Patterns of Collaboration in Co-authorship Networks in Chemistry - Mesoscopic Analysis and Interpretation

Theresa Velden, and Carl Lagoze

tav6, cjl2 @cornell.edu
Information Science, Cornell University

Abstract
We are investigating the manner in which different scientific fields differ in their communication cultures. Our research methodology combines quantitative (graph-theoretic) and qualitative (ethnographic) techniques, which mutually inform each other. In this paper, we present the results of a case study on the identification and classification of scientific collaborations based on structural patterns in co-authorship networks. We investigate three co-authorship networks in specialised subfields of organic chemistry and physical chemistry using an information-theoretic clustering algorithm by Rosvall et al. to extract the modular structure of the networks and thereby investigate the mesoscopic structures of these networks. The clusters found mostly correspond to hierarchically organised research groups or multi-centre research networks. Based on the co-author links between these clusters we analyse group and network interactions, using interviews with participants from the fields in question to inform our interpretation. Structurally one can distinguish two broad classes of linking patterns between clusters: 'exclusive' connections where a single node (author) connects two clusters, and 'many-to-many' connections where substantial fractions of the nodes of both clusters are involved. Within these broad classes one can find further subclasses that correspond to typical collaboration and mobility events.

Introduction

Scientific fields differ in their intellectual and social organisation, and consequently in their communication practices (Whitley 2000, Cetina 1999, Fry & Talja 2007). Investigating these differences via ethnographic studies alone is insufficient because evidence is gathered from only a small, local fraction of scientists in a field. At the opposite end of the scale, large sets of publication data promise access to aggregate behavioural patterns of a research community. However, because of the standardization of formal scientific publishing across science, a bibliometric approach alone may fail to uncover underlying, field specific differences in scientific communication practices. In our comparative research into field-specific scientific communication cultures we combine qualitative ethnographic field studies with structural analysis of publication networks. In this manner, we can use small-scale, nuanced evidence from field studies to inform our interpretation of large-scale publication networks and investigate to what extent field specific practices are reflected in structural features of publication networks. This evolves a tradition of close-up analysis of scientific networks started by Crane's work (1972) on invisible colleges, which was taken up more recently by Zuccala (2004).

Our specific concern is to understand the relevance and meaning of certain network artefacts (such as co-author links and inter-citations) in the context of specific scientific fields, and to discover how they relate to concepts of community, and to communication practices that we observe in our field studies. The field specificity of global bibliometric indicators such as citations and journal impact factors (Moed 1985, Althouse 2008), the number of co-authors on a paper (Liberman & Wolf 1998) or percentage of self-citations (Snyder & Bonzi 1998) is well known, although the delineation of scientific fields, and hence the appropriate normalisation of such measures for comparisons across scientific fields remains problematic (Zitt 2005, Adams 2008). Recently, the characteristics of co-author networks as social networks have been highlighted (Kretschmer 1994, Newman 2001). With other social networks they share global topological properties such as small world-property, clustering,

and assortative degree mixing as well as a long-tail degree distribution and a scaling law of clustering coefficient (Sen 2006). In the recent surge of work on the analysis of complex networks, and in particular on clustering algorithms to extract the modular structure of real world networks, co-author networks rank among the most prominent examples (e.g. Radicchi et al 2004, Newman 2004, Palla 2005). Most work in this area does not focus on analysis of network features specific to a scientific field. Instead, existing work uses co-author networks as a way to demonstrate algorithmic advances by comparing global network properties of co-authorship networks to other types of 'real world' networks (such as transportation or biological networks).

Those works in information science and bibliometrics that investigate the characteristics of co-author networks in specific fields from a social network perspective tend to focus on global topological network properties, or on the position and ranking of individual actors within the network, see e.g. Kretschmer (2004), or Acedo (2006), and Liu (2005) for an exception that at least touches on the group level organization. However, as Guimera et al. (2007) argue, for modular networks global properties fail to capture important structural and functional distinctions between networks. To discover those one has to investigate the mesoscopic structure of networks that takes into account module interconnectivity and differences in connectivity patterns between nodes based on their structural position in the network. Accordingly the study presented in this paper focuses on the *mesoscopic level of analysis* - that is we analyze connectivity patterns between modules of closely interconnected authors in co-authorship networks, in order to explore the field-specificity of community structures and communication patterns.

