RESEARCH ARTICLE

Finding scientific communities in citation graphs: Articles and authors

Shreya Chandrasekharan[1], Mariam Zaka[2], Stephen Gallo[2], Wenxi Zhao[1], Dmitriy Korobskiy[1], Tandy Warnow[3], and George Chacko[1,3,4]

[1]Netelabs, NET ESolutions Corporation (an NTT DATA Company), McLean, VA, USA
[2]American Institute of Biological Sciences, Herndon, VA, USA
[3]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA
[4]Grainger College of Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Keywords: citation graph, clustering, community finding, Invisible College, scientific organization

ABSTRACT

Understanding the nature and organization of scientific communities is of broad interest. The "Invisible College" is a historical metaphor for one such type of community that refers to a small group of scientists working on a problem of common interest. The scientific and social behavior of such colleges has been the subject of case studies that have examined limited samples of the scientific enterprise. We introduce a metamethod for large-scale discovery that consists of a pipeline to select themed article clusters, whose authors can then be analyzed. A sample of article clusters produced by this pipeline was reviewed by experts, who inferred significant thematic relatedness within clusters, suggesting that authors linked to such clusters may represent valid communities of practice. We explore properties of the author communities identified by our pipeline, and the publication and citation practices of both typical and highly influential authors. Our study reveals that popular domain-independent criteria for graphical cluster quality must be carefully interpreted in the context of searching for author communities, and also suggests a role for contextual criteria.

1. INTRODUCTION

In this article, we report on an effort to use citation data to identify groups of scientific articles that may reflect small-scale organization in the scientific enterprise. We are inspired by the "Invisible College" concept, which appears to originate from a group of intellectually active individuals who held meetings around 1660 and eventually formed the Royal Society of London in 1663 (Price & Beaver, 1966; Royal Society, 2020) but more generally refers to a relatively small self-assembled group of scientists with common scientific interests.

Importantly, there is a sense that these colleges are "in groups" with influence over prestige, research funding, and the scientific ideas of their community (Price & Beaver, 1966). Thus, these groups may advocate for or exhibit resistance to new ideas within their domains of interest (Barber, 1961). Furthermore, while such groups may espouse idealized norms of science (Merton, 1957), they are also driven by social interests, such as personal recognition, that influence both individual and collective behavior (Barber, 1962; Crane, 1972; Hagstrom, 1965).

Crane (1972) has described studies in rural sociology and mathematics, and referenced others in the biological sciences, psychology, and physics. An important distinction has also been made

an open access journal

Check for updates

Citation: Chandrasekharan, S., Zaka, M., Gallo, S., Zhao, W., Korobskiy, D., Warnow, T., & Chacko, G. (2020). Finding scientific communities in citation graphs: Articles and authors. *Quantitative Science Studies, 2*(1), 184–203. https://doi.org/10.1162/qss_a_00095

DOI: https://doi.org/10.1162/qss_a_00095

Received: 27 July 2020
Accepted: 2 November 2020

Corresponding Author:
George Chacko
chackoge@illinois.edu

Handling Editor:
Ludo Waltman

The MIT Press

between local (small) and global groups, with small groups being credited as the locus of rapid change and innovation (Hull, 1988, p. 112). Furthermore, Price has noted, in apparent reference to Invisible Colleges, that communications are likely challenged in groups larger than 100 members (Price, 1963) and that the small "strips" at the research front of science may be at most the work of "a few hundred" persons (Price, 1965); this number is also cited in a clustering study (Small & Sweeney, 1985). Some small groups may also represent the coalescence of researchers around new ideas; therefore, they are of additional interest.

Since the 1960s, the scientific enterprise has grown considerably, experienced greater globalization (Wagner, 2008) in the 21st century, and exhibits new features (e.g., large international collaborations such as the Human Genome Project [Lander et al., 2001] and the Advanced LIGO project [Harry & the LIGO Scientific Collaboration, 2010]). Even so, the tendency of scientists to form small groups that collaborate to advance their scientific and social interests is unlikely to have vanished; these groups may well exist as specializations within larger structures. Thus, we seek to understand organizational structures in the modern scientific enterprise and reconcile our observations, to the extent possible, with those from the 1960s.

Price and Beaver (1966), in a case study of the Information Exchange Group No 1 (Green, 1965) organized by the U.S. National Institutes of Health to focus on electron transfer and oxidative phosphorylation, described a group of 517 members, 62% of whom were from the United States and the rest from 27 different countries. This social group assembled around a scientific question and exchanged memos to discuss their interests. Price and Beaver used these memos as proxies for research articles and citations, and noted 1,239 authorships in 533 memos with two-author memos being the mode that was stable across a 5-year period. The majority of these authors were associated with only a single memo, and the top 30 authors each contributed to six or more memos. Three conclusions were drawn from this study: First, that there existed a small nucleus of highly active researchers and others who collaborated with them only once; second, that separate groups existed within this college; and third, that collaboration was a key feature. This valuable case study of Price and Beaver is, however, limited by examining a tiny sample of the enterprise as it existed in the first half of the decade 1960–1970. It is far more likely that a range of group sizes and behaviors exists now (perhaps even then).

As modern bibliographies and accessible computing make large-scale studies possible, a natural question is whether small communities of researchers can be identified using bibliographic data. Whether citations are an adequate proxy for social communications between scientists remains an open question. In our use of citation data to identify and characterize putative colleges or communities of practice, our working hypothesis is that such groups can be detected by identifying clusters of articles that are citation-dense, as common interests will result in citation of relevant documents, and especially of those authored by the "in group." We also recognize that data other than citation links could provide excellent insight into our question: for example, invited attendance records at small focused conferences and peer review of journal articles or applications for funding. However, with perhaps a few exceptions, such data are not always easily available.

Rather than attempting to directly identify author communities of practice, we first construct article clusters, and then examine the authors within the article clusters. The rationale for this approach is that each community of practice is by its nature formed around a specific research question or area, whereas individual scientists may participate in multiple communities of practice based on different scientific and social interests. Therefore, we use clustering for the purpose of identifying groups of articles that reflect interactions between members of small communities.

We reason that converging results, where similar clusters are generated by more than one method, may help identify clusters with citation signal high enough that the clusters themselves

are relatively independent of the method used to cluster them (convergence between two clustering techniques should reduce the false discovery rate). We approach this problem in two steps. First, we build article clusters using citation links and select a subset of the clusters using specific criteria; second, we study author citation patterns in the resultant clusters. The first step has six modules: (a) assemble a citation graph, (b) construct article clusters based on direct citation links between articles, (c) repeat with a different clustering method, (d) select clusters using converging results from the two different methods, (e) select clusters of a given size range, and (f) restrict to those clusters that have elevated intracluster citation activity indicative of community behavior. Hence, the first step is best seen as a pipeline with flexible choices for each module, making it a *metamethod*.

Our use of article-level citation patterns is motivated by the argument that clusters defined by articles that cite each other are more informative than clusters derived from journal-derived categories (Milojevic, 2020; Shu, Julien, Zhang, Qiu, Zhang, & Larivière, 2019; Waltman & van Eck, 2012) or from searches for words or phrases that recur in articles (Klavans & Boyack, 2017). We note that recent work on clustering articles using weighted and unweighted citation patterns suggests that improved similarity of documents within clusters can be achieved (Ahlgren, Chen, Colliander, & van Eck, 2020). However, our focus is not to select clusters of the most similar documents—rather, we seek to identify documents linked to each other by citation, an expression of the social behavior of authors. Further, we restrict our examination to moderately sized clusters because of our interest in relatively small communities.

We use three different clustering algorithms to generate clusters of interest: Markov Clustering (Dongen, 2000) referred to as *MCL*, the Leiden Algorithm (Traag, Waltman, & van Eck, 2019) referred to as *Leiden*, and Graclus (Dhillon, Guan, & Kulis, 2007). MCL is an unsupervised clustering algorithm for community detection that is based on stochastic flow simulation and has been extensively used for a variety of clustering applications: It is scalable, does not require prespecification of the number of clusters to be generated, and has tunable parameters that control breadth of search and granularity of output. Leiden has similar properties and has been used with citation data. Graclus is a spectral clustering method that we have used to construct article clusters from citation data (Devarakonda, Korobskiy, Warnow, & Chacko, 2020) and has been evaluated by others (Almeida, Guedes, Meira, & Zaki, 2011; Šubelj, van Eck, & Waltman, 2016). Graclus requires that the number of clusters to be formed is specified at runtime.

