Mapping the Evolution of Scientific Community Structures in Time

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

Shiyan Yan
School of Information
University of Michigan
105 S. State Street
Ann Arbor, MI 48109, USA
shiyansi@umich.edu

Kan Yu
School of Information
University of Michigan
105 S. State Street
Ann Arbor, MI 48109, USA
kanyu@umich.edu

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

ABSTRACT

The increasing online availability of scholarly corpora promises unprecedented opportunities for visualizing and studying scholarly communities. We seek to leverage this with a mixed-method approach that integrates network analysis of features of the online corpora with ethnographic studies of the communities that produce them. In our development of tools and visualizations we seek to support the going back and forth between views of community structures and the perceptions and research trajectories of individual researchers and research groups. We here present results from tracking the temporal evolution of community structures within a research specialty. We explore how the temporal evolution of these maps can be used to provide insights into the historical evolution of a field as well as extract more accurate snapshots of the community structures at a given point in time. We are currently conducting qualitative interviews with experts in this research specialty to assess the validity of the maps.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*information networks*

Keywords

Scientific Communities, Network Analysis, Maps of Science, Temporal Evolution

 WWW'15 Companion, May 18–22, 2015, Florence, Italy. ACM 978-1-4503-3473-0/15/05. http://dx.doi.org/10.1145/2740908.2742020

.

1. INTRODUCTION

The mapping of scientific communities and their collaboration networks [9] is flourishing for a number of reasons. The now pervasive publication of scholarly results in digital libraries and institutional and disciplinary repositories provides a comprehensive, machine-readable corpus for this analysis. Complementing this is the emergence of sophisticated algorithms for the analysis of complex networks [10] and the wide availability of advanced user-friendly analysis and visualization tools like pajek [2] and gephi [1] for generic networks, or specialized tools for the analysis and visualization of scholarly networks, such as VOSviewer [13], CiteSpace [4] or Rexplorer [11]. This makes it possible to develop science mappings that visualize characteristic features and the evolution of scientific communities. Such mappings are of interest to researchers like ourselves who study those communities, funding agencies who seek to understand outcomes of funding, policy makers who aim to understand the effect of science policy, and given increasingly interdisciplinary research contexts also scientists who seek to orient themselves about research trends and collaboration opportunities in other fields.

In our previous work [18], we generated static community maps based on bibliographic data from 20 years of publications in a field. In this paper, we explore how the temporal evolution of these maps can be used to provide insights into the historical evolution of a field as well as extract more accurate snapshots of the community structures at a given point in time.

2. BACKGROUND

As described in [16] we take a mixed method approach to studying social behaviors in scientific communities that integrates ethnographic field studies with network analytic methods. The ethnographic field studies involve visits of several weeks length to the research sites (laboratories, offices) of research groups to observe their research practices and social behaviors, and to interview group members about their experiences and perceptions. The ethnographies serve several functions. Importantly they allow us to study a phe-

nomenon of interest, e.g. openness and sharing behaviors between research groups, in great detail. They provide nuanced evidence to help interpret behaviors and develop hypothesis about causal relationships. At the same time, they also support the quantitative, network analytic study of scientific communities that complement the in-depth study of individual groups. Insights gained during these observations and interviews of study participants allow us to fine-tune and optimize the delineation of publication data sets to represent research specialties, and to calibrate and correctly interpret network analytic measures of community patterns such as the co-author links between research groups [18]. Our mixed method approach evolves a tradition of close-up analysis of scientific networks and communication practices started by Crane's work [5] on invisible colleges and taken up more recently, by Zuccala [19] and Cambrosio et al.[3].

The work presented in this paper extends our hitherto static approach to the mapping of community structures of research fields [18] by adding time sensitivity to the analysis. Our focus is on the evolution of topics within a research field along with the evolution of collaborative links between research groups, since scientific research specialties are a complex social and cognitive phenomenon [6]. Sociologically, research specialties can be characterized as collective production communities that emerge from the indirectly coordinated activity of autonomous actors (research groups) who aim to contribute to a shared knowledge base [8, 14]. In our tool development we seek to facilitate the going back and forth between aggregate community-level structures and the experiences and research trajectories of individual researchers and research groups.