Early on in our research we found interpretation of co-authorlinks between co-author groups as indicators of collaboration between those groups to be problematic: we realised that in the weakly interconnected field that we were studying, co-authorship links between groups did not indicate direct inter-group collaboration but were just residues of the fact that individuals migrated between groups on their career path, e.g. from phd student to postdoc, an observation made also by Nepusz (2008). Hence, to understand the different types of connectedness that co-author links between groups represent, we need to classify such linking patterns, and to match them with different types of collaboration or exchange. In this paper we present a qualitative validation of this approach based on visualization of graph structures and interpretation of various network metrics.

In our research, qualitative and quantitative approaches are closely interlinked. The study presented here is exemplary for the way in which qualitative understanding of research practices and context help us to evaluate quantitative outcomes, and how quantitative results guide our attention for further qualitative study. Take for example the issue of analyzing the community structure of social networks that is an important aspect of our research into communication practices. As described further below, a wide variety of clustering algorithms exist that will all deliver quite different partitions of a network. To decide whether clustering provides us with a research unit that is useful for the analysis of scientific communication and interaction patterns of a research field, we need validation and interpretation of those clusters in the context of this specific field (Caruana 2006, Schaeffer 2007). Given such a clustering we can then look at the network of clusters that shows the interactions between author groups and direct our qualitative research efforts at further understanding those interaction patterns – that is to uncover those motivations and processes that underlie the structural patterns (Lievrouw 1990, Zuccala 2004).

Methodology

To develop a deep qualitative understanding of communication practices in different fields of chemistry we are conducting ethnographic field studies in selected research groups. We will call these groups *seed groups* in the remainder of this paper[i]. A precondition for selecting such a group for our study is that it is a relevant recognized player in some specialised field of research. The publication data sets used in this study represent those research specialties in which our seed groups are actively engaged. As a result, for each data set we have access to informants with whom we can check the interpretation and significance of features that we detect in our network analysis.

Data

We have obtained our data sets from Web of Science (WoS, Thomson Reuters) using a lexical query (see appendix) to capture the literature published in small specialities within organic chemistry (Field B) and physical chemistry (Field A and C). For the construction of the co-author networks we have omitted those (transient) authors that have published only a single paper within the data set. The original number of authors and the reduced number of authors in the co-author networks are given in table 1.

Table 1. Data Sets

	Field A	**Field B**	**Field C**
Sub-discipline	Physical chemistry	Organic chemistry	Physical chemistry
Time span	1991-2008	1991-2008	1987-2008
Average # of authors per paper (median)	4.0 (4)	3.3 (3)	3.8 (3)
# authors	7,823	20,683	47,471
# authors (reduced)	2,241	7,472	18,440
# authors in giant (relative size)	2,014 (89.9%)	6,660 (89.1%)	17,269 (93.6%)

We note that the average number of authors on a paper is lower in organic chemistry than in the two data sets from physical chemistry.

All three networks have a giant component that includes about 90% of the co-authors in the network[ii]. Below we show a temporal analysis of the evolution of author numbers in the data set and the growth of the relative size of the giant component by slicing the data set into the following three chunks of increasing size: 1991-1996, 1991-2002, 1991-2008 (Fields A,B), and 1987-1993, 1987-2000, 1987-2008 (Field C). Figure 1 shows that the largest growth of the relative size of the giant component happens when moving from 6 years (Fields A, B), respectively 7 years (Field C) to 12 years (Fields A, B), respectively 14 years (Field C). As the absolute number of authors in the network doubles or even triples in the last time step, the relative size of the giant component seems to reach saturation (clearly for Field C, presumably also for Fields A,B), indicating that the time window is sufficient for all basic relations between actors to have developed (Newman 2001a, Barabasi 2002).

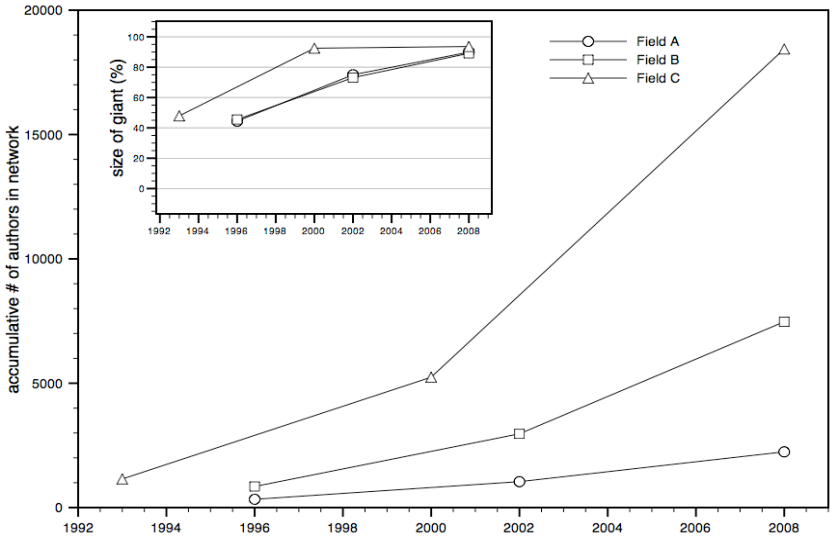

Figure 1. Accumulative temporal evolution of co-author numbers in network (large diagram) and of relative size of giant component (inset).

