Journal of Informetrics 12 (2018) 181–190

Contents lists available at ScienceDirect

Journal of Informetrics

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

Regular article

Network assembly of scientific communities of varying size and specificity

Daniel T. Citron[a],*, Samuel F. Way[b]

[a] Cornell University, United States
[b] University of Colorado Boulder, United States

ARTICLE INFO

Article history:
Received 22 May 2017
Received in revised form
18 December 2017
Accepted 18 December 2017

Keywords:
Collaboration networks
Network assembly
Social network analysis
Topic modeling
Scientometrics

ABSTRACT

How does the collaboration network of researchers coalesce around a scientific topic? What sort of social restructuring occurs as a new field develops? Previous empirical explorations of these questions have examined the evolution of co-authorship networks associated with several fields of science, each noting a characteristic shift in network structure as fields develop. Historically, however, such studies have tended to rely on manually annotated datasets and therefore only consider a handful of disciplines, calling into question the universality of the observed structural signature. To overcome this limitation and test the robustness of this phenomenon, we use a comprehensive dataset of over 189,000 scientific articles and develop a framework for partitioning articles and their authors into coherent, semantically related groups representing scientific fields of varying size and specificity. We then use the resulting population of fields to study the structure of evolving co-authorship networks. Consistent with earlier findings, we observe a global topological transition as the co-authorship networks coalesce from a disjointed aggregate into a dense giant connected component that dominates the network. We validate these results using a separate, complimentary corpus of scientific articles, and, overall, we find that the previously reported characteristic structural evolution of a scientific field's associated co-authorship network is robust across a large number of scientific fields of varying size, scope, and specificity. Additionally, the framework developed in this study may be used in other scientometric contexts in order to extend studies to compare across a larger range of scientific disciplines.

1. Introduction

A co-authorship network outlines the professional connections between scientific researchers and their collaborators. Co-authorship networks are important objects of study, as they are a measurable representation of the communities that assemble in order to work in an particular area of research. Such communities allow for the transfer of knowledge and skills and sharing of resources required for researching complex problems (Börner et al., 2010; de Solla Price, 1986; Guimera, Uzzi, Spiro, & Amaral, 2005; Kaiser, 2005). The assembly of co-authorship networks represents one aspect of the more general problem of understanding the process through which social or collaborative networks attract new members and evolve structurally over time (Backstrom, Huttenlocher, Kleinberg, & Lan, 2006; Jacobs, Way, Ugander, & Clauset, 2015).

* Corresponding author.
E-mail addresses: dtc65@cornell.edu (D.T. Citron), samuel.way@colorado.edu (S.F. Way).

https://doi.org/10.1016/j.joi.2017.12.008
1751-1577/© 2017 Elsevier Ltd. All rights reserved.

The recent availability of electronic publishing and online repositories of scientific articles has enabled large-scale studies of scientific research practices (Börner & Shiffrin, 2004; Ginsparg, Houle, Joachims, & Sul, 2004; Tabah, 1999). In particular, these repositories provide record of collaborations between the authors of each paper, making it possible to construct comprehensive co-authorship networks and analyze their assembly over time. Two recent studies have investigated the development of a small group of research fields (9 and 12 fields, respectively), by measuring the assembly of each field's co-authorship network using a large electronic collection of articles (Bettencourt & Kaiser, 2015; Bettencourt, Kaiser, & Kaur, 2009). Expanding upon historiographical surveys, they search for patterns in the growth and development of co-authorship networks across different scientific fields. These studies argue that while each field differs in size and publishing practices (differing in rate of publication, size of collaborations, etc.), nevertheless there appear to be common patterns in how each field's co-authorship network develops. Specifically, each co-authorship network undergoes a topological transition in which a densely connected giant component of researchers forms over time. This dramatic structural change has been compared to the emergence of a giant component seen in a percolation transition (Newman, 2010), and serves as an empirical indication that the research community undergoes large-scale social reorganization as more researchers join and collaborate with others (Bettencourt et al., 2009; Bettencourt & Kaiser, 2015; Guimera et al., 2005).

