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
Available online 3 January 2018

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/

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

convention errs on the side of splitting individual authors into multiple entities. That is to say, authors who inconsistently report their names in publications will be counted as multiple separate nodes for the purposes of this study. This convention also decreases the possibility of many different entities becoming combined into a single composite node, which would artificially collapse together many different nodes in our co-authorship networks. We verify that our results are robust to changing the author labeling convention by repeating all subsequent analysis using "[First Initial] [Last Name]" in Appendix C. Larger-scale analyses involving a broader reach of disciplines will require additional steps to disambiguate author identities (such as the tools described in Bhattacharya and Getoor (2007) and Song, Huang, Councill, Li, and Giles (2007). After preprocessing author names in this way, the arXiv data set includes 96,000 unique authors.

For the purposes of text mining and topic modeling we focus on each article's title and abstract under the assumption that authors write titles and abstracts with the intention of concisely summarizing an article's contents. Past studies have argued that focusing on article abstracts has the additional benefit of minimizing the amount of "structural" text processed by the topic model, allowing the inferred topic structures to focus on field-specific content, rather than commonalities in presentation of the English language (Ginsparg et al., 2004; Joachims, 2002).

3. Methods

3.1. Topic model

Past studies exploring the formation of co-authorship networks have relied on manual annotation to determine which authors contribute to and are therefore considered part of a scientific field (Bettencourt et al., 2009; Bettencourt & Kaiser, 2015; Lee et al., 2010). This approach, however, requires a great deal of human effort and, consequently, has been applied to only a few disciplines and with somewhat arbitrary definitions of which publications and authors belong to the community in question. It therefore remains unclear how robust past results are to varying the criteria for selecting communities, and for varying levels of specificity governing the breadth and size of such communities.

To address these limitations, we introduce an approach that uses topic modeling to automate the process of identifying groups of semantically-related documents and partitioning their authors into fields corresponding to their areas of expertise (Boyack, Klavans, & Börner, 2005). As a consequence of the number of documents belonging to a given subfield and the commonality of its language, the topics and thus the fields extracted by this technique will vary in terms of size and specificity, yielding a population of corresponding co-authorship networks. That is, we can test whether the reported structural patterns are robust to varying definitions of sub-community. At the same time, we explore the usefulness of topic modeling as an automated, scalable means for partitioning the global network of all researchers into co-authorship networks organized around specific fields.

Topic modeling is an unsupervised machine learning technique that characterizes the underlying thematic content of a given corpus by identifying groups of semantically-related, co-occurring words—the "topics"—while simultaneously identifying the proportion of each topic present in each document in the corpus. Here, we use latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003; Griffiths & Steyvers, 2004), a popular topic model that produces static definitions for topics, formalized as probability distributions over all words in a given vocabulary. Accordingly, for each document the model infers a distribution over these topics.

Prior to applying topic modeling, we utilize several common natural language processing techniques to preprocess the corpus text. In particular, we combine the text from each article's title and abstract into a single document, remove all non-alphabetic characters, and convert all letters to lowercase. Common English stop words ("the," "and," "of," etc.) are also removed, as well as certain words that appear very commonly in the arXiv data set but that contain no scientific content (numbers, names of publishers, "thank you," etc.). The document text is also lemmatized in order to increase the likelihood of discovering overlaps in the word usage within and between documents

After preprocessing all articles, we use MALLET (McCallum, 2002), an open-source implementation of LDA, to train a series of topic models, varying the number of topics between $k = 25$ and $k = 100$. As expected, for small k, LDA produces broadly-defined topics, and for large k, more narrowly-defined topics. For our purposes, $k = 50$ provides sufficient resolution for the model to recover topics that resemble established subfields within condensed matter physics. We emphasize that we do not intend to use this topic model to represent the optimal or definitive partition of arXiv according to subject matter. Rather, our model provides a large set of readily-interpretable topics, varying in both size and specificity, allowing us to test the robustness of past claims against a heterogeneous population of fields and their corresponding authors. We present our analysis of the $k = 50$ topic model below and note that our results are robust to small changes in k. That is, the results that we report below do not change significantly if we repeat our subsequent analyses using a model with $k = 45$ or $k = 55$ topics.