Procedure

The nature of the node 'communities' that are extracted from a network depends on the clustering algorithm that is used. Some clustering procedures are hierarchical by design, so that they offer a hierarchy of groupings, starting from single nodes to the total set. Newman (2004) shows how his algorithm that is not hierarchical by design can still be used to repeatedly drill down into a large, 56,000 author data set, first detecting clusters at the level of entire fields (such as high-energy physics, or condensed matter physics), then smaller specialities, and eventually after a fourth iteration, clusters corresponding to research groups (such as his own with 28 members). Other algorithms have a very restrictive definition of clusters as very tight groupings of nodes and therefore produce clusters with little variation in group internal structures, e.g. the concept of k-cliques in Palla et al (2005). See Fortunato & Castello (2007) for an informative overview on recent clustering algorithms for the extraction of community structures from complex networks and the issue of implied community definitions.

Given the co-author networks as described above, we use an information-theoretic clustering algorithm[iii] for undirected, integer valued networks (Rosvall & Bergstrom 2007) to partition the co-author network into clusters of closely interconnected co-authors. We have chosen this algorithm because we have found that it extracts very well what Seglen (2000) has called '*functional research groups*' – that is, those research performing units that not only contain a collocated group of researchers in a laboratory led by a principal investigator (PI), but also closely cooperating domestic or international colleagues and visiting scientists. We represent the results of this clustering as an *author-level network* where nodes represent individual co-authors and where the cluster membership of nodes is indicated by the colouring of nodes. We will focus here on the study the fine structure of group-to-group co-authorship links.

For further analysis we extract the giant component of the co-author network. We use pajek (Batagelj & Mrvar 2003) to load the network, to partition it into strongly connected components, and to extract the largest partition, i.e. the giant component. Then, for each field,

our analysis proceeds as follows:

1. We use pajek to extract the cluster that includes the PIs of our seed groups. Using standard network centralization indices (degree, closeness, betweenness) calculated with pajek, we compare the internal structure of selected clusters. We validate the interpretation of these clusters with our informants.
2. For analysis of the linking patterns between clusters we extract the neighbourhoods of the seed group from the original giant component – that is those clusters that are linked by at least one co-author link to the cluster of the seed group. By visualizing each neighbour cluster and its connection pattern with the seed cluster we then characterize its visually most obvious features.
3. We ask our informants to scan through the list of neighbours (highlighting the most productive authors in each cluster) and to tell us about their scientific relationship to the neighbouring cluster and how the co-author links can be explained. Typical scenarios and structures emerged that we describe in the next section.

Results

Interpretation of Clusters

We validated with the PIs of the three seed groups that the cluster extracted from the co-author network captures their immediate, collocated research groups (researchers of various seniority levels from PhD students over postdocs to their own deputies or subgroup leaders) plus relevant external cooperation partners (from their own institution, or national and international cooperations). The PIs were slightly surprised that some individuals were subsumed into their cluster and not independently represented. This particularly seems to be the case for those individuals that have a strong institutional identity distinct from the academic group of the PI – as is the case for research group leaders from industrial companies. Note that our data sets capture only publication activity in very specialised areas of research – most PIs and their groups in our field study engage in several, non-related research areas. Consequently the representation of research groups by co-author clusters in this study is only a partial, as it includes only those functional group members involved in the specific area of research targeted by our data set.

The seed groups of the PI-led research groups all show a centralized hierarchical structure with the most productive author in the centre (figure 2). They differ considerably in size, but each seed lab cluster has a size well above the average size of clusters extracted from the giant component (see table 3 below). The degree, closeness, and betweenness centralization indices of the seed group clusters indicate a high level of centralization (see table 2). The seed lab of field B scores highest, confirming the visual impression of a distinct star-like structure. Possible explanations are the lower average number of co-authors on papers in this field that reduces cross-links between co-workers of the PI, as well as lack of a subgroup structure in the organization of the group.

