Searching for temporal patterns in the time series of publications of authors in a research specialty

Theresa Velden
School of Information
University of Michigan
309 S. State Street
Ann Arbor, MI 48104, USA
tvelden@umich.edu

Carl Lagoze
School of Information
University of Michigan
309 S. State Street
Ann Arbor, MI 48104, USA
clagoze@umich.edu

Kan Yu
School of Information
University of Michigan
309 S. State Street
Ann Arbor, MI 48104, USA
kanyu@umich.edu

ABSTRACT
In this paper we report results of our investigation of temporal patterns in the publication activity of authors in a research specialty. We base our analysis on Web of Science data for a field in the physical and chemical sciences from 1991-2012. We determine the research groups in the field by clustering the co-author network and generate our sample for this analysis by selecting the most productive author of each co-author cluster to represent the activity of that group. Whereas a statistical time series analysis did not reveal any specific patterns, a time series clustering approach generated a grouping of time series that correlates with the structural network position ('node role') of the respective authors in the clustered co-author network. This work is part of a long-term research project employing a mix of qualitative and network analytic methods to investigate field-specific differences in collaborative patterns.

Keywords
Research specialty, temporal publication activity, co-author network, node role analysis, statistical time series analysis, time series clustering.

INTRODUCTION
Apart from studies of high level trends such as the increase in publication output for individual fields or across fields [see e.g. Tabah 1999, Morris 2005] or the study of variations in individual productivity over time [e.g. Diamond 1984, Levin & Stephan 1991, Hall et al 2007], there are few studies on temporal patterns in publication activity of scientific communities

Our goal in this explorative study has been to determine whether characteristic temporal patterns exist in the publication activities of research groups in a field. By temporal patterns we are looking in particular for regular 'rhythmic' patterns that repeat over time or any other systematic temporal patterns that may be related to specific work or collaboration practices of a field. Such patterns could be indicative of the work and collaboration practices in a field, e.g. a group's publishing activity might exhibit a burst-like pattern in a field that requires considerable time investment into the design and realization of custom built instruments for experiments, or dependence on the availability of scarce observation times at a shared facility (telescope, synchrotron). Field differences, if detected, could be informative for the study and comparison of field-specific behavioral patterns. In our broader research context we seek to leverage quantitative results such as those described here to inform and guide qualitative investigations that reveal nuanced features of field-specific behavior e.g. of openness and sharing (Velden 2013) which are of interest to information system design and science policy.

Our initial approach to this problem was to conduct a statistical analysis of time series that represent the publication activity of the five most productive individuals in a topic area within a research field or of topic-specific aggregates of such time series. However, the ARIMA (Autoregressive Integrated Moving Average) analysis of those time series failed to detect any interesting temporal patterns in the form of autocorrelation of the data (systematic trends, 'seasonal' periodicity, self-dependency

77th ASIS&T Annual Meeting, October 31- November 5, 2014. Seattle, WA, USA.

with a specific lag time etc.). Hence we broadened our search and decided to use time series clustering to explore whether we could discover groupings of time series based on their similarity to one another. Such groupings may indicate characteristics of temporal publication patterns that distinguish scientists and their role (form of contribution) within a scientific community. Our findings are described in this paper.

DATA

The data we used to construct the time series of publication activity for authors in a research specialty were retrieved from the Web of Science (WoS) database using a lexical query previously developed during an ethnographic field study. The lexical query combines relevant terms to describe core research objects and methods in a research specialty at the boundary between chemistry and physics. The query was developed during an ethnographic field study [Velden et al 2010] and the method we developed for improving recall and precision of the query is described in [Velden & Lagoze 2013]. One issue of WoS data is the identification of unique authors. To deal with name homonymy we use the author name disambiguation approach introduced in Velden et al [2011]. The disambiguated data set used in this study comprises 87,198 publications and spans a period of 22 years (1991-2012).

METHOD

Time series construction

From the bibliographic records retrieved we extracted author and publication year information to construct time series of individual authors' publication activity (publication counts) at an annual level of granularity. We dismissed choosing a monthly resolution since in the research field studied publications in journals are more relevant than conference publications (the latter often are only paper abstracts). Hence a monthly resolution would seem excessive, especially if one takes into account the variation in latency between article submission and actual publication in a scientific journal (see e.g. Kravitz & Baker 2011, Dióspatonyi et al. 2001).