3. METHOD

The basic entities in our analysis are research groups or research networks as represented by co-author clusters in a co-author network [16], and topic areas, as represented by clusters of documents in a direct citation network [18]. For these entities we then construct two types of visualizations that complement each other in informational content. First, a topic affinity network shows the cognitive affinity between topics based on citation links. Second, we extract the group level network of collaborative links between research groups in the field. An overlay map shows for each research group in this network what topic area it is most active in, thereby providing a visualization of the socio-cognitive community structure within the research specialty. By slicing our data set into three 8-year time periods, we investigate the evolution of these structures over time[1].

3.1 Data

The data used in this study was extracted from bibliographic records of the Thomson Reuters Web of Science retrieved using the advanced webservice API in October 2013 using a lexical query on the title field of publication records. The specific database queried was the Science Citation Index Expanded (SCI) edition of the Web of Science Core Collection. We searched and retrieved publications for the 22-year period from 1991 to 2012. The lexical query was developed during ethnographic field studies between 2007-2009 and optimized using methods described in [18] to capture and delineate 'cluster science', a research specialty in the physical and chemical sciences. We only included records of document type 'article' in the final data set. The author name data was disambiguated using a co-author based approach described in [17]. Before the data is used for the construction of the co-author network and the direct citation network it is cleaned and reduced, e.g. to remove one-time authors, as described in [15]. The resulting data set has 80,760 records and 99,228 unique authors.

[1]This time window is a compromise between time resolution and structural cohesiveness of the network to be analyzed, i.e. allowing sufficient time for relevant connections to be made. From previous work with this data set [15] we know that certain features of interest, such as group internal structures, get expressed in the co-author network only after publications accumulate over a time frame of at least 5 years. We could have chosen this lower bound but opted here for the 8-year time window. Depending on the publication rates of a field, a different time window size will be appropriate, however what criteria to use to chose the optimal time resolution is still an open research question.

3.2 Topic Area Affinity Networks

We derive topic areas from clustering the direct citation network generated from the publications in a given time period (here: 1991-1998, 1998-2005, 2005-2012). We cluster the direct citation network using *infomap*, an information theoretic algorithm that models information flows [12] twice. This way we obtain clusters of document clusters to represent research topics in the field. The distribution of cluster sizes is fairly uneven with a few larger clusters containing the majority of documents. For pragmatic reasons (ease of interpretation of the visualized network) we select the eleven largest clusters for the construction of the topic affinity network. This way we capture, depending on the time window, between 80% and 93% of all publications in the giant component of the direct citation network.

We operationalize the affinity (or antagonism) between topic areas by comparing the total strength of citation links between two areas to a null model that assumes that documents get assigned to topic areas at random with a probability proportional to relative topic area size [18]. The affinity between a source topic area and a target topic area is calculated as follows:

Assume:
A_{11-i}: Largest 11 areas except area i
$N_{p(j)}$: Number of papers in topic area j
C_{ij}: Number of citations from topic area i to topic area j

We define the citation based affinity A between two topic areas i and j as the residual:

$$A_{ij} = \frac{\text{Actual Count}_{ij} - \text{Expected Count}_{ij}}{\sqrt{\text{Expected Count}_{ij}}}$$

where:

$$\text{Actual Count}_{ij} = C_{ij}$$

$$\text{Expected Count}_{ij} = \frac{N_{p(j)}}{\sum_{k \in A_{11-i}} N_{p(k)}} \times (\sum_{k \in A_{11-i}} C_{ik})$$

In the affinity networks, the existence of a link indicates a surplus of connectivity between the two topic areas in question, whereas the absence of a link may either mean normal (random) background connectivity or a negative affinity

value ('antagonism'). We note that affinity as defined here is a relative property. It expresses the relative preference given by documents in one topic area to citing documents in another area given the choice between the ten other topic areas included in the affinity calculation. Even if a seemingly very strong link is present to one of the other topic areas in the affinity network, it is possible that the affinity to document clusters outside the set of topic areas selected for this analysis or even outside of the data set (external citations) is significantly greater than to the ones in the set.