After training our topic model, we manually inspect each topic to determine whether it resembles a field of condensed matter physics. As an example, the most probable words associated with Topic 5 include keywords such as "quantum," "state," "qubit," "entanglement," and "decoherence." Looking at the set of articles to which the topic model assigns a high probability ($P(\text{Topic} = 5) > 0.6$), we find articles such as "Demonstration of Two-Qubit Algorithms with a Superconducting Quantum Processor" (0903.2030) and "Controllable coupling between flux qubits" (cond-mat/0507496). Together, these observations suggest that articles strongly associated with Topic 5 are related to quantum computing and quantum information. We also check that the articles identified by the topic model do not merely reflect clusters of articles specific to arXiv by inferring topics on the articles belonging to the Web of Science (WoS) data set. In the case of Topic 5, we find articles such as "Flexible

two-qubit controlled phase gate in a hybrid solid-state system" and "Two-electron coherence and its measurement in electron quantum optics," which confirms that articles associated with Topic 5 appear to be related to quantum computing.

In addition to quantum computing, LDA recovers topics resembling other established subfields of condensed matter physics, including spin glasses (Topic 1); Bose–Einstein condensates (Topic 3); magnetic materials (Topic 19); glassy physics (Topic 28); topological phases (Topic 30); and cuprate superconductors (Topic 43). (Refer to Appendix A to see each topic's interpretation.)

3.2. Co-authorship network generation

We use our topic model to construct a set of co-authorship networks, where each network represents the set of authors that produced the articles strongly associated with one of the topics discovered by the topic model. We emphasize that the topic modeling algorithm is only given information related to the textual content of the articles and receives no information about authorship, authors' collaborative relationships, or publication dates. While there are topic modeling algorithms that do take into account other links between documents (e.g. Guo, Zhu, Chi, Zhang, & Gong, 2009; Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2004), we want to determine whether textual content is sufficient to reproduce patterns in how groups of researchers in the same related form a collaborative community.

We find the articles that are primarily associated with each topic t by selecting the subset of articles assigned a probability weight $P(t) > 0.6$. The cutoff at 0.6 selects articles that are strongly associated with one particular topic, but is not so strict that it excludes too many articles. With $P(t) > 0.6$, we associate between 100 and 3000 arXiv articles with each topic. We also use an alternative thresholding criterion to check whether the choice of thresholding biases our results. We repeat all subsequent analysis using a second method of categorizing articles whereby each article is assigned to the smallest set of topics that account for 50% of its subject matter. All reported results are robust to varying the thresholding scheme.

We construct a co-authorship network by identifying the authors of each topic's associated articles. Each author is represented in the topic's network as a node. Two author nodes are linked by an edge if they have written an article together (Newman, 2001, 2004). Hence, a group of authors who collaborated on an article together appears in the network as a fully connected clique, and two articles with multiple authors in common will appear in the network as overlapping cliques that share nodes. (We also use a modularity score to measure the extent to which authors associated with different topics connect to one another. We find that our topic model does tend to sort authors into distinct communities in D.)

We reconstruct each co-authorship network's assembly and growth over time using each month of arXiv's operation from April 1992 through June 2015 as a discrete time step. At each time step we include in the network all author nodes that have written articles at or prior to the current time step. We also connect all pairs of author nodes that have collaborated on one or more articles at or prior to the current time step.

4. Results

4.1. Co-authorship network measurements

Fig. 1 shows the network growth for four different example topics: quantum computing (Topic 5), magnetic material properties (Topic 19), transport measurements (Topic 12), and mechanical properties of materials (Topic 41). For the first two topics in Fig. 1 there appear to be three separate stages through which the giant component develops. Each network begins as a disjointed set of cliques, as the authors who share a field publish in separate groups. Next, a few of the cliques join together, forming a loosely connected, almost tree-like backbone of connected cliques. In the final stage, enough cliques overlap with one another such that the largest connected component becomes densely connected. This characteristic three-stage pattern is consistent with what has been reported previously (Lee et al., 2010). By contrast, the largest component of Topic 12's network only grows to reach the treelike stage, and Topic 41's network has no giant component.

We confirm this interpretation of the network visualizations by measuring various properties of each topic's co-authorship network. We measure the fraction of nodes belonging to the largest connected component ("giant component size"). We also measure the giant component's mean geodesic path length between all pairs of nodes belonging to the giant connected component ("mean path length"). The mean path length ranges between a minimum for fully connected networks and a maximum for treelike networks, and so serves as a measure of how closely connected the individuals belonging to the giant component are to one another (Bettencourt et al., 2009; Leskovec, Kleinberg, & Faloutsos, 2005; Newman, 2010).