Sampling

In order to create a sample of authors that represent the activity of a research group in the field we made use of information from the co-author network constructed from the data [Velden et al 2010], specifically the cluster membership of authors. We only considered the most productive author of each co-author cluster for inclusion in the sample. We excluded authors whose publication activity spanned less than 15 years (i.e. the time period from first occurrence to last occurrence of a publication by this author in the data set) reasoning that shorter time periods will be less effective in revealing temporal patterns. We obtained 1128 time series, ranging between 15 to 22 years of length.

We further determined possible overlap between those time series, i.e. what proportion of publications in a pair of time series coincided because both authors are co-authors on the same publication. To avoid spurious clustering results that would simply reflect overlap of time series due to co-authorship we included only time series with less than 10% overlap with any other time series. This reduced our sample to 659 time series. We chose this cut off after considering that only 478 of the 1128 time series did not show any overlap. Reducing our sample to that number would have limited the statistical power of our analysis (especially testing for correlations with author node role, explained below).

Normalization

We normalized each author's time series of simple raw counts of annual publications by dividing by the total count of publications of the author over the entire time period. Hence, the y-axis of the normalized time series shows the proportion of total publications of the author published in the respective year. Further, we aligned the time series to one another by starting the x-axis of each time series with the first active year of the author. Two characteristic examples of normalized author time series are shown in figure 1.

Time series clustering

Three fundamental approaches can be taken for time series clustering [Warren Liao 2005]: clustering based on the raw data of the time series, clustering based on a feature vector describing specific properties of the time series, and finally clustering based on parameters derived from statistical time series modeling of the time series. Given our failure to find a good match between our time series data and ARIMA models and the paucity of theoretical guidance from the literature on what features of time series might be particularly meaningful for a comparison, we opted for using the raw time series data clustering approach.

Similarity measure

To compare and determine the similarity of time series we decided to use the Dynamic Time Warping (DTW) algorithm [Giorgino 2009] to look for global similarities of the time series[1]. Without having any strong theoretical guidance on the differences to expect between the time series and what would signify interesting patterns regarding the alignment of publication events[2], we opted for a measure that matches time series of different lengths and with a certain tolerance toward relative stretching of the time axis. DTW is an elastic similarity measure. This

[1] The corresponding R package is available at http://dtw.r-forge.r-project.org/

[2] Had we dealt with a scientific community oriented towards conference publications and had we specifically searched for signals of large conferences we would have chosen a measure that accurately resolves and matches events time point by time point.

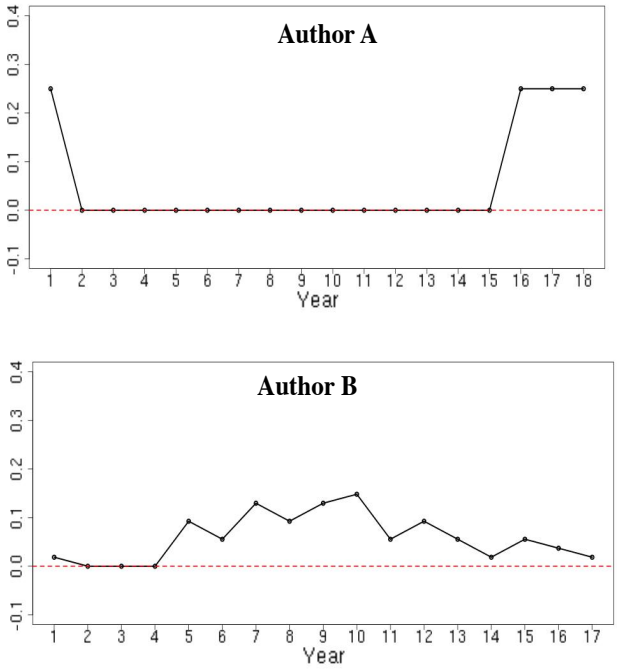

Figure 1: Examples of Normalized Time Series Of Publications, Author A) Activity period = 18 years, total publications = 4. Author b) Activity period = 17 years, total publications = 54.

distinguishes it from inelastic measures such as the Euclidean distance, where distance between two time series is calculated only for value pairs at the same point of the time-axis. Wang et al [2013] found in an experimental comparison of 9 different similarity measures on 38 time series data sets from a range of different domains that, for small data sets, elastic measures outperform inelastic measures such as Euclidean distance or lock-step measures. They also found that many newly suggested edit distance-based measures such as LCSS, EDR and ERP didn't provide significantly more accurate classification results than the simpler, 40-year old DTW measure.