We further find that some of the clusters show a quite different internal structure, as can be seen from a comparison of figures 2 and 3. In figure 3 we show two exemplary clusters from field C. They include several central nodes none of which clearly dominates the network. The clusters are still hierarchically organized in the sense that some nodes are clearly larger than most other nodes, and that those smaller nodes are mainly linked to the large nodes. Their centralization indices are substantially lower than those for the seed group clusters, confirming this visual impression. From investigation of institutional affiliations given in the underlying WoS records and participant feedback we identify (as indicated in the figure

caption) one of these clusters as an international collaboration network, and the other as a network of closely cooperating colleagues from a major research institution in this specialty area.

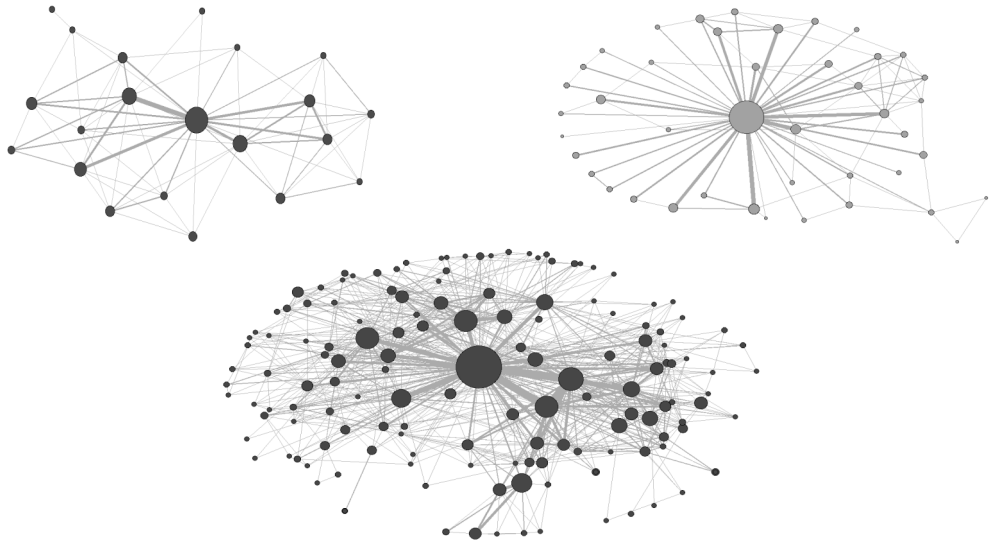

Figure 2. Seed group clusters with centralized internal structure (top left: field A, top right: field B, bottom: field C). Nodes represent authors; node sizes the number of papers they have authored. Links represent co-author links, thickness of links the number of co-author events.

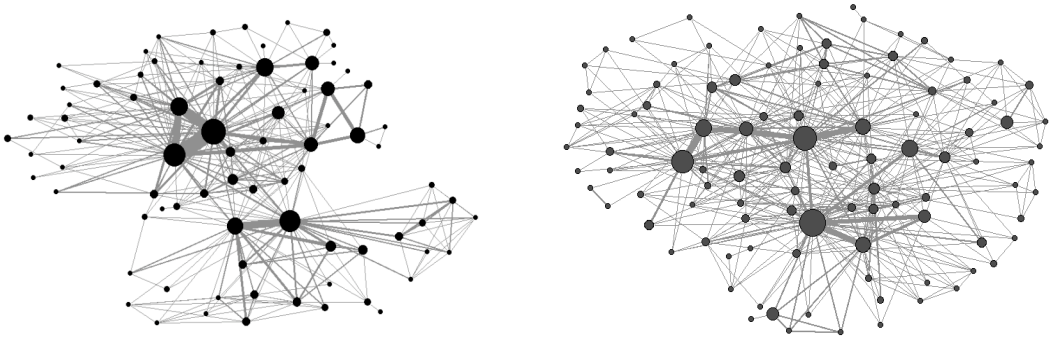

Figure 3. Examples from field C for clusters that correspond to multi-centred networks. Left: international research network (predominantly Japanese, Korean, and French) (C_1), right: cooperative network of a research Institute (C_2). Nodes represent authors; node sizes the number of papers they have authored. Links represent co-author links, the thickness of links the number of co-author events.

