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.