The time line stretching tolerance of the DTW measure can be bounded by specifying a warping time window size, that is the time interval that is admissible when comparing and searching for a best match between points of two time series. We selected a slanted warping window of three years' size. A window size of three seemed a reasonable tolerance given the length of a typical research project (project funding, dissertation duration) and possible delays in the publication of results. An example of how the distance between two time series is calculated using DTW with a slanted window size of three is provided in figure 2.

Clustering algorithm

Using the DTW similarity measure we generated a similarity matrix of the pairwise comparison of the 659 time series in our sample. To group the time series into clusters based on this similarity matrix we applied affinity propagation clustering, a clustering method that in an

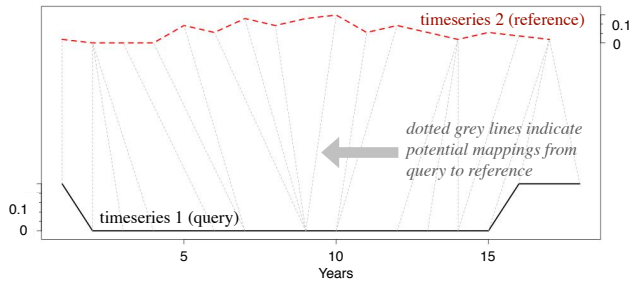

Figure 2: Time axis mapping example for the calculation of DTW distance between two time series with a slanted window of size 3. The solid black line at the bottom represents the first time series, the so called 'query', and the red dotted line at the top represents the second time series, called the 'reference'. The origin of time axes is aligned and the first point is matched by default. The dotted grey lines indicate which mappings of points (time line distortions) are considered in each step of the algorithm. The time window size restricts how often a once chosen point on the time axis of the reference curve can be mapped to again and the monotonicity of the mapping functions ensures that loops are avoided. For further details consult [Giorgino 2009].

iterative process of passing messages between data points identifies exemplars among those data points and forms clusters of data points around these exemplars [Frey & Dueck 2007, Bodenhofer et al. 2011]. We used the default setting for the input parameter in the R package[3]. The value of this parameter does not specify a specific number of clusters to be returned but it does influence the granularity of the clustering, i.e. whether few or many clusters will be returned. The affinity propagation clustering method is very efficient. Based on the heat map representation of the clustering it delivers for our data and the visualization of ten randomly sampled time series for each cluster, it seems to do a reasonably good job (see Results section).

Network node role

One of the characteristics of each author in our data set is her or his structural position in the co-author network that can be constructed from the publication data for the entire field. This co-author network is highly modular, with co-author clusters representing authors who frequently publish together. These are usually the members of the same research groups and their close associates, or small sub-networks of closely collaborating research groups [Velden et al. 2010]. Authors in such a modular network can be classified by a node role classification [Guimera et al 2007] that captures the extent to which an author is connected within his or her own module, making it a hub or non-hub node, and the extent to which its links are distributed among many clusters or just concentrated on its own

[3] Available for download at http://cran.r-project.org/web/packages/apcluster/index.html.

Node Role	Descriptive name	Frequency in Sample
R1	Ultra-peripheral Non-hub	175
R2	Peripheral Non-hub	126
R3	Satellite connector Non-hub	29
R4	Kinless Non-hub	1
R5	Provincial Hub	185
R6	Connector Hub	135
R7	Global Hub	1

Table 1. Classification of node Roles in a modular network. Roles R1-R4 refer to non-hub nodes, and roles R5-R7 refer to hub nodes.

cluster. This classification assigns seven node roles to nodes in a modular network. Our sample contains 331 non-hub nodes and 321 hub nodes and 7 nodes for which no node role could be determined for technical reasons[4]. Table 1 provides an overview on the node roles included in our sample. We use node role to determine which authors were grouped together in the time series clusters.

RESULTS

The affinity propagation-clustering algorithm groups our sample of 659 normalized time series into nine clusters of different sizes (see table 2). A visualization of the clustering results provided by the R program is given in figure 3. The heat map is a visualization of the similarity matrix organized by the nine clusters identified by the algorithm. Both the heat map and the tree structure of the cluster similarities indicate a high level partition of the data. Partition one is indicated by the large dark square in the upper left of the heat map and includes clusters 2,5,7,8,9 with a total of 307 time series. Partition 2 is indicated by the large dark square in the lower right of the heat map and includes clusters with a total of 352 time series.