Table 2. Centralization indices of selected clusters

	Seed group A	Seed group B	Seed group C	Cluster C_1	Cluster C_2
degree	0.67	0.88	0.79	0.41	0.44
closeness	0.77	0.91	0.80	0.45	0.50
betweenness	0.43	0.86	0.79	0.21	0.31

Fine structure of cluster-to-cluster linking patterns

Next we extract the neighbourhood networks of each of the seed labs; that is, those clusters that have one or several co-authorship links between any of their authors and any of the authors of the seed group cluster. The number of clusters in the neighbourhood networks of the seed labs in field A and B are comparable in size, but the neighbourhood network of the seed group in field C with 315 neighbour clusters stands out in comparison – the total number of authors in this neighbourhood (including the seed cluster) corresponds to almost 30% of the entire author population of the giant component (see table 3).

Table 3. Clusters and Neighbourhood Networks

	Field A	**Field B**	**Field C**
Sub-discipline	Physical chemistry	Organic chemistry	Physical chemistry
Time span	1991-2008	1991-2008	1987-2008
# clusters in giant	173	573	1,190
Average # of authors in cluster (in giant)	11.6	11.6	14.5
# of authors in seed group cluster	21	43	159
# clusters in seed group neighbourhood	14	18	315
# of authors in neighbourhood (percentage of total authors in giant)	296 (14.7%)	590 (8.9%)	4,929 (28.5%)

On visual inspection of the linking patterns between the seed clusters to their neighbour clusters we can identify two major types of structural patterns: *'exclusive'* connections where a single author who is either member of the neighbour cluster or the seed cluster connects two clusters, and *'many-to-many'* connections where substantial fractions of the nodes of both clusters are involved. Whereas the seed group in field A has only 'exclusive' type connections to its neighbours, the seed labs in fields B and C show both connection types. The observation for seed group A matches well with the fact that its PI bemoans a lack of inter-group collaboration in this specialty. He observes that he and his colleagues focus on building independent research profiles in this area and avoid formal cooperation.

Figure 4 shows a variety of 'exclusive' and 'many-to-many' type connections between the seed group cluster in field C and a subset of its neighbour clusters. We distinguish two structural subtypes of exclusive connection types: 1-1 connections and 1-to-many. We also distinguish two structural subtypes of 'many-to-many' connections: those connection patterns that include direct PI-PI co-author links and seem to have a rather deep penetration into the cluster, and those that involve only interaction with a subgroup of the seed cluster but not (or only minimally) the PI. We match these connection patterns to real world scenarios from the accounts of our informants about these relationships in table 4. Numbers in parentheses in the table identify the corresponding instances in the networks in figure 4.

Table 4. Linking patterns

Pattern Type	**Definition**	**Underlying Real World Scenarios** (Examples from seed group field C)
Exclusive 1-to-many	One single node connects two clusters	Visiting scientist with links home into national research network (1) career migration (2) repeated instances of career migration (2b) career migration of now closely collaborating colleague (2*) 1-off commissioned work (3) 1-off cooperations on independent topics (3b) funded project collaboration of subgroup leader (7)
Exclusive 1-to-1 "bridge"	Two nodes, one from each cluster, are exclusively linked	'unauthorized' collaboration by postdocs (4) provision of synthesis samples (5) exclusive cooperation by closely collaborating colleague (6)
Many-to-many	interlinked node groups from each cluster participate in the connection	intensive thematic and methodological international (Swiss) cooperation for since PI's postdoc time (8) very broad, thematic based, many faceted international (UK) cooperation (career migration, temporal exchange of postdocs and PhD students) (9) many faceted methodological international (Dutch) cooperation (10) cooperation on many topics due to complementary knowledge, joint PhD students and exchange of staff, many funded projects (11) PI colleague at same institute - intensive topical and methodological cooperation; much stronger informal cooperation than formal in form of publications (12) multi faceted cooperation with national institute, exchange of senior staff who bring along their networks (13)
Many-to-many (peripheral)	interlinked node groups from each cluster participate in the connection with no or minimal PI-PI links	Thematic cooperation, nurtured by EC funding, no exchange of PhD students; people come for measurements and leave again; disciplinary very distant (electrical engineers) (14) Seed group conducts measurement services to this group, only subgroup of instrument involved, institutionally supported by partner agreement between institutions, PI former postdoc in seed group (15) Russian institute, seed group is partially fused with this institute as several members are partially funded by seed group, repeated exchange of coworkers (16)

