The Extraction of Community Structures from Publication Networks to Support Ethnographic Observations of Field Differences in Scientific Communication

Theresa Velden and Carl Lagoze
School of Information, University of Michigan, 4365 North Quad, Ann Arbor, MI 48109-1285.
E-mail: {tvelden, clagoze}@umich.edu

The scientific community of researchers in a research specialty is an important unit of analysis for understanding the field-specific shaping of scientific communication practices. These scientific communities are, however, a challenging unit of analysis to capture and compare because they overlap, have fuzzy boundaries, and evolve over time. We describe a network analytic approach that reveals the complexities of these communities through the examination of their publication networks in combination with insights from ethnographic field studies. We suggest that the structures revealed indicate overlapping subcommunities within a research specialty, and we provide evidence that they differ in disciplinary orientation and research practices. By mapping the community structures of scientific fields we increase confidence about the domain of validity of ethnographic observations as well as of collaborative patterns extracted from publication networks thereby enabling the systematic study of field differences. The network analytic methods presented include methods to optimize the delineation of a bibliographic data set to adequately represent a research specialty and methods to extract community structures from this data. We demonstrate the application of these methods in a case study of two research specialties in the physical and chemical sciences.

Introduction

This article introduces a method for the analysis of scientific publication networks that supports ethnographic observations of field differences in scientific communication. This work is part of a larger research program to develop a mixed-method approach that combines qualitative-ethnographic and quantitative-network analytic methods to systematically study and develop theoretical explanations for differences in scientific communication practices across research specialties. The emergence of the World Wide Web and adoption of web-based technologies to support scientific communication have revived research interest into field differences in scientific communication practices (Fry & Talja, 2007; Kling & McKim, 2000). However, comparative studies of scientific fields with an ethnographic depth of understanding of research practices and social behaviors are rare (Gläser, 2006), and bibliometric studies provide only narrow insights into field specific practices because of the limitations of publication data to reflect underlying research and communication processes (Lievrouw, 1990). The current scarcity of empirical material comparing scientific communities at the field level inhibits theory development to answer fundamental questions such as how field-specific characteristics of scientific knowledge interact with the social ordering of scientific communities (Gläser, 2006), and how epistemic cultures shape communication practices (Cronin, 2003; Fry & Talja, 2004) and attitudes toward openness and sharing (Velden, 2011a).

To advance research on these questions, we bridge the methodological gap with an integrated approach that combines the analysis of publication networks with ethnographic field studies. Comparative studies of scientific communities at field level face several challenges that this approach is designed to overcome. Whereas bibliographic data and the social networks constructed from them are popular objects of study in bibliometrics (e.g., Guimera et al., 2005; Jansen et al., 2010; Kretschmer, 1994; Lambiotte & Panzarasa, 2009; Melin & Persson 1996), quantitative structural and statistical analyses suffer from a lack of explanatory power

with regard to the underlying processes and their meaning to the actors involved (Gläser & Laudel, 2001; Lievrouw, 1990). On the other hand, ethnography is well suited for generating understandings of actors and their practices and motivations, but difficult to scale. This suggests combining the two approaches (Lievrouw, 1990) that hitherto have been rarely used in a truly synergistic manner. However, how local observations scale-up to and tie in with behavioral patterns at the field level is not obvious: how do behavioral characteristics at the field level emerge from local practices, and what is the domain of validity for observations made?

The challenges posed by these open questions are exacerbated by the complexity of scientific fields as units of analysis. The salience of the **research specialty**[1] as a self-organizing unit of researchers that is responsible for the production of new scientific knowledge was recognized in sociology of science in the 1960s and 1970s (Becher & Trowler 2001; Morris & Van der Veer Martens 2008). Those studies also highlighted the overlapping, layered configuration of research specialties, their fuzzy boundaries, and their dynamic temporal evolution. This structural complexity and temporal fluidity provide challenges for the delineation of research specialties as analytic units for comparative studies across fields (Zitt, 2005), and for the valid extrapolation of local ethnographic observations of individuals and groups to the collective level of scientific communities within a research specialty.

Rather than simply aim for the delineation of a research specialty, we suggest the development of methods that help us navigate the complexity of layered scientific communities that make up a research specialty and transcend its boundaries. In this article we report first results using network analytic methods to reveal community structures within scientific fields, thereby contributing to the literature about the mapping of research specialties recently reviewed by Morris and Van der Veer Martens (2008). We identify scientific communities as units of analysis for comparative studies of the influence of material and epistemic characteristics of research on collaboration and communication behaviors. The identification of such units of analysis supports the scaling-up of ethnographies and the extraction of typical collaboration patterns from publication networks for meaningful comparisons of research specialties.

This article is organized as follows: The Background section provides a review of related literature and a summary of our overall research strategy that uses a mixed-method approach to study field differences in scientific communication. The network analytic methods used in this article are described in Methods. They include methods to extract community structures from bibliographic data and methods to support field delineation, that is, the definition of an adequate bibliographic data set from which to extract community structures. The extraction of community structures is achieved through the quantitative analysis of topical and social affinities between topic areas within a research specialty (Topic Area Association and Affinity Networks),

and the visualization of community structures by geographic and topical overlay maps on fieldwide group collaboration networks (Group Collaboration Network Overlays). The methods supporting the definition and preparation of the bibliographic data set consist of an approach to monitor the quality of author name disambiguation and the resulting reduction in network distortion because of name homonymy (Author Name Disambiguation), and a set of heuristic methods to optimize field delineation by controlling recall and precision in the lexical query (Field Delineation). We demonstrate in the Extraction of Community Structures Using Network Analysis section the application of these methods by reporting results from a case study of two fields in the physical and chemical sciences and discuss implications of the empirical results for the methodological approach in the Discussion section.

Background

In this section, we review related work that contextualizes and motivates our efforts to develop a mixed-method approach to study field differences in scientific communication, and we provide an overview of the approach as developed so far.

Related Work

We focus on three central themes in reviewing the related literature: efforts to generate a theoretical understanding of how epistemic differences between research specialties interact with the social organization of scientific communities, leading to field differences in communication practices; the challenges of capturing and describing scientific communities as unit of analysis in comparative studies; and motivations and strategies for a mixed-method approach that combines network analysis with ethnographic field studies.

The emergence of the World Wide Web has triggered a renewed and intensified research interest in differences between research specialties and their communication practices, and how such differences explain variation in the form and speed of take-up of new technologies to support knowledge sharing in scientific communities (e.g., Birnholtz & Bietz, 2003; Cana, 2010; Cronin, 2003; Fry & Talja, 2007, Hine, 2008; Kling et al., 2003, 2004; Kling & McKim, 2000, Nentwich, 2005; Walsh & Bayma, 1996). These scientific inquiries face substantial theoretical challenges, such as the open question of how to conceptualize the dynamics of change in complex socio-technical systems (discussed, e.g., by Geels, 2007) and the lack of theoretical understanding of differences in the intellectual and social organization of research specialties and how these differences shape scientific communication practices. What are the salient characteristics of research specialties that influence communication practices, and what is the appropriate level of analysis (disciplines, subdisciplines, research specialties)

for meaningful comparison and contextualization for observed communication practices?

To describe how the intellectual and the social organization of a research community are intertwined, Knorr Cetina (1999) developed the notion of "epistemic cultures," a concept that has become popular in the information science literature concerned with differences in scientific communication practices across fields (e.g., Baus, 2010; Beaulieu, et al., 2007; Cronin, 2003; VanHouse, 2002). However, Knorr Cetina's ethnographic case study of high-energy physics and molecular biology derives salient cultural features bottom-up, delivering only loosely coupled descriptions of the two epistemic cultures (Gläser, 2006). The concept of "epistemic cultures" does not provide an analytical framework of the scientific research process that would support a systematic comparison between fields (Gläser, 2006).

In a more systematic approach to capture the socio-intellectual characteristics of intellectual fields, Whitley (2000) introduces a taxonomy that distinguishes disciplines by degree of practical and strategic dependence and by degree of task and strategic uncertainty. Empirical studies have used the taxonomy to explain field specific practices of data sharing, the creation of web-based information and communication resources, and the use of open access repositories (Birnholtz & Bietz, 2003; Cana 2010; Fry & Talja, 2007). Problematic for the application of Whitley's taxonomy however is the determination of the appropriate unit of analysis. Fry and Talja (2007) observe that Whitley has developed his taxonomy at the level of well-established disciplines (e.g., "20th-century physics") and that it falls short of capturing more dynamic, possibly trans-disciplinary research collectives.

Another related problem is the conceptualization of scientific fields as reputational work organizations (Gläser, 2006). The organizational sociology perspective applied by Whitley provides only a partial insight into the role of scientific communication in the process of knowledge production. Therefore, the link between the taxonomic classification of a research field and how that classification explains field-specific communication practices remains vague. Empirical studies in information science and bibliometrics oftentimes fall back on coarse disciplinary or subdisciplinary divisions for field comparisons. However, smaller, more specialized entities within (sub)disciplines expose significant differences in social organization and research culture (Becher & Trowler, 2001; Galison, 1997; Mulkay, 1977), and it has been suggested that an appropriate level of analysis for the comparison of field specific communication behaviors has to be sought at a finer level of granularity, such as the research specialty (e.g., Fry & Talja, 2004; Jamali & Nicholas, 2008).

Theoretical support for this view is provided by Gläser's theory for the collective production of scientific knowledge in research specialties. The theory is derived from a thorough review of empirical studies in science studies, and it ascribes an indirect but crucial coordinating role to the shared knowledge base of the *scientific community*[2] in a research specialty (Gläser, 2006). By the orientation of the self-directed, autonomous research activities of community members toward contributing to this shared knowledge base, a social ordering emerges that enables the collective knowledge production process (Gläser, 2006). This model implies that scientific communication plays a key role in the coordination (social ordering) of the collective production of knowledge within a research specialty. Gläser hypothesizes that differences in the social ordering of communities can be linked to differences in the nature of their respective knowledge bases. We would expect such differences to be reflected also in the scientific communication practices of a community.

Unfortunately, research specialties are difficult to capture as a unit of analysis, last but not least due to the self-similarity of science (Zitt, 2005). The boundaries of research specialties are fuzzy, as even field members do not necessarily agree on where to draw the line because research specialties overlap and evolve in time. Bibliometric field delineation, that is, the extraction of a subset of publications from a bibliographic database to represent either all the authors or all the relevant publications in a research field, is a difficult task, with no simple generic solution available.

Journal-level field delineation that rests on the selection of core journals in a field has been shown to be problematic because of the wide scope of most journals and the wide spread of specialized literature across many different journals (Aksnes et al., 1999; Huang et al., 2011). Instead, lexical queries that retrieve individual publications by matching query terms to their bibliographic metadata support a more precise retrieval of relevant literature. However, they tend to suffer from a lack of recall, especially for newly emerging and interdisciplinary fields (Mogoutov & Kahane, 2007). Sophisticated hybrid approaches have been developed to delineate and monitor such fields. They start from a high precision and low recall seed of publications generated by a lexical query, and extend this seed either through citation coupling (Laurens et al., 2010; Zitt & Bassecoulard, 2006) or keyword-based retrieval (Mogoutov & Kahane, 2007). The quality of the field delineation is further affected by the coverage of databases available to the analyst, the extent of access to those databases, and the quality and granularity of the data. Depending on the scientific domain, different databases or a combination of databases provide better results (Strotmann & Zhao, 2010).

In this study we retrieve our data from the Thomson Reuters Web of Science database (WoS) because it has reasonable coverage in established fields in physics and chemistry (Moed, 2005), especially for the 20-year time range we were interested in. Because of economic restrictions in access to the complete WoS, we could not use the hybrid approach developed by Zitt and Bassecoulard (2006) that requires access to the "cited by" links from a cited publication to the citing publication, a restriction pointed out by Huang et al. (2011) and Mogoutov and Kahane (2007). Instead we start from a lexical query and then use heuristic network analytic methods to improve recall and precision. The new methods that we introduce in Field Delineation for

the optimization of the lexical query are unique due to their transparency with regard to the social topology of the field, as discussed in the Field Delineation section.

We suggest a mixed-method approach is essential to produce the empirical base critically needed to develop a theoretical understanding of the field-specific shaping of communication practices. By combining ethnography with the network analysis of bibliographic data, we can deliver nuanced and situated understandings of communication practices along with insights into community structures and behavioral patterns at the collective level, as suggested by (Lievrouw, 1990). These structural patterns indicate the domain of validity for local observations and further support the quantification of some aspects of field differences, such as collaboration intensity or group sizes and structures.

Strategically, our motivation for combining ethnography with network analysis has strong parallels with the "network ethnography" approach advocated by Howard (2002).[3] He perceives the combination of ethnographic fieldwork with network analysis as a synergistic, transdisciplinary method with distinct advantages for the study of communication in modern organizations over new media. Aiming to develop more rigorous methods to obtain generalizable qualitative data, he values this particular method combination because it improves the construction and strategic sampling of field sites for ethnographic research.

