Toward a Mesoscopic Analysis of the Temporal Evolution of Scientific Collaboration Networks

Syed Ishtiaque Ahmed sa738@cornell.edu
Scott Allen Cambo sac355@cornell.edu
Department of Information Science, Cornell University

Carl Lagoze clagoze@umich.edu
Theresa Velden tvelden@umich.edu
School of Information, University of Michigan

Abstract

This poster reports on our latest results in a multiyear project that employs a mixed network analytic and ethnographic approach to understand the factors underlying field-specific attitudes towards openness and sharing of scholarly data. We report initial results of adding a temporal dimension to an analysis of scientific collaboration networks that provide evidence for comparative study of community structures and collaboration patterns across scientific fields. The addition of a temporal dimension to the analysis allows us to study the dynamic processes involved in the evolution of a scientific community and to determine field specific patterns. Further, it improves the accuracy with which the internal structures of scientific collectives can be resolved. This ongoing work advances an ethnographically grounded approach to the mesoscopic analysis of collaboration networks. Supported by ethnographic insights, we can connect mesoscopic network features to notions of research groups, group leadership and implied seniority, inter-group collaboration, between group migration, and ephemeral one-off exchanges. Eventually, a mesoscopic perspective should allow us to significantly improve the validity of models to explain network evolution.

Keywords: co-author networks, mixed methods, temporal evolution, scientific collaboration, community structures

Introduction

This poster reports on our latest results in a multiyear project that employs a mixed network analytic and ethnographic approach to understand the factors underlying field-specific attitudes towards openness and sharing of scholarly data. We report initial results of adding a temporal dimension to an analysis of scientific collaboration networks that provide evidence for comparative study of community structures and collaboration patterns across scientific fields. The addition of a temporal dimension to the analysis allows us to study the dynamic processes involved in the evolution of a scientific community and to determine field specific patterns. Further, it improves the accuracy with which the internal structures of scientific collectives can be resolved. The results can then be used to guide the strategic sampling of field sites for comparative ethnographic field studies.

This work aims at advancing an ethnographically grounded approach to the mesoscopic analysis of collaboration networks (Velden, Haque, & Lagoze, 2010; Velden & Lagoze, 2012). Supported by ethnographic insights, we can connect mesoscopic network features to notions of research groups, group leadership and implied seniority, inter-group collaboration, between group migration, and ephemeral one-off exchanges. Previous work has oftentimes conceptualized co-author nodes as autonomous actors

---

Acknowledgments: National Science Foundation Grant No OCI-1025679, International Fulbright Science and Technology Fellowship for S.I. Ahmed.
Ahmed, S. I., Cambo, S. A., Lagoze, C., & Velden, T. A. (2013). Toward a mesoscopic analysis of the temporal evolution of scientific collaboration networks. *iConference 2013 Proceedings* (pp. 878-881). doi:10.9776/13446

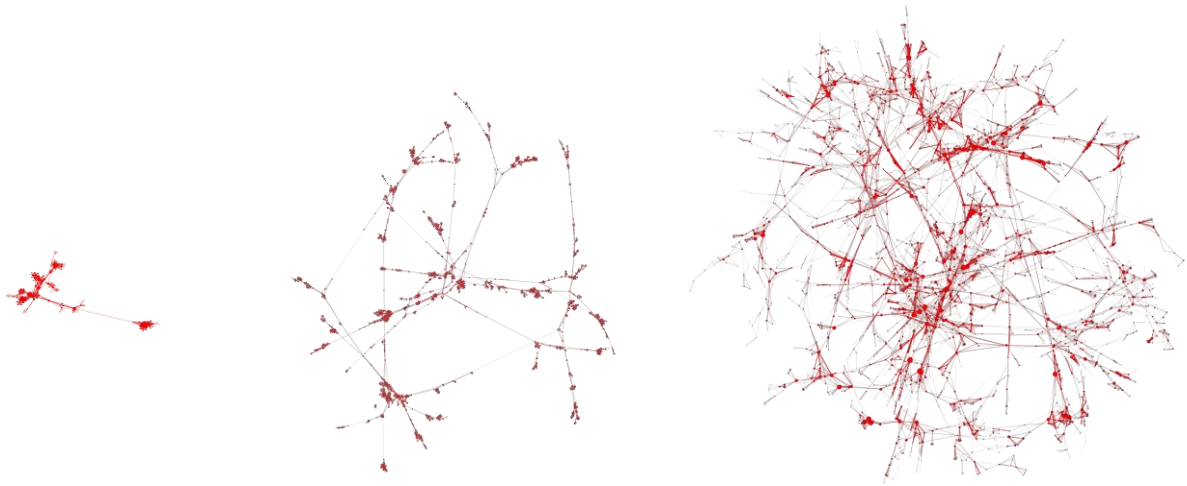
driven by individualistic mechanisms such as preferential attachment, ignoring the actual social composition of research collectives and the various socially distinct processes contributing to global network growth and densification. Eventually, a mesoscopic perspective should allow us to significantly improve the validity of models to explain network evolution.