To shed some light on the shape and commonalities between time series grouped together into the same cluster, figure 4 depicts ten randomly chosen time series for each cluster. Overall the examples visually suggest that the clustering is working well, grouping together time lines that look similar to each other in overall shape. The example time series in each of the clusters belonging to partition 1 (clusters 2,5,7,8,9) seem to be closer to one another, which corresponds to the more homogenous dark shading in the similarity matrix in figure 3. For partition 2 (clusters 1,3,4,6) there seems to be greater inhomogeneity within clusters, corresponding in the similarity matrix in figure 3 to a mix of some very high similarities (dark color) and several lighter lines indicating low similarity. Closer inspection of individual time lines in figure 4 reveals that timelines in partition 1 tend to have many more points of publication activity whereas time lines in partition 2 tend to have one or more extended periods of inactivity. Because of the normalization chosen for the timelines, having many years with publication activity has a homogenizing effect. This is because a broad distribution of the publication output over many years brings all proportions down such that, even if a year without publication is compared to one with publications, the difference between the points is small. For timelines with only a few years of publication activity, the difference between active years and inactive years is high. This effect is ameliorated by the elasticity of the distance measure but obviously not entirely suppressed.

To make sense of the clustering and determine what authors whom have been grouped together may have in common, we investigate the correlation between cluster membership of time series and node roles. In our previous research we have found node roles informative, as oftentimes they can be associated with certain social roles that are reflected by the network position of the respective author. For example, hub nodes tend to represent research group leaders (PIs), or R4 nodes (non-hub nodes with a distribution of links across several co-clusters) can represent experts that collaborate to provide specialized services to research groups (e.g. crystallographers who conduct complex structural analysis for a community of synthetic chemists).

The table of adjusted residuals of our correlation analysis (shown in the appendix) shows a strong and significant correlation between cluster membership and the hubness of a node. The residuals indicate that for hub nodes there is a strong positive correlation with being included in a cluster

Cluster	Cluster Size (Number of Time Series)
Cluster 1	62
Cluster 2	27
Cluster 3	117
Cluster 4	26
Cluster 5	15
Cluster 6	147
Cluster 7	196
Cluster 8	23
Cluster 9	46

Table 2. Cluster Sizes

[4] The variance of within cluster degree in their co-author cluster is zero and a normalization step in the calculation of node role requires division by this variance.

of partition 1 (clusters 2,5,7,8,9) and a negative correlation with being included in a cluster of partition 2 (clusters 1,3,4,6). For non-hub nodes we find the opposite, a negative correlation with being included in one of the clusters in partition 1 and a positive correlation with being included in one of the clusters in partition 2.

Finally, we note a basic partition in our data that corresponds to the shape of timelines that are characterized by the presence versus the absence of extended periods of inactivity. Based on this, we test the authors in our sample for a correlation between node role and the proportion of inactive years within the entire activity span of an author. Figure 5 shows that there is a distinct (and significant) jump between non-hub nodes and hub nodes in the median of the proportion of inactive years.