We have chosen the field of immunology as our primary test case, as it has existed for many years, grown and diversified over time, and exchanges discovery and methods with other biomedical areas. We also examined a second, more recent, data set from immunology and another from the field of ecology. While recognizing that some of these article clusters we construct using citations will spill across disciplinary boundaries, we also expect that many of the clusters will contain articles that reasonably represent membership of the parent field (immunology or ecology). We would not expect, for example, that the vast majority of clusters in the immunology data sets represent communities from condensed matter physics or medieval history.

We explored two cluster quality measures, conductance and coherence, for the resulting clusters, interpreting them in the light of our purpose for clustering (Almeida et al., 2011; Emmons, Kobourov, Gallant, & Börner, 2016; von Luxburg, Williamson, & Guyon, 2012). For one of these data sets, we also asked experts to evaluate a small sample of the MCL clusters for thematic relatedness (Figure 1). Lastly, we identified the author communities associated with each of the selected publication clusters, and specifically identified influential authors. We consider this study a first step in designing and testing a metamethod that could enable large-scale identification of communities of different sizes and types, based on different search criteria, data collection

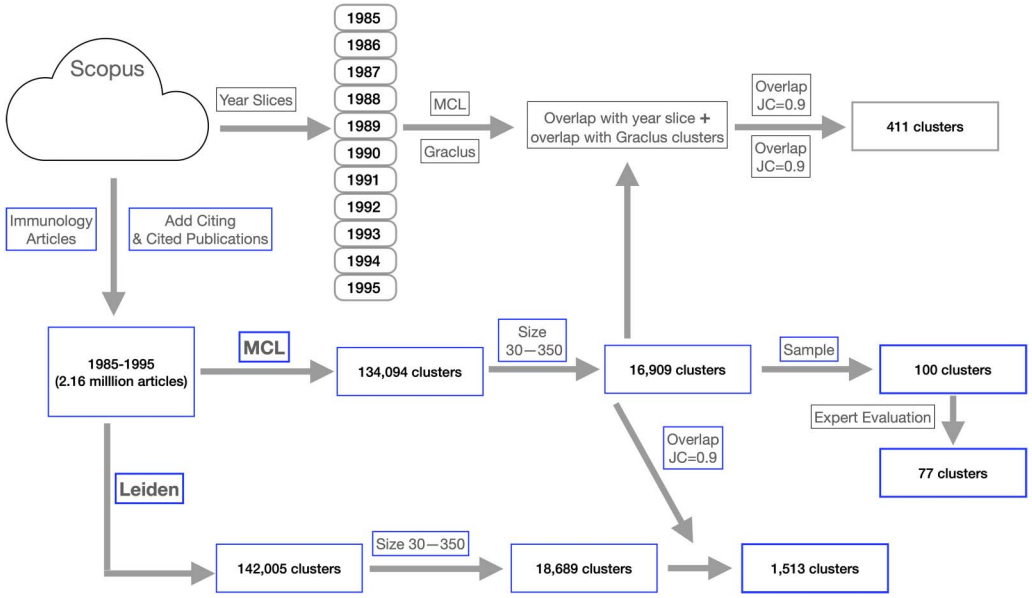

Figure 1. Initial workflow to calibrate community detection. The schematic shows how MCL, Leiden, and Graclus clustering were used to cluster articles. Authors for the publications in clusters selected by this pipeline were subsequently analyzed to discern communities of practice. The imm_1 data set, which consists of 2.16 million publications from Scopus for the years 1985–1995, is used as an example. Boxes with blue borders represent our preferred MCL-Leiden approach: (a) For MCL, this produced a set of 134,094 clusters. We then restricted attention to those clusters containing between 30 and 350 publications, which resulted in 16,909 clusters. From these, a sample of 100 clusters was provided to two evaluators who rated 77 of them as strongly themed. (b) Leiden clustering produced 142,005 clusters, which reduced to 18,689 clusters after size restriction (30–350). Convergence was measured using the intersection/union ratio (JC) between members of a pair of clusters. Convergence between MCL and Leiden clustering resulted in 1,513 clusters when a JC of 0.9 was used as a filter. Boxes with grey edges indicate an alternate protocol using MCL and Graclus that relies on fractionation of data by year of publication (year-slices) and the use of proxy clusters. In parallel, each of the 11 year-slices was clustered with MCL and with Graclus. Every cluster from the set of 16,909 was matched, using overlap as the matching criterion, to a single MCL cluster from all clusters generated in the 11 year-slices. The acceptance criterion for matching of two clusters was also a JC of 0.9. Each MCL cluster from the year-slices was also matched to a Graclus cluster from the same year using this criterion. Thus, the box containing the text "Overlap JC = 0.9" is shown twice (top right) to indicate two selection steps.

techniques, choice of clustering algorithm, adjustable parameters, and cluster selection criteria. We present our findings in two major sections focused on pipeline construction and author analysis, respectively.

2. MATERIALS AND METHODS

2.1. Data

We assembled 3 data sets (Table 1) along with their cited references and the articles that cite them: (a) imm1985-1995 (*imm_1*) representing 11 years of immunology data, (b) imm2000-2004 (*imm_2*) representing 5 years of immunology data, and (c) eco2000-2010 (*eco*) representing 11 years of ecology data. Of these three, we chose the year range for imm1985–1995, as seed publications would have been able to accumulate citations over at least 20 years, to manage the size of the working data sets, and to be able to reuse a curated data set used in two publications. We subsequently added the other two data sets in response to critique. As a source of bibliographic data for this study, we used Scopus (Elsevier BV, 2020), as implemented in the ERNIE project (Korobskiy, Davey, Liu, Devarakonda, & Chacko, 2019). At the time of this analysis, our Scopus data consisted of ~95 million publication records plus their cited references. From these data, we selected publications in English, of type "article" with publication type

Table 1. Comparison of data sets and clusters generated by MCL and Leiden. med_cond refers to the median conductance of the clusters (e.g., the 134,094 clusters generate by MCL from the imm_1 data set)

Data set	num_nodes	num_seed_nodes	num_edges	num_clusters	med_cond
imm_1_mcl	2,163,683	147,015	6,846,323	134,094	0.741
imm_1_leiden	2,163,683	147,015	6,845,768	142,005	1.0
imm_2_mcl	2,358,152	85,673	7,079,790	84,215	0.705
imm_2_leiden	2,358,152	85,673	7,079,790	110,994	1.0
eco_mcl	4,662,774	419,310	19,918,316	337,503	0.838
eco_leiden	4,662,774	419,310	19,908,278	291,583	1.0

"core," and Scopus All Science Journal Classification (ASJC) codes corresponding to the fields of interest. For immunology data we used ASJC code 2403 (Immunology) and for ecology data we combined codes 1105 (Ecology, Evolution, Behavior and Systematics), 2302 (Ecological Modelling), and 2303 (Ecology). Articles thus extracted were referred to as "seed articles."

We then amplified the set of seed articles, in each of these data sets, with those articles that directly cited them (through 2020) as well as by articles cited in the seeds. The only constraints we imposed on the cited or citing articles were to require that they were English publications of type "core"; in particular, the cited and citing articles were not constrained by ASJC codes. This process resulted in node (article) and edge (direct citation links) counts as follows (with the count of seed articles in parentheses): (a) *imm_1*: 2,163,683 (147,015) articles and 6,846,323 edges, (b) *imm_2*: 2,358,152 (85,673) articles and 7,079,790 edges, and (c) *eco*: 4,662,774 (419,310) articles and 19,918,316 edges.

We mapped article identifiers to author identifiers in Scopus so that citation activity localized to clusters could be linked to authors. We treated each author identifier as a unique person, but we also observed cases where individuals had more than one author identifier on account of slightly different spellings of names, different names, and different institutional affiliations. Typically one profile per person dominated in terms of publications being assigned to it, but our total author counts will overestimate the actual number of individuals involved. Interestingly, 1,034,537 articles and 1,874,331 authors are common across the two immunology data sets.

2.2. Clustering Software

For Markov Clustering analysis, we downloaded and compiled source code for the MCL-edge software (Dongen, 2000). After evaluating different runtime parameters, we clustered test sets using an expansion parameter of 2 (default) and an inflation parameter of 2.0 to minimize the number of large aggregated clusters. For a random graph comparison, we performed one million reciprocal citation exchanges between randomly selected pairs of publications on these data and then ran MCL on the resultant data. For analysis with Leiden, we downloaded and installed the Java implementation of this algorithm (Traag, 2020). We chose runtime parameters that generated an approximation of the size distribution of clusters generated using MCL. The resolution parameter was set to 0.002 under the Constant Potts Model (CPM) quality function. We used the default settings for the other parameters: 10 iterations, randomness parameter of 0.01, and 1 random start. For reproducibility, we used a seed value of 2020. For analysis with Graclus, we used a previous installation of the source code as described in Devarakonda et al. (2020).