However, border crossings that combine ethnographic studies with quantitative network analysis are rare in the study of research communities. There exists a strong disciplinary divide between the quantitative focus of fields such as bibliometrics, scientometrics, and the science of networks,[4] and the predominance of qualitatively oriented work in fields such as the sociology of science and science and technology studies (Gläser & Laudel, 2001, 2007; Van den Besselaar, 2000). Some of the very few exceptions are Geiger and Ribes (2011), who develop the idea of trace ethnography to study the use of information technology-based systems in cyberscience; Cambrosio et al. (2004), who analyze large-scale collaborative research in an area of bio-medicine by making use of heterogeneous networks that link entities such as humans, laboratories, and molecules; and Zuccala (2006), who explores invisible college structures in a field of mathematics extending work by Lievrouw et al. (1987) and Crane (1972) who used mixed-method approaches to study invisible colleges. This article contributes to this latter strand of work by revealing community structures within scientific fields with the specific aim to support the scaling-up of ethnographic analysis for comparative studies of scientific communication practices.

A Mixed-Method Approach to Study Field Differences

Our approach to the comparative study of scientific communities can be characterized as "integration" of network analysis and ethnographic studies, following the terminology by Moran-Ellis (2006) to distinguish and characterize mixed-method approaches. This means we use both methods in an interdependent way such that results obtained by one method feed directly into the other.

Figure 1 depicts schematically the interdependency of important steps in the research process, showing how insights generated with one method feed into the other method, and vice versa. Our starting point has been ethnographic field studies of five research groups from two research specialties in the physical and chemical sciences. One of the authors visited each group for several weeks, conducting observations and semistructured interviews with most of the members of each research group (Step 1). Based on the understanding of the research field and local research practices gained during these visits and supported by feedback from knowledgeable study participants, we developed a lexical query for retrieval of data from a bibliographic database to capture the publication output of the field (Step 2). Throughout this process, observations from the field studies together with network analytic data inform decisions on how to adapt and optimize the lexical query, a process that is described in detail in Recall of Lexical Query and Precision of Lexical Query below. The coauthor network is generated from the publication data retrieved, and various network features are extracted and visualized. These include the structural composition of coauthor clusters and between-cluster interlinking patterns. The interpretation of network features in terms of real-life group organization patterns and collaboration scenarios is derived from interviews and joint examination of network visualizations with senior researchers at the field sites (Step 3). The network features are quantified for comparisons across research specialties (Velden et al., 2010), generating further research questions. Finally, community structures within the fields are extracted (Step 4), as detailed in Topic Area Association and Affinity Networks and Group Collaboration Network Overlays. Observations from the field studies support the interpretation of those structures that in return contextualize some of the observations and may guide the strategic sampling of further field sites or interview partners. Eventually, findings are synthesized to identify field specific patterns of communication and to analyze epistemic, material, and social factors shaping the underlying practices (Step 5). Preliminary findings on field differences in openness and sharing behaviors in two groups are reported in Velden (2011a, 2013). The insights into community structures reported in this article will guide the refinement of future research into field-specific communication practices.

The significance of this combined ethnographic and network-analytic approach is that it allows us on the one hand to develop in-depth contextual understandings of researchers' behaviors through ethnographic observation and analysis. The ethnographic component produces nuanced insights into processes underlying the abstracted patterns of coauthor and citation networks, making actors' motivations and every-day practices susceptible to analysis
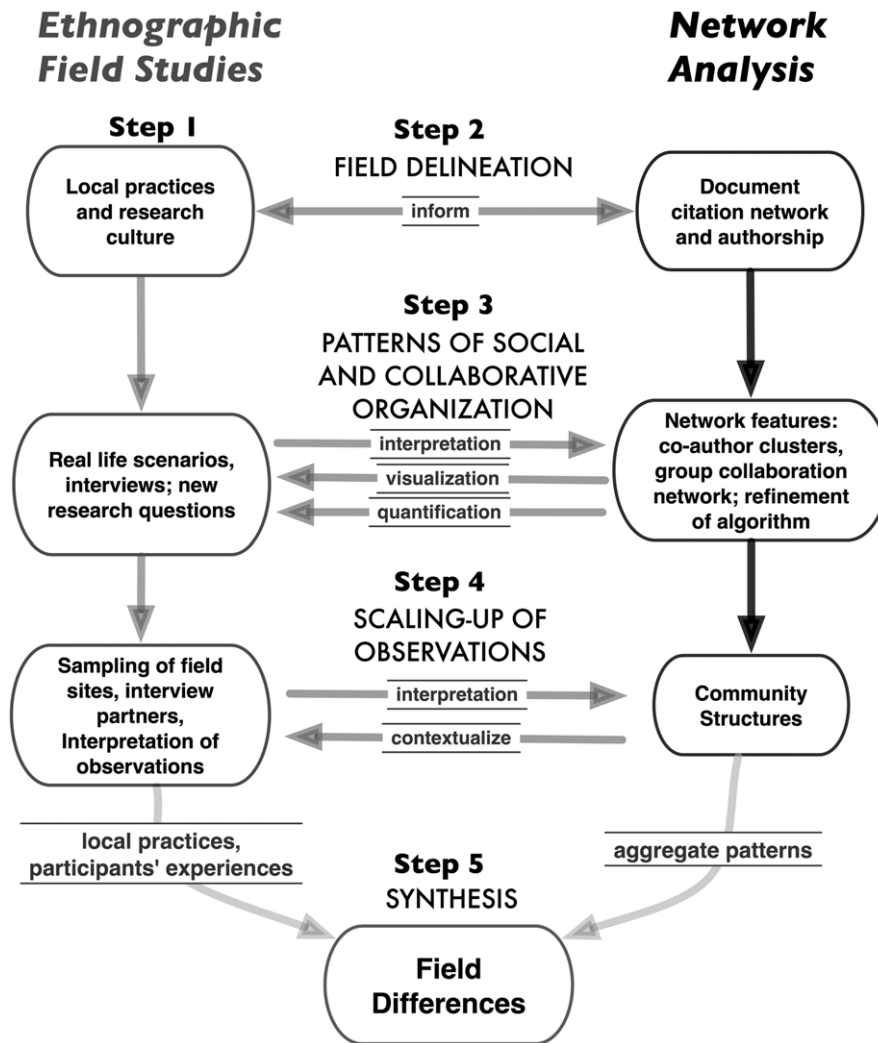
Ethnographic Field Studies

Step 1

Local practices and research culture

Step 2
FIELD DELINEATION

inform

Network Analysis

Document citation network and authorship

Step 3
PATTERNS OF SOCIAL AND COLLABORATIVE ORGANIZATION

Real life scenarios, interviews; new research questions

interpretation

visualization

quantification

Network features: co-author clusters, group collaboration network; refinement of algorithm

Step 4
SCALING-UP OF OBSERVATIONS

Sampling of field sites, interview partners, Interpretation of observations

interpretation

contextualize

Community Structures

local practices, participants' experiences

Step 5
SYNTHESIS

aggregate patterns

Field Differences

FIG. 1. Schematic representation of the combination of qualitative and quantitative methods in a mixed-method approach to study and compare social practices in scientific communities across scientific fields. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

and interpretation. The network analysis of the publication data, on the other hand, provides us with quantifiable, aggregate data on community-wide behaviors to the extent that they are signified by citation or coauthorship. This way we can capture emergent community-wide patterns of group organization, collaboration, and knowledge integration, which are only partially visible from qualitative, locally specific observations.

Beyond providing complementary data, the integration of the two methods is a dynamic process and ameliorates shortcomings of either method in the following ways: The data visualizations generate new questions such as what is the role of small coauthor clusters in a field that sustain their publication activity over the entire time period? Simultaneously, the ethnographic analysis of research processes and social organization critically supports the interpretation of the network features. It thereby ensures that quantifiable, structural differences between fields can be linked to specific behavioral patterns, and it further guides meaningful

refinements of the network analysis (e.g., the distinction between transfer and collaboration links between coauthor clusters). Furthermore, the network analytic extraction of community structures reported in this article supports the scaling-up of meaningful units of analysis. By making community structures visible, it supports the strategic sampling of research sites and increases confidence in the validity of ethnographic observations and their extrapolation to larger collective aggregates at the community level.

Methods

This article focuses on the development of network analytic methods to reveal community structures within a scientific field that can then feed into the mixed-methods approach described previously. We construct two basic network types from bibliographic data of the publication output in a scientific field, the undirected coauthor network and the directed document citation network. The nodes in
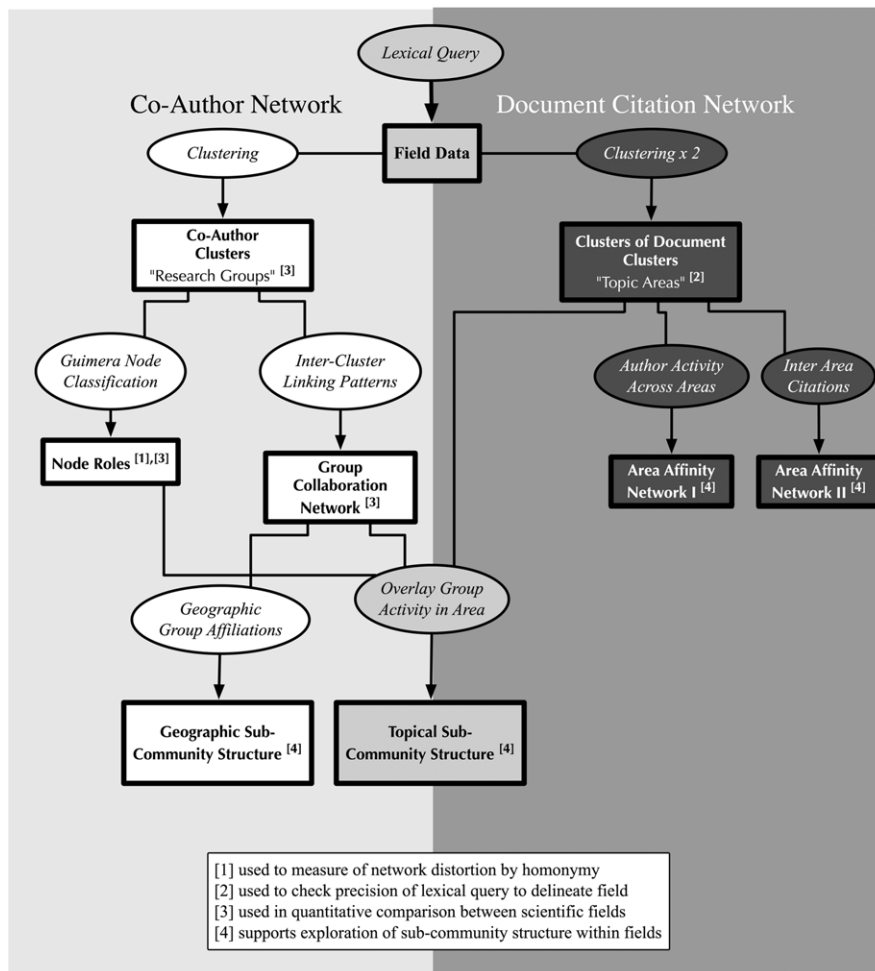
Co-Author Network

Document Citation Network

Lexical Query

Clustering — Field Data — Clustering x 2

Co-Author Clusters "Research Groups" [3]

Clusters of Document Clusters "Topic Areas" [2]

Guimera Node Classification

Inter-Cluster Linking Patterns

Author Activity Across Areas

Inter Area Citations

Node Roles [1],[3]

Group Collaboration Network [3]

Area Affinity Network I [4]

Area Affinity Network II [4]

Geographic Group Affiliations

Overlay Group Activity in Area

Geographic Sub-Community Structure [4]

Topical Sub-Community Structure [4]

[1] used to measure of network distortion by homonymy
[2] used to check precision of lexical query to delineate field
[3] used in quantitative comparison between scientific fields
[4] supports exploration of sub-community structure within fields

FIG. 2. Overview of network analysis steps. Details on network clustering are given in the Network Clustering section, on the Guimera node role classification in the Node Role Classification section, and on the group collaboration network with geographical and topic area overlays in the Group Collaboration Network Overlays section. Details on the creation of topic area affinity networks are provided in the Topic Area Association and Affinity Networks section. The way we use node role classification to assess network distortion is explained in Author Name Disambiguation. The way we use topic areas to support field delineation is described in Precision of Lexical Query. Results on the area affinity networks are reported in Mapping Community Structures as Overlays on Collaboration Network. The quantitative comparison of the two fields with regard to the size of group collaboration networks and results on the extraction of subcommunity structure of the two fields are reported in the Mapping Community Structures as Overlays on Collaboration Network section.

the coauthor network represent authors and links represent joint coauthorship of a publication. Node size indicates the number of publications by the author in the data set and link strength indicates the number of publications in the data set that two authors have coauthored. In the document citation network nodes represent publications, and a directed link represents a reference made from one publication to another. Figure 2 provides an overview of the main steps of processing the two networks, indicating how the network constructs retrieved are used in the analysis. The following sections describe in more detail the basic network analytic constructs used in this study.

Network Clustering

In a first step, each network is clustered to extract its modular substructure. The clustering algorithm used is an information theoretic algorithm that models information flows (Rosvall & Bergstrom, 2008). It has shown superior behavior in retrieving clusters from benchmark networks with clusters of a range of different sizes when compared with other popular clustering algorithms (Lancichinetti & Fortunato, 2009). However, any clustering method based on the optimization of a global measure will have an intrinsic resolution limit (Lancichinetti & Fortunato, 2009) such that the clusters retrieved need to be considered with caution. Our mixed-method approach mediates this caution. Judging from the insights we got in our ethnographic field studies about the identity of and relationships between groups, we have found that the coauthor clusters that this algorithm retrieves for the coauthor networks included in this study are very plausible representations of those social collectives that act as primary contributors of original work in a field. For example, the algorithm retrieved from our data set for Field 2 was a large experimental physics group (82 coauthors during 20 years) as well as a much smaller theoretical

physics group (16 coauthors). On the other hand, in an example from Field 1, two chemistry groups led by two **Principal Investigators**[5] (PIs) who are friends and have been working very closely over an extended period of time in this particular field were retrieved as a single cluster of coauthors. Furthermore, by manual lookup of homepages on the World Wide Web we have found that many small coauthor clusters[6] in Field 2 correspond to theory groups. This observation also suggests that the clustering algorithm is capable of retrieving socially relevant units from the network. Hence, the clusters retrieved can be identified as **functional research groups**[7] (Seglen & Aksnes, 2000), either research groups led by a single PI or closely collaborating teams of research groups with several PIs (Velden et al., 2010).

