig. 9(b)) reveals that methods are most frequently ranked second in their contribution to link formation.

Chemicals are often ranked highest, but sometimes third or last, which results in the average rank of methods being slightly higher than that of chemicals. How can we interpret these results? In the research articles of MEDLINE, diseases are the primary research "topics", with an average of one-per-paper. Chemicals can be a sort of "little topic" in some cases (e.g., an examination of the genetic etiology of disease), but are more often "little methods" through which a disease or other biological phenomena is probed, diagnosed or treated. In this way, explicit and implicit methods, which represent the *actions* involved in a research project—analogous to the verbs in a sentence—dominate the pathways through which individual scientists think and science as a whole evolves. There are fewer methods than chemicals, with a higher average degree in the hypergraph of MEDLINE, but taken together chemicals and methods form the major shuttle through which things weave together and form the fabric of science.

6. Conclusion

In this paper, we built on the work of Latour, who argued for the diverse network of scientific *things* involved in the scientific process, and the work of Newell and Simon, who proposed that problem solving and discovery occur by wandering over complex conceptual networks. We then developed a multi-mode hypergraph model of science that takes into account the higher-order complexity and heterogeneity of science as a system. This framework enables novel insight into several aspects of the evolving structure of science. We find that the majority of new links formed every year draw on things that are already neighbors or are of distance two. We also find that the hypergraph picture provides a different perspective on the local structure than a one-mode projection of the hypergraph. Even though the full complement of possible paths through which scientific things *could* recombine is broader than the paths through which they *become* combined in published research, the hypergraph is hyper-small, with the substantial majority of disconnected pairs connectable through one of many two-step paths.

Our model of science posits that scientific things (authors, chemicals, diseases, methods) combine within projects, formalized as sets, through a random walk process. By wandering across a mental map (or mesh) of science, new associations are woven between things, subsequently influencing what can be conceived, investigated, and published in the future. Our analysis of millions of scientific articles in MEDLINE shows that this model has considerable predictive power regarding what scientists and indeed science as a whole can imagine, discover and publish over time. Moreover, the paths we estimated as most influential are disproportionately likely to leave a trace in the resulting papers. As such, the local structure of this complex hypergraph appears to be a primary substrate on which science as a system evolves.

Logistic regression also reveals patterned dispositions or preferences through which scientists and science as whole deviate from the random walk model—preferring some paths and avoiding others. Most strikingly, scientists connect things through things of a different type. They tend to connect methods through non-methods, diseases through non-diseases and chemicals through non-chemicals. Scientists begin to study new methods, diseases and chemicals by thinking through other types of things. This may reflect the structure of scientific papers: by passing through paths of a different type, a scientist might minimize the time required to assemble all of the components required for a publishable study. This may also partially reflect a cognitive phenomenon called "retrieval-induced forgetting." Once a scientist is thinking about the focal scientific thing associated with an experiment, neighboring things of the same type are effectively suppressed and the likelihood for her to think along paths inscribed by other types of things is enhanced. This finding has striking implication for the study of human discovery in complex networks. Adding new types of thing to an evaluation of network discovery will likely offer not diminishing marginal returns or even linear improvement in understanding and predictive power. Rather, in science and likely many other spheres (e.g., technological invention, artistic production, new venture creation), the addition of new types of nodes will result in a superlinear increase in understanding, because creative actors connect things through *other* types of things.

We foresee several extensions of this study. First, by extracting further data from publications we can extend our understanding to more dimensions. For example, if we overlay the hypergraph of science with scientific disciplines, our model may reveal differences in scientific practices *across* disciplines. With information on author institutions and funding sources, our model can be generalized to identify how institutional forces shape the evolution of science (e.g., connecting to new diseases or methods through funding sources). Furthermore, the importance of certain people, institutions and topics in the network may disproportionately influence the things on which scientists and science lavish attention. Second, competition in science is fundamentally oriented toward innovation. Variation in the probability that a given scientific or technological discovery is made (i.e., how "innovative" it is) may relate to success of that discovery (e.g., how highly cited, theoretically integrated, patented, licensed, built and bought it is) (Foster et al. , 2013). Last, these hypergraph and random walk models can be used to analyze search processes and dynamics in other systems like technological invention and human group or team formation that have a native hypergraph structure (Lindelauf et al., 2012; Zhu et al., 2013; Aitkin et al., 2014). In short, our hypergraph investigation of how science evolves could cast light on the evolution of many social and technical systems.