2.3. Convergent Clustering

Our initial experiment (Figure 1) was to cluster the 2.16 million publications in the imm_1 data set, which resulted in 134,094 clusters using MCL and 142,005 clusters using Leiden. Graclus was not able to compute comparable numbers of clusters on these data. Consequently, we used fractionated data (11 year-slices corresponding to each year in 1985–1995) as input to Graclus and MCL and successfully clustered these smaller data sets, and used clusters from them as proxies for clusters from the complete data set. Due to our focus on small communities, we selected clusters of interest by first restricting them to those containing between 30 and 350 nodes (articles). Then we selected every cluster from one technique that matched a cluster from another technique using the relatively stringent filter of a Jaccard coefficient (JC) of 0.9 between two clusters.

In the schematic representation of this workflow (Figure 1), a reduced yield for the MCL-Graclus protocol, which also involves an extra matching step, is evident. Consequently, we did not use the MCL-Graclus combination further. We do not exclude the use of Graclus in future studies, but until the software is able to generate comparable distributions of clusters to MCL, Leiden, and other clustering tools that might be used, its use may have to be restricted to smaller data sets and fewer clusters.

In our subsequent analyses, we restricted ourselves to the use of MCL-Leiden only for convergent clustering. Using our initial settings (Figure 1), we harvested 1,513 clusters of low conductance (median of 0.159), whose contents we subsequently analyzed and found relatively peripheral to a broad definition of immunology. Consequently, we lowered the stringency of selection to a JC of 0.20 (corresponding to the 25th percentile value of JCs for cluster pairs) with the intention of including clusters of greater conductance and used this 25th percentile for selection in all three data sets.

We then applied a further selection criterion to capture intracluster citation activity. For each cluster, we computed IENR, the intracluster edge-to-node ratio. Due to our interest in detecting author communities, we restricted our attention to the 99th percentile of IENR values (Figure 3); this selection identifies clusters with high intracluster citation activity.

To evaluate clusters and shuffled-citation clusters generated by MCL, Leiden, or Graclus, we examined edge (citation) densities by intracluster conductance (Devarakonda et al., 2020; Emmons et al., 2016). The conductance of a cluster S, denoted $\phi(S)$, is defined using the formula:

$$\phi(S) = \frac{|\partial(S)|}{\min(vol(S), 2m - vol(S))}$$

where $|\partial(S)|$ is the size of the boundary (i.e., number of edges with exactly one endpoint in S), vol(S) is the volume of S (i.e., the sum of the degrees of the vertices in S), and m is the number of undirected edges in the entire network (Shun, Roosta-Khorasani, Fountoulakis, & Mahoney, 2016).

Textual coherence was measured by using the Jensen-Shannon Divergence (JSD) (Boyack et al., 2011; Colliander & Ahlgren, 2011; Endres & Schindelin, 2003), which is used to compute the distance between two probability distributions. To compute textual coherence, we used the titles and abstracts of all the articles in our study. On average, roughly 7% of the publications across the three data sets were missing titles and/or abstracts. We first concatenated these titles and abstracts and preprocessed them by lemmatization. We then removed stop-words using a list of 510 tokens comprising basic NLTK stop-words, PubMed stop-words, and a select list of tokens from the top 500 most frequent words in our data set. For each cluster of size greater than 10 (after removing missing values), we performed a second preprocessing by filtering out those tokens that occurred only once in the entire cluster. Next, we converted all the remaining tokens

by article in the cluster into a matrix of term frequencies (i.e., for each article, we had a vector of counts of all the tokens). We also obtained a vector of counts for all the unique tokens in the cluster.

JSD was then computed between the vector of term frequencies of the cluster and each article in the cluster using the following:

$$JSD_{p,q} = \frac{1}{2}D_{KL}(p,m) + \frac{1}{2}D_{KL}(q,m)$$

where $m = \frac{p+q}{2}$, p is the probability of a term in a document, q is the probability of the same term in the cluster, and D_{KL} is the Kullback-Leibler divergence, given by:

$$D_{KL}(p,m) = \sum p_i log\left(\frac{p_i}{m_i}\right)$$

We computed the textual coherence for a given cluster X of size n (after removing missing values) as follows. Letting JSD_X denote the arithmetic mean of all article JSD values in X, we define the textual coherence of X to be $JSD_X - JSD_{random(n)}$ (Boyack et al., 2011), where $JSD_{random(n)}$ denotes the JSD of a random cluster of the size n.

$JSD_{random(n)}$ is the arithmetic mean of all the JSD values computed from random selected sets of size n from all the articles in our study. For each value n, we estimate $JSD_{random(n)}$ by selecting 50 article subsets of size n at random, and averaging the results. The method of computing each iteration of $JSD_{random(n)}$ is exactly the same as the method described for JSD_X above.

After completion of MCL clustering and computing conductance and coherence values, we compared a random sample of 1,000 clusters from the year 1990 and visualized the effect of shuffling citations on conductance and coherence compared to the original citation data (Figure 2).

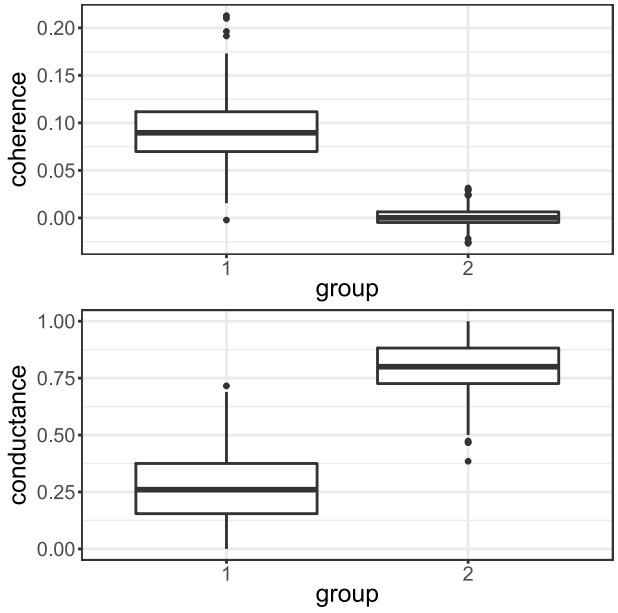

Figure 2. Conductance and coherence profiles of 1,000 MCL clusters (group 1) compared to random clusters (group 2), showing that MCL clusters have lower conductance and greater coherence compared to random clusters of the same size. The 1990 immunology year-slice was either clustered (*x*-axis: group 1) or subjected to 1 million citation shuffling (group 2) operations and then clustered using MCL-edge software with an expansion parameter setting of 2 and inflation parameter setting of 2.0. From each of the resultant data sets, a sample of 1,000 clusters of size 30–350 publications were randomly selected and analyzed for conductance and coherence.

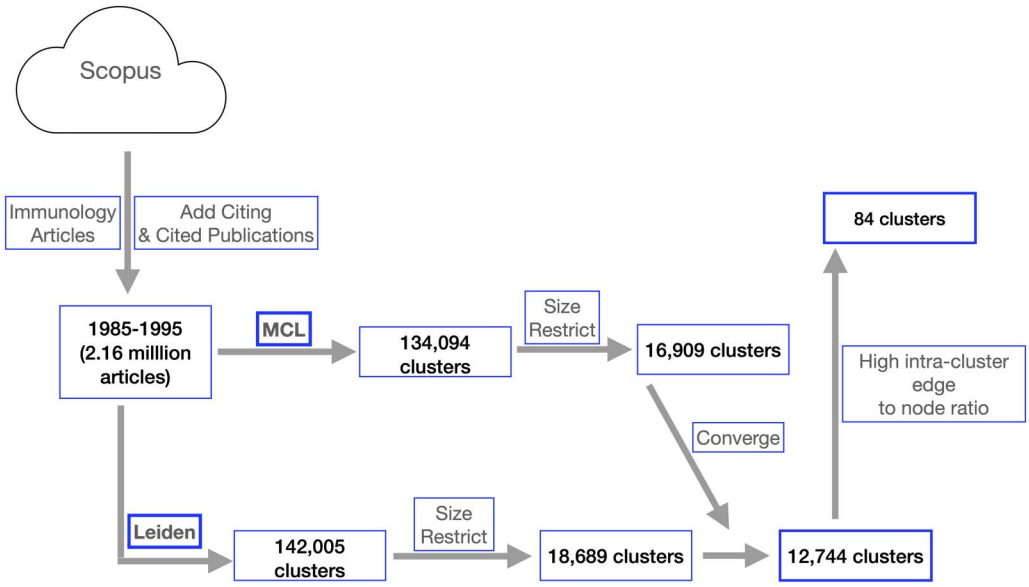

Figure 3. Pipeline for community detection. Authors of the publications in clusters selected by this pipeline were subsequently analyzed to discern communities of practice. Data from the imm_1 data set, which consists of 2.16 million publications from Scopus for the years 1985–1995, are shown. Using MCL, a set of 134,094 clusters was generated, which was restricted to 16,909 clusters by selecting those containing between 30 and 350 publications each. Leiden clustering of the same data produced 142,005 clusters, which reduced to 18,689 clusters after size restriction (30–350). Converging results were measured using the intersection/union ratio (JC) between members of a pair of clusters. Convergence between MCL and Leiden clustering resulted in 12,744 clusters when a JC of 0.20 was used as a filter (and 1,513 clusters when the JC was set at 0.9). Subsequent selection at the 99th percentile value of the intracluster edge-to-node ratio (IENR) resulted in 84 clusters.