Using Rosvall's clustering algorithm twice on the citation network results in a high concentration of documents in several large clusters, such that the clusters of minimum size of 1 to 2% of all documents capture around 80 to 90% of documents in the field data set. We take these document clusters to represent topic areas within the research specialty. In this study, we select 2% as a minimum size for topic areas and focus our analysis on those major topic areas within each field.

Reference Inclusion Rate

The degree to which publications within one of the topic areas within the field cite work within or outside the field could be meaningful to assess the integration and overlap of that topic area with the field. By the reference inclusion rate of a topic area we refer to the proportion of references from publications published within that topic area that point to previously published publications in the entire field data set. We suggest that calculating the reference inclusion rates of a topic area can be a useful measure of how comprehensively a topic area is represented by the field data set. Assuming that the data set captures a chosen field well, a lower reference inclusion rate could then be interpreted as an indication that a topic area overlaps with another field or several other fields, whereas topic areas with higher reference inclusion rates would seem to be more strongly contained within the field. Specifically, to track the temporal evolution of reference inclusion rates, we used a moving 5-year window, calculating for each year between 1996–2010 the reference inclusion rate considering only references to publications in the previous 5 years.

Node Role Classification

To capture structural differences between modular networks, Guimera et al. (2007) introduce a classification of nodes in clustered networks that is derived from their linking patterns. At the top level, this classification distinguishes between hub nodes and nonhub nodes. Hub nodes have disproportionally many cluster internal links relative to the average inside-cluster degree of the nodes in the respective cluster, whereas nonhubs have average or below average inside cluster links. Based on their outside links to nodes in other clusters, hubs are further subdivided into "provincial hubs," "connector nodes," and "satellite connector nodes" where the former have been least outside links, and the latter have links to many other clusters. Similarly, nonhubs are subdivided into "ultraperipheral nodes," "peripheral nodes," "connector nodes," and "satellite connector nodes."

We apply this classification to the nodes in the clustered coauthor network. The stratification of nodes in the coauthor network by their node role can then be used to detect network distortions because of author name ambiguity, as explained in more detail in the Author Name Disambiguation section.

Topic Area Association and Affinity Networks

To explore patterns of affinity or antagonism between topic areas, we quantify the connectivity between topic areas in terms of citations and in terms of author activity overlap, that is, share of publications of authors who contribute to several topic areas. For the null model of expected connectivity we assume that an author randomly selects a document in another topic area to cite or randomly selects another topic area to publish in. According to this null model, the probability of a citation from an article in one area (source area) to an article in another area (target area) is proportional to the relative proportion of numbers of articles in the target area when compared with the other major topic areas that a citation may point to. Similarly, the null model probability that an author publishing in a topic area (source area) chooses for his or her out-of-area publications a specific topic area (target area) is proportional to the relative size of that target area when compared with the other potential target areas.[9]

We quantify the association between areas as the empirical (actual) deviation from this null model. We treat the null model as providing the expected values for a multinominal distribution. We apply a chi square goodness-of-fit test to evaluate whether, for a specific source area, the actual distribution of out-of-area citations or out-of-area publications across potential target areas deviates from the hypothesized null model distribution. The residuals quantify for each target area the positive or negative deviation from the expectation values. They are given by:

$$residuals = (count\_actual - count\_expected)/count\_expected$$

where the expected counts are calculated according to the null model as:

$$count\_expected := relative\ target\ area\ size * count\_total$$

where

Relative target area size: Proportional size of target area relative to all potential target areas.

Count_total (citations): Sum over all publications in the source area of the sum of all their citations to any publication in any of the target areas.

Count_total (authors): Sum over all authors in the source area of the sum of all their publications in any of the target areas.

We interpret positive residuals as affinities, pointing to a positive and disproportionally strong association between two topic areas. To visualize the affinities between areas we create a directed weighted network where nodes represent topic areas and the strength of links between nodes is given by those positive residuals. Note that zero affinity does not mean the absence of any links between two areas, but could either represent the assumed background of area connectivity that scales with area sizes or a negative deviation from this assumed background (antagonism).

Group Collaboration Network Overlays

In Velden et al. (2010), we derive a conceptual and algorithmic distinction between "transfer" links and "group collaboration" links for a coauthor network. The latter linking patterns are understood as bibliometric traces of intensive collaborative relationships among research groups in contrast to "transfer links," which indicate ephemeral exchange or transfer processes, such as group-to-group career migration of individuals or one-off, service-type collaborations. Applying this distinction, we extract from the clustered coauthor network a group collaboration network where nodes represent coauthor clusters and links represent actual intergroup collaboration. These collaboration networks represent the collaborative core of a field. To study community substructures in a field, we overlay those collaboration networks with geographic affiliation information or topical affiliation information on the groups as described below and illustrated, for example, by Figure 10 in the Results section.

The geographic affiliation of groups is extracted from the institutional affiliations listed for most publications in the WoS database. We evaluate this information only at the continent level. Each cluster is represented by all the publications coauthored by at least one of its authors. For this set of publications, we determine the country affiliation that is most often listed. In case the second-placed country is listed at least 50% as many times as the most often listed country, and if these two countries belong to different continents, we assign a mixed, two continent geographical affiliation.

The topical overlay maps show for each topic area how active the groups in the collaboration network are in the respective areas. To increase precision, we only consider the publications of hub nodes in the cluster;[10] if a cluster has no hub nodes we consider the publication output of the entire collective of authors in the cluster. A group's activity in a topic area is calculated as the proportion of its publications that is part of the document cluster representing that topic area. We use node color to represent group activity in the topic overlay on the group collaboration network.

Data

The bibliographic data used in this study have been selected to represent two scientific fields: a research specialty in synthetic chemistry (Field 1), and an experimentally oriented research specialty that overlaps with chemistry and physics (Field 2). Field 1 is defined by a chemical reaction mechanism, the creation of molecular catalysts that enable this type of reaction, and the application of these reactions in chemical syntheses. Field 2 is driven by solving fundamental conceptual questions and is defined by a physically defined class of objects that provides opportunities to study these fundamental questions. Both fields are dominated by basic research conducted at universities and research institutes rather than applied research. These fields were selected in the context of our mixed-method research into field differences in communication behaviors.

We retrieve the bibliographic data from the WoS for a time period of 20 years (1991–2010). We use a lexical query developed to identify the publication output of the two fields, an iterative process of field delineation described in Field Delineation. Because authors are identified in the WoS database by their last name and initials only, name homonymy[11] is problematic, distorting the coauthor networks that can be constructed. To improve the accuracy of the network we apply an author disambiguation algorithm, described below in the Author Name Disambiguation section.

Author Name Disambiguation

To address the problem of name homonymy we apply an algorithm, which was developed and evaluated in Velden et al. (2011), to disambiguate author names. We use descriptive network statistics to monitor the improvement made in resolving the undistorted network structure, a method that avoids the expensive investment in creating a representative ground truth sample (Velden et al., 2011). This approach to measuring network distortion rests on the assumption that, given perfect resolution of authors' identities, the structural position of an author in the network and the commonality of his or her last name are uncorrelated. Therefore, an uneven distribution of common last names across the seven classes of node roles in the coauthor network (introduced in the Reference Inclusion Rate section) signifies network distortion because of name ambiguity.

The cumulative probability distributions of last name commonalities by node role for all nodes in the giant component of the coauthor network are shown before and after disambiguation in Figure 3. Successful disambiguation is reflected by close agreement between the curves for the different node roles. Most curves still reflect a long tailed distribution of last name commonalities, but without the very heavy tails the curves showed before disambiguation. Very common names have become more equally distributed between node roles and disambiguation has been effective (though not perfect) for both data sets.

The effect of the disambiguation on the number of unique authors identified in the data set is shown in Table 1. As
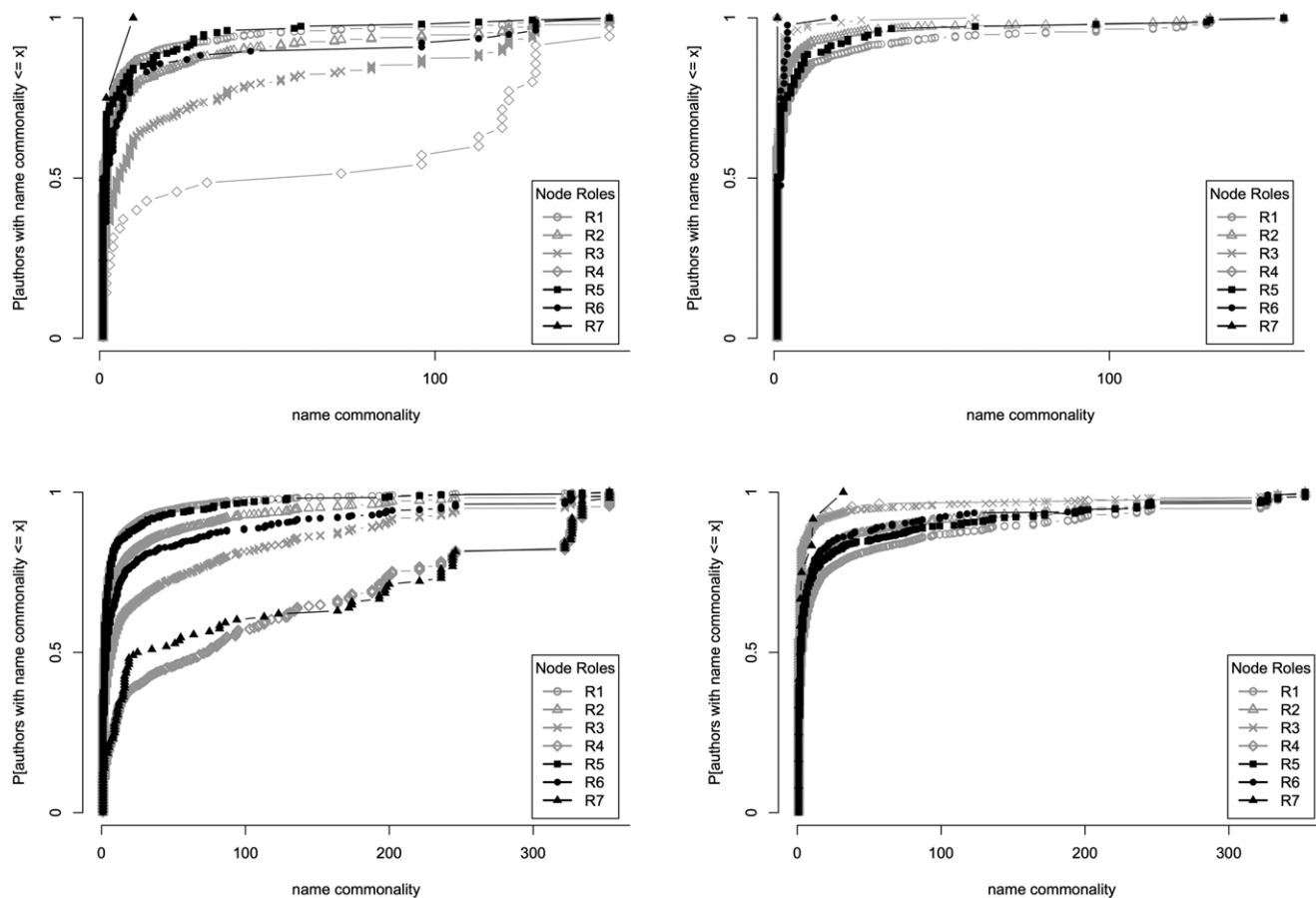
FIG. 3. Cumulative probability distributions of name commonalities for Field 1 (top) and Field 2 (bottom) before disambiguation (left) and after disambiguation (right).

TABLE 1. Data set before and after author name disambiguation.

| | Field 1 | | Field 2 | |
	Nondisambiguated	Disambiguated	Nondisambiguated	Disambiguated
No. of publications	14,599		65,003	
No. of authors	25,500	27,946	105,776	131,285
No. of 1-time authors	15,937	18,664	64,869	90,006

could be expected we obtain a higher number of authors after disambiguation, and the number of authors with only one publication increases substantially. However, because we filter out 1-time authors from the data to obtain more concise coauthor networks (Velden et al., 2010), the net result in terms of total number of authors in the coauthor networks is roughly neutral.