Methods

We are developing an open source code base (http://github.com/tvelden/communities) that allows us to flexibly generate co-author networks following different time-slicing schemes: 'accumulative' for tracking the accumulative growth of the network, and 'sliding' for generating a dynamic view of the evolution of network structures by considering only publications in a specific time window. This sliding window can move across the entire time range covered by the available data. We have integrated into the code methods that support the mesoscopic analysis of networks, such as the network clustering code by (Rosvall & Bergstrom, 2008) and our own implementation of a node classification algorithm for clustered networks by (Guimera, Sales-Pardo, & Amaral, 2007). The latter classification scheme allows us to distinguish types of nodes by their structural embedding into their surrounding co-author cluster as well as by their out of cluster connectivity. For example, hub nodes extracted from our networks by this classification scheme can be identified as research group leaders (Velden et al., 2010).

Data

We have developed a lexical query to extract from the Web of Science (WoS) of Thomson Reuters the publication output of two fields in the physical and chemical sciences between 1991- 2010, one in synthetic chemistry (field 1), and one at the boundary of physics and physical chemistry (field 2). An important step is the cleaning of data. To improve the accuracy of the co-author networks we apply an author disambiguation algorithm (Velden, Haque, & Lagoze, 2011). We further use a statistical approach to define hyper-authorship in a data set-specific way and use it to exclude a small set of papers (1-3%) that are not representative of the research style in the long-tail science fields that we study here. A manual analysis finds that in many cases those hyper-authorship papers represent out-of-scope papers that the lexical query mistakenly captured. In a few cases we also find large-scale collaborations that contribute to the specific field we study, but represent only a marginal sub-community within the field. Finally, we exclude authors who have co-authored only a single paper. About two-thirds of authors are removed in this step. We found that metrics for the global network topology were not affected by this latter reduction step. The reduced data is much more manageable for analysis and visualization purposes, such as the visualization of the giant component of the network of field 2 in figure 1.



(a) 1991-1992 diameter = 15 (b) 1991-1993 diameter = 59 (c) 1991-1995 diameter = 41

Figure 1. Network structure of giant component of field 2 at different stages of its (accumulative) evolution. Between 1992 (a) and 1993 (b) the initial core of the giant component is formed and the network diameter suddenly increases in size. In 1995 (c), even though the network has been further growing in size, the network diameter decreases, indicating the increasing densification of the network over time.

Initial Findings

We calculate global network metrics (table 1) and compare them to the characteristics of co-author networks investigated by (Bettencourt, Kaiser, & Kaur, 2009) to confirm that our network data represent scientific fields within a common range of topological characteristics and that we are not dealing with extreme outliers. We find that the global metrics conform broadly to the characteristics of other fields. For both our fields, the scaling parameter for network densification is similar e.g. to the field of carbon nanotubes and in accordance with a non-pathological fields of a community of researchers that share concepts and techniques (Bettencourt et al., 2009).

Table 1
Basic Network Properties and Global Metrics for Preprocessed Data

	# papers	# authors	edges (weighted)	diam. giant	size giant (nodes)	size giant (edges)	densification (scaling param.)
Field 1	12,641	13,397	58,375	ca. 35	62.3%	76.7%	1.14
Field 2	56,122	60,457	315,491	ca. 40	65.1%	84.2 %	1.18