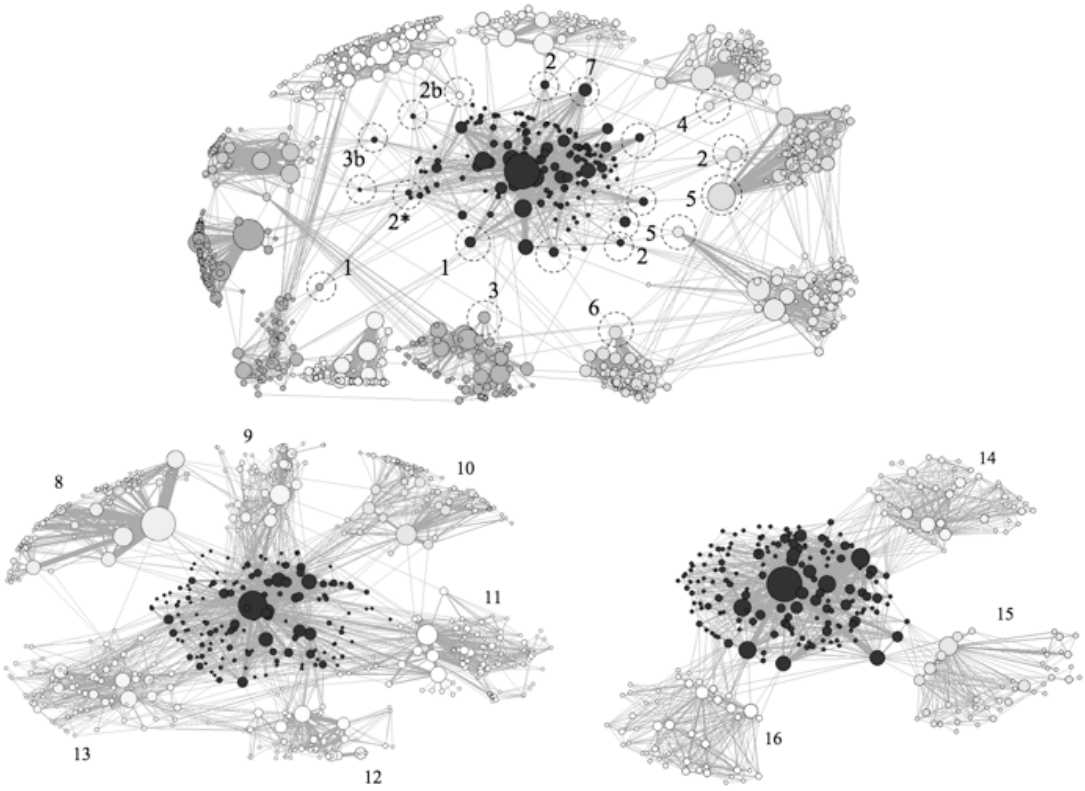

Figure 4. Selected instances of how the seed group (field C) connects to neighbouring clusters. A node represents an author, node size number of publications, node shade cluster membership, a link co-authorship, and link thickness number of joint publications. Top: exclusive type connections (dotted circles highlight involved nodes). Bottom: many-to-many type connections; left: with direct PI-to-PI links, right: without or with weak PI-PI link. The numbers given refer to the description of the underlying collaboration scenario derived from interviews (see table 4).

Discussion

We match topological co-author network features at the mesoscopic level of analysis to informants' accounts, and thereby deliver an interpretation of these features and a validation of their significance.

Specifically we observe that the basic research performing units in the specialties we are studying differ substantially in structure and composition. First, our seed labs and several of their neighbours expose the uni-centred hierarchical footprint of a PI-led group with its periphery (corresponding well to the functional research groups Seglen (2000) defined for his case study on microbiology). We speculate that differences between the physical chemistry and organic chemistry seed labs in the network centralization indices given in table 2 may point to sub-discipline specific differences. They seem to correlate with the field specific average number of co-authors per paper, but other factors, such as the organization of a group into subgroups, may also contribute. Secondly, we also find in all three fields multi-centred research clusters characterized by lower centralization indices. They correspond to multiple-PI collaborations representing intra-institute, national, or international research networks.

Our analysis of structural inter-cluster connection patterns and how they match to underlying real world scenarios (table 4) points to a fundamental difference between 'exclusive' 1-1 or 1-many co-authorship links, and 'many-to-many' co-authorship links. These basic types imply

different forms of interaction and collaboration. The former are footprints of people moving between labs and of one-off cooperations involving sample exchange or measurement services[iv]. According to our informants' accounts, migration events can take three forms: 1) reciprocal exchange between groups, 2) strategic training where a young researcher returns to his home lab after a postdoc stint at another lab, or 3) a non-symmetric career migration flow that reflects differences in reputation and resources of groups or research networks. Hence exclusive type inter-group connections constitute a group-level network of exchange of people, skill, material, and measurement services within a specialty. The many-to-many connection patterns on the other hand correspond to substantial inter-group collaborations. By distinguishing between those two basic structural types we will be able to investigate and compare collaboration practices in scientific fields in two different dimensions.