2.4. Expert Evaluation

When testing our pipeline, we evaluated small samples of 10 clusters, generated from different parameter settings, for relevance to a broad interpretation of immunology. The clusters in these samples were subjectively annotated as "boutique," "intermediate," and "relevant."

For assessing intracluster thematic relatedness, we used the expertise of two of the authors of this article (Zaka and Gallo), who are professional peer review specialists in the biomedical sciences, and highly experienced at clustering proposals for funding based on multiple criteria such as subdisciplines, methods, disease, and researchers. In preliminary experiments, we provided a small number of training clusters to these evaluators, representing a range of conductance values, to assist in developing a common set of principles by which they would evaluate a test set. The clusters in this development set were not considered further.

We then randomly selected 90 MCL clusters, each with 30–350 publications, and with conductance values of no more than 0.5. As smaller clusters occur more frequently, the sample of 90 was constructed from two strata based on size to ensure representation of the larger cluster sizes. An additional 10 clusters with conductance values greater than 0.5 were added to the sample. The two evaluators were each asked to evaluate 50 selected clusters (45 from the set of 90 and 5 from the set of 10) for thematic relatedness, given only the titles and abstracts for each publication in each cluster. Using intuitive sensibility (Salas, Rosen, & DiazGranados, 2009) based on their expertise and experience, they assigned scores on a simple scale of 1–4, where 1 represented a well-themed cluster exhibiting a single discernible scientific theme, 2 a moderate level of thematic relatedness, 3 poor thematic relatedness, and 4 "unable to evaluate." The evaluators

also annotated each cluster with keywords such as "hemophilia" or "adenosine deaminase" to indicate the theme that they discerned (see Supporting Information).

2.5. Author Analysis

We computed an initial array of descriptive measures that included the number of authors within clusterings and within clusters, the number of articles per author within clusterings and clusters, the number of in-graph citations per author within clusterings and clusters, the number of clusters that an author had contributed to through her or his articles, and coauthorship counts. From these data we also generated tier assignments for authors within clusters (as defined below) to categorize them by local influence.

For each cluster with at least one internal edge, we selected the top 10% of articles by citations received. The citation count at the tenth percent was set as a threshold value and the number of articles at the tenth percent (10% of cluster size) was assigned as the threshold count. We then calculated the total number of articles in the cluster that have received citations greater than or equal to the threshold value. If the number of articles exceeded the threshold count, we increased the threshold value by 1 and denoted it as the final threshold value.

Any article that received citations greater than or equal to the final threshold value was labeled as Tier 1, any article that received no citation at all in the cluster is labeled as Tier 3, and all other articles were labeled as Tier 2. Note that in this protocol, we may have articles that receive only one citation under Tier 1. These labels were transferred to authors. For each author in every cluster, we took the minimum tier value received by the author as their tier value for that cluster. In other words, if an author had more than one article in a cluster, the best tier assignment for the author was chosen to represent the author's status in the cluster. Each author received a single tier label from one cluster. We then count all instances of an author being in Tiers 1, 2, and 3.

3. RESULTS AND DISCUSSION

We present our results in two stages. The first stage describes our pipeline, initial observations, and adjustments to it. The second stage addresses analysis of authors from clusters selected through this pipeline.

3.1. Pipeline Construction

For clustering analysis, we assembled three data sets (Section 2 and Table 1), two from the field of immunology, and one from ecology: (a) imm1985–1995 (*imm_1*), representing 11 years of immunology data from the years 1985–1995, (b) imm2000-2004 (*imm_2*), representing 5 years of more recent immunology data, and (c) eco2000-2010 (*eco*), representing 11 years of ecology data.

We also clustered the other two data sets, imm_2 and eco, using a combination of MCL and Leiden as depicted in Figure 3. The results of these clusterings are summarized in Table 2.

Interestingly, both MCL and Leiden generated a large number of singleton clusters on all three data sets, with Leiden consistently generating more singletons than MCL on the same data under the conditions being used (Table 2). Of the singletons, an article by Fraker and Speck (1978) that describes a chemical technique for labeling the surface of cells with iodine is cited 351 times by articles in 340 different clusters. This is an example where a publication that is relevant to several areas of investigation within a field is not placed within any of the clusters for these different areas, and is instead placed in a singleton cluster by itself. The example makes the point that small

Table 2. Conductance of size-restricted (30–350) MCL and Leiden clusterings. MCL and Leiden clusterings of imm_1, imm_2, and eco data sets were restricted to clusters ranging in size from 30–350 nodes each. The median conductance (med_cond) is shown for the size-restricted clusters. The counts of singletons, clusters of size less than 30, and greater than 350 are also shown.

Data set	num_clusters	med_cond	singletons	< 30	> 350
imm_1_mcl	16,909	0.503	16,534	11,7013	172
imm_1_leiden	18,689	0.42	116,621	122,750	566
imm_2_mcl	27,062	0.52	5,954	57,017	136
imm_2_leiden	20,617	0.425	87,070	89,775	602
eco_mcl	38,903	0.521	59,049	298,395	205
eco_leiden	38,771	0.457	241,691	251,306	1,506

clusters (and even singletons) generated by MCL can be highly connected to other clusters—even while not having any internal edges—and that they merit future investigation.

We also present two other examples of edge-cases that illustrate citation behavior captured in clusters. A cluster containing Cucala, Bauerfeind, et al. (1987), focused on nonsteroidal anti-inflammatory drugs (NSAIDs), consists of 125 publications, of which 123 articles are cited by a single article from the cluster, which in turn, is cited by another article in the cluster. Another cluster containing Michel, Hunder, et al. (1992) is also interesting: Michel et al. (1992) is a comparison of hypersensitivity vasculitis and the Henoch:Schonlein purpura, and is cited by the other 107 publications in the cluster.

Such edge cases are not very relevant to our search for interactive author communities and should be removed, but are a consequence of using a clustering technique that forces disjoint clusters (which most clustering methods accomplish). In comparison to these edge cases, a cluster of interest (cluster #8661) consisted of 44 publications focused on immunoglobulin genes with 19 of them receiving citations from 31 nodes within the cluster.

As potentially useful evaluators of cluster quality for all three data sets, we calculated intracluster conductance and textual coherence (Section 2). Conductance is based on the citation graph, and small values have been considered desirable (Shun et al., 2016). By its formula (see above), conductance values decrease as the size of the boundary (number of edges connecting the cluster to other clusters) decreases; hence, clusters that have low connection to other clusters have very low conductance values, while a dense cluster that nevertheless has a large number of edges to other clusters will not have a particularly low conductance value.

A comparison of Leiden and MCL clusters across the three data sets is informative. Both MCL and Leiden produce similar size and conductance distributions for all three data sets (Figures 4 and 5), with the Leiden distribution shifted slightly right of the MCL distribution with respect to size and slightly left with respect to conductance. Furthermore, for all three data sets, median conductance progressively reduces from the full MCL clustering to the size-restricted MCL set (30–350) to the MCL clusters selected by convergence with Leiden clustering (Tables 1–3), even after relaxing the stringency of selection by convergence (Section 2).

Coherence, in contrast, is based on text within the documents and not on citations between documents. As we note in Section 1, our purpose for clustering by direct citation is to be able to identify expressions of citation behavior, and hence clustering on the citation network is more directly relevant than clustering based on coherence.