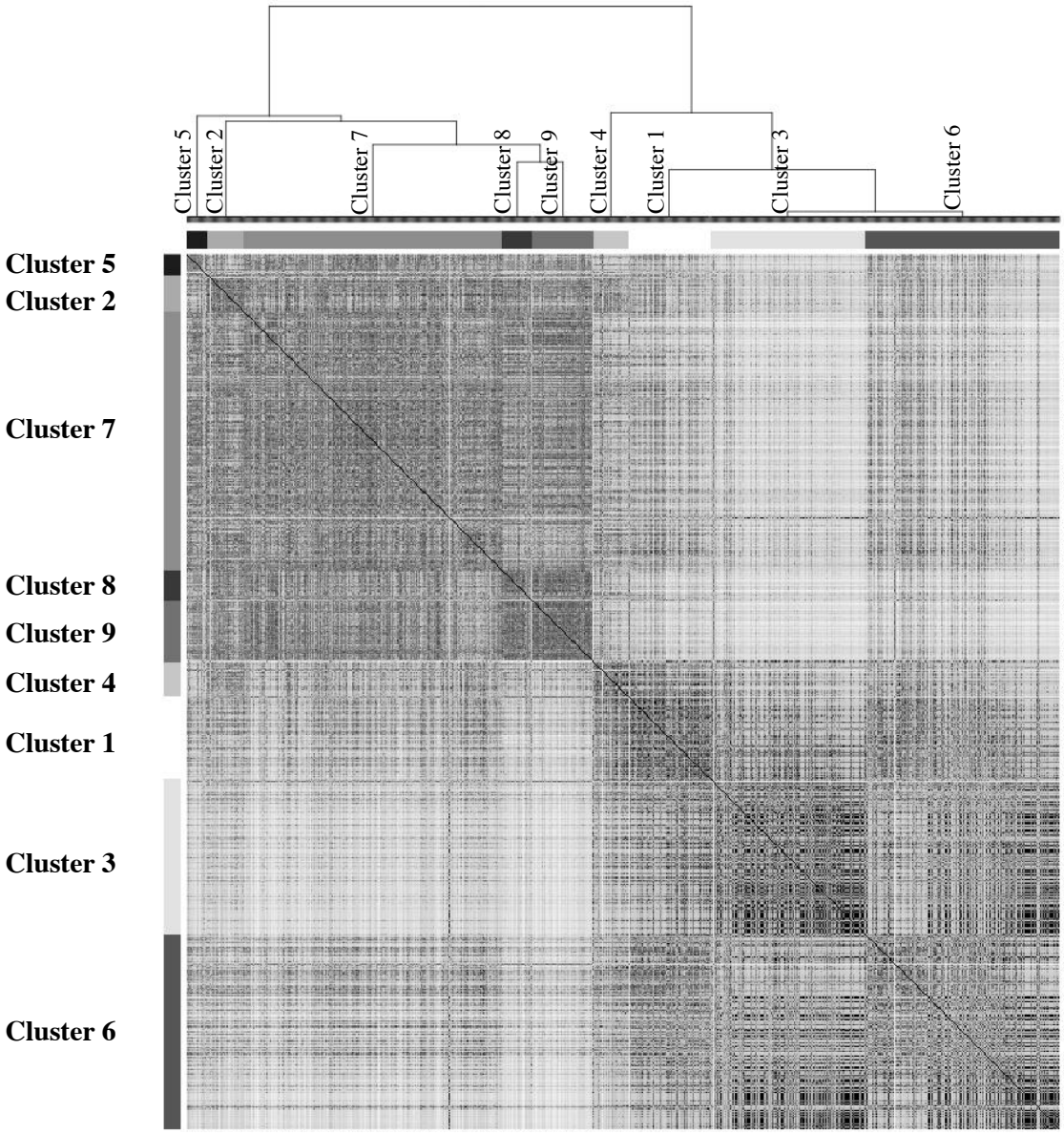

Figure 3: Heat map visualization of clustering result. Each element in the matrix represents the similarity between two time series with darker colors indicating stronger similarities. The color blocks in the top and the side bar indicate the grouping of the time lines into nine clusters. The sizes of these clusters are given in table 2. The tree structure at the top of the visualization shows the similarity between the clusters with the distance of a merge point from the bottom indicating distinctiveness of the two clusters getting merged.