We then focus on the question of how new authors join the network and the role of preferential attachment, initially by replicating for our data the analyses done by (Abbasi, Hossain, & Leydesdorff, 2012) on a data set for the field of 'steel structure' (1999-2008) research and by (Milojevic´ , 2010) on a data set on 'nano science' (2000-2004). Abbasi et al. suggest that betweenness centrality is the driver of preferential attachment in the evolution of research collaboration networks, more so than degree centrality, which traditionally has been the focus of preferential attachment models. They observe for their data set that the betweenness centrality of existing nodes in a co-author network correlates more strongly with the number of new authors they attract and link to than degree centrality. However, we consider the 0.23 to 0.32 Spearman correlation strengths that they report as relatively low, having limited explanatory value for the variation observed in the number of links existing authors form with new authors. Our data show an even lower correlation (between 0.1 and 0.2 for field 1, and 0.05 and 0.15 for field 2), and the correlation values we obtain for degree centrality and betweenness centrality are almost the same. We have conducted an additional analysis focused specifically on hub nodes (i.e. research group leaders) and the number of new authors that link to them. We find that the correlation between centrality of a hub node and number of new authors linking to them vanishes, suggesting that new authors entering the field do not discriminate their attachment to hub nodes by the respective centrality score of the hub node in the network. Hence we cannot corroborate the finding of a dominance of betweenness centrality over degree centrality as driver of the evolution of collaboration networks. Further, the low correlation values contradict claims that preferential attachment based on network centrality plays a major role in explaining attachment dynamics of new authors.

As pointed out by Milojevic power law scaling of the distribution of number of collaborators (associated with a hypothesized preferential attachment mechanism at work), is not the dominant feature characterizing such distributions in co-author networks. Instead, the majority of authors (88% in the 2000-2004 data set of nano-science publications studied by Milojevic) are included in the log-normal hook of the distribution. Milojevic interprets the hook and its peak as suggestive of a characteristic mode of collaboration corresponding to the typical number of collaborators needed in a research field to produce a publishable result. Our data for both fields display the same log-normal hook feature with a peak at 2 collaborators. These peak values are slightly smaller than those in nano-science subfields that could be comparable to our fields. This could be due to differences in preprocessing of the data. We have started analyzing the extent to which these features persist if we consider only specific (Guimera) classes of nodes.

Finally, we are investigating how the network evolves at the level of network components. One phenomenon in our data that caught our attention has been the temporal oscillation of the size of the second largest component. Whenever its size drops, this indicates a 'feeding event' in which its nodes join the giant. We have checked whether the second largest component acts as a major 'staging ground' for nodes to join the giant. Testing this hypothesis we find that after 20 years only 8% (field1) or 2% (field

2) of the nodes in the giant component have at some point in the past been members of the second largest component. The largest annual influx of nodes to the giant component are new nodes entering the network (typically 50-70% for both fields). For field 1, however, we find a potentially interesting 3-year phase of successive mergers of the giant with the second largest component such that in 1999 eventually 28% of all nodes in the giant component have passed at some point through the second largest component.

Conclusions

We suggest that attempts at explaining the dynamics of network growth need to distinguish more carefully between the different types of nodes and social processes underlying co-authorship collaborations. We expect valuable insights into the evolution of collaboration networks in scientific communities and field specific collaboration patterns form an ethnographically grounded and time-sensitive analysis of collaboration networks. We here focus on co-author networks, however in future work we anticipate to include in the temporal analysis of layered citation and co-author networks for the mapping of community structures within scientific fields.

We also note that our experiences with the replication of other authors' results revealed a number of critical issues that underline the potential benefit of an open data approach, allowing routine sharing of the data sets underlying published analyses, for developing a strong reliable empirical base for field comparisons.

References

Abbasi, A., Hossain, L., & Leydesdorff, L. (2012, July). Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks. *Journal of Informetrics*, *6*(3), 403–412. Available from http://dx.doi.org/10.1016/j.joi.2012.01.002

Bettencourt, L. M., Kaiser, D. I., & Kaur, J. (2009, July). Scientific discovery and topological transitions in collaboration networks. *Journal of Informetrics*, *3*(3), 210–221. Available from http://linkinghub.elsevier.com/retrieve/pii/S1751157709000285

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

Milojevic´ , S. (2010, March). Modes of collaboration in modern science: Beyond power laws and preferential attachment. *Journal of the American Society for Information Science and Technology* , *61*(7), 1410–1423. Available from http://doi.wiley.com/10.1002/asi.21331

Rosvall, M., & Bergstrom, C. (2008). Maps of information flow reveal community structure in complex networks. *PNAS*, *105*, 1118.

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

Velden, T., Haque, A., & 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., & Lagoze, C. (2012). Mapping scientific communities to scale-up ethnographies. In *iconference* (p. 563-564).