Fig. 2 shows measurements of the size and mean path length of the giant component for each of the topics shown in Fig. 1. For Topics 5 and 19 (two leftmost columns), the giant component's size increases steadily as more and more nodes are added to the network. At the same time, the mean path length first increases as the giant component grows initially and then peaks and decreases (Leskovec et al., 2005). This non-monotonic behavior suggests two stages in the development of the giant component: initial growth as cliques first start to overlap with one another, and densification when enough "long-range" edges form to reduce the average distance between authors (Lee et al., 2010; Newman, 2010; Watts & Strogatz, 1998) These two growth stages are consistent with a treelike cluster of cliques that becomes a densely connected cluster. As a point of comparison, the largest component in Topic 12 does grow to include a large fraction of the nodes in the network, but its mean path length increases steadily over time.

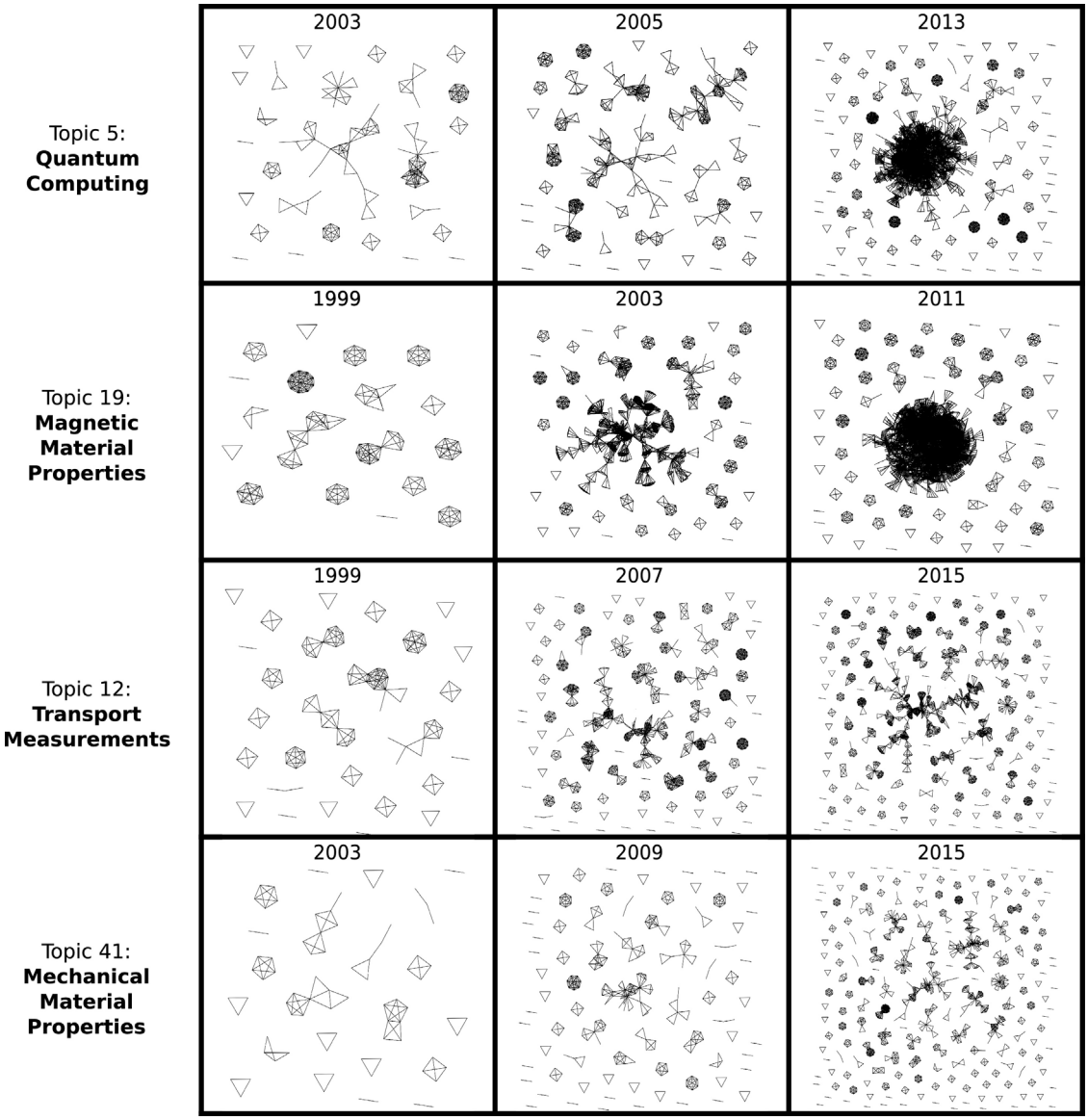

Fig. 1. Examples of different network behaviors. Each row shows how a topic's co-authorship network develops over time, with network snapshots labeled by the year observed. Each node represents an author, and each edge represents a collaboration between the two authors. Disconnected single nodes are not pictured. The top two rows (Topics 5 and 19) illustrate three stages of network assembly: disjointed group of cliques; treelike connected cluster of cliques; densely connected giant component. The third row (Topic 12) is an example of a network that only forms a treelike largest component. The bottom row (Topic 41) is an example of a network that forms no single giant component.