3.3 Group Collaboration Networks

The co-author network is clustered using the *infomap* algorithm [12], the same algorithm we use for clustering the direct citation network. The co-author clusters retrieved constitute the nodes of the group-level collaboration network. Links between these group clusters are filtered to extract only those (more complex and strong links) that indicate actual inter-group collaboration links, filtering out the residuals due to author migration from one to another group or the more temporary links of one-off collaborations (oftentimes service collaborations). We accept links as inter-group collaboration links if and only if the removal of two nodes that are members of either cluster does not eliminate all the links between the two clusters. This algorithm is calibrated by our previous ethnographic field study of groups in this field where we found that the resulting network corresponds very well to the actual inter-group collaborations of our study participants [16].

The group collaboration networks of the 8-year time windows reuse the groups as defined for the accumulative inter-group collaboration network. It merely adapts the group size to the number of authors actually active during the respective time period and the strength of inter-group links to the number of co-author links from articles published during that time period.

3.4 Topic Overlay Maps

The topic overlay maps used in this article are generated by creating a network partition file that indicates for each node (group) in the group collaboration network the topic are that its research is focused on during the given 8-year time frame. We establish research focus by requiring more than 50% of a group's publications to be published in the respective topic area.

3.5 Temporal Dynamics

To obtain insights into the temporal evolution of topics and collaboration links in the research specialty, we use slightly different strategies for research groups and for topics, due to the differing tractability of topics in the much smaller topic affinity network (11 nodes in the topic affinity network) versus research groups in the group collaboration network (629 nodes in the giant component of the collaboration network). For the group collaboration network we delineate the basic entities of our mapping (research groups) by clustering the accumulative 1991-2012 co-author network. Then, we introduce temporal dynamics to the group collaboration network by varying the size of topics and their affinity links, as well as the size of research groups and their collaboration links, according to the publications published within the 8-year window of a given time period.

Depending on the specific purpose of the mapping, a more dynamic approach would be desirable to account for cases where within the 22-year period covered by our data, researchers change group membership and groups split or merge. However, such an approach poses challenges such as how to track the continuity of a research group (e.g. how to establish the identity of a research group across time slices) and how to visualize the resulting trajectories and their interconnections for networks of several hundreds nodes in a user friendly manner. To our knowledge these challenges have not yet been satisfactorily solved [7].

We apply a more dynamic approach to the smaller topic affinity network. Instead of delineating topics based on the accumulative 22-year data set and then merely adjusting node sizes and links by only counting publications in a given 8-year time window, we re-cluster the document citation network for each 8-year time window. We support the tracking of continuity between topics across time slices by generating annotations that indicate at a high-level the content of the topics. We provide two types of annotations to describe and distinguish the subject matter of the topic areas, one derived from the titles of the most popular journals in each topic area, and another derived from specific keywords and their relative frequency in the titles of articles the respective topic area.

4. RESULTS

As a baseline, we show the 'static' topic affinity network that is obtained from accumulative (1991-2012) data set, see Figure 1. It depicts the research specialty as consisting of an almost linear alignment of research topics that can be associated - according to journal titles - with specific (sub)disciplines in physics, chemistry and materials science. It stretches from cluster physics, through surface science, materials chemistry to cluster chemistry. The map shows a weak link between the extreme ends of this alignment, between topic area 1 and topic area 6 (a new development not yet visible in an earlier 1991-2010 version of this map that is provided in [18]) that only weakly foreshadows one of our main findings from the temporally resolved maps discussed below.

The temporal evolution of the topic affinity network depicted in Figure 2 provides further insights into the underlying dynamics. These networks capture changes in the prominence, alignment and interrelatedness of research topics in the field. Note that for each time slice, topic areas are strictly numbered by relative size, ranging from A1 (largest), to A11 (smallest). We observe the following developments::

1. The emergence of a separate topic area focused on the interaction of clusters with radiation from advanced light sources such as synchrotrons, x-ray free electron lasers or femto and atto second lasers.

   It first appears in the 1998-2005 map as fifth largest topic area (A5) and remains visible in the 2005-2012 map as a distinct topic area (A6) whereas in the earlier map (1991-1998) this kind of work was inseparable from the large 'cluster physics' topic area (A1).