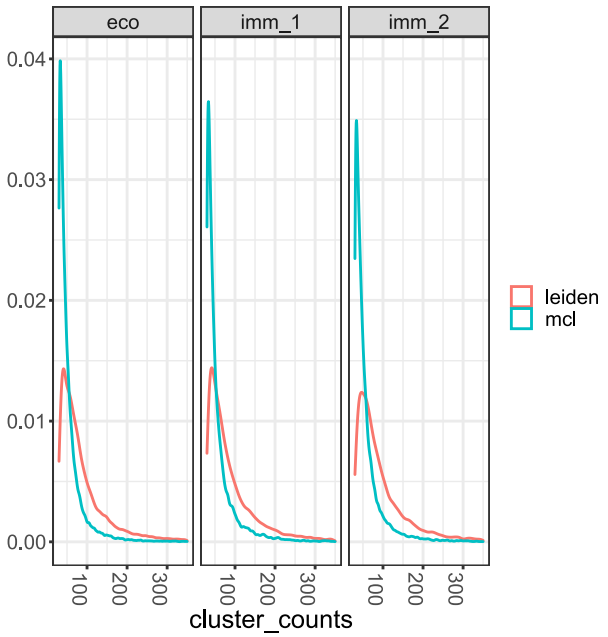

Figure 4. Size distribution of MCL and Leiden clusterings. Density plots are shown for the distribution of cluster sizes arising from MCL and Leiden clusterings of the three data sets after a size restriction of 30–350 was applied to the initial clustering.

We performed textual coherence measurements (Table 4), noting that textual coherence, while clearly valuable as a measure of cluster quality, may be useful if "good" clusters are expected to exhibit similarity in text content. We note that MCL clusters appear to have slightly better coherence than Leiden clusters and that the coherence measured for MCL clusters is

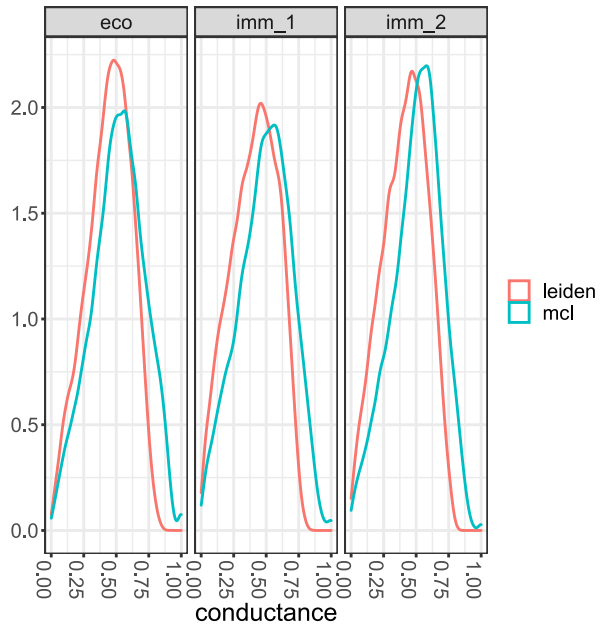

Figure 5. Conductance profiles of MCL and Leiden clusterings. Density plots are shown for the distribution of conductance values of clusters generated by MCL and Leiden for the three data sets after a size restriction of 30–350 was applied to the initial clustering.

Table 3. Convergent selection of clusters using MCL and Leiden in combination. *Upper panel*: Selection with JC of 0.9; *Lower panel*: Selection with JC at 25th percentile

Convergent_clustering	num_clusters	num_articles	med_size	med_cond
imm_1_mcl_leiden	1,513	101,226	50	0.159
imm_2_mcl_leiden	3,169	201,118	52	0.207
eco_mcl_leiden	4,268	280,898	53	0.204
imm_1_mcl_leiden	12,744	793,030	47	0.443
imm_2_mcl_leiden	20,130	1,249,653	49	0.465
eco_mcl_leiden	29,271	1,696,420	46	0.463

comparable to the Pubmed Related Article (PMRA) clustering reported by Boyack et al. (2011, Figure 2), which performed best among five techniques evaluated in that study.

Given roughly similar trends across the three data sets, we chose to analyze just one data set to explore the potential value of this pipeline and manage the scope of investigation. Hence, we only analyzed results from the imm_1 data set in detail, although we provide some comparative data for all three data sets (Tables 1–3, Figures 4 and 5).

Of the 100 MCL clusters given to the two experts to evaluate (Section 2), 77 clusters were rated 1 (well themed), 18 clusters were rated 2 (moderately themed), and five clusters were rated 3 (poorly themed), suggesting that the evaluators considered roughly three-quarters of the clusters to be strongly themed given their knowledge and experience.

Of these, 93/100 and 71/77 clusters were also selected by convergence of MCL and Leiden using the 25th percentile of JC values as a filter, suggesting that some correspondence may exist between expert selection and the size-restricted convergence pipeline. It is important to note, however, that thematic relatedness is a measure that is intracluster and does not speak to whether a cluster is related to other clusters in the graph being analyzed.

We observe that the median conductance of MCL clustering tends to be high (~0.7 averaged across three data sets). When a size restriction is applied, the main effect is to remove the small clusters. In particular, the singletons (of which there are many: see Table 2) are deleted, which results in the median conductance decreasing (singletons have conductance of 1). With the size restriction, the median conductance of the Leiden clustering was lower than the unrestricted MCL clustering, in the case of all three data sets. If the additional constraint of selection by convergence

Table 4. Textual coherence profiles of clusters selected by convergence of MCL and Leiden clusterings. *Upper panel*: Selection with JC of 0.9; *Lower panel:* Selection with JC at 25th percentile

measure	imm1_mcl_leiden	imm2_mcl_leiden	eco_mcl_leiden
min_coherence	−0.006	−0.009	0.009
median_coherence	0.092	0.089	0.113
max_coherence	0.249	0.337	0.391
min_coherence	−0.006	−0.009	−0.000
median_coherence	0.096	0.093	0.112
max_coherence	0.267	0.337	0.391

is applied, then median conductance decreases further (Tables 1–3). Given that low conductance values have been considered desirable in the graph clustering literature (Shun et al., 2016), clusters selected by serial application of size restriction and convergence could be considered of high quality. However, as we have discussed above, groups of publications that are highly disconnected from the rest of the literature being studied are likely to have very low conductance values, raising the possibility that the pipeline we used could have the potential to produce only a small subset of the publication clusters of interest to us.

Therefore, we examined the publication clusters we obtained in this pipeline with respect to two additional criteria: (a) quantitatively through the ratio of internal edges to the number of nodes (IENR) in each cluster (i.e., as a measure of intracluster citation activity) and (b) qualitatively by estimating the extent to which the research topic represented by a cluster harmonizes with a broad definition of the field of immunology.

As we have noted, the serial application of size restriction and MCL-Leiden convergence using a JC of 0.9 or greater reduces 134,094 clusters to 1,513 clusters. The median conductance of the MCL clustering is 0.74, which is reduced to 0.50 after size restriction and then to 0.159 after convergent selection (Table 3). We examined a sample of this output for content (Section 2). Of a sample of 10 clusters drawn from this convergent population, seven of 10 clusters were labeled "boutique" relative to immunology. For example, two of these boutique clusters were interpreted as focusing on "fetal brain injury" and "anxiety and counseling," respectively. These examples may provide insight into why low convergence values are potentially associated with a higher incidence of boutique clusters. Intracluster conductance is low whenever the cluster is clearly discrete, with few edges between it and any other cluster, especially if there is relatively high intracluster density. Boutique publication clusters would have this kind of graphical representation: Even if the cluster itself has high thematic relatedness and high intracluster citation behavior, there will be relatively few edges to other clusters, because of a lack of interaction between the boutique research and other topics. Furthermore, such discrete clusters are easily found by clustering methods—it is what they are designed to detect—so they will survive stringent convergent selection. We do not exclude that such groups could signal early sites of innovation, but the samples identified as boutique did not suggest that this was the case.

Thus, the initial protocol we used had the interesting outcome of producing very clearly defined clusters with high thematic relatedness but few references to the other articles in the data set being analyzed. Furthermore, such clearly defined clusters are likely to be the clusters that will easily be found by most clustering methods, and so may be preferentially selected for by convergence.

Therefore, we considered two alternative pipelines (both using the size-restricted population) to see if we could produce clusters that were more likely to be labeled as "relevant." Because low conductance values were counterproductive, we first examined an approach wherein we sampled from the median conductance range of the size-restricted population: In this sample, six of 10 were rated relevant and three were rated "boutique." Second, we examined the effect of lowering the threshold for convergence from a JC of 0.9 to one of 0.5: This protocol produced five clusters rated "relevant." Thus, both approaches seem to succeed in increasing the frequency of "relevant" clusters, compared to the original pipeline. Therefore, we adjusted the JC down to the 25th percentile value for matched clusters with the aim of enabling the selection of higher conductance clusters while maintaining some convergent selection.