We expect that a key ingredient for distinguishing between collaboration scenarios beyond the two basic structural types is to identify node roles. PIs, sub-group leaders, collaboration partners in the periphery of a group, and migrating PhD students differ in their representation within the network if one takes node size, their structural position within the group, the strength of their binding to the group, and the distribution of their external links over all neighbour clusters into account. Further, the collaboration scenarios differ in terms of the kind of people involved. Hence it should be possible to infer the different roles authors have from their structural position in the network, and to use this information to further refine the identification of collaboration scenarios. Guimera et al. (2007) have suggested for complex modular networks a distinction between seven node roles. They distinguish between hubs with many module internal links, and non-hubs with few module internal links, and further subtypes of hubs and non-hubs according to the distribution of their links to module-internal and module-external nodes. Given such a categorization of nodes based on their structural characteristics, we can not only calculate and compare the distribution of node roles between our co-author networks, but also visualize their role specific connectivity patterns, and investigate the correspondence of those role specific patterns to collaboration scenarios.

Finally, we point to a number of limitations of this study: we develop interpretations of structural features only from the context of a single research group in each field (the seed group). To avoid distortions, we plan to extend our qualitative field studies to include other groups in each field to check the interpretations we develop against other participants' views. Secondly, our analytic method is getting calibrated in the research environment of certain fields in chemistry and physics. We believe that conceptually it can be transferred to other fields and disciplines as well. But it may have to be significantly altered to be effective, since in other research contexts different data and interaction forms will be relevant (e.g. in computer science the predominant role of conferences, and secondary role of journal publications). Finally, we are capturing here exclusively collaboration forms that manifest themselves in co-authorship - other forms of informal collaboration are not captured, although informants underline their importance to their work. There are additional traces of interaction and influence, such as inter-citation, or workshop participation, to complement the picture, but there will remain important forms of informal exchange that are not documented.

Conclusions

Through combination of qualitative and quantitative methods we have arrived at fundamental insights about collaboration patterns in scientific specialties in chemistry. A mesoscopic analysis of linking patterns in co-authorship networks has pointed us to different structural types of research performing units as well as different types of collaborative relations. Our field study results allow us to match these structural features with real world scenarios that

clarify the meaning and validate the relevance of those structural differences. The next step will be to refine the analysis of linking patterns by taking node roles and node role specific connectivity into account, and to develop an algorithmic method that delivers a quantification of the distribution of collaboration types for systematic comparison between specialties.

Acknowledgments

We are indebted to our field study participants. Further, this research has been made possible through financial support by the National Science Foundation through grants IIS-738543 SGER: Advancing the State of eChemistry and DUE-0840744 NSDL Technical Network Services: A Cyberinfrastructure Platform for STEM Education. Support also came from Microsoft Corporation for the project ORE-based eChemistry.

References

Adams, J.; Gurney, K. & Jackson, L. (2008), Calibrating the zoom - a test of Zitt's hypothesis, *Scientometrics* 75(1):81-95.

Althouse, B.; West, J.; Bergstrom, T. & Bergstrom, C. (2008), Differences in Impact Factor Across Fields and Over Time. *Arxiv* preprint arXiv:0804.3116.

Barabasi, A, Jeong, H., Neda, Z. et al. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311 (3-4): 590–614.

Batagelj, V. & Mrvar., A. (2003). Analysis and visualization of large networks. *In Graph Drawing Software*, (pp. 77–103). Berlin: Springer.

Caruana, R., Elhawary, M., Nguyen, N. & Smith, C. (2006), Meta Clustering, in '. In *Proceedings of the Sixth International Conference on Data Mining* (ICDM'06)'.

Fortunato, S. & Castellano, C. (2007). Community structure in graphs. arxiv: 0712.271.

Fry, J. & Talja, S. (2007), 'The intellectual and social organization of academic fields and the shaping of digital resources', *Journal of Information Science* 33(2):115.

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.

Knorr Cetina, K. (1999), *Epistemic Cultures*. Havard University Press.

Kretschmer, H. (1994), Coauthorship networks of invisible colleges and institutionalized communities. *Scientometrics* 30(1):363-369.

Kretschmer, H. (2004), Author productivity and geodesic distance in bibliographic co-authorship networks, and visibility on the Web. *Scientometrics*, 60(3):409–420.

Liberman, S. & Wolf, K. B. (1998), Bonding number in scientific disciplines, *Social Networks* 20(3): 239-246.