2. The growth of the surface science orientation within cluster science and its strengthening link with materials chemistry. In the 1998-2005 time window the strengthening of the affinity link between the two areas

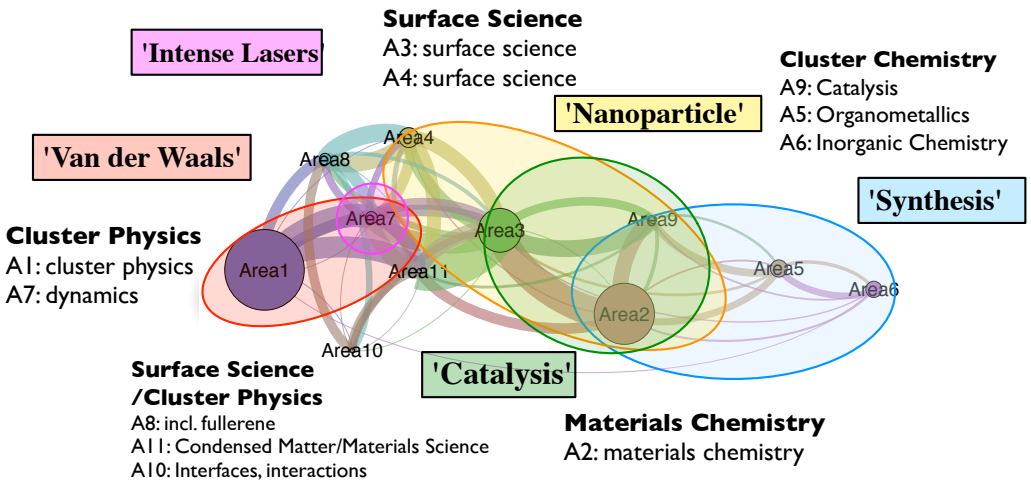

Figure 1: 'Static' view on the topic affinity network. This network is based on the accumulative 1991-2012 publication data. Nodes represent the 11 largest topic areas that were extracted from the direct citation network. Directed links represent topic area affinity based on surplus of direct citations relative to a random null model. Link color indicates source node of the link. Annotations indicate the substantive matter of the topic areas based on the disciplinary orientation of the most frequent journal titles in each topic area and on distinctive keywords in the article titles in each topic area (boxed terms).

A2 and A3 is very visible; In the time window 2005-2012 surface science has become the dominating topic area among the eleven largest topic ares within the field (A1), and the materials chemistry areas A5 and A7 have tight affinity links with it.

3. The emergence and strengthening of a link between the extreme ends of the cluster science topic area alignment such that the topology is no longer long stretched but a closed circle.The dynamic topic clustering shows for the 2005-2012 time window the emergence of A3, a large (3rd in size) hybrid topic area that combines (Inorganic) Cluster Chemistry and Cluster Physics and connects Cluster Physics with Materials Chemistry without the Surface Science intermediary. This is amplified by a corresponding dense clustering of groups working in A3 in the group collaboration network in Figure 2.

Of particular interest to us, as we design studies of field differences in behavioral patterns among and within scientific communities, is how these changes in the cognitive structures in the field project onto the social network of inter-group collaborations and what we may learn about important actors in the field and how topic focus and collaborative interconnectivity interact.

As shown on the right hand side of Figure 2 an overall feature of the group level collaboration network structure through all three time periods is its division into two parts. This division seems to reflect the disciplinary orientation of groups towards either chemistry or physics. Initially the separation is evidenced by the network having two unconnected large network components. In the later two time periods, an interconnected giant component has formed, however it still exposes a structural subdivision into two parts. This division seems to be slowly diminished by growing collaborative connections between the two parts of the network.

In addition, the topic affiliation of some of the groups seems to transcend the boundary, that is groups in certain topic areas can be found on both sides of the collaboration network. In the early time window (1991-1998) groups focused on research in surface science topic areas (A3, A6 and A11) appear in both network components, thereby transcending this division. In the second time window (1998-2005) groups contributing mainly to the materials chemistry and cluster chemistry topic areas A2 and A3 are visible in both parts of the network. Finally, in the most recent period (2005-2012) there seems to be some intermixing by groups mainly active in A4 (cluster physics/surface science).