Field Delineation

We start with a standard approach to publication-level field delineation by using a lexical query to retrieve all publications during a defined time frame that have specific terms in the title, abstract, or keyword field of the bibliographic record. In addition we filter the records retrieved by subject categories offered by WoS.[12]

The task of developing the lexical query of terms characteristic for each research field is rather unproblematic for Field 1 because this field is defined by work on a class of catalyzed chemical reactions that is known by a specific name that is highly standardized and commonly used in titles or abstracts of publications in this field. However, for Field 2 the field delineation process is challenging and we had to go through several iterations. We checked the progress made by a combination of participant feedback and network analyses. In the next two sections, we introduce two
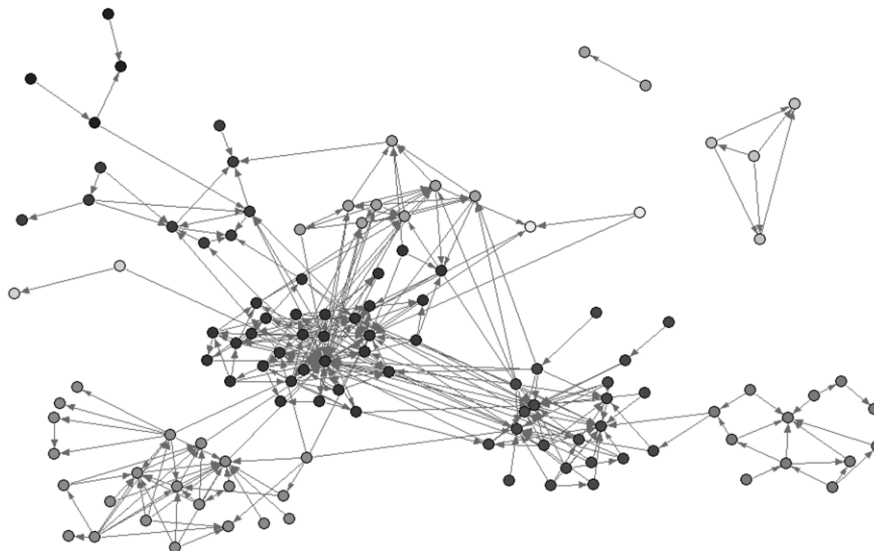
FIG. 4.   Self-citation network of a research group leader. Nodes are documents, node colors indicate cluster membership (interpreted as research topics), and links between documents represent citations between documents (self-citations). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

heuristic network analytic methods to guide the optimization of the lexical query, one for checking recall, Recall of Lexical Query, and one for checking precision, Precision of Lexical Query.

Recall of lexical query.   The recall of the lexical query—whether one is capturing all of the relevant publications—can be tracked by checking whether the publications of renowned scientists in the research field are actually included in the data set retrieved.

To this end, one needs to carefully select well-known researchers who span the breadth of the field. For several individuals we retrieve all their publications as indexed by the WoS database. These individuals are the leaders of the research groups included in the ethnographic field study as well as selected researchers who were designated by field study participants as being important figures in the field.

Because the research interests of prolific authors are often spread over several research specialties, we need to identify the subset of publications of a researcher that is relevant to the research field of interest. Hellsten et al. (2007) found that document clusters in the self-citation network of an author's entire work distinguish the research topics that a researcher has worked on. Hence, we construct for each selected researcher the directed, unweighted self-citation network of his or her publication output. Using a clustering algorithm[13] by Rosvall and Bergstrom (2008) we extract document clusters to identify research topics within a researcher's work. An example of such a clustered self-citation network is provided in Figure 4.

We verified the appropriateness of the clustering provided by this algorithm by reviewing the resulting clustered network of their publications with the research group leaders of the five groups in our ethnographic field study. Each of

them felt it was returning a comprehensive representation of their research interests, as exemplified by the following quote:

> I think that if you look at this, that's my life. Yeah, I think you've got virtually everything. Any subarea (missing?) . . . no, not really. Nope. [PI, field 1]

Next we need to identify those document clusters from a researcher's self-citation network that pertain to the targeted research field. To this end we manually inspect publication titles, and assign a tentative topic description to each cluster. These decisions whether a cluster belongs to the research specialty or not seems relatively unproblematic at least for those clusters of publications that can be considered core contributions of that researcher to the research specialty. We then can calculate the overlap between the field data and those clusters. Whenever such a cluster is only weakly represented we inspect titles and abstracts to search for plausible, sufficiently standardized terms to add to the lexical query to improve recall.

Precision of lexical query.   To check the precision of the lexical query means to check whether it captures irrelevant publications that belong to another research field with no or minimal overlap with the scientific field of interest. Because research specialties are overlapping and scientists often-times work in several research specialties it is not always straightforward to delineate a research field based on topical distinctions alone. To control for extensions into neighboring fields we investigate how well the various topic areas within a field (retrieved from clustering the citation network, see the Network Clustering section) connect to one another in terms of authors contributing to several of them.
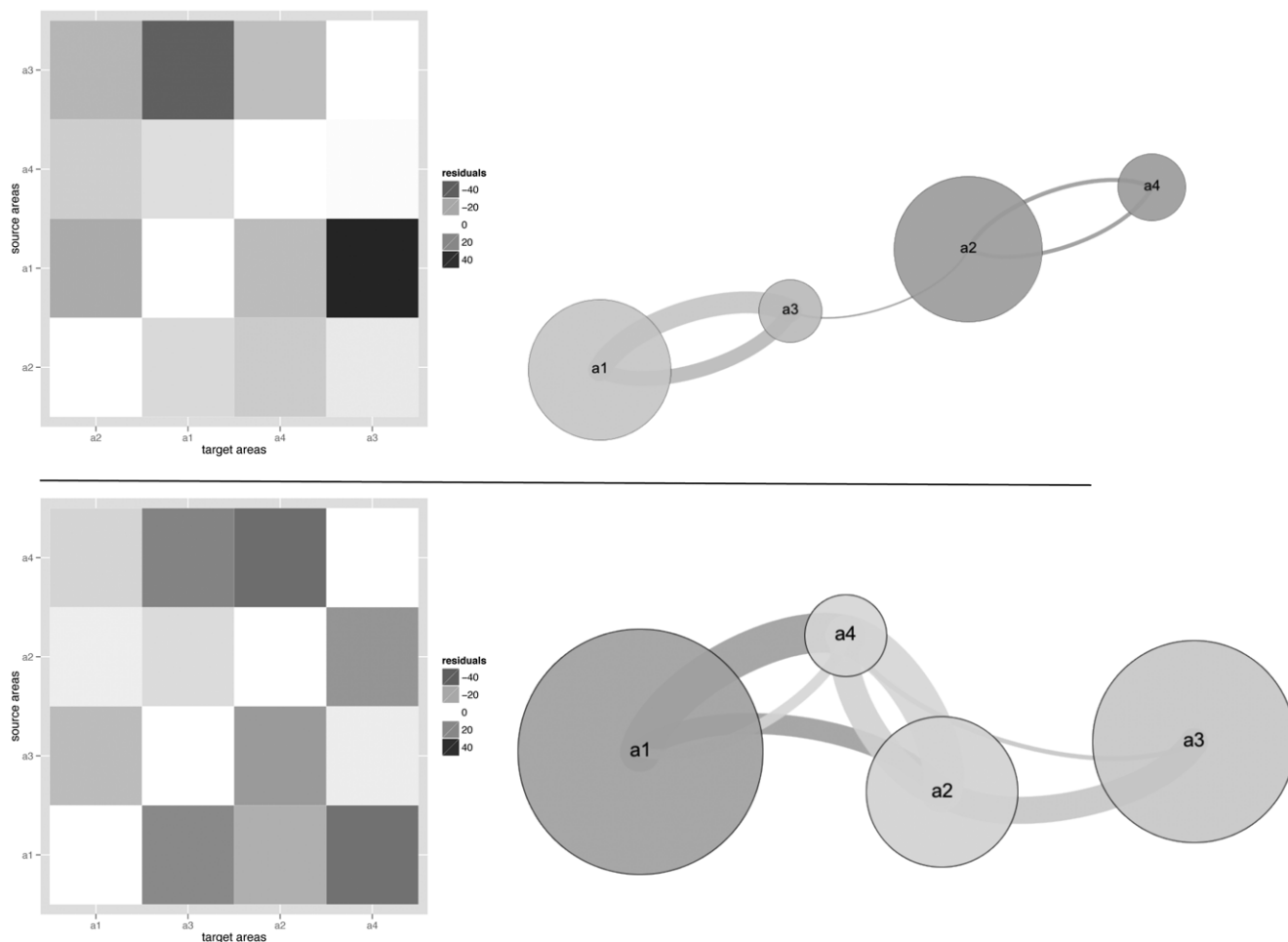
FIG. 5.   Heat map and affinity network of author activity based area associations between the largest four topic areas retrieved by two different versions of the lexical query for Field 2, an early one (left) and the final one (right). Heatmap colors indicate strength of positive (green) or negative (red) associations between a source area and a target area. Link colors indicate source node of directed links. Link directionality is indicated by link curvature: The link is curved with the flow in clockwise direction. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Figure 5 shows two heat maps used to visualize and compare interarea connectivity for two variants of the lexical query for Field 2. It depicts the relative participation of authors across the four largest topic areas for an early version of the lexical query and the final version of the query that generated the data set used in this study. The matrix values represent the residuals of the expected association of topic areas derived from author activity, introduced in the Topic Area Association and Affinity Networks section. To the left the resulting affinity networks are shown, highlighting positive associations between topic areas. The heat map for the historic lexical query for Field 2 shows a checkerboard pattern that indicates that the four largest topic areas form two almost disjoint sets of areas, as shown also by the affinity network. For the final lexical query for Field 2, the heat map pattern indicates that although the two largest areas 1 and 3 are antagonistic, areas 1, 2, and 4 are connected and a crucial connection between those three areas and area 3 is made by area 3 authors contributing disproportionally also to area 2, and to a lesser extent, also to area 4, as shown in the affinity network. Because this author activity analysis for the historical query indicated that two of the four

largest areas had minimal overlap with the other two largest areas, the question arose whether this disconnect signaled the unintended inclusion of another research field through the lexical query.

On inspection of the titles of publications in areas 2 and 4 of the network for the historical query, we found that they were mostly associated with a term referring to particles of a dimension that went beyond the focus of those studied in the research specialty initially identified. A participant whose research spans both areas had suggested it in the course of the field study. These particles are produced with chemical methods and require a quite different skill set, and it looked like few researchers had ventured (yet?) to bridge those areas.

The evidence of social discontinuity depicted in the heat map of the intermediary query prompted us to reconsider and to opt against inclusion of this more chemically oriented research direction in the data set for Field 2. This kind of judgment requires a deeper understanding of the distinctions made by study participants and highlights the value of a mixed-method approach.

TABLE 2. Network sizes and properties.

	Field 1		Field 2	
Document citation network				
No. of documents excl. singletons	13,910		55,648	
No. of document clusters	509		2,704	
No. of clusters of document clusters	102		1,069	
Coauthor network				
No. authors after filtering	9,116		40,808	
No. of clusters	1,132		4,270	
No. of nodes in giant (proportion)	6,645	(72.9%)	33,203	(81.4%)
No. of clusters in giant (proportion)	532	(47.0%)	2,086	(48.6%)
Average cluster size in giant (median)	19.7	(11)	27.5	(13)
No. of linked clusters in collaboration network (proportion)	48	(9.0%)	477	(22.9%)

Extraction of Community Structures Using Network Analysis

Basic properties of the coauthor networks and the document citation networks constructed from the bibliographic data for Fields 1 and 2 are given in Table 2. In the supplementary material we show the literature growth in Fields 1 and 2 between 1991 and 2010. Field 1 has a distinct s-shaped growth curve with a strong growth over a ten-year time period, more than quadrupling the annual publication output between 1996 and 2005. The growth curve in Field 2 shows a weaker s-shape, just about doubling its annual publication output in the ten-year period between 1995 and 2004.

In the following, we report on the extraction of topic areas and their relatedness from the document citation networks (Topic Areas Derived from Citation Network section) and how we combine them with group collaboration networks to visualize community structures in geographic and topical overlay maps (Mapping Community Structures as Overlays on Collaboration Network).

Topic Areas Derived From Citation Network

For Field 1 we obtain 6 major topic areas that fulfill the 2% size criterion (see the Network Clustering section). Together they represent 96.1% of documents in the document citation network. For Field 2 we obtain 11 major topic areas that represent 84.9% of documents in the document citation network. Figure 6 shows for each field the network of citations between the major topic areas. We find that Field 1 is dominated by one large topic area that is more than twice as large as the other five topic areas taken together.

The area sizes in Field 2 are more balanced, with three rather large topic areas and five mid-size topic areas, and three small topic areas.

Topic area labeling. We manually assign topic descriptions to the topic areas extracted from the citation network based on a combination of resources. First, we familiarize ourselves with research topics within the field through ethnographic observation and interviews with researchers in the field as well as the reading of review articles of the field. Next, we extract and browse the publication metadata for each of the document sets in a cluster, specifically the most frequent author names, the titles of articles, as well as the titles of the most frequent journals published in.[14] Furthermore, we use string searches on text files with the bibliographic data for the publications in each document cluster to double check the plausibility of assignments made by counting the prevalence of technical terms that have a certain topic specific relevance and discriminatory power. The tables depicted in Figures 9 and 10 provide an overview of the resulting topic assignments for the major document clusters in each field. We find that, frequently, topic areas indicate disciplinary orientations, along with specific research foci. For example, in Field 1 there are three topic areas associated with polymer chemistry, however each with a different research theme.