Lastly, we examined samples of clusters from the 99th percentile of IENR scores in the size-restricted population and noted that eight of 10 clusters in a sample were labeled "relevant" and only one cluster was labeled "boutique." This single cluster was of low conductance. Examining a

sample at the 50th percentile had the opposite effect, with seven of 10 clusters labeled "boutique."

We inferred that our pipeline can be modified to address the specific question of interest. For example, if the objective is to find publication clusters that are central within the larger scientific community, then moderate to high conductance values will be best, whereas if the objective is to find new or boutique research problems, then low conductance values may be better. In contrast, high IENR values may always be beneficial to identify loci of citation activity (as they indicate high intracluster citation behavior). The process of optimizing the selection criteria for a given study would necessarily be intensive, but the benefits of introducing domain-specific criteria might justify such expenditure of effort.

We think of this pipeline as a metamethod because of its modular nature, which allows practitioners to select the clustering method of their choice, add or drop modules, and tune the parameters of individual methods to be most suitable to the question of interest. Finally, although conductance and coherence have both been proposed as quality measures for evaluating clusters, they are not directly relevant to thematic relatedness, which requires human expert evaluation for reliable assessment. In our study, we used human experts to evaluate 100 MCL clusters, who observed that MCL clusters tend to have strong thematic relatedness. However, as this expert evaluation was limited to only 100 MCL clusters, and each cluster was only evaluated by one reviewer, it is premature to draw definitive conclusions about the thematic relatedness of MCL clusters produced by this pipeline.

3.2. Author Analysis

We use a cluster consisting of 92 nodes and 386 authors to illustrate relationships between articles and authors (Figure 6).The articles within span 35 years and concern the drug Alefacept intended to treat psoriasis (Hodak & David, 2004). Of the 386 authors, 25 are Tier 1, none are Tier 2, and 361 are Tier 3 authors. Similar to Price and Beaver's (1966) report, we observed that 361 of these authors had contributed one or two articles each, nine authors contributed five or more articles, and one had contributed 10 articles. The author with 10 articles was also linked to three awards from the U.S. National Institutes of Health for research on skin disease. Five large coauthor groups can be observed along with smaller ones, again consistent with Price and Beaver (1966).

On a larger scale, we examined author patterns across clusters in all three data sets (Table 5) after applying size restrictions (30–350) and convergence at the 25th percentile of JC scores, and observed similar trends across them: (a) A few authors were assigned as Tier 1 in each cluster, (b) most authors were designated Tier 3 in a cluster, and (c) a few authors were also present in many clusters, and these could be differentiated by their Tier 1 and Tier 3 counts. The first two observations are again consistent with Price and Beaver (1966), but the third suggests that authors of influence tend to exert their influence across multiple clusters.

At an even higher level, we examined the authors from clusters produced by MCL on the imm_1 data set before any subsequent down-sizing (size-restriction, convergence with Leiden, and high intracluster density). This network has 2,163,683 articles and 134,094 clusters. Of the 2,821,931 authors contributing to these articles, roughly 61% authored documents in only one cluster, 28% authored documents in two to five clusters, and 11% authored documents in six or more clusters; the average number of clusters per author was 3.1. We discovered several authors associated with very large numbers of clusters, and 0.08% of 2,821,931 authors were found in 100 clusters or more.

However, even at this level, there are different contexts that can result in large values. For example, one author had 1,005 articles distributed across 623 clusters, which resolved into 42

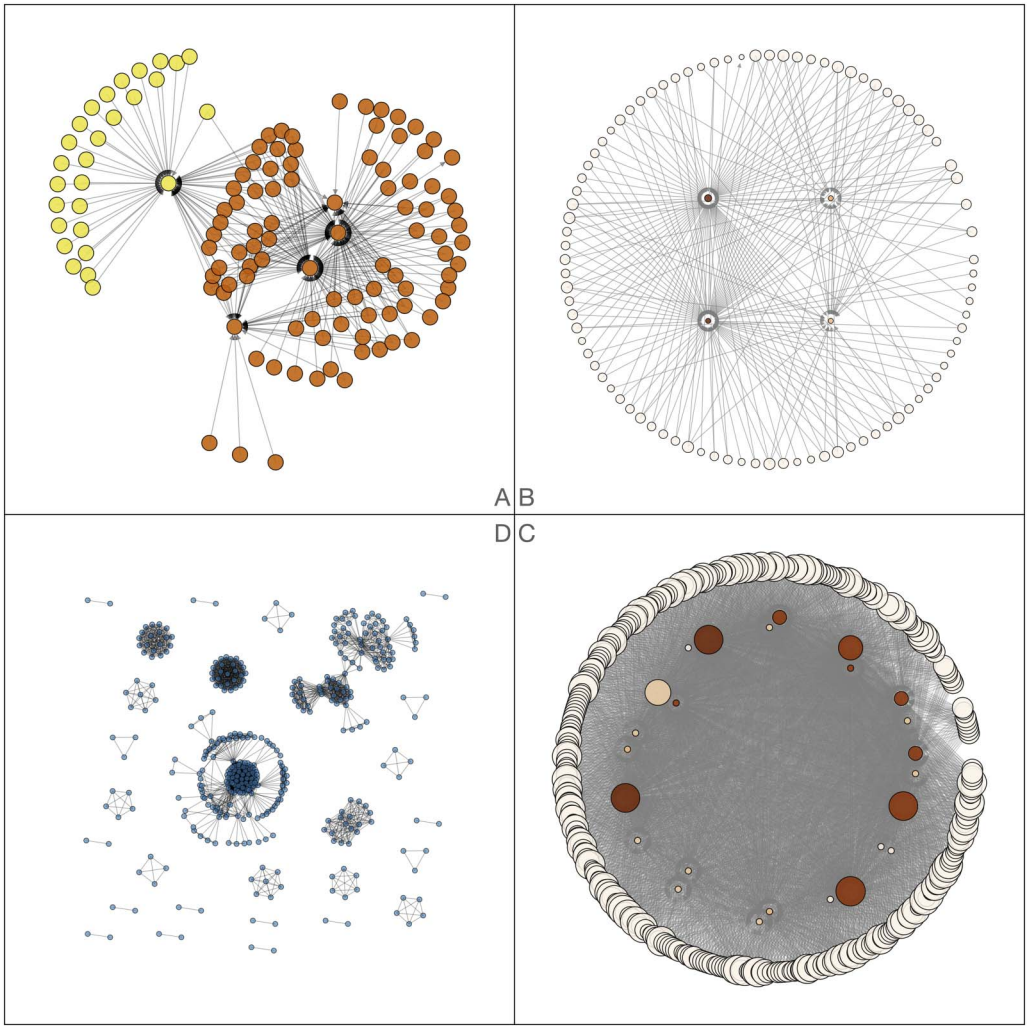

Figure 6. Article and author citations of a 92 node cluster with 386 authors. The visualization shows citations between clusters, articles, and authors as well as coauthorship relationships. *A*. Inter- and intracluster citation activity of a 92 node cluster (orange-brown) with its best connected cluster (yellow). Each node is an article and each edge is a citation. *B*. Intra-cluster citation activity of the 92 node cluster. Each node is an article and each edge is a citation. Four articles receive most of the citations. Node size is proportional to citing activity. Color intensity is proportional to number of citations received. *C*. Author network from the same 92 node cluster. Each node is an author. Node size is proportional to citing activity. Color intensity is proportional to number of citations received. The cluster comprises 25 Tier 1 authors and 386 Tier 3 authors. Tier 1 authors are seen as nodes located inside the circumference of the plot. *D*. Co-authorship network. Each node is an author. Each edge is a coauthor instance. Color intensity is proportional to the frequency of coauthorship between two authors. Five large coauthor groups can be seen.

Tier 1, 78 Tier 2, and 460 Tier 3 assignments; 43 of these clusters did not have any internal edges. A second example is the case of a researcher who had 642 publications in 519 clusters that resolved into 64 Tier 1, 257 Tier 2, and 121 Tier 3 assignments; the remaining 77 clusters also did not have internal edges. Rating one of these authors higher than the other in terms of accomplishment and influence involves some degree of subjectivity, although we are inclined to endorse the second.