The co-authorship network development patterns are not merely the result of sampling a large number of articles that join together by chance. For comparison, we consider a null model in which articles are grouped together at random, rather than grouped together according to topic modeling, to test whether the topic modeling is responsible for identifying the clusters of authors. For each instance of the null model, thousands of articles are selected from the arXiv cond-mat data set at random. The co-authorship network of this randomly-selected group of articles is then constructed, and the properties of the largest connected component are measured. The results of this null model are plotted in gray in Fig. 2, where the vertical height of the gray region represents the mean $\pm$ one standard deviation across 100 instances of the null model. The null model's average behavior contrasts dramatically with the measurements of the scientific co-authorship networks identified using the topic model. These results strongly suggest that the aggregation of authors to form a giant, densely connected component is not merely the result of sampling an arbitrary subset of arXiv. Rather, it appears that the topic model, which was given no information about authorship or other such links between documents, was able to identify clusters of researchers based on their textual content alone. The nonrandom grouping of authors further validates the topic model's meaningful clustering of articles: the articles represent the output of an association of researchers with similar interests.

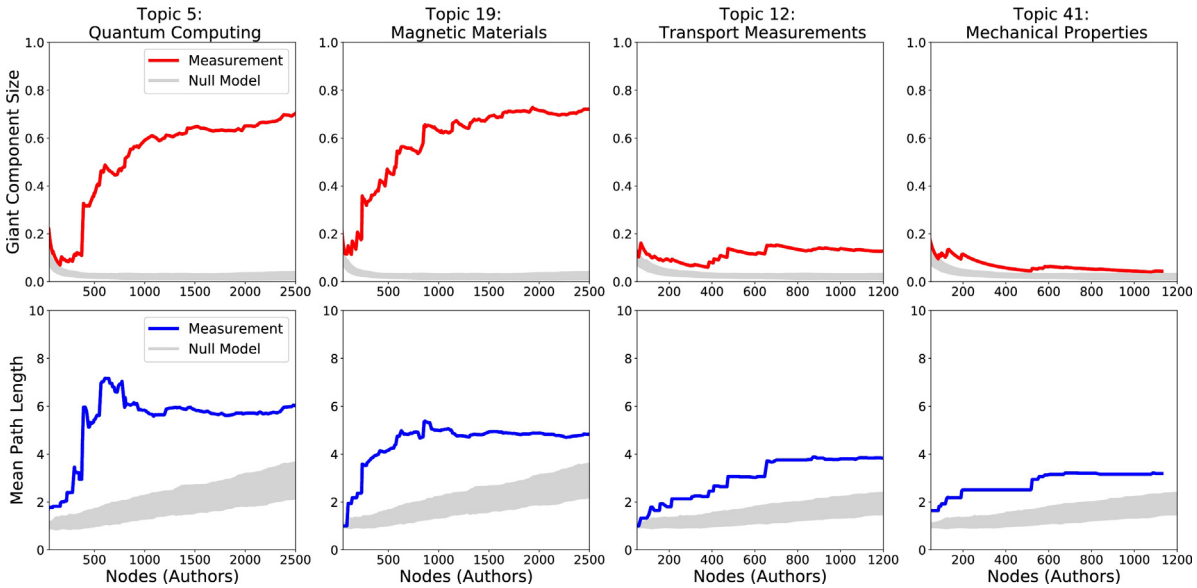

Fig. 2. Quantitative measurements of co-authorship networks. The top row shows the fraction of nodes belonging to the largest component as a measure of network size, plotted vs. the total number of nodes in the network. The bottom row shows the mean geodesic path length of the largest component ("mean path length") vs. the total number of nodes in the network. For Topics 5 and 19, the largest component grows to dominate the network. As the largest component grows, its mean path length increases quickly at first and then begins to decrease. For Topic 12, a single large component grows, but remains treelike and its mean path length only continues to increase. For Topic 41, no giant component forms.

The example topics shown in Figs. 1 and 2 exemplify three general types of network assembly observed for the other topics. Out of the 50 topics, 22 have co-authorship networks that undergo the transition from a scattered collection of cliques; to an extended, treelike connected group of cliques; to a densely connected giant component. These results are qualitatively consistent with those obtained earlier for groups of articles annotated by human experts (Bettencourt et al., 2009; Bettencourt & Kaiser, 2015). From the remaining topics, 17 form a single large component that occupies a small fraction of nodes in the network, but have not yet formed enough long-range ties that the network mean path length stops growing monotonically. The remaining 9 topics show little or no sign that they form any giant connected component. (Refer to Appendix B for a summary of all co-authorship networks' behavior.)