Acknowledgements

The authors thank OSDC (opensciencedatacloud.org) for providing computing infrastructure and the swift team (swift-lang.org) for their support on high performance computing. The authors also acknowledge undergraduate assistant Taweewat Somboonpanyakul for his help with parallel computations, the National Library of Medicine and Jane Rosov for assistance with MEDLINE, and generous financial support from the John Templeton Foundation ("Metaknowledge Network") and NSF grant #1158803 ("Tracing Influence and Predicting Impact in Science"). Comments from Northwestern's Institute for Complex Systems seminar, Stanford's Work, Technology and Organizations seminar and UCLA's Innovation and Creativity workshop greatly improved the manuscript.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.socnet.2015.02.006

References

Adamic, L.A., Adar, E., 2003. Friends and neighbors on the web. Soc. Netw. 25 (3), 211–230, http://dx.doi.org/10.1016/S0378-8733(03)00009-1

Aitkin, M., Vu, D., Francis, B., 2014. Statistical modelling of the group structure of social networks. Soc. Netw. 38, 74–87.

Anderson, M.C., Bjork, E.L., Bjork, R.A., 2000. Retrieval-induced forgetting: evidence for a recall-specific mechanism. Psychon. Bull. Rev. 7 (3), 522–530.

Borgatti, S.P., Everett, M.G., 1997. Network analysis of 2-mode data. Soc. Netw. 19 (3), 243–269.

Burt, R.S., 1992. Structural Holes: The Social Structure of Competition. Harvard University.

Bishop, C.M., 1995. Neural networks for pattern recognition. Clarendon Press, Oxford University Press, Oxford, New York.

Bourdieu, P., 2004. Science of science and reflexivity. Polity.

Bullock, J., Armstrong, S., Shear, J., Lies, D., McIntosh, M., 1990. Formation of ion channels by colicin B in planar lipid bilayers. J. Membr. Biol. 114 (1), 79–95, http://dx.doi.org/10.1007/BF01869387

Backstrom, L., Huttenlocher, D., Kleinberg, J., Lan, X.,2006. Group formation in large social networks: membership, growth, and evolution. In: in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 44–54.

Callon, M., 1986. Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay. In: Law, J. (Ed.), Power, Action, and Belief: A New Sociology of Knowledge? Routledge & Kegan Paul, pp. 196–229.

Cohen, M.D., March, J.G., Olsen, J.P., 1972. A garbage can model of organizational choice. Adm. Sci. Q. 17 (1), 1–25.

Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., Suri, S., 2008. Feedback effects between similarity and social influence in online communities. In: in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 160–168.

Cooper, C., Frieze, A., Radzik, T., 2011. The cover times of random walks on hypergraphs. In: Kosowski, A., Yamashita, M. (Eds.), Structural Information and Communication Complexity, no. 6796 in Lecture Notes in Computer Science. Springer Berlin Heidelberg, URL http://link.springer.com/chapter/10.1007/978-3-642-22212-2_19, pp. 210–221.

Evans, J.A., Foster, J.G., 2011. Metaknowledge. Science 331 (6018), 721–725.

Foote, R., 2007. Mathematics and complex systems. Science 318 (5849), 410–412.

Foster, J.G., Rzhetsky, A., Evans, J.A., 2013. Tradition and innovation in scientists' research strategies. arXiv:1302.6906

Fawcett, T., 2006. An introduction to roc analysis. Pattern Recognit. Lett. 27 (8), 861–874.

Faust, K., 1997. Centrality in affiliation networks. Soc. Netw. 19 (2), 157–191.

Foster, J.A., Rich, C.B., DeSa, M.D., Jackson, A.S., Fletcher, S., 1980. Improved methodologies for the isolation and purification of tropoelastin. Anal. Biochem. 108 (2), 233–236.

Guimerà, R., Uzzi, B., Spiro, J., Amaral, L.A.N., 2005. Team assembly mechanisms determine collaboration network structure and team performance. Science 308 (5722), 697–702.

Goodreau, S.M., 2007. Advances in exponential random graph (p*) models applied to a large social network. Soc. Netw. 29 (2), 231–248, http://dx.doi.org/10.1016/j.socnet.2006.08.001

Hasan, M.A., Zaki, M.J., 2011. A survey of link prediction in social networks. In: Aggarwal, C.C. (Ed.), Social Network Data Analytics. Springer, US, pp. 243–275, http://dx.doi.org/10.1007/978-1-4419-8462-3_9

Hinton, G.E., 1992. How neural networks learn from experience. Sci. Am. 267 (3), 144–151, PMID: 1502516.

Kossinets, G., Watts, D.J., 2006. Empirical analysis of an evolving social network. Science 311 (5757), 88–90.

Kossinets, G., Watts, D.J., 2006. Empirical analysis of an evolving social network. Science 311 (5757), 88–90, http://dx.doi.org/10.1126/science.1116869

Latour, B., 1987. Science in Action: How to Follow Scientists and Engineers Through Society. Harvard University Press.