Topic area evolution. Next we look at the growth of topic areas over time as well as the reference inclusion rates for each topic area (defined in the Reference Inclusion Rate section). The results are shown in Figure 6. We make the following observations:

Growth in Field 1 is driven by the dramatic growth of topic area *1:Catalyst Development and Organic Chemistry*. This observation conforms to popular narratives about the history of the field. Accordingly, it experienced a significant breakthrough in the late 1980s and early 1990s when a few pioneers succeeded in producing a specific type of molecular catalysts that turned out to have great application potential for syntheses in organic chemistry. The research group leader of an organic chemistry group in our field study describes how he got interested in this new type of reactions and catalysts very early in the development of the field:

> "So, during the last years this [reaction] has been the main research area [of the group] and it still is the most important area in terms of deployment of PhD students and co-workers [...] When we started it was not obvious what one could do with it and the catalysts were not there yet. There existed isolated works and it was simply the personal assessment, that this

FIG. 6. Citation network of the six largest topic areas in Field 1 (top), and 11 largest topic areas in Field 2 (bottom). Nodes represent document clusters extracted from document citation networks for Field 1 and Field 2. Node size is determined by number of documents in a topic area, link strength by the number of citations between the documents in the respective pair of topic areas, and link color indicates the source node of the directed links. Network visualization performed with gephi using the Force Atlas 2 algorithm. Below the network visualizations, the evolution of reference inclusion rates between 1996 and 2010 for Field 1 (top) and Field 2 (bottom) is shown, and to the right, the evolution of area sizes over the 20-year time frame from 1991 to 2010 for Field 1 (top) and Field 2 (bottom). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]
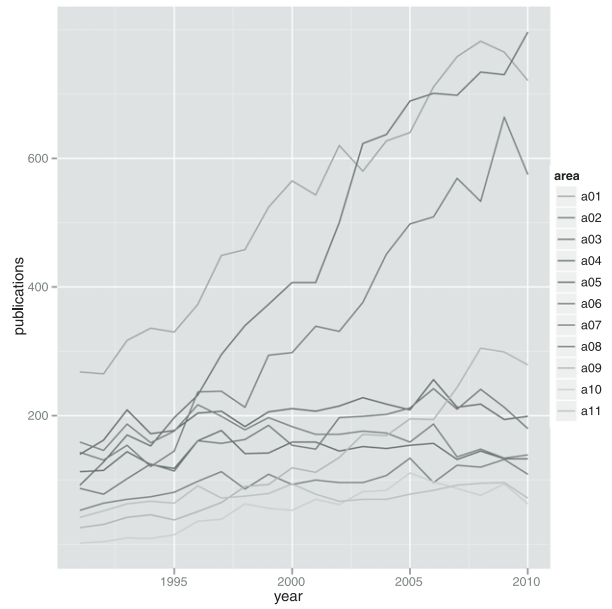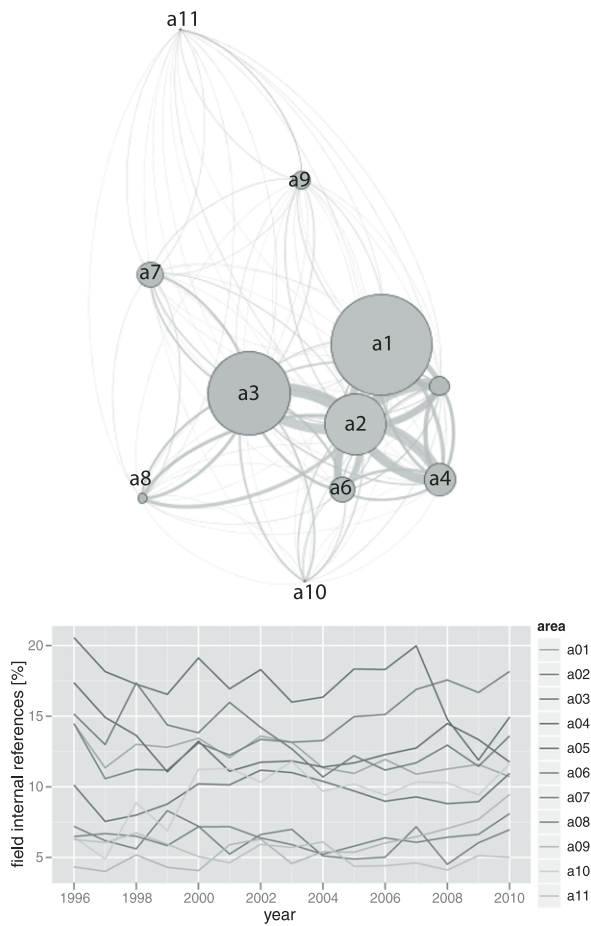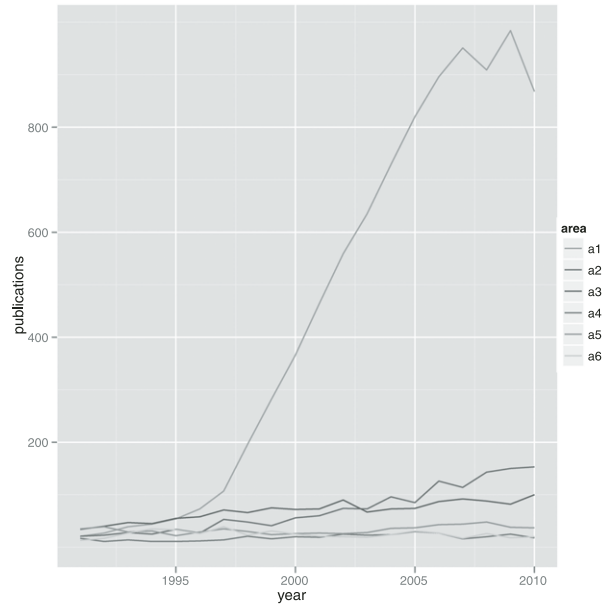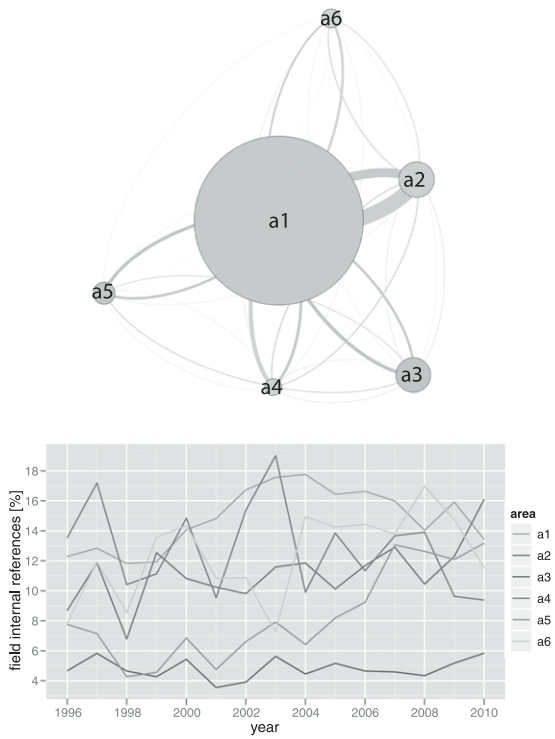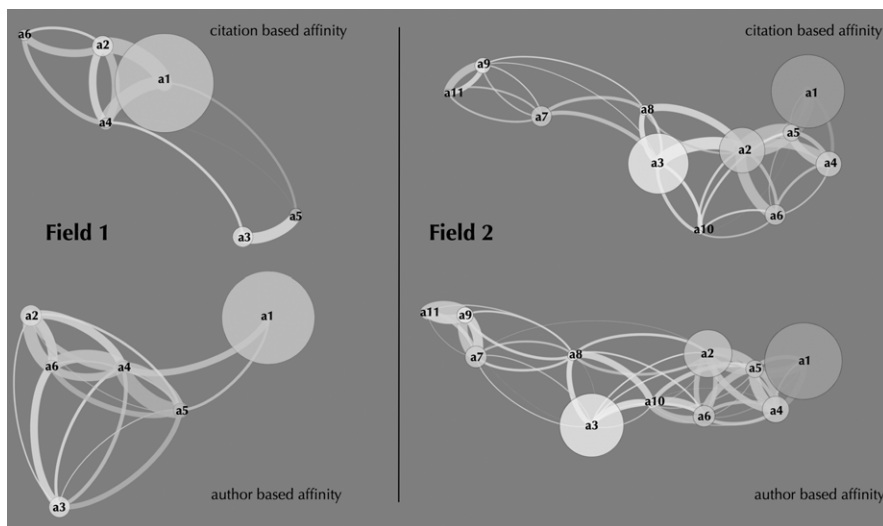
FIG. 7. Affinity diagrams for citation-based association (top) and author-activity-based association (bottom) between topic areas in Field 1 (left) and in Field 2 (right). The affinity networks highlight positive associations between areas. Link colors indicate the source node of the directed links. Link directionality is indicated by link curvature: The link is curved with the flow in clockwise direction. The lay-out was produced using gephi and the Force Atlas 2 algorithm. Data tables with the exact values are provided in the supplementary material. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

is something that could be very useful for organic chemistry. That's why we started that. [. . .] That was, as I already mentioned, at the beginning of the 90s. At this time no broadly useable catalysts existed. At the time when we started, there existed no well-defined catalyst that had a high tolerance for functional groups. [. . .] the reason to do first studies [was] motivated by the application, applicability in natural product synthesis. So, that was the beginning."

This group has become one of the major contributors to the field, active both in catalyst development and applications of those catalysts in organic synthesis. The major driving force for the strong growth of field 1 after the mid-1990s has been the availability of a couple of effective catalysts and their application in organic synthesis.

Furthermore, we note that in Field 1 area *3:Organometallics & Applications in Inorganic Chemistry* has a particularly low reference inclusion rate, setting it apart from the other areas.

The growth dynamics in Field 2 are more complex as four major topic areas (*1:Physical Chemistry & Chemical Physics*, *2:Surface Science*, *3:Materials Science*, and *9:Inorganic Chemistry*) show significant growth and share the majority of publication output. Relative growth is most pronounced for areas 2 and 3. The three research groups included in our ethnographic field study were active primarily in area 1 and area 5, hence we heard accounts about the growth of research areas and associated community dynamics based on experiences of researchers active in those two topic areas. Work in topic area 1 has a tradition reaching back into the 1980s and was fairly established by the beginning of the 1990s with almost 300 articles per annum in our data set (Figure 6). It has seen significant growth in the 1990s, as research groups in the community have increasingly made use of synchrotron radiation to study geometric and electronic properties of free

clusters. Those involved in research in topic area 5 focus on dynamic processes when clusters are submitted to high-intensity laser light and pulsed laser light. Based on our data set this area is much smaller than area 1, and is stagnating in size. However, "on the ground" excitement at the end of the first decade in the 2000s is great due to a new generation of radiation facilities (x-ray free electron lasers, or XFELs). A handful of these expensive facilities are under construction worldwide and coming online. At this time the user base is still very small and access very competitive, but all of the three groups in our field study had been involved in first pioneering experiments at these new facilities. This suggests that this area of research within Field 2 is viable and further evolving.

The reference inclusion rates are highest for topic areas *2:Surface Science*, *4:Physical Chemistry & Chemical Physics*, and *5:Dynamic Processes*, and lowest for areas *7:Inorganic Chemistry*, *8:Catalysis*, *9:Inorganic Chemistry*, and *11:Bio-Inorganic Chemistry*.

In the next two sections we investigate whether we can detect positive associations or affinities between topic areas based on their topical relatedness or their social connectedness (see the Topic Area Association and Affinity Networks section for definitions).

Topical relatedness between topic areas. An affinity between two topic areas in terms of citation indicates some form of topical relatedness beyond a common background of topical connectivity present in the field. The respective affinity networks for Field 1 and Field 2 are depicted in Figure 7, and the full data tables of positive or negative association values (residuals) are provided in the supplementary material. We find the following patterns of topical relatedness for the two fields summarized in the affinity diagrams:

In Field 1, area *4:Polymer Chemistry/Organometallics* and area *2:Polymer Chemistry* stand out as receiving disproportionally many citations from other areas. Area 1, which dominates all other topic areas in size, disproportionally often cites those two areas, and detailed inspection of the document clusters in area 1 reveals that some polymer chemistry document clusters have been subsumed into area 1, offering an explanation for this relative preference for polymer topic areas for citations from documents in area 1. The smaller areas seem to fall into two distinct sets:

[Set 1]   2:Polymer Chemistry, 4:Polymer Chemistry/Organometallics, 6:Polymer Chemistry/Molecular Catalysis
[Set 2]   3:Organometallics & Applications in Inorganic Chemistry, 5:Development of Supported Catalysts

In Field 2, area *2:Surface Science* stands out as a central knowledge resource in the field receiving disproportionately many citations from six other topic areas. Overall the network has an elongated structure of overlapping groupings of areas. One can view it as two sets of areas at the extreme ends of the structure, that connect in the middle through areas 2, 3, and 8.

[Set 1]   1:Physical Chemistry & Chemical Physics, 4:Physical Chemistry & Chemical Physics, 5:Dynamic Processes, 6:Applied Surface Science, 10:Applied Physics
[Set 2]   7:Inorganic Chemistry, 9:Inorganic Chemistry, and 11: Bio-Inorganic Chemistry
[Set 3 (connecting)]   2:Surface Science, 3:Materials Science, and 8:Catalysis

Social connectedness of topic areas.   In this section, we explore the preference that authors who publish in one area have for publishing also in other areas. We suggest that this kind of association between two topic areas implies a form of social connectedness as those authors interact with their colleagues, for example, at conferences and workshops or through peer review, and thereby facilitate the informal flow of knowledge and social mobility between topic areas. The resulting association heat maps and affinity networks for Field 1 and Field 2 are depicted in Figure 7.