Finding that a topic's corresponding co-authorship network does not form a densely connected GCC does not necessarily suggest that the research field is not well-established. There are several possible reasons why a dense giant component does not form in all cases. The existence of a giant component only indicates that there are a great many researchers that have collaborated with one another. Inter-group collaborations may be more frequent in some fields than in others, and a giant component is only likely to form when there are many collaborations between research groups. Additionally, the arXiv does not necessarily represent a comprehensive sampling of articles from all subfields of science. The arXiv's coverage of some fields may be incomplete, such as microscopy (Topic 15) and surface chemistry (Topic 47).

4.2. Validation across corpora

The characteristic growth patterns seen for the co-authorship networks of authors from the arXiv remain consistent when we repeat the same analysis using another corpus. We use the topic model trained on the arXiv data to infer topics for the condensed matter physics articles from the Web of Science (WoS). The same procedures for generating and measuring the co-authorship networks for the WoS articles reveals that the topic model trained on the arXiv is still able to identify large connected clusters of articles in the WoS. Fig. 3 compares the behavior of the co-authorship networks that occur within both the arXiv and WoS.

In the majority of cases, the co-authorship networks identified from the WoS articles behave similarly to the ones identified on arXiv. For example, the co-authorship networks for research on quantum computing and magnetic material properties (Topics 5 and 19, the two leftmost columns of Fig. 3) form a dense giant component for both arXiv and for WoS. There is also a group of topics whose networks form only a treelike giant component or no giant component in the arXiv data but do form a dense component with a shrinking mean path length in the WoS data. Topics that do this include transport measurements and mechanical material properties (Topics 12 and 41, shown in the two rightmost columns of Fig. 3), as well as nanoscale devices (Topic 16) and inelastic scattering experiments (Topic 33). We note that these topics have an experimental focus. Experimental research subjects are known to have less coverage on arXiv, but are covered more comprehensively in the WoS (Larivière et al., 2014). There are also a few topics with decreased coverage on WoS because the WoS does not categorize them as condensed matter. For example, articles on ultracold atoms (Topics 3 and 20) may be categorized separately

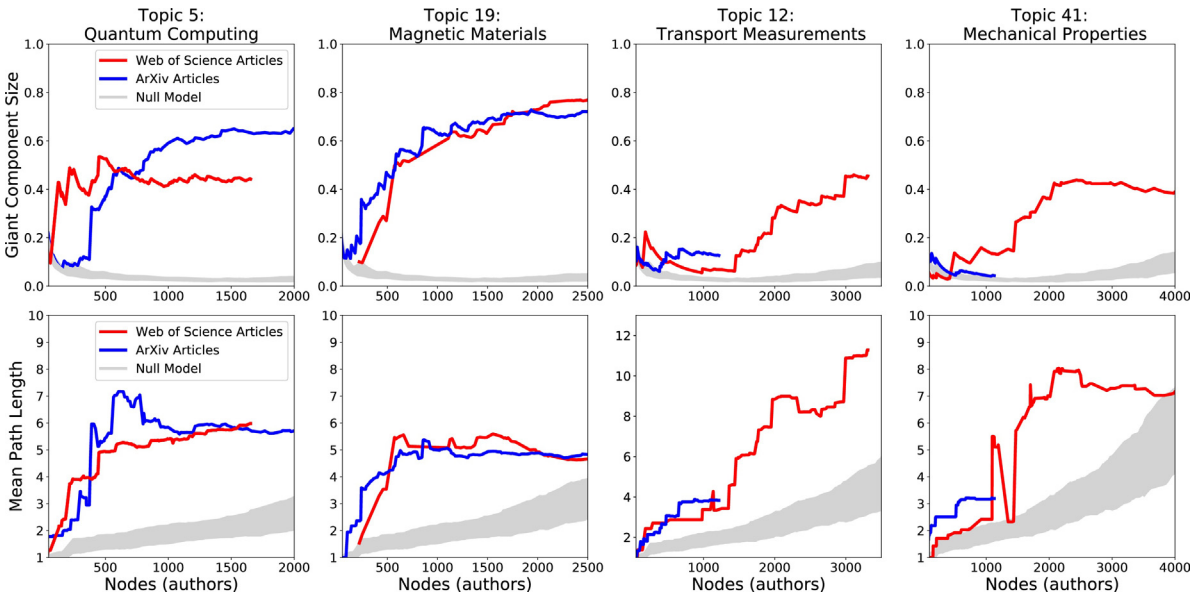

Fig. 3. Comparison between co-authorship networks generated from arXiv and Web of Science. Each column corresponds to a different topic. The top row shows the fraction of nodes belonging to the largest component as a measure of network size vs. the total number of nodes in the network. The bottom row shows the mean geodesic path length of the largest component ("mean path length") vs. the total number of nodes in the network. Each plot shows the measurements made of the co-authorship network from the Web of Science (in red), from arXiv (in blue), as well as co-authorship networks generated from randomly chosen articles from Web of Science (null model, in gray). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