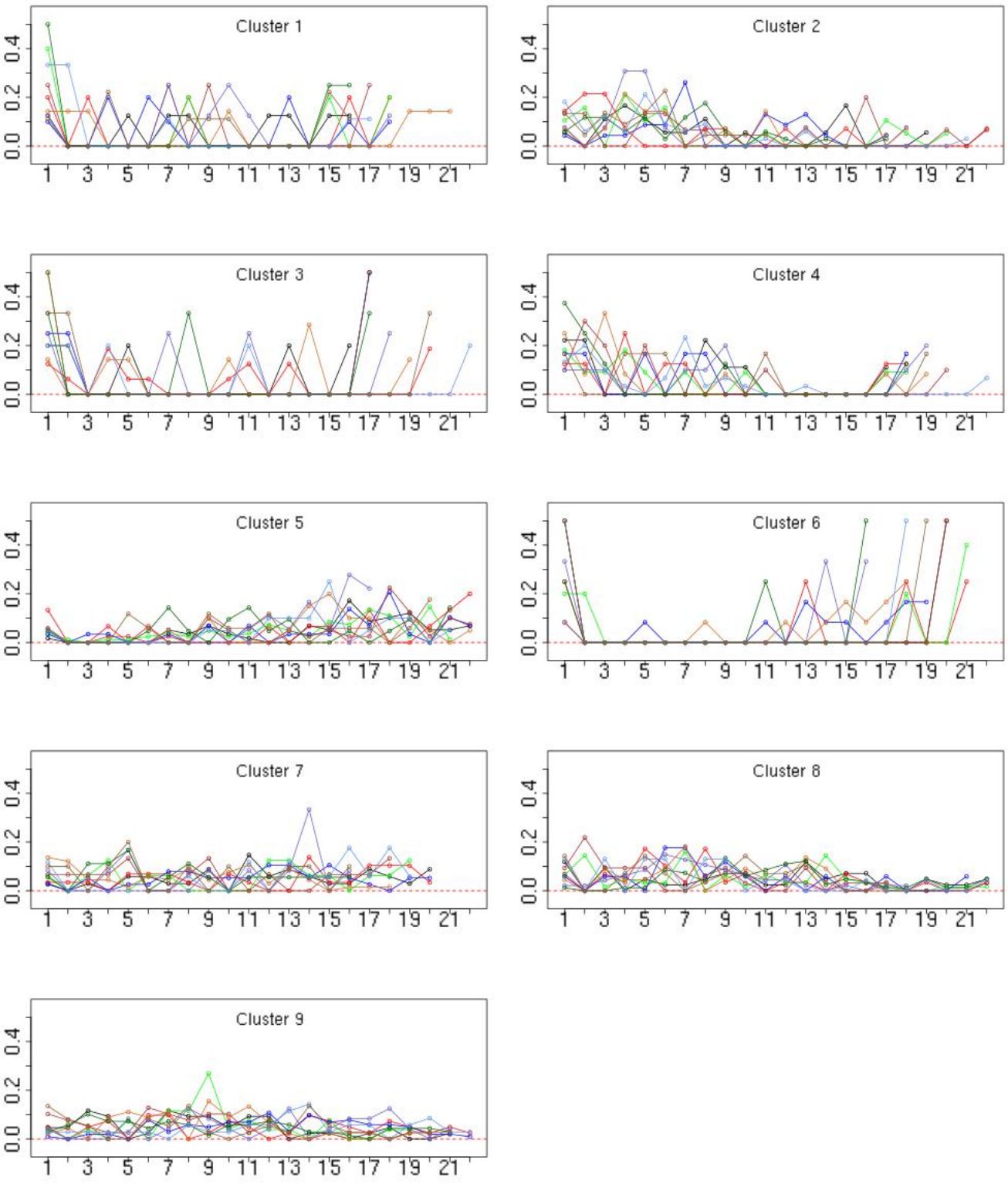

Figure 4: Ten examples of author publication activity time series randomly selected from each cluster.

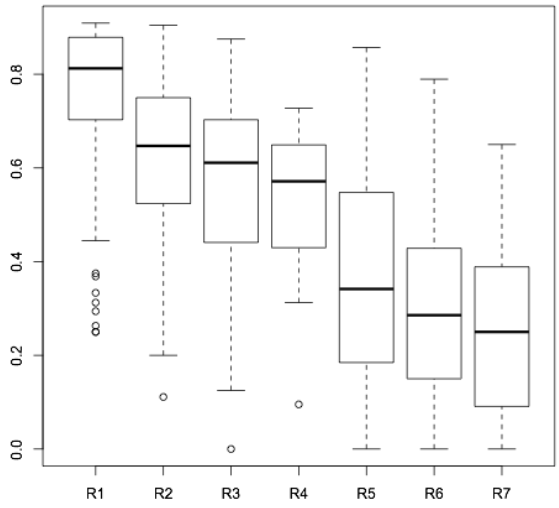

Figure 5: Distribution of the proportion of inactive years in an author's activity period by node role.

DISCUSSION

There are different ways to look at the significance of finding a correlation between the time series clusters of the authors in our sample and their hubness as nodes in the co-author network:

1. In hindsight, the clustering of time series by node hubness seems straightforward if one considers two factors: First, a visually quite striking feature of the shape of time series are extended durations of inactivity. Based on the similarity metric we chose for the clustering algorithm, we can expect similar looking time series to be grouped together. Second, as shown in figure 5, there exists a correlation between the hubness of nodes in our sample and the proportion of inactive years. Hence, the time series clustering will tend to group the time series of either hub nodes or non-hub nodes together.