Our observation that highly cited authors are members of many different clusters is also consistent with prior findings. For example, using bibliographic coupling to link articles, Zeng et al. (2019) analyzed a set of authors from the physics literature and observed topic switching to be

Table 5. Author statistics for each citation network: Number of clusters per author, and number of clusters in which they are Tier 1, Tier 2, or Tier 3 authors after selection of MCL clusters by size restriction and the 25th percentile of JC scores

	num_clusters	Tier 1	Tier 2	Tier 3
imm_1 (min)	1	0	0	0
imm_1 (median)	1	0	0	1
imm_1 (max)	156	37	50	139
imm_2 (min)	1	0	0	0
imm_2 (median)	1	0	0	1
imm_2 (max)	256	33	127	172
eco (min)	1	0	0	0
eco (median)	1	0	0	1
eco (max)	340	37	87	317

more common in later career authors. In an analysis of 100,000 authors, Ioannidis, Baas, et al. (2019) observed cases where publication profiles were split across fields, which implies that they would likely be found in many clusters.

Further down the pipeline, we examined authorship of the clusters that were selected by (a) size restriction, (b) convergence at the 25th percentile of the JC for matched clusters, and (c) 99th percentile of high intracluster citation activity. At this final level of selection, the number of clusters selected is small (less than 1%) of the previous stage (i.e., size restriction and convergence). This final selection optimizes for clusters with high intracluster citation activity and excludes many others. Relaxing this constraint should be useful in future studies. It should be noted, however, that the number of authors in this smaller number of clusters is approximately 29,000 (imm_1), 45,000 (imm_2), and 36,000 (eco). The number of authors found in multiple clusters, however, is small, with maximum values of 4, 6, and 15 respectively for imm_1, imm_2, and eco. An alternate explanation is that only a small number of clusters fits the model of the Information Exchange Group 1.

4. CONCLUSIONS

We set out to study whether the concept of an "Invisible College" was relevant to the modern scientific enterprise. Here, using an approach that combines network construction, clustering, and analysis we find, consistent with Price and Beaver (1966), that a few authors tend to be influential within author communities focused around specific topics. These small communities tend to have a few Tier 1 and many Tier 3 members, and variable numbers of Tier 2 members (usually fewer than Tier 3). Another observation in this study, which held true for all three networks we examined, was that the most highly cited authors are generally members of many different clusters. This is an intriguing trend that may suggest that the most highly cited scientists today work on many different questions. On the other hand, another explanation may be that science itself has become more diversified, so that these small publication clusters are on different but nevertheless related questions, so that highly cited scientists still tend to work on closely related questions. Distinguishing between these two hypotheses merits further research.

Our observations above were based on a particular pipeline that we used to identify possible communities of practice: (a) Use convergent clustering using a primary clustering method (here, MCL) with a secondary clustering method (here, Leiden) to cluster the citation network, (b) find those primary clusters within an appropriate size range (here, 30–350) that have high intracluster density, and (c) extract the author community for each article cluster. Our preliminary analysis revealed the negative consequences of using too stringent a convergence criterion, and also showed that using Leiden was better (for our purposes) than using Graclus as the secondary clustering method. However, evaluating other variants on this pipeline will help us assess the degree to which these trends hold under other settings. For example, it is possible that making Leiden primary and MCL secondary could lead to different insights, or that using soft clustering (which allows overlapping clusters) could produce new insights.

Changes to the pipeline would also be naturally beneficial if the question of interest changes. In particular, our interest was in small author communities, but other investigators could be interested in larger communities, which could be addressed by changing the size range limits. Furthermore, changes to the size range would also enable insight into how smaller communities are contained within (or span between) larger communities, and thus shed light on community organization. Thus, as research questions change, the specific settings for the pipeline will need to change.

One of the most interesting observations from this study is that restricting attention to clusters with low conductance is not necessarily helpful: As we observed, publication clusters that have very low conductance values seem to be disconnected from the rest of a citation network, thus indicating research areas that are not central to the overall research field being studied. Instead, clusters with high intracluster density but moderate, or possibly even high, conductance values may represent research communities that are more central (i.e., "relevant") to the research area. Note that at the extreme case, a singleton cluster will have maximum possible conductance (1.0), and yet can be highly connected to many other clusters, thus reflecting a significant publication that may well be central to the field. Thus, both types of clusters are of interest, but reflect different types of communities. This observation is consistent with the argument (von Luxburg et al., 2012) that the choice of clustering methodology should be based on the domain and particular research question and also makes a case for careful design of cluster selection criteria.

Despite these encouraging results, we are well aware of the limitations of this study, and of using citation and cluster analysis to identify communities of practice. The best techniques would ideally incorporate expert evaluation at scale, which is unfortunately not feasible; we were able to conduct an expert evaluation for only 100 of the clusters produced by this pipeline. Our limited expert analysis suggests that this pipeline could produce article clusters derived from a scientific theme, and hence that the author communities detected using the pipeline represent likely communities of practice, as was our objective. Our limited characterization of small samples with different conductance values also merits further investigation.

A criticism of our study could be that the data collection approach, seed articles from journal classification labels plus citing and cited articles, may result in sparse citation links that affect subsequent clustering, author tier calculations, and interpretation of results. Our focus on small communities may, at least partially, offset this concern. However, a future study could also include all the cited references of the articles that cite seed articles as well as citation links between these cited references and the other articles we already collect.

In closing, this study primarily concerned examining immunology publications and others connected to them by citation. Other studies have shown that citation behavior can depend significantly on the field (Bradley, Devarakonda, et al., 2020; Wallace, Larivière, & Gingras, 2012), making extrapolation of trends from one field to another premature. Thus, the trends in this study

may not be consistently found in other research disciplines or time frames. Our future work will elucidate these initial observations, evaluate additional clustering techniques, and focus on elucidating interactions between authors within and across clusters to refine the pipeline we envision.

ACKNOWLEDGMENTS

We thank Vladimir Smirnov for helpful discussions on Markov clustering. The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or Elsevier. The ERNIE project involves a collaboration with Elsevier. We thank our Elsevier colleagues for their support of the ERNIE project. We thank the handling editor and three anonymous reviewers for extremely helpful critiques and suggestions.

AUTHOR CONTRIBUTIONS

Shreya Chandrasekharan: Conceptualization, Methodology, Investigation, Writing—original draft; Writing—review & editing. Mariam Zaka: Investigation, Writing—review & editing. Stephen Gallo: Investigation,Writing—Review and Editing. Dimitriy Korobskiy: Methodology. Wenzi Zhao: Methodology. Tandy Warnow: Conceptualization, Methodology; Writing—original draft; Writing—Review and Editing. George Chacko: Conceptualization, Methodology, Investigation, Writing—original Draft, Writing—review & editing, Funding acquisition, Resources, Supervision.

COMPETING INTERESTS

The authors have no competing interests. Scopus data used in this study was available to us through a collaborative agreement with Elsevier on the ERNIE project. Elsevier personnel played no role in conceptualization, experimental design, review of results, or conclusions presented. George Chacko's present affiliation is with the University of Illinois Urbana-Champaign. The majority of his contributions to this article were made as an employee of NETE Solutions Corporation.

SUPPORTING INFORMATION

Supplementary material and code used in this study is available on our Github site (Korobskiy et al., 2019) with the filename SM2.pdf.

FUNDING INFORMATION

Research and development reported in this publication was partially supported by federal funds from the National Institute on Drug Abuse (NIDA), National Institutes of Health, U.S. Department of Health and Human Services, under Contract Nos. HHSN271201700053C (N43DA-17-1216) and HHSN271201800040C (N44DA-18-1216). Tandy Warnow receives funding from the Grainger Foundation.

DATA AVAILABILITY

Access to the bibliographic data analyzed in this study requires a license from Elsevier. Code generated for this study is freely available from our Github site (Korobskiy et al., 2019).

REFERENCES

Ahlgren, P., Chen, Y., Colliander, C., & van Eck, N. (2020). Enhancing direct citations: A comparison of relatedness mea sures for community detection in a large set of pubmed publications. *Quantitative Science Studies*, *1*(2), 714–729. **DOI:** https://doi.org/10.1162/qss_a_00027

Almeida, H., Guedes, D., Meira, W., & Zaki, M. J. (2011). Is there a best quality metric for graph clusters? In *Machine learning and knowledge discovery in databases* (pp. 44–59). Berlin, Heidelberg: Springer. **DOI:** https://doi.org/10.1007/978-3-642-23780-5-13

Barber, B. (1961). Resistance by scientists to scientific discovery. *Science*, *134*, 596–602. **DOI:** https://doi.org/10.1126/science.134.3479.596, **PMID:** 13686762

Barber, B. (1962). *Science and the social order*. New York: Collier Books.