as "atomic, molecular, and optical physics" and articles on soft condensed matter (Topics 25 and 50) may be categorized separately as "fluids." Consequently, these topics' decreased inclusion in the WoS data set leads to smaller and less densely connected co-authorship networks.

Overall, 27 out of 50 topics have co-authorship networks that develop similarly for both the WoS data and the arXiv data (Appendix B). Additionally, 10 experimentally-focused topics have co-authorship networks that grow to form large giant components on account of having better coverage on the WoS compared to the arXiv. Another three topics (Topics 9, 10, and 42) have very low coverage on the arXiv (fewer than 100 associated articles) and do not form giant connected components with either the arXiv or the WoS. Given that, across both corpora, none of these three topics has many strongly associated articles, it is likely that Topics 9, 10, and 42 are actually "junk topics," meaning that they do not reflect coherent themes and so are not useful for the purposes of the present study. The consistency of the behavior of these co-authorship networks measured across different corpora suggests that the collaborative communities identified using the model are reflected in multiple data sets.

4.3. Robustness to edge removal

Finally, we address the question of whether the co-authorship network development patterns seen in our data and in previous studies are robust to relaxing the assumption that all edges in the co-authorship network are maintained indefinitely after they are established. Previous studies have constructed co-authorship networks wherein that collaborative link, once established, are maintained forever (Bettencourt et al., 2009; Bettencourt & Kaiser, 2015; Lee et al., 2010). In practice, when such a collaborative relationship requires significant efforts to maintain, this assumption is not necessarily valid.

We re-assemble the co-authorship networks for each of the topics, this time allowing edges to expire after a fixed number of months. That is to say, if two authors do not repeat a collaboration after a certain amount of time, the edge representing their relationship is removed from the network. The results are plotted in Fig. 4, where the uppermost curve (gray; "no limit") shows how the giant component grows if edges survive indefinitely, while the other curves show how those measurements change if the edges are removed after 2 (blue), 5 (green), or 10 (red) years.

Limiting the lifetime of edges to a few years causes giant components to develop much more slowly, or to not develop at all. For Topics 5 and 19, the network measurements for 5 and 10 years are very close to the indefinite lifetime limit. This suggests that these networks are particularly robust to edge removal, reflecting a very densely connected giant component where edges are frequently renewed (Lee et al., 2010). For Topic 38, the giant component forms much more slowly, and actually begins to disassemble for edge lifetimes of 2 or 5 years. For Topic 12, finite edge lifetime only suppresses the component formation of a large component. (Appendix E contains additional visualizations of these graphs, comparable to those appearing in Fig. 1.)