Liu, X. & Bollen, J. & Nelson, M. & Van de Sompel, H. (2005), 'Co-authorship networks in the digital library research community', *Information Processing and Management* 41(6):1462-1480.

Lievrouw, L. A. (1990), Reconceiling Structure and Process in The Study of Scholarly Communication. In C.L. Borgman,(Ed.) *Scholarly Communication and Bibliometrics*, London: Sage Publications.

Liu, X. & Bollen, J. & Nelson, M. & Van de Sompel, H. (2005), 'Co-authorship networks in the digital library research community', *Information Processing and Management* 41(6), 1462--1480.

Moed, H. & Burger, W. & Frankfort, J. & Van Raan, A. (1985). The application of bibliometric indicators: Important field-and time-dependent factors to be considered. *Scientometrics* 8(3):177.

Palla, G., Derenyi, I., Farkas, I. & al.(2005) Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–8.

Nepusz, T.; Petróczi, A.; Négyessy, L. & Bazsó, F. (2008), 'Fuzzy communities and the concept of bridgeness in complex networks', *Physical Review E* 77(1):16107.

Newman, M. (2001). The structure of scientific collaboration networks. *PNAS*, 98(2):404-409.

Newman, M. (2001a). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, 64(1):16131.

Newman, M. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E*, 69: 066133.

Radicchi, F. et al. (2004) Defining and identifying communities in networks. *PNAS*, 101(9):2658–2663.
Rosvall, M. & Bergstrom, C. (2007). An information-theoretic framework for resolving community structure in complex networks. *PNAS*, 104(18):7327.
Schaeffer, S. (2007), 'Graph clustering', *Computer Science Review* 1(1):27-64.
Seglen, P. & Aksnes, D. (2000), Scientific Productivity and Group Size: A Bibliometric Analysis of Norwegian Microbiological Research, *Scientometrics* 49(1):125-143.
Sen, P. (2006). Complexities of social networks: A Physicist's perspective. *Arxiv preprint physic*s, 0605072.
Snyder, H. & Bonzi, S. (1998), 'Patterns of self-citation across disciplines (1980-1989)', *Journal of Information Science* 24(6):431.
Whitley, R. (2000), *The intellectual and social organization of the sciences*, Clarendon Press.
Zitt, M.; Ramanana-Rahary, S. & Bassecoulard, E. (2005), Relativity of citation performance and excellence measures: from cross-field to cross-scale effects of field-normalisation. *Scientometrics* 63(2), 373-401.
Zuccala, A. (2005), 'Modeling the invisible college'*, Journal of the American Society for Information Science and Technology* 57(2):152-168.

Appendix

Guided by our informants we developed lexical queries to retrieve from Web of Science the literature covering a specific speciality– sometimes a single term, sometimes a 2-3 line long query using Boolean operators. Our informants validated the data sets checking whether crucial authors whom they expected to see were actually represented, and whether out-of-scope topics had been accidentally introduced. The data was then further processed to eliminate articles with only one author or unknown authors (indicated by 'Anon' in the WoS records), or with only a single reference as the latter records typically do not refer to original research articles or reviews.

This study is part of a larger dissertation research project**. To ensure the anonymity of the human participants of the field study** part of this research in accordance with the IRB[v] approved protocol for the dissertation research **we cannot disclose the lexical queries** that would provide access to the exact data sets discussed in this report. If access to the networks build from the data used in this study is needed for a particular research or validation purpose we are happy to work out a solution that retains the anonymity of our study participants.

Data Field A:
Time span: 1991-2008, number of records: 2,835 (retrieved on 25 Sep 2008)

Data Field B:
Time span: 1991-2008, number of records: 12,817 (retrieved on 14 Nov 2008)

Data Field C:
Time span: 1987-2008, number of records: 30,636 (retrieved on 17 Dec 2008).

---

[i] We call these 'seed' groups as we regard them as entry points into scientific communities, and plan to extend our field studies following links (of cooperation or competition) of these seed labs.

ii When comparing this number to other co-author networks in the literature, remember that this number is calculated for a reduced co-author network (after excluding transient authors). Hence this number will overestimate the relative size of the giant component for the unreduced network.

iii Available from Martin Rosvall's home page at http://www.tp.umu.se/~rosvall/code.html

[iv] Our field studies confirm that in areas of synthetic chemistry the task of having to find someone with the instrumental equipment to conduct certain measurements on your sample is very common.

[v] Institutional Review Board, http://en.wikipedia.org/wiki/Institutional_review_board