2. However, this latter correlation between node hubness and proportion of inactive years would seem trivial only if one considered all nodes in a coauthor network and calculated the proportion of inactive years relative to the entire time period covered by the data. This is because the majority of nodes in the network are transient non-hub nodes of role types R1 and R2. They participate only for a short time period in the network (most of them presumably students who leave the field after they graduate). Interestingly, as we show above, the correlation between node hubness and inactive years holds even if one restricts the sample to only the most productive author for each co-author cluster with an activity span of 15 years and more and calculates the proportion of inactive years relative to each author's specific activity period.

3. Therefore, another way to look at this result is that it emphasizes that about half of the co-author clusters represented by the authors in our sample are characterized by a distinctively different mode of participation in the network. In addition to their most productive members being non-hub nodes in the network structure we find a strikingly different pattern in their time series of publications. This observation provides a starting point for a future qualitative investigation to understand the presence and role of these groups within the field.

4. Finally, the tendency of the time series of hub nodes and non-hub nodes to be grouped into separate clusters calls attention to those nodes who are outliers in that their time series are grouped together with the other type of nodes. Again, these outliers represent a curiosity and call for a qualitative investigation into their mode of participation and role within the field.

SUMMARY AND CONCLUSIONS

Temporal patterns of publication activity within specific scientific communities have been little researched. We set out to search for temporal patterns in the time series of authors in a research specialty. Our focus is on authors that represent co-author groups (research groups) in order to look for characteristics of publishing activity that might be indicative of group-level differences in research practices or forms of participation in the field. Whereas a test of autocorrelation patterns came up negative, time series clustering has provided some additional insight into the composition of the field and distinct differences between research groups and their behavioral patterns. To follow up on these results we plan to conduct a qualitative study into the presence and role of co-author groups in the network that either are characterized by having a non-hub group leader and a time series falling into the non-hub partition, or for whom the structural node role of the research group leader makes them an outlier within the clustering.

ACKNOWLEDGMENTS

We acknowledge funding support from two grants: 1) OCI 1301874 Understanding Conditions for the Emergence of Virtual Orgs, and 2) SMA 1258891 EAGER: Collaborative Research: Scientific Collaboration in Time.

REFERENCES

Bodenhofer, U., Kothmeier, A., & Hochreiter, S. (2011). APCluster: an R package for affinity propagation clustering. *Bioinformatics*, *27*(17), 2463-2464.

Diamond, A. M. (1984). An economic model of the lifecycle research productivity of scientists. *Scientometrics*, *6*(3), 189-196.

Dióspatonyi, I., Horvai, G., & Braun, T. (2001). Publication speed in analytical chemistry journals. *Journal of chemical information and computer sciences*,*41*(6), 1452-1456.

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, *315*(5814), 972-976.

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: the dtw package. *Journal of statistical Software*, *31*(7), 1-24.

Guimera, R., Sales-Pardo, M., & Amaral, L. A. (2007). Classes of complex networks defined by role-to-role connectivity profiles. *Nature physics*, *3*(1), 63-69.

Hall, B. H., Mairesse, J., & Turner, L. (2007). Identifying age, cohort, and period effects in scientific research productivity: Discussion and illustration using simulated and actual data on French physicists. *Econ. Innov. New Techn.*, *16*(2), 159-177.

Kravitz, D. J., & Baker, C. I. (2011). Toward a new model of scientific publishing: discussion and a proposal. Frontiers in computational neuroscience, 5.

Levin, S. G., & Stephan, P. E. (1991). Research productivity over the life cycle: Evidence for academic scientists. *The American Economic Review*, 114-132.

Morris, S. A. (2005). Manifestation of emerging specialties in journal literature: A growth model of papers, references, exemplars, bibliographic coupling, cocitation, and clustering coefficient distribution. *Journal of the American Society for Information Science and Technology*, *56*(12), 1250-1273.

Tabah, A. N. (1999). Literature dynamics: Studies on growth, diffusion, and epidemics. *Annual review of information science and technology (ARIST)*, *34*, 249-86.

Velden, T. (2013) Explaining field differences in openness and sharing in scientific communities. In Proceedings of the 2013 conference on Computer supported cooperative work, pages 445–458. ACM.

Velden, T. and Lagoze, C. (2013). The extraction of community structures from publication networks to support ethnographic observations of field differences in scientific communication. Journal of the American Society for Information Science and Technology, 64(12):2405–2427.