Latour, B., 1999. Pandora's Hope: Essays on the Reality of Science Studies. Harvard University Press.

Latour, B., Woolgar, S., 1986. Laboratory Life: The Construction of Scientific Facts. Princeton University Press.

Latour, B., 2005. Reassembling the Social-An Introduction to Actor-Network-Theory. Oxford University Press.

Leydesdorff, L., Rotolo, D., Rafols, I., 2012. Bibliometric perspectives on medical innovation using the medical subject headings of PubMed. J. Am. Soc. Inf. Sci. Technol. 63 (11), 2239–2253, http://dx.doi.org/10.1002/asi.22715

Liben-Nowell, D., Kleinberg, J., 2007. The link-prediction problem for social networks. J. Am. Soc. Inf. Sci. Technol. 58 (7), 1019–1031, http://dx.doi.org/10.1002/asi.20591/full

Lindelauf, R., Borm, P., Hamers, H., 2012. One-mode projection analysis and design of covert affiliation networks. Soc. Netw. 34 (4), 614–622.

Martin, T., Ball, B., Karrer, B., Newman, M.E.J., 2013. Coauthorship and citation patterns in the physical review. Phys. Rev. E 88 (1), 012814, http://dx.doi.org/10.1103/PhysRevE.88.012814

Menon, A.K., Elkan, C.,2011. Link prediction via matrix factorization. In: in: Machine Learning and Knowledge Discovery in Databases. Springer, pp. 437–452, http://dx.doi.org/10.1007/978-3-642-23783-6_28

Newman, M.E.J., 2001. The structure of scientific collaboration networks. Proc. Natl. Acad. Sci. 98 (2), 404–409, http://dx.doi.org/10.1073/pnas.98.2.404 http://www.pnas.org/content/98/2/404, PMID: 11149952.

Newman, M.E.J., 2004. Coauthorship networks and patterns of scientific collaboration. Proc. Natl. Acad. Sci. 101 (Suppl. 1), 5200–5205, http://dx.doi.org/10.1073/pnas.0307545100 http://www.pnas.org/content/101/suppl_1/5200, PMID: 14745042.

Newell, A., Simon, H.A., 1972. Human Problem Solving. Prentice-Hall.

Newman, M.E.J., 2001. Scientific collaboration networks, II. Shortest paths, weighted networks, and centrality. Phys. Rev. E 64 (1), 016132, http://dx.doi.org/10.1103/PhysRevE.64.016132

Norman, K.A., Newman, E.L., Detre, G., 2007. A neural network model of retrieval-induced forgetting. Psychol. Rev. 114 (4), 887.

Rapoport, A., 1953. Spread of information through a population with socio-structural bias: I. Assumption of transitivity. Bull. Math. Biophys. 15 (4), 523–533.

Robins, G., Snijders, T., Wang, P., Handcock, M., Pattison, P., 2007. Recent developments in exponential random graph (p*) models for social networks. Soc. Netw. 29 (2), 192–215, http://dx.doi.org/10.1016/j.socnet.2006.08.003

Ripley, B.D., 2007. Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge, New York.

Snijders, T.A.B., 2002. Markov chain Monte Carlo estimation of exponential random graph models. J. Soc. Struct. 3.

Snijders, T.A.B., Pattison, P.E., Robins, G.L., Handcock, M.S., 2006. New specifications for exponential random graph models. Sociol. Methodol. 36 (1), 99–153, http://dx.doi.org/10.1111/j.1467-9531.2006.00176.x

Taramasco, C., Cointet, J.-P., Roth, C., 2010. Academic team formation as evolving hypergraphs. Scientometrics 85 (3), 721–740, http://dx.doi.org/10.1007/s11192-010-0226-4

Torvik, V.I., Smalheiser, N.R., 2009. Author name disambiguation in MEDLINE. ACM Trans. Knowl. Discov. Data 3 (3), http://dx.doi.org/10.1145/1552303.1552304

Uzzi, B., Spiro, J., 2005. Collaboration and creativity: the small world problem. Am. J. Sociol. 111 (2), 447–504, http://dx.doi.org/10.1086/ajs.2005.111.issue-2 http://www.jstor.org/stable/10.1086/432782

Weibel, P., Latour, B., 2005. Making things public: atmospheres of democracy: [exhibition], ZKM, Center for art and media Karlsruhe, 20.03.-03-10.2005, (Mass.). MIT press.

Zhu, M., Huang, Y., Contractor, N.S., 2013. Motivations for self-assembling into project teams. Soc. Netw. 35 (2), 251–264.