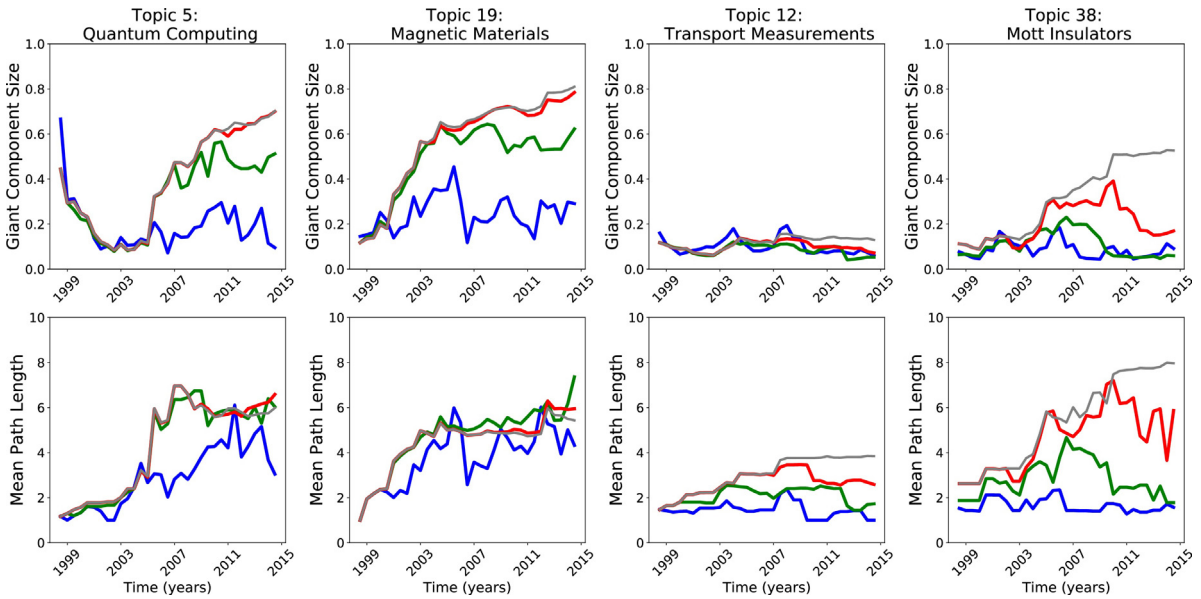

Fig. 4. Network Robustness to Edge Removal. For four topics, we show how the network assembly changes when edges only remain in the network for a limited amount of time. Each plot shows the network's giant component size over time for four different edge lifetimes. For short edge lifetimes (2 years in blue; 5 years in green), the giant connected component fails to develop or develops much more slowly compared to the permanent edge ("no limit," gray) case. For longer edge lifetimes (10 years, red), the giant component approaches the no limit case. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Currently, it is unknown what criteria for including and excluding nodes and edges from co-authorship network models best reflect the reality of authors entering and exiting different fields. What is clear, however, is that the assumption that the relationships represented by edges between authors last forever is important for obtaining the quantitative results that reflect a topological transition in the co-authorship network. Shortening the lifetime of edges can dramatically change a co-authorship network's evolution over time.

5. Discussion

This study expands upon previous research exploring the growth and development of co-authorship networks using topic modeling to algorithmically identify and study a large population of scientific fields, along with their associated articles and authors. Our results show that, for the topics determined using LDA, a large majority of co-authorship networks undergo a topological transition to form a densely-connected giant component characterized by three stages of development. These patterns corroborate findings from earlier studies that focused on small numbers of (often manually assembled) co-authorship networks. Our results demonstrate that the characteristic topological transition is robust to variations in the definition of a scientific field, both in terms of size and specificity. Additionally, our methods employ algorithmic clustering and require no input from human experts, yet the results are largely consistent with previous studies. We also found that the patterns in co-authorship network development are consistent across corpora, which we demonstrate by repeating our analysis using data from both the arXiv and the Web of Science. One notable difference between the two corpora is reflected in how arXiv's selections of articles related to certain experimentally-focused topics are under-populated: in these cases, the co-authorship networks constructed using the larger WoS data set undergo a topological transition, while the corresponding networks drawn from the arXiv data do not.

Topic modeling is a rich and actively growing area of research within the statistical modeling and natural language processing communities. In our study, we used latent Dirichlet allocation, one of the most popular yet simplest forms of topic modeling. This model assumes a static definition for topics and thus scientific communities, which are known evolve with time. Additionally, the model does not directly incorporate other, non-semantic relationships between documents (such as co-authorship or citations), which may signal alternate forms of cohesion within a scientific community. For our purposes, we consider the assembly and development co-authorship networks over relatively short periods of time and thus favor LDA's straightforward approach. Future work in this area, however, should explore more sophisticated algorithms that consider topic dynamics (e.g. Blei & Lafferty, 2006; Wang & McCallum, 2006) and additional measures of community cohesion in order to more thoroughly address the co-evolution of scientific fields.

Our method for algorithmically generating and analyzing a large number of fields can also be used as a framework for further exploring the claims made in a wide variety of bibliometric contexts. For example, one could also perform a comparison of the micro-scale dynamics of individual authors many different fields. Recent studies have used agent-based

models of author behavior to explain the patterns in publishing behavior that one sees in different fields of science (e.g. Boyack et al., 2005; Sun, Kaur, Milojević, Flammini, & Menczer, 2013). Once again, most of these studies have relied on manually annotated data sets, and as such, they have historically been limited to only a handful of fields. The approach that we develop in this study, however, enables future work, in conjunction with comprehensive data sets like the arXiv or Web of Science, to further test the accuracy of these models of author behavior across a large and diverse population of scientific fields.

Author contributions

Conceived and designed the analysis: Daniel T. Citron.
Collected the data: Daniel T. Citron. Contributed data or analysis tools: Samuel F. Way
Performed the analysis: Daniel T. Citron.
Wrote the paper: Daniel T. Citron, Samuel F. Way. Other contribution: Samuel F. Way.

Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1144153, and NSF award SMA 1633747. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors would also like to acknowledge Michael W. Macy, Paul H. Ginsparg, Alexandra Schofield, and Haofei Wei, as well as Brent Schneeman, Laurence Brandenberger, Richard Barnes, and the other attendees of the Santa Fe Institute's 2015 Complex Systems Summer School for helpful discussions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.joi.2017.12.008.

References

arXiv (2016). arXiv submission rate statistics, 2016. http://arxiv.org/help/stats/2016_by_area/index.
Börner, K., & Shiffrin, R. M. (2004). Mapping knowledge domains. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 5183L 5185.
Börner, K., Contractor, N., Falk-Krzesinski, H. J., Fiore, S. M., Hall, K. L., Keyton, J., et al. (2010). A multi-level systems perspective for the science of team science? *Science Translational Medicine*, *2*(49), 49cm24–49cm24
Backstrom, L., Huttenlocher, D., Kleinberg, J., & Lan, X. (2006). Group formation in large social networks: Membership, growth, and evolution. *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, ACM*, 44–54.
Bettencourt, L. M. A., & Kaiser, D. I. (2015). *Formation of scientific fields as a universal topological transition.* , arXiv.org.
Bettencourt, L. M. A., Kaiser, D. I., & Kaur, J. (2009). Scientific discovery and topological transitions in collaboration networks? *Journal of Informetrics*, *3*(3), 210–221.
Bhattacharya, I., & Getoor, L. (2007). Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *1*(1), 5.
Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd international conference on machine learning, ACM*, 113–120.
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993L 1022.
Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science? *Scientometrics*, *64*(3), 351–374.
Certain data included herein are derived from Clarivate Analytics Web of Science TM. © Copyright Clarivate Analytics 2017. All rights reserved.
de Solla Price, D. J. (1986). *Little science, big science. . . and beyond*. Columbia University Press.
Ginsparg, P., Houle, P., Joachims, T., & Sul, J. H. (2004). Mapping subsets of scholarly information. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 5236L 5240.
Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 5228L 5235.
Guimera, 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.
Guo, Z., Zhu, S., Chi, Y., Zhang, Z., & Gong, Y. (2009). A latent topic model for linked documents. *Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval, ACM*, 720–721.
Jacobs, A. Z., Way, S. F., Ugander, J., & Clauset, A. (2015). Assembling the facebook: Using heterogeneity to understand online social network assembly. *Proceedings of the ACM Web Science Conference, ACM*, 18.
Joachims, T. (2002). *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers.
Kaiser, D. (2005). *Drawing theories apart: The dispersion of Feynman diagrams in postwar physics*. University of Chicago Press.
Larivière, V., Sugimoto, C. R., Macaluso, B., Milojević, S., Cronin, B., & Thelwall, M. (2014). arxiv e-prints and the journal of record: An analysis of roles and relationships? *Journal of the Association for Information Science and Technology*, *65*(6), 1157–1169.
Lee, D., Goh, K.-I., Kahng, B., & Kim, D. (2010). Complete trails of coauthorship network evolution. *Physical Review E*, *82*(2), 026112.
Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005). Graphs over time: Densification laws, shrinking diameters and possible explanations. *Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining, ACM*, 177–187.
McCallum, A. K. (2002). *Mallet: A machine learning for language toolkit*. http://mallet.cs.umass.edu
Newman, M. E. J. (2001). The structure of scientific collaboration networks? *Proceedings of the National Academy of Sciences of the United States of America*, *98*(2), 404–409.
Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(Supplement 1), 5200–5205.
Newman, M. E. J. (2010). *Networks: An introduction*. Oxford University Press.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. *Proceedings of the 20th conference on uncertainty in artificial intelligence, AUAI Press*, 487–494.
Song, Y., Huang, J., Councill, I. G., Li, J., & Giles, C. L. (2007). Efficient topic-based unsupervised name disambiguation. *Proceedings of the 7th ACM/IEEE-CS joint conference on digital libraries, ACM*, 342–351.
Sun, X., Kaur, J., Milojević, S., Flammini, A., & Menczer, F. (2013). *Social dynamics of science*. pp. 3. Scientific Reports.
Tabah, A. N. (1999). Literature dynamics: Studies on growth, diffusion, and epidemics. *Annual Review of Information Science and Technology*, *34*, 249L 86.
Wang, X., & McCallum, A. (2006). Topics over time: A non-Markov continuous-time model of topical trends. *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, ACM*, 424–433.
Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks? *Nature*, *393*(6684), 440–442.