We observe for Field 1 a relative isolation of area *1:Catalyst Development and Organic Chemistry*. Also, authors in area *5:Development of Supported Catalysts* show a particularly strong preference for contributing also to area *4:Polymer Chemistry/Organometallics*.[15] Overall the global structure of topic area relationships depicted in the affinity network shows two sets of areas, one oriented toward polymer chemistry, and another oriented towards organometallics and inorganic chemistry. They are strongly coupled via the two polymer chemistry areas 4 and 6:

[Set 1]   2:Polymer Chemistry, 4:Polymer Chemistry/Organometallics, 6:Polymer Chemistry/Molecular Catalysis
[Set 2]   3:Organometallics & Applications in Inorganic Chemistry, 5:Development of Supported Catalysts, 4:Polymer Chemistry/Organometallics, 6:Polymer Chemistry/Molecular Catalysis

For Field 2 we obtain a more dense affinity network compared to the citation-based affinity network. We find especially strong mutual affinities between areas *9: Inorganic Chemistry* and *11:Bio-Inorganic Chemistry*, as well as between areas *7:Inorganic Chemistry* and *9:Inorganic Chemistry*, and between areas *4:Physical Chemistry & Chemical Physics* and *5:Dynamic Processes*, and between areas *2:Surface Science* and *5:Dynamic Processes* and between *6:Applied Surface Science* and *10:Applied Physics*. The network shows a similar pattern to the citation network, however suggesting a slightly different subdivision of areas, namely into three overlapping sets:

[Set 1]   1:Physical Chemistry & Chemical Physics, 2:Surface Science, 4:Physical Chemistry & Chemical Physics, 5:Dynamic Processes, 6:Applied Surface Science
[Set 2]   2:Surface Science, 3:Materials Science, 6:Applied Surface Science, 8:Catalysis, and 10:Applied Physics
[Set 3]   7:Inorganic Chemistry, 8:Catalysis, 9:Inorganic Chemistry, and 11:Bio-Inorganic Chemistry

Mapping Community Structures as Overlays on the Collaboration Network

The basic structural differences between the group collaboration networks in Field 1 and Field 2 have already been discussed in Velden et al. (2010), albeit for networks generated from nondisambiguated data. Those structural features persist after author name disambiguation, as can be seen in Figure 8. Specifically these are the fragmentation of the group collaboration network in Field 1 versus the integration of coauthor clusters in Field 2 into one large network component; and the lower rate of group collaboration in Field 1 versus Field 2. In Field 1 only 9.8% of clusters in the giant component of the coauthor network have one or more collaborative relationships to other clusters, whereas in Field 2 this number is more than twice as high, namely 22.9%. In the remainder of the section we create geographic and topical overlay maps by projecting information about the coauthor clusters onto the group collaboration network to visualize community structures.

Geographic overlay.   The geographic overlay maps in Figure 8 depict the dominant geographic affiliations for each coauthor cluster in the group collaboration network (see the Group Collaboration Network Overlays section). For both fields we observe geographical ordering in the group collaboration networks. A quantitative evaluation shows a preference for collaborations between groups with specific geographic affiliations (details and data are provided in supplementary material). The geographical ordering is most pronounced in Field 1, but it is present also in Field 2. For both fields we find that Asian groups and North American groups each show a strong preference for collaborating with groups from their own continent. European groups, on the other hand, show neutral propensity to collaborate with other European groups. In Field 1, European groups show a
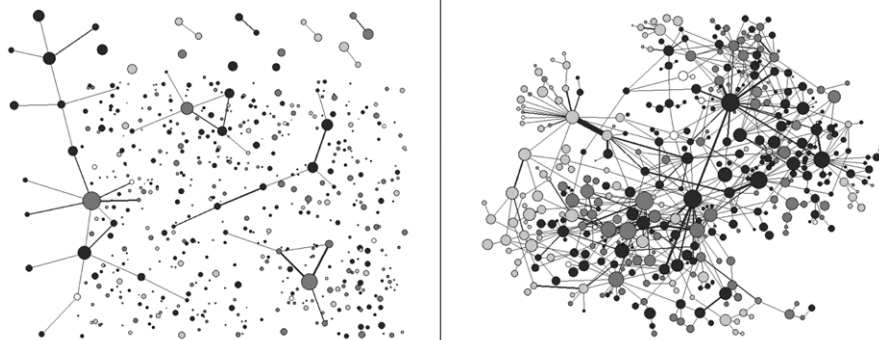
FIG. 8.   Intergroup collaboration network built from the giant component of the coauthor network in each field. For Field 1 (left) all coauthor clusters, with and without group collaboration links are shown. For Field 2 (right) only the giant component of the group collaboration network is shown, omitting 118 coauthor clusters with no or few collaboration links to other coauthor clusters. Node sizes represent number of coauthors in a cluster and link strength the number of coauthor relationships between authors in the respective clusters. Node color indicates the dominant geographical affiliation of a coauthor cluster (light yellow = Asia, blue = Europe, red = North America, light green = Asia-Europe, violet = Europe-North America, white = Other). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

strong preference for collaboration with mixed European/North American groups, and in Field 2, European groups show a preference for collaborating with Asian/European groups and European/North American groups.

Inspecting the geographic overlay maps in Figure 8 we observe that in Field 2 the overlay map highlights the internal clustering of Asian groups in two larger and three smaller clusters while there is a more cohesive interlinking of European groups with North American and mixed European/North American groups. The map suggests bridge-building role of European/North American affiliated groups, which represent 10% of the coauthor clusters in the group collaboration network. Indeed, the European/North American affiliated groups in the network have on average disproportionally many collaboration links (3.4), versus European groups (2.0), North American groups (1.9), or Asian groups (1.4), along with a preference for interlinking with European groups, North American groups, and other European/North American groups (data included in supplementary material).

Topical overlay.   The topic overlay maps show for each major topic area in a field the varying intensities with which groups are publishing in the respective area (see the Group Collaboration Network Overlays section). We make the following observations:

The topic overlay map in Figure 9 highlights the fact that most of the groups in field 1 are active in the topic area *1:Catalyst Development and Organic Chemistry*. Collaborating groups often have their activity foci in the same topic areas, but occasionally groups collaborate that have their activity focus in different fields. However, the collaboration network is too small and fragmented for a meaningful comparison to the topic area affinity structure presented in the previous section.

The topic overlay map in Figure 10 echoes the topic area substructure revealed for Field 2 with the help of the affinity
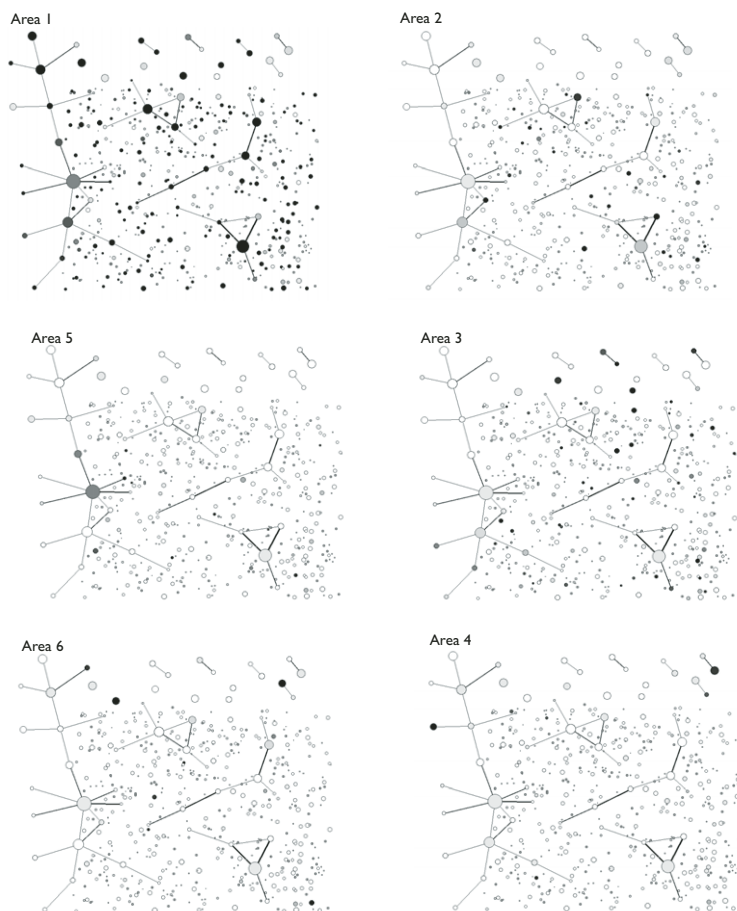
networks. Coauthor clusters with high activity in topic area *2:Surface Science* bridge two disparate regions of the group collaboration network: one region in the upper right side that includes on the upper left the majority of coauthor clusters with high activity in topic areas *7:Inorganic Chemistry*, *9:Inorganic Chemistry*, and *11:Bio-Inorganic Chemistry* and to the right the majority of coauthor clusters with high activity in topic areas *3:Materials Science*, *6:Applied Surface Science*, *8:Catalysis*, and *10:Applied Physics*. On the opposite, lower left side of the collaboration network is a region that includes the majority of coauthor clusters with high activity in topic areas *1:Physical Chemistry & Chemical Physics*, *4:Physical Chemistry & Chemical Physics*, and *5:Dynamic Processes*. There are deviations from this overall pattern, however, in the sense that there is some "fuzziness" in the mapping of some of the topic areas onto the collaboration network. Some coauthor clusters that are very active in an area are located at opposite ends of the collaboration network, see, for example, areas 2, 3, 6, and 7. On inspection of the geographically labeled collaboration network in Figure 8, this fuzziness could be explained with geography as a competing ordering principle in the collaboration network that is particular strong for Asian research groups in the field.

Discussion

Below we discuss the results, focusing on field delineation (Challenges of Field Delineation), on the extraction of community structures of research specialties (Community Structures Within Research Specialties), and on the utility of this for scaling-up ethnographies (Scaling-up of Ethnographies).

Challenges of Field Delineation

Depending on the field, defining a lexical query that captures relevant publications with high accuracy is

FIG. 9. Topic overlay on collaboration network of Field 1. Nodes represent coauthor clusters, links intergroup collaboration. Node color indicates relative intensity of publication activity of a cluster in the topic area ranging from white (= 0%) to black (> 90%). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

challenging. One of the two fields in our study is defined by a single, standardized term that is very reliably used in publications in this field. Hence, we consider the data set of Field 1 as a rather accurate representation of the publication output in this field (to the extent that it is captured by the WoS database). Indeed, we find that the growth curve of annual publication numbers and the topic area-specific growth curves match the popular historical account of the field and its evolution, in particular the onset of rapid growth in the mid-1990s and the dynamics of growth driven by the topic area within the field, which focuses on the design of catalysts and their application in organic chemistry.