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., … Börner, K. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLOS ONE*, *6*(3), e18029. **DOI:** https://doi.org/10.1371/journal.pone.0018029, **PMID:** 21437291, **PMCID:** PMC3060097

Bradley, J., Devarakonda, S., Davey, A., Korobskiy, D., Liu, S., … Chacko, G. (2020). Co-citations in context: Disciplinary heterogeneity is relevant. *Quantitative Science Studies*, *1*(1), 264–276. **DOI:** https://doi.org/10.1162/qss_a_00007

Colliander, C., & Ahlgren, P. (2011). Experimental comparison of first and second-order similarities in a scientometric context. *Scientometrics*, *90*(2), 675–685. **DOI:** https://doi.org/10.1007/s11192-011-0491-x

Crane, D. (1972). *Invisible colleges: Diffusion of knowledge in scientific communities*. Chicago: University of Chicago Press.

Cucala, M., Bauerfeind, P., Emde, C., Gonvers, J. J., Koelz, H. R., & Blum, A. L. (1987). Is it wise to prescribe NSAIDs with modern gastroprotective agents? *Scandinavian Journal of Rheumatology*, *16*(sup65), 141–154. **DOI:** https://doi.org/10.3109/03009748709102193, **PMID:** 3317804

Devarakonda, S., Korobskiy, D., Warnow, T., & Chacko, G. (2020). Viewing computer science through citation analysis: Salton and Bergmark Redux. *Scientometrics*, *125*, 271–287. **DOI:** https://doi.org/10.1007/s11192-020-03624-0

Dhillon, I., Guan, Y., & Kulis, B. (2007). Weighted graph cuts without eigenvectors: A multilevel approach. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* (Vol. 29:11, pp. 1944–1957). New York: ACM Press. **DOI:** https://doi.org/10.1109/TPAMI.2007.1115, **PMID:** 17848776

Dongen, S. (2000). A cluster algorithm for graphs. *CWI (Centre for Mathematics and Computer Science)*. Retrieved from https://micans.org/mcl/src/mcl-05-090.tar.gz (accessed May 2020).

Elsevier BV. (2020). *Scopus*. https://www.scopus.com/home.uri, accessed July 2020.

Emmons, S., Kobourov, S., Gallant, M., & Börner, K. (2016). Analysis of network clustering algorithms and cluster quality metrics at scale. *PLOS ONE*, *11*(7), e0159161. **DOI:** https://doi.org/10.1371/journal.pone.0159161, **PMID:** 27391786, **PMCID:** PMC4938516

Endres, D., & Schindelin, J. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory*, *49*(7), 1858–1860. **DOI:** https://doi.org/10.1109/tit.2003.813506

Fraker, P. J., & Speck, J. C. (1978). Protein and cell membrane iodinations with a sparingly soluble chloroamide, 1, 3, 4, 6-tetrachloro-3a, 6a-diphenylglycoluril. *Biochemical and Biophysical Research Communications*, *80*(4), 849–857. **DOI:** https://doi.org/10.1016/0006-291x(78)91322-0

Green, D. (1965). Information Exchange Group No. 1. *Science*, *148*, 1543. **DOI:** https://doi.org/10.1126/science.148.3677.1543-b, **PMID:** 17769904

Hagstrom, W. O. (1965). *The scientific community*. New York: Basic Books.

Harry, G., & the LIGO Scientific Collaboration. (2010). Advanced LIGO: The next generation of gravitational wave detectors. *Classical and Quantum Gravity*, *27*(8), 084006. **DOI:** https://doi.org/10.1088/0264-9381/27/8/084006

Hodak, E., & David, M. (2004). Alefacept: A review of the literature and practical guidelines for management. *Dermatologic Therapy*, *17*(5), 383–392. **DOI:** https://doi.org/10.1111/j.1396-0296.2004.04041.x, **PMID:** 15379773

Hull, D. L. (1988). *Science as a process*. Chicago: University of Chicago Press. **DOI:** https://doi.org/10.7208/chicago/9780226360492.001.0001

Ioannidis, J. P. A., Baas, J., Klavans, R., & Boyack, K. W. (2019). A standardized citation metrics author database annotated for scientific field. *PLOS Biology*, *17*(8), e3000384. **DOI:** https://doi.org/10.1371/journal.pbio.3000384, **PMID:** 31404057, **PMCID:** PMC6699798

Klavans, R., & Boyack, K. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology*, *68*, 984–998. **DOI:** https://doi.org/10.1002/asi.23734

Korobskiy, D., Davey, A., Liu, S., Devarakonda, S., & Chacko, G. (2019). *Enhanced Research Network Informatics Environment (ERNIE)* (Github Repository). NET ESolutions Corporation. https://github.com/NETESOLUTIONS/ERNIE

Lander, E., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. **DOI:** https://doi.org/10.1038/35057062, **PMID:** 11237011

Merton, R. K. (1957). *Social theory and social structure*. Glencoe, IL: Free Press.

Michel, B., Hunder, G., Bloch, D., & Calabrese, L. (1992). Hypersensitivity vasculitis and Henoch-Schönlein purpura: A comparison between the 2 disorders. *Journal of Rheumatology*, *19*(5), 721–728.

Milojevic, S. (2020). Practical method to reclassify Web of Science articles into unique subject categories and broad disciplines. *Quantitative Science Studies*, *1*(1), 183–206. **DOI:** https://doi.org/10.1162/qss_a_00014

Price, D. d. S. (1963). *Little science, big science*. New York: Columbia University Press. **DOI:** https://doi.org/10.7312/pric91844

Price, D. d. S. (1965). Networks of Scientific Papers. *Science*, *149*, 510–515. **DOI:** https://doi.org/10.1126/science.149.3683.510, **PMID:** 14325149

Price, D. d. S., & Beaver, D. D. (1966). Collaboration in an invisible college. *American Psychologist*, *21*(11), 1011–1018. **DOI:** https://doi.org/10.1037/h0024051, **PMID:** 5921694

Royal Society. (2020). *History of The Royal Society*. https://royalsociety.org/about-us/history, accessed July 2020.

Salas, E., Rosen, M. A., & DiazGranados, D. (2009). Expertise-based intuition and decision making in organizations. *Journal of Management*, *36*(4), 941–973. **DOI:** https://doi.org/10.1177/0149206309350084

Shu, F., Julien, C.-A., Zhang, L., Qiu, J., Zhang, J., & Larivière, V. (2019). Comparing journal and paper level classifications of science. *Journal of Informetrics*, *13*(1), 202–225. **DOI:** https://doi.org/10.1016/j.joi.2018.12.005

Shun, J., Roosta-Khorasani, F., Fountoulakis, K., & Mahoney, M. W. (2016). Parallel local graph clustering. *Proceedings of the VLDB Endowment*, *9*(12), 1041–1052. **DOI:** https://doi.org/10.14778/2994509.2994522

Small, H., & Sweeney, E. (1985). Clustering the science citation index® using co-citations. *Scientometrics*, *7*(3), 391–409. **DOI:** https://doi.org/10.1007/BF02017157

Šubelj, L., van Eck, N. J., & Waltman, L. (2016). Clustering scientific publications based on citation relations: A systematic comparison of different methods. *PLOS ONE*, *11*(4), e0154404. **DOI:** https://doi.org/10.1371/journal.pone.0154404, **PMID:** 27124610, **PMCID:** PMC4849655

Traag, V. (2020). *CWTSLeiden/networkanalysis* (Github Repository). CWTS Leiden. https://bit.ly/3j1DX9H, accessed September 2020.

Traag, V., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, *9*(1), 1–12. **DOI:** https://doi.org/10.1038/s41598-019-41695-z, **PMID:** 30914743, **PMCID:** PMC6435756

von Luxburg, U., Williamson, R. C., & Guyon, I. (2012). Clustering: Science or art? In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning, JMLR Workshop and Conference Proceedings* (Vol. 27, pp. 65–79).

Wagner, C. (2008). *The new invisible college: Science for development*. Washington DC: Brookings Institution Press.

Wallace, M. L., Larivière, V., & Gingras, Y. (2012). A small world of citations? The influence of collaboration networks on citation practices. *PLOS ONE*, *7*, e33339. **DOI:** https://doi.org/10.1371/journal.pone.0033339, **PMID:** 22413016, **PMCID:** PMC3296690

Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392. **DOI:** https://doi.org/10.1002/asi.22748

Zeng, A., Shen, Z., Zhou, J., Fan, Y., Di, Z., … Havlin, S. (2019). Increasing trend of scientists to switch between topics. *Nature Communications*, *10*(1). **DOI:** https://doi.org/10.1038/s41467-019-11401-8, **PMID:** 31366884, **PMCID:** PMC6668429