Groups that connect major parts of the network tend to be white in this set of overlay maps indicating their publication output is shared between several topic areas. One example is the large white node in the middle of the left component of the 1991-1998 group collaboration network in Figure 2. It connects three major branches in this early collaborative network and seems to be well positioned to provide insights into the evolution of the field from the specific sub-disciplinary perspective represented by this group. We can trace this group's activity through the entire time period covered by our data. It follows a trajectory from A5 (also A4 and A9) in 1991-1998, and A9 (also A11 and A3) in 1998-2005, to A11 (also A1) in 2005-2012. Its major affiliation at the level of accumulative data is A5 (+ 2 and 9), reflecting a sub-disciplinary orientation towards organometallics and inorganic chemistry, a specialization that is confirmed by the professional website of one of its lead authors.

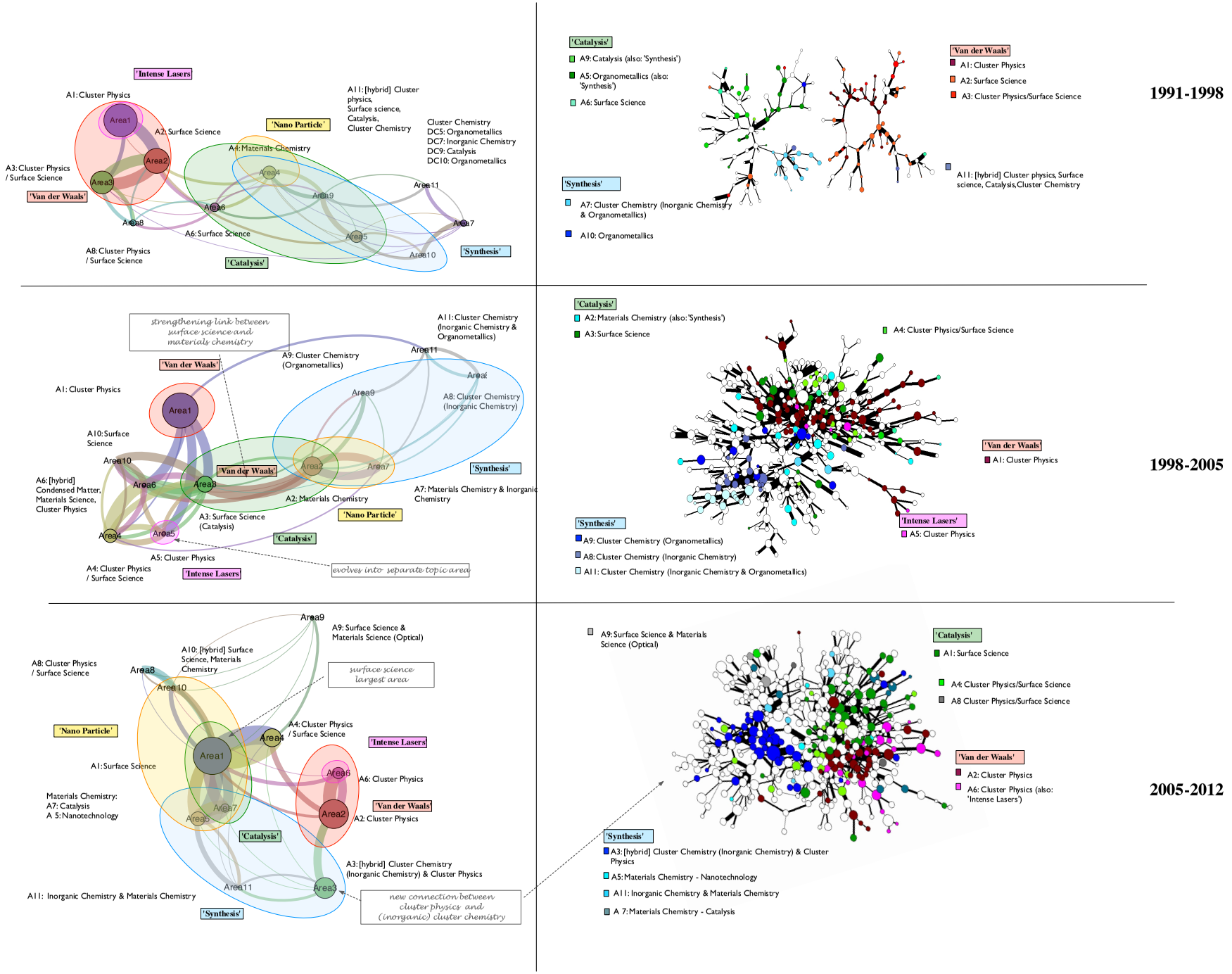

Figure 2: Left: Evolution of topic area affinity network. The grey text boxes highlight new developments in the field that are described in the text. Right: Overlay of topic area activity on the giant component of the group collaboration network. Colors indicate the topic area that a group is focused on (i.e. it publishes more than 50% of its publications in that topic area). No color (white) indicates groups that either distribute activity more evenly across several topic areas or that focus on a (smaller) topic area not included in the network.