In Field 2, the temporal evolution of the field is dominated by growth in four major topic areas located across the entire area affinity network. So far our ethnographic field study provides only a preliminary account of field growth enabled by the technical advancement of expensive, shared experimental facilities that research groups in this field have access to. Overall, Field 2 grew less dynamically and especially the growth curve of area 1 shows saturation suggesting a mature field. We would expect to be able to identify additional subcommunity specific factors that shape the growth of the field in future research led by insights into the community structures in this field.
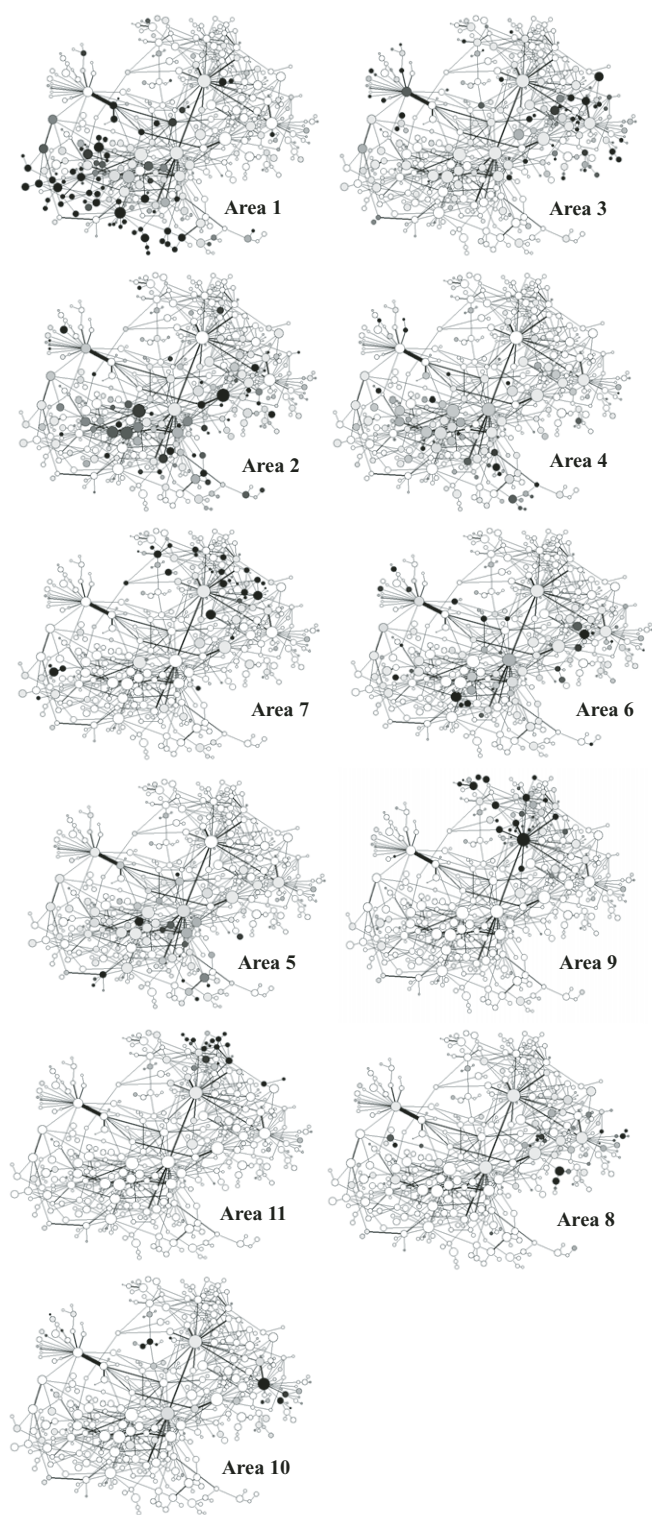
Field 2	topic area description	most popular journals (# of articles)
Area 1 (10632 articles)	Physical Chemistry Chemical Physics	J OF CHEMICAL PHYSICS (2435) J OF PHYSICAL CHEMISTRY A (1262) CHEMICAL PHYSICS LETTERS (864) PHYSICAL CHEMISTRY CHEMICAL PHYSICS (390) J OF PHYSICAL CHEMISTRY B (277)
Area 3 (8902 articles)	Materials Science	J OF PHYSICAL CHEMISTRY B (503) LANGMUIR (416) JACS (363) CHEMISTRY OF MATERIALS (299) J OF PHYSICAL CHEMISTRY C (264)
Area 2 (6819 articles)	Surface Science	J OF CHEMICAL PHYSICS (860) PHYSICAL REVIEW B (535) SURFACE SCIENCE (367) J OF PHYSICAL CHEMISTRY A (361) CHEMICAL PHYSICS LETTERS (340)
Area 4 (4028 articles)	Physical Chemistry Chemical Physics	J OF CHEMICAL PHYSICS (512) CHEMICAL PHYSICS LETTERS (319) J OF PHYSICAL CHEMISTRY A (244) PHYSICAL REVIEW B (237) J OF MOLECULAR STRUCTURE-THEOCHEM (95)
Area 7 (3394 articles)	Inorganic Chemistry	INORGANIC CHEMISTRY (464) ORGANOMETALLICS (442) J OF ORGANOMETALLIC CHEMISTRY (391) J CHEM SOC - DALTON TRANSACTIONS (191) INORGANICA CHIMICA ACTA (172)
Area 6 (3370 articles)	Applied Surface Science	NUCL. INSTR. & METH. IN PHYS RES SECT. B (341) PHYSICAL REVIEW B (253) SURFACE SCIENCE (196) APPLIED SURFACE SCIENCE (136) J OF APPLIED PHYSICS (121)
Area 5 (2853 articles)	Dynamic Processes	J OF CHEMICAL PHYSICS (250) PHYSICAL REVIEW A (227) PHYSICAL REVIEW B (216) Z. F. PHYS. D-ATOMS MOL. & CLUSTERS (201) PHYSICAL REVIEW LETTERS (179)
Area 9 (2704 articles)	Inorganic Chemistry	INORGANIC CHEMISTRY (345) JACS (157) PHYSICAL REVIEW B (117) DALTON TRANSACTIONS (96) INORGANICA CHIMICA ACTA (93)
Area 8 (1954 articles)	Catalysis	J OF CATALYSIS (199) J OF PHYSICAL CHEMISTRY B (136) APPLIED CATALYSIS A-GENERAL (98) CATALYSIS LETTERS (81) J OF PHYSICAL CHEMISTRY (76)
Area 11 (1501 articles)	Bio-Inorganic Chemistry	BIOCHEMISTRY (228) J OF BIOLOGICAL CHEMISTRY (141) JACS (124) INORGANIC CHEMISTRY (95) J OF BIOLOGICAL INORGANIC CHEMISTRY (64)
Area 10 (1113 articles)	Applied Physics	NUCL. INSTR. & METH. IN PHYS RES SECT. B (141) PHYSICAL REVIEW B (83) APPLIED PHYSICS LETTERS (75) J OF APPLIED PHYSICS (67) J OF NON-CRYSTALLINE SOLIDS (48)

FIG. 10. Topic overlay on collaboration network of Field 2. Nodes represent coauthor clusters, links intergroup collaboration. Node color indicates relative intensity of publication activity of a cluster in the topic area ranging from white (= 0%) to black (> 90%). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

We have put forward the suggestion that the comparison of reference inclusion rates between topic areas may be helpful to identify potential areas of overlap with other fields. The fact that in both fields most topic areas with low reference inclusion rates are related to inorganic chemistry may point to a discipline-specific referencing pattern (e.g., higher rates of references to older literature or application of specific methods developed in another field) rather than indicating overlap with another closely related field. On the other hand, if we look at the topic overlay on the group collaboration network for Field 2 we observe that those topic areas with lowest reference inclusion rates are mapped to the outer regions of the group collaboration network, whereas the topic areas with higher reference inclusion rates are mapped onto the center of the network. This would suggest that lower reference inclusion rates do indicate the relative marginality of a topic area within the field. At this point, we regard the evidence for the interpretation of variations in reference inclusion rates as inconclusive.

Finally, the methods we have developed to optimize precision and recall of the lexical query are heuristic in character. The method for optimizing recall relies on subjective decisions whether a cluster of publications in the self-citation network of a renown researcher should be considered part of the targeted field or not, and whether the set of individuals selected represents the field with sufficient breadth. Similarly, the decision whether to exclude certain terms in the lexical query to increase precision is ambiguous. However, in contrast to approaches to field delineation that require the ad-hoc setting of numerical cut-off parameters (e.g., Zitt & Bassecoulard, 2006), the decision is supported by the ability to trace back and link the antagonism between topic areas in a data set to those topic areas' topical focus and hence to terms included in the lexical query. Hence, the field delineation is guided by the social topology of the field that emerges from researchers' choices to cross field boundaries with their research and publication activities. A mixed-method approach further supports this process by providing an understanding of actors' distinctions with regard to conceptual definitions of research topics developed in the ethnographic field study.

Community Structures Within Research Specialties

The topic areas extracted from the two research specialties by clustering of the citation network reflect (sub)disciplinary orientations, along with specific research foci. This multi-disciplinary composition of each of the two fields corroborates insights from our ethnographic studies (from participant accounts, observations and scrutiny of conference programs) that scientists from a range of different (sub)disciplinary backgrounds[16] are active in the same research specialty.

The analysis of topical relatedness and social connectivity between topic areas reveals pronounced patterns of affinity and antagonism. More clearly than a mere citation network (as depicted in Figure 6) the affinity networks

suggest topical and social closeness and distance between topic areas within a field. These patterns vary depending on whether we consider intercitation between areas or author activity across areas; however, the global structures are very similar. They point to subcommunities within research specialties that reflect differences in disciplinary orientation and we suggest also in research practices. We found that some of the antagonisms detected between topic areas match with sentiments expressed by our field study participants. For example, members of an experimental physics group in Field 2 who are mostly active in areas 1 and 5 commented on their collaboration with a physical chemistry group whose work is focused on in area 3. They expressed reservations about the kind of research pursued by the other group and invoked a "dirty chemistry" trope referring to chemical synthesis and ill-defined research objects in contrast to the pure, well-defined objects ("every atom counts") that they themselves, as physicists, preferred to study.

Finally, for a field such as Field 2, where a substantial proportion of the groups is part of the collaboration network, the topic overlay map provides a compelling insight into the ordering of the group collaboration network. The dominant feature of the topic overlay map for Field 2 is the ordering of the group collaboration network first by topic area, and then by topic area affinities. The distribution of coauthor clusters in the collaboration network as depicted in Figure 10 mirrors the global structure of the affinity networks obtained from citations and author activity. This implies that groups with a research focus in an area tend to collaborate with groups that have their focus of activity in the same area or a related, positively associated area.

For some topic areas we find that geographic ordering of the group collaboration network competes with this topical ordering. Taking into account that data about the growth of Field 2 show a recent dramatic increase in share of publication output in particular by researchers in China, we hypothesize that this competing geographic ordering may indicate a yet incomplete integration of these research groups into the global science system (Veugelers, 2009).

The affinity networks together with the topical overlay map suggest a subdivision of Field 2 into three smaller subcommunities and it could be argued that they represent research specialties in their own right. However, these subcommunities are all but disjoint. Almost 50% of groups and the large majority of authors are joined together in the giant component of the coauthor network. This implies at least overlap between those subcommunities in the form of either direct collaboration between groups in different subcommunities or individuals participating in multiple communities either concurrently or over time. Furthermore, as evidenced by the topical overlay map as well as ethnographic observations, groups collaborate across even distant topic areas, conferences in the field bring these different subcommunities together, and field study participants acknowledge the similarity of research questions pursued by researchers across these communities. Hence, we suggest reserving the term research specialty for this largest group of

researchers that collectively produce new knowledge in a domain, and refer to these smaller entities as topic areas (if defined based on the citation network of publications) or subcommunities (when referring to the researchers involved).

Scaling-Up of Ethnographies

Ethnographic field research is expensive in terms of time. Instead of relying on a statistically representative sample, ethnography relies on the strategic selection of field sites. Network analysis of the relationships between actors in the domain of interest can support this selection process (Howard, 2002). We suggest that the community structures revealed in this analysis relate to differences in disciplinary orientation and research practices, underlining the fact that fields at the specialty level cannot be assumed to be homogeneous in this regard. Therefore, network analysis cannot only highlight particular actors of interest, but through the mapping of community structures as demonstrated in this article, support the strategic selection of field sites for study and comparison.

To zoom in on one of the subcommunities revealed in our network analysis will usually reduce the diversity of (sub)disciplinary research practices of the research groups included. However, to assume in turn that these subcommunities are largely homogeneous in their research practices would be gravely erroneous. Instead, the analytic unit we are offered is that of a subcommunity that can be distinguished in its (sub)disciplinary orientation from other subcommunities within the field, but whose core of closely communicating research groups still includes a range of (sub)disciplinary backgrounds that come along with distinct research practices. For example, from our field studies we know that the subcommunity in Field 2 that is composed of research groups active in topic areas 1, 4, and 5 brings together groups with backgrounds in physical chemistry, experimental atomic and molecular physics, and theoretical physics. This finding echoes observations in an exploratory study of the extent and limitations of cross-disciplinarity within a research specialty studying aspects of molecular motors in bio-nanotechnology (Rafols & Meyer, 2006). To the extent we could determine, such a subcommunity does not further subdivide into subcommunities that would be more homogeneous with regard to disciplinary backgrounds and research practices. This seems to be either because of lack of critical mass (there are just too few theoretical physics groups working on this type of research questions to constitute a community) or because the interaction of groups with complementary skills is essential to produce results that are considered as legitimate and meaningful (such as the need to combine theoretical models with empirical data).

This implies that the ethnographer is still faced with a complex unit of analysis to study and develop a deeper understanding of. However, we suggest that the kind of community analysis introduced here supports better-informed choices of field sites supporting the selection of units of analysis that correlate with a closely interacting community, and the design of comparative studies that may throw distinct differences into relief. The scaling-up of ethnographic observations is a critical step for comparative science studies that aim for a systematic understanding of the interaction of research practices and the social organization of scientific communities. Behaviors such as openness and sharing of knowledge between research groups are shaped by the competition dynamics within a field that in turn are shaped by a number of research-specific epistemic and material factors (Velden 2013). The advantage and relevance of having network analytic tools that help identify subcommunity structures is that we can control for epistemic and social differences introduced by subcommunity structures.

Conclusions

The scientific community that collectively produces the knowledge base of a research field is an important unit of analysis for understanding how epistemic and social factors influence, for example, communication behaviors, but it is challenging to capture. We demonstrate here that publication networks can be "mined" as part of a mixed-method approach to reveal and map the complexities of a scientific community. We argue that insights into the community structures within a research field are crucial for the strategic selection of ethnographic field sites. By better understanding community structures of scientific fields and hence the context of local ethnographic observations, we increase confidence about the domain of validity of ethnographic results, and hence their value for systematic comparisons across scientific fields.

Future work will be directed at adding temporal resolution of community structures for improved guidance in the selection of field sites and to track the evolution of community structures over time. Also, based on strategic sampling of field study sites supported by insights into community structures, we plan to investigate variations of openness and sharing behaviors within fields to systematically study the influence of epistemic factors.

Acknowledgments

References

Aksnes, D., Olsen, T., & Seglen, P. (1999, July). Inadequacy of a journal-based research field delineation. In Proceedings of Seventh Conference of the International Society for Scientometrics and Informetrics. Colima, Mexico.

Baus, D. (2010). Cultural exchange in a heterogeneous research field: Approaching scientific culture with anthropological thought. Spontaneous Generations: A Journal for the History and Philosophy of Science, 3(1), 95–104.

Beaulieu, A., Scharnhorst, A., & Wouters, P. (2007). Not another case study: A middle-range interrogation of ethnographic case studies in the exploration of e-science. Science, Technology & Human Values, 32(6), 672.