Another study (Lee, Goh, Kahng, & Kim, 2010) takes three example fields (complex networks research; ADS/CFT; Randall–Sundrum model) and describes three stages of development characteristic to co-authorship network assembly in science. Each network begins as a set of disconnected groups, which then join together to form a large treelike component. As the research community grows and mixes further, the large component becomes densely connected to itself through the formation of long-range ties. This general pattern is consistent with what was reported in Bettencourt and Kaiser (2015) and Bettencourt et al. (2009), which also emphasized how the long-range ties between authors created a densely connected community with very short distances between different authors.

Together, these previous studies suggest the existence of common patterns in how scientific communities assemble over time. However, they rely on manual annotation of their data, which requires a great deal of labor in order to assemble a co-authorship network. This in turn limits the number of examples studied and reported on, making it difficult to justify the claim that the patterns observed for a few examples are universal across all scientific fields.

In the present study, we propose a framework for analyzing a large population of example topics in order to verify that the development of co-authorship networks, as characterized by earlier studies, is robust across many scientific fields. Specifically, we use techniques from natural language processing and machine learning to generate a larger set of example co-authorship networks from the arXiv, a large scientific corpus. We use topic modeling to cluster articles together based on their semantic content, and interpret the clusters of articles as representing different fields of science. We measure the algorithmically-generated co-authorship networks to determine whether they develop in a manner similar to the manually-annotated co-authorship networks studied previously. We aim to facilitate a larger survey of co-authorship networks across scientific fields first by testing the efficacy of topic modeling as a way to rapidly detect a large number of fields, and then by comparing the assembly behavior of each field's co-authorship network for the purposes of testing whether their growth patterns remain consistent for a large set of fields of varying size and specificity.

2. Data set

The arXiv is an open-access repository of scientific preprints accessible online at www.arxiv.org. The site was founded in 1991 and, as of the end of 2016, hosts over 1.1 million articles, primarily in the areas of Physics, Mathematics, and Computer Science (arXiv, 2016). Here, we take as our data set the 189,000 articles categorized as Condensed Matter Physics ("cond-mat" on the arXiv) by the submitting author (or by the arXiv's administrators) during the period starting in April of 1992 and ending in June 2015.

The arXiv data have several important advantages for the purposes of the present study. The articles' full texts and relevant metadata are available to the public. Additionally, arXiv has been well studied from a scientometric perspective (Larivière et al., 2014), and has been used to test techniques for algorithmically categorizing scientific articles according to their content (Ginsparg et al., 2004).

The set of arXiv articles is only a sample of all published works, and, due to differences in the site's adoption across communities, arXiv's coverage varies from one subfield to the next. We therefore test that our results obtained by measuring the arXiv actually represent real-world co-authorship networks and not an artifact of the arXiv's incompleteness. Specifically, to validate our results, we also analyze a subset of the condensed matter articles found on the Web of Science (WoS). WoS is a database of scientific articles maintained by Clarivate Analytics. We use the 660,000 articles classified as Condensed Matter Physics published between April 1992 and June 2015, requiring that all have titles, abstracts, and author names available in the database (Certain data included herein are derived from Clarivate Analytics Web of Science TM., 2017). The set of articles from Web of Science partially overlaps with the arXiv data set and represents a complementary data set with non-uniform coverage of the subfields contained on arXiv (Larivière et al., 2014). Using the WoS as a secondary data set makes it possible to verify whether the arXiv contains a truly representative sample of Condensed Matter Physics articles, as well as to check whether the results obtained using the articles from the arXiv are not merely an artifact of the arXiv's incomplete coverage of certain scientific subfields.

To track the contributions of individual authors, we adopt the convention of labeling each author with their uppercase full names as reported in the publication metadata. In the context of co-authorship network measurement, this author naming