Velden, T., Haque, A., and Lagoze, C. (2011) Resolving author name homonymy to improve resolution of structures in co-author networks. In JCDL'11, June 13-17, 2011, Ottawa, Ontario, Canada.

Velden, T., Haque, A. and Lagoze, C. (2010). A new approach to analyzing patterns of collaboration in co-authorship networks: mesoscopic analysis and interpretation. Scientometrics, 85(1):219–242.

Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., & Keogh, E. (2013). Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, *26*(2), 275-309.

Warren Liao, T. (2005). Clustering of time series data—a survey. *Pattern Recognition*, *38*(11), 1857-1874.

APPENDIX

To determine whether there exists a correlation between an author's node role in the co-author network and the cluster membership of the time series of his or her publications we calculate the adjusted residuals. Values higher than 2 or lower than -2 suggest a positive or negative dependency, respectively. For some node roles the expected count in more than 20% of the cells are below 5 (highlighted in light red in the color version of the table). Hence we used Fisher's exact test to confirm that there is dependence of cluster membership on node role, with a p-value < 0.001.

The residuals suggest a positive correlation of clusters 7,8, and 9 (and to some extent cluster 2) with hub nodes, and a negative correlation with non-hub nodes. Clusters 6 and 3 (and to some extent clusters 1 and 4) show a positive correlation with non-hub nodes and a negative correlation with hub nodes.

		NA	R1	R2	R3	R4	R5	R6	R7	Σ
Cluster 1	Count (expected)	0 (1)	21(17)	22(12)	5(3)	0(0)	11(17)	3(13)	0(0)	**62**
	adjusted residual	*-0.86*	*1.37*	***3.44***	*1.48*	*-0.32*	*-1.90*	***-3.21***	*-0.32*	
Cluster 2	Count (expected)	0 (0)	3(7)	4(5)	0(1)	0(0)	13(8)	7(6)	0(0)	**27**
	adjusted residual	*-0.55*	*-1.86*	*-0.58*	*-1.14*	*-0.21*	***2.37***	*0.72*	*-0.21*	
Cluster 3	Count (expected)	6 (1)	65(31)	28(22)	10(5)	0(0)	7(33)	1(24)	0(0)	**117**
	adjusted residual	*4.73*	***7.83***	*1.46*	*2.41*	*-0.46*	***-5.86***	***-5.80***	*-0.46*	
Cluster 4	Count (expected)	0(0)	8(7)	9(5)	1(1)	0(0)	6(7)	2(5)	0(0)	**26**
	adjusted residual	*-0.54*	*0.50*	***2.05***	*-0.14*	*-0.20*	*-0.58*	*-1.65*	*-0.20*	
Cluster 5	Count (expected)	0 (0)	0(4)	3(3)	0(1)	0(0)	5(4)	8(3)	0(0)	**16**
	adjusted residual	*-0.42*	*-2.43*	*-0.04*	*-0.87*	*-0.16*	*0.29*	*2.96*	*-0.16*	
Cluster 6	Count (expected)	1(2)	65(39)	38(28)	6(6)	0(0)	26(41)	11(30)	0(0)	**147**
	adjusted residual	*-0.51*	***5.50***	***2.35***	*-0.21*	*-0.54*	***-3.18***	***-4.43***	*-0.54*	
Cluster 7	Count (expected)	0(2)	12(52)	22(37)	6(9)	1(0)	83(55)	70(40)	1(0)	**195**
	adjusted residual	*-1.72*	***-7.69***	***-3.32***	*-1.07*	*1.54*	***5.37***	***6.35***	*1.54*	
Cluster 8	Count (expected)	0(0)	0(6)	0(4)	0(1)	0(0)	11(6)	12(5)	0(0)	**23**
	adjusted residual	*-0.51*	***-2.94***	***-2.37***	*-1.05*	*-0.19*	***2.15***	***3.83***	*-0.19*	
Cluster 9	Count (expected)	0(0)	1(12)	0(9)	1(2)	0(0)	23(13)	21(9)	0(0)	**46**
	adjusted residual	*-0.73*	***-3.88***	***-3.42***	*-0.76*	*-0.27*	***3.43***	***4.38***	*-0.27*	
	Σ	**7**	**175**	**126**	**29**	**1**	**185**	**135**	**1**	