Becher, T., & Trowler, P.R. (2001). Academic tribes and territories. Buckingham: Society for Research into Higher Education (SRHE) & Open University Press.

Birnholtz, J.P., & Bietz, M.J. (2003). Data at work: Supporting sharing in science and engineering In "GROUP"03 (pp. 339–348). New York: ACM Press.

Börner, K., Sanyal, S., & Vespignani, A. (2007). Network science. Annual Review of Information Science and Technology, 41(1), 537–607.

Boyack, K.W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? JASIST, 61(12), 2389–2404.

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

Cana, M. (2010). Open access repositories in the cultural configuration of disciplines: Applying actor-network theory to knowledge production by astronomers and philosophers of science. PhD thesis, Rutgers University-Graduate School-New Brunswick.

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

Cronin, B. (2003). Scholarly communication and epistemic cultures. New Review of Academic Librarianship, 9(1), 1–24.

Fry, J., & Talja, S. (2004). The cultural shaping of scholarly communication: Explaining e-journal use within and across academic fields. Proceedings of the American Society of Information Science and Technology, 41, 20–30.

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.

Galison, P. (1997). Image and logic: A material culture of microphysics. University Of Chicago Press.

Geels, F.W. (2007). Feelings of discontent and the promise of middle range theory for STS: Examples from technology dynamics. Science, Technology, & Human Values, 32(6), 627–651.

Geiger, R., & Ribes, D. (2011). Trace ethnography: Following coordination through documentary practices. In System Sciences (HICSS), 2011 44th Hawaii International Conference pp. 1–10.

Gläser, J. (2006). Wissenschaftliche Produktionsgemeinschaften—die soziale Ordnung der Forschung, Vol. 906. Frankfurt/New York: Campus Verlag.

Gläser, J., & Laudel, G. (2007). The social construction of bibliometric evaluations In R. Whitley & J. Glдser (Eds.), The changing governance of the sciences (pp. 101–123). The Netherlands: Springer.

Gläser, J., & Laudel, G. (2001). Integrating scientometric indicators into sociological studies: methodical and methodological problems. Scientometrics, 52(3), 411–434.

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.

Guimera, R., Uzzi, B., Spiro, J., & Amaral, L. (2005). Team assembly mechanisms determine collaboration network structure and team performance. Science, 308(5722), 697–702.

Hellsten, I., Lambiotte, R., Scharnhorst, A., & Ausloos, M. (2007). Self-citations, co-authorships and keywords: A new approach to scientists' field mobility? Scientometrics, 72, 469–486.

Hine, C. (2008). Systematics as cyberscience: Computers, change, and continuity in science, Cambridge: MIT Press.

Howard, P. (2002). Network ethnography and the hypermedia organization: New media, new organizations, new methods. New Media & Society, 4(4), 550–574.

Huang, C., Notten, A., & Rasters, N. (2011). Nanoscience and technology publications and patents: A review of social science studies and search strategies. The Journal of Technology Transfer, 36(2), 145–172.

Jamali, H., & Nicholas, D. (2008). Information-seeking behaviour of physicists and astronomers. In "Aslib Proceedings," pp. 444–462.

Jansen, D., von Gиrtz, R., & Heidler, R. (2010). Knowledge production and the structure of collaboration networks in two scientific fields. Scientometrics, 83(1), 219–241.

Kling, R., & McKim, G. (2000). Not just a matter of time: Field differences and the shaping of electronic media in supporting scientific communication. Journal of the American Society for Information Science, 51(14), 1306–1320.

Kling, R., McKim, G., & King, A. (2003). A bit more of it: Scholarly communication forums as socio-technical interaction networks. Journal of the American Society of Information Science and Technology, 54, 47–67.

Kling, R., Spector, L.B., & Fortuna, J. (2004). The real stakes of virtual publishing: The transformation of E-Biomed into PubMed central. Journal of the American Society for Information Science and Technology, 55(2), 127–148.

Knorr Cetina, K. (1999). Epistemic cultures—How the sciences make knowledge. Harvard University Press.

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

Lambiotte, R., & Panzarasa, P. (2009). Communities, knowledge creation, and information diffusion. Journal of Informetrics, 3(3), 180–190.

Lancichinetti, A., & Fortunato, S. (2009). Community detection algorithms: a comparative analysis. Physical Review E, 80(5), 056117.

Laurens, P., Zitt, M., & Bassecoulard, E. (2010). Delineation of the genomics field by hybrid citation-lexical methods: interaction with experts and validation process. Scientometrics, 82(3), 647–662.

Lievrouw, L.A. (1990). Reconceiling structure and process in the study of scholarly communication. In Borgman C.L. (Ed.), Scholarly communication and bibliometrics. Newbury Park, CA: Sage Publications.

Lievrouw, L., Rogers, E., Lowe, C., & Nadel, E. (1987). Triangulation as a research strategy for identifying invisible colleges among biomedical scientists. Social Networks, 9(3), 217–248.

Melin, G., & Persson, O. (1996). Studying research collaboration using co-authorships. Scientometrics, 36(3), 363–377.

Moed, H. (2005). Citation analysis in research evaluation. Dordrecht, the Netherlands: Springer.

Mogoutov, A., & Kahane, B. (2007). Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracking. Research Policy, 36(6), 893–903.

Moran-Ellis, J. (2006). Triangulation and integration: Processes, claims and implications. Qualitative Research, 6(1), 45–59.

Morris, S., & Van der Veer Martens, B. (2008). Mapping research specialties. Annual Review of Information Science and Technology, 42(1), 213–295.

Mulkay, M. (1977). Sociology of the scientific research community. In I. Spiegel-Rиsing & D. de Solla Price (Eds.), Science, technology, and society: A cross-disciplinary perspective (pp. 93–148). Beverly Hills, CA: SAGE Publications Inc.

Nentwich, M. (2005). Cyberscience: Modelling ICT-induced changes of the scholarly communication system. Information, Communication & Society, 8(4), 542–560.

Rafols, I., & Meyer, M. (2006). Knowledge-sourcing strategies for cross-disciplinarity in bionanotechnology. SPRU Electronic Working Paper Series.

Rafols, I., Porter, A.L., & Leydesdorff, L. (2010). Science overlay maps: A new tool for research policy and library management. Journal of the American Society of Information Science, 61(9), 1871–1887.

Rosvall, M., & Bergstrom, C. (2008). Maps of information flow reveal community structure in complex networks. Proceedings of the National Academy of Sciences, 105, 1118.

Seglen, P., & Aksnes, D. (2000). Scientific productivity and group size: A bibliometric analysis of norwegian microbiological research. Scientometrics, 49(1), 125–143.

Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2009). Comparative study on methods of detecting research fronts using different types of citation. Journal of the American Society for Information Science and Technology, 60(3), 571–580.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for information Science, 24(4), 265–269.

Strotmann, A., & Zhao, D. (2010). Combining commercial citation indexes and open-access bibliographic databases to delimit highly interdisciplinary research fields for citation analysis. Journal of Informetrics, 4(2), 194–200.

Van den Besselaar, P. (2000). Communication between science and technology studies journals: A case study in differentiation and integration in scientific fields. Scientometrics, 47(2), 169–193.

Van House, N. (2002). Digital libraries and practices of trust: Networked biodiversity information. Social Epistemology, 16(1), 99–114.

Velden, T. (2013, February). Explaining field differences in openness and sharing in scientific communities. CSCW 2013: Computer Supported Cooperative Work, San Antonio, TX.

Velden, T. (2011a). A methodology for studying field differences in scientific communication—Explaining openness and sharing in two scientific communities in the chemical and physical sciences. PhD thesis, Cornell University.

Velden, T., Haque, A., & Lagoze, C. (2011, June). Resolving author name homonymy to improve resolution of structures in co-author networks. In Joint Conference on Digital Libraries (JCDL)'11, Ottawa, Ontario, Canada.

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.

Veugelers, R. (2009). Towards a multipolar science world: Trends and impact. Scientometrics, 82(2), 439–456.

Walsh, J.P., & Bayma, T. (1996). Computer networks and scientific work. Social Studies of Science, 26(3), 661–703.

Whitley, R. (2000). The intellectual and social organization of the sciences. New York: Oxford University Press.

Zitt, M. (2005). Facing diversity of science: A challenge for bibliometric indicators. Measurement, 3(1), 38–49.

Zitt, M., & Bassecoulard, E. (2006). Delineating complex scientific fields by a hybrid lexical-citation method: An application to nanosciences. Information Processing & Management, 42(6), 1513–1531.

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

Appendix: Explanation of Terms Used

Research field / research specialty: By *field* we here refer to a research field at the level of a research specialty. Whereas the terms *discipline* (such as "chemistry" or "physics") or *subdiscipline* (such as "organic chemistry") refer to larger, bureaucratized units of teaching and academic employment, *research specialty* refers to the knowledge base generated collectively by a self-organizing, loosely coordinated community of researchers (Gläser 2006). The scientific community of researchers that is active in a research specialty can have a variety of sizes, from a few researchers to hundreds or perhaps even thousands of researchers (depending on the size of the smallest collective unit needed to produce new knowledge claims). As discussed in the article, research specialties (as a cognitive concept) overlap, as do the communities of scientists contributing to the knowledge bases of research specialties.

Scientific community: This is the collective of researchers active in a research specialty. Membership in a scientific community is based on self-perception and expressed in research activities oriented towards contributing to the shared knowledge base of that research specialty (Gläser, 2006).

(Institutional) research group: A *research group* is an organizational unit at a university or research institute, usually of collocated researchers led by a principal investigator (PI). Typically, the PI is the only "constant" presence in the group over time. In our experience, research groups oftentimes actively contribute to several research specialties. Because not all group members of an institutional research group are doing research in the same research field, a coauthor cluster extracted from a field specific data set will capture only a partial view of an institutional research group. Instead, in the kind of sciences that we have been studying so far—characterized by a median of around four coauthors per publication—the coauthor clusters that we extract with Rosvall's algorithm come closest to capturing functional research groups (defined below).

Functional research group: To be distinguished from a research group as an organizational unit of an academic institution is the collective of researchers that Seglen and Aksnes (2000) have termed *functional research groups.* These typically are research collectives including a collocated PI-led group of researchers extended by closely cooperating domestic or international colleagues or visiting scientists. Most coauthor clusters extracted in our study will not correspond to institutional research groups, but represent functional research groups. For lack of temporal resolution, the coauthor clusters in our study are not a direct representation of a functional research group as instantiated at any particular point in time, but represent its accumulated bibliographic footprint during a 20-year time frame.

Principal investigator (PI): This is a term borrowed from U.S. American funding agency parlance. By this term we refer to the senior leader of an institutional research group or a functional research group; in an institutional academic setting this is typically a researcher at a professorial rank.

Endnotes

1. Terms in bold, italicized fonts are explained in the appendix.
2. See explanation of term in the appendix.
3. We are hesitant, however, to adopt the term *network ethnography* as it suggests a misleading identification of the research object in the ethnography, scientific communities, with networks. Empirical evidence suggests that scientific communities may contain networks, but are not identical with and constituted by networks (Gläser, 2006).
4. For a review of this newly emerging field see, for example, Börner et al. (2007).
5. See explanation of term in the appendix.
6. Given they were continuously active over the entire time period, and not just small because of a temporally limited participation in the field.
7. See explanation of term in the appendix.
8. The clustering of citation networks to study research specialties has been pioneered by Small (1973) based on a subset of highly cited documents and using co-citation instead of direct citations. Recently, Shibata et al. (2009) conducted a comparison of direct citation, co-citation, and bibliographic coupling to detect research fronts, and concluded that direct citation outperformed the other two approaches in detecting research fronts as it identified large and emerging clusters earlier. In addition, they observed that

direct citation networks had the largest clustering coefficients, an indication of higher content similarity. At the time we developed our approach this finding suggested that direct citation is a good choice for revealing the topical substructure in a research specialty. Since that time a comprehensive analysis of a large bio-medical corpus by Boyack and Klavans (2010) has delivered conflicting results, concluding that bibliographic coupling and co-citation analysis outperform direct citation in identifying research fronts. In future work it may be of interest to assess the difference that the exact construction of the network to be clustered makes for the extraction of topic areas and subsequent community structure analysis.

9. Because our focus here is on teasing out the interrelatedness of areas, we do not include the source area as one of the potential target areas. Hence, we here disregard self-referentiality although it may represent a potentially interesting property of topic areas.

10. In the fields studied here, these hub nodes can be typically identified as the PIs of research groups and represent the research continuity of those research groups. If we included all publications coauthored by any member of a coauthor cluster we would introduce noise in the form of publications that cluster members made while they were members of other groups.

11. Homonymy occurs when multiple authors with the same last name and initials get merged into a single author entity.

12. This subject filtering is used to exclude irrelevant document sets retrieved from other fields of research because of alternate meaning of terms used in the query.

13. Infomap_undir, available from the homepage of Martin Rosvall (http://www.tp.umu.se/~rosvall/code.html).

14. An alternative way for topic area identification would be the cognitive mapping approach that has been recently proposed by Rafols et al. (2010) that projects publication data onto the science map constructed from inter-citation of ISI WoS subject categories. However, these subject categories are one-step removed from the journal titles as they classify and group journals. Most journal titles are fairly descriptive, so we here relied on a direct interpretation of these journal titles rather than the derived subject categorization.

15. A plausible link given that polymers are one of the popular supports for catalysts in Field 1.

16. We identify disciplinary background with a scientist's formal training and nominal faculty appointment.