5. DISCUSSION

The picture that emerges is one in which, on the one hand, (sub-)disciplinary orientations of the groups that publish in the research specialty remain a rather stable feature. A group that is specialized in one sub-disciplinary area rarely changes its disciplinary orientation entirely. This is in agreement with our ethnographic field studies of research specialties in the physical and chemical sciences. We witnessed occasional strategic hires into a group to extend its local skill set, e.g. to add a synthetic chemistry capability into a largely experimental physics group. However, the foundation of the group remains the sub-disciplinary training of its leader that provides continuity to the research trajectory of the group. Further, for some topic areas, large parts of the inter-group collaboration network are constituted by collaboration links within that topic area. Therefore a significant part of the collaborative work in the field seems to not transcend topic areas that generally correlate with sub-disciplinary orientations.

That said, there are also clear 'connectors', groups that work across two or more areas. Due to their activity we see an overall integration of the collaboration network. Also, there are those (fewer) topic areas that are hybrid in their (sub)disciplinary orientations such as A 3 in the 2005-2012 time window. These topic areas and the research groups contributing to them may be of particular interest for the study of interdisciplinary collaboration and exchange in the field.

6. CONCLUSIONS

We have generated time resolved views into the co-evolution of topics and collaborative links between research groups in a research specialty. We have developed an interpretation of these time resolved maps regarding trends in the field. Further, the maps provide a valuable resource that suggests individuals and groups that would be of particular interest to include in future qualitative studies that aim to develop a deeper understanding of the community structures within this field and their evolution over time. A next step in our research will be to validate these maps in interviews with researchers in the field.

7. 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.

8. REFERENCES

[1] M. Bastian, S. Heymann, M. Jacomy, et al. Gephi: an open source software for exploring and manipulating networks. *ICWSM*, 8:361–362, 2009.

[2] V. Batagelj and A. Mrvar. Analysis and visualization of large networks. In *Graph Drawing Software*, pages 77–103. Springer, Berlin, 2003.

[3] A. Cambrosio, P. Keating, and A. Mogoutov. Mapping collaborative work and innovation in biomedicine a computer-assisted analysis of antibody reagent workshops. *Social Studies of Science*, 34(3):325–364, 2004.

[4] C. Chen. Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3):359–377, 2006.

[5] D. Crane. *Invisible Colleges - Diffusion of Knowledge in Scientific Communities*. The University of Chicago Press, 1972.

[6] Y. Ding. Community detection: Topological vs. topical. *Journal of Informetrics*, 5(4):498–514, 2011.

[7] M. Giatsoglou and A. Vakali. Capturing social data evolution using graph clustering. *Internet Computing, IEEE*, 17(1):74–79, 2013.

[8] J. Gläser. *Wissenschaftliche Produktionsgemeinschaften - die soziale Ordnung der Forschung*, volume 906 of *Campus Forschung*. Campus Verlag, Frankfurt / New York, 2006.

[9] S. Morris and B. Van der Veer Martens. Mapping research specialties. *Annual review of information science and technology*, 42(1):213–295, 2008.

[10] M. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[11] F. Osborne, E. Motta, and P. Mulholland. Exploring scholarly data with rexplore. In *The Semantic Web–ISWC 2013*, pages 460–477. Springer, 2013.

[12] M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[13] N. J. Van Eck and L. Waltman. Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2):523–538, 2010.

[14] T. Velden. 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, 2013.

[15] T. Velden, S. Cambo, S. Ahmed, and C. Lagoze. Toward a time-sensitive mesoscopic analysis of co-author networks: A case study of two research specialties. In *ISSI 2013, 15-19 July, Vienna, Austria*, 2013.

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

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

[18] T. Velden and C. Lagoze. 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, 2013.

[19] A. Zuccala. Modeling the invisible college. *Journal of the American Society for Information Science and Technology*, 57(2):152 – 168, 2006.