Karlsruhe Reports in Informatics 2012,11
Edited by Karlsruhe Institute of Technology,
Faculty of Informatics
ISSN 2190-4782

Static and Dynamic Aspects of Scientific Collaboration Networks

Christian Staudt, Andrea Schumm, Henning Meyerhenke,
Robert Görke, Dorothea Wagner

2012

KIT – University of the State of Baden-Wuerttemberg and National Research Center of the Helmholtz Association

Static and Dynamic Aspects of Scientific Collaboration Networks

Christian Staudt christian.staudt@student.kit.edu
Andrea Schumm andrea.schumm@kit.edu
Henning Meyerhenke meyerhenke@kit.edu
Robert Görke robert.goerke@kit.edu
Dorothea Wagner dorothea.wagner@kit.edu

Institute of Theoretical Informatics, Karlsruhe Institute of Technology (KIT), Am Fasanengarten 5, 76131 Karlsruhe, Germany

***Abstract*—Collaboration networks arise when we map the connections between scientists which are formed through joint publications. These networks thus display the social structure of academia, and also allow conclusions about the structure of scientific knowledge. Using the computer science publication database *DBLP*, we compile relations between authors and publications as graphs and proceed with examining and quantifying collaborative relations with graph-based methods. We review standard properties of the network and rank authors and publications by centrality. Additionally, we detect communities with *modularity*-based clustering and compare the resulting clusters to a ground-truth based on conferences and thus topical similarity. In a second part, we are the first to combine *DBLP* network data with data from the *Dagstuhl Seminars*: We investigate whether seminars of this kind, as social and academic events designed to connect researchers, leave a visible track in the structure of the collaboration network. Our results suggest that such single events are not influential enough to change the network structure significantly. However, the network structure seems to influence a participant's decision to accept or decline an invitation.**

I. Introduction

In scientometrics, the quantitative study of science, network analysis has become a prominent tool. *Coauthorship networks* have attracted interest both as *social networks* and as *knowledge networks*: They display the social structure of academia, while their bibliographic aspect allows conclusions about the structure of scientific knowledge. Accordingly, networks of this kind are the objects of ongoing research: Newman ([1], [2], [3], [4]), for example, studies properties of coauthorship networks in the realm of physics (*Los Alamos e-Print Archive*, *SPIRES*), mathematics (*Mathematical Reviews*), biomedical research (*Medline*) and computer science (*NCSTRL*), summarizing many statistical properties of coauthorship networks. Aspects like *connectedness*, *distance*, *degree distribution*, *centrality* and *community structure* are recurring themes in such studies. Where we follow up on these topics, we cite relevant related work in the respective sections of this paper.

Based on the extensive publication database *DBLP* [5], we model relations between authors and publications as graphs, mapping almost the entire field of computer science. This allows us to examine and quantify the collaborative relations between researchers using graph-based methods. We compile a graph in which edges link coauthors, as well as a bipartite author-paper graph. In the first part, we review standard properties of the network, rank authors and publications by centrality, and detect communities with *modularity*-based clustering. In the second part, we combine the network with seminar data provided by the *Schloss Dagstuhl* [6] conference center: The *Dagstuhl Seminars* assemble researchers with the goal of fostering (collaborative) work in cutting-edge areas of computer science. We examine whether such events leave a track in the structure of the collaboration network. For this purpose, we apply appropriate measures to a time-resolved version of the *authorship graph*.

We are the first to perform a joint analysis of the *Dagstuhl* and *DBLP* datasets, which allows us to study the impact of social/academic events on the time-evolution of the network structure. Our results suggest that a participant's decision to accept or decline an invitation can be predicted from the network data to some extent. While our analysis of the *DBLP* data mostly confirms properties of similar networks, the distribution of the number of coauthors differs from data reported in [4]. We also describe an approach to finding central researchers based on *eigenvector centrality* in the bipartite authorship graph, a combination that to our knowledge has not been used before. Additionally, we apply *modularity* clustering and compare the detected communities to a ground-truth defined by conferences, from which we infer distinct areas of research.

II. Preliminaries

A. Collaboration Network Model

As of 2011, *DBLP* covers about 1.5 million publications by 0.8 million authors. The earliest work dates from 1936, and we include all works up to 2009 in our analysis. We describe briefly how a coauthorship network is extracted from the publication database and represented as different types of graphs. The database associates publications and authors and thus provides two main relations, *authorship* and *coauthorship*, formalized as follows:

Def. 1. *Given the sets of authors* $\mathbf{A}$ *and publications* $\mathbf{P}$*, the* authorship *relation is defined as*

$$\forall \{a, p\} \in \mathbf{A} \times \mathbf{P} : \quad a \smile p \iff a \text{ is author of } p$$

The coauthorship *relation between two authors from* $\mathbf{A}$ *is defined as*

$$\forall \{a, b\} \in \mathbf{A} \times \mathbf{A} : a \frown b \iff \exists p \in \mathbf{P} : a \smile p \wedge b \smile p$$

From these, two graph representations of the network follow: A bipartite *authorship graph* (or author-paper graph) $G_{\mathbf{PA}}$, in which each publication is connected by edges to its authors; and a *coauthorship graph* $G_{\mathbf{A}}$, in which two authors are connected by an edge if they are coauthors of a joint publication.

Def. 2. *The* authorship graph *is a mapping from the sets of publications* $\mathbf{P}$ *and authors* $\mathbf{A}$ *to the node sets* $V_{\mathbf{P}}$ *and* $V_{\mathbf{A}}$*, resulting in a bipartite graph* $G_{\mathbf{PA}} = (V_{\mathbf{A}}, V_{\mathbf{P}}, E)$*, where*

$$\{v_a, v_p\} \in E \iff a \smile p$$

Def. 3. *The* coauthorship graph *is a mapping from the set of authors* $\mathbf{A}$ *to the node set* $V_{\mathbf{A}}$*, resulting in the graph* $G_{\mathbf{A}} = (V_{\mathbf{A}}, E)$*, where*

$$\{v_a, v_b\} \in E \iff a \frown b$$

While $G_{\mathbf{A}}$ is sufficient when focusing only on the social network of coauthors, $G_{\mathbf{PA}}$ preserves the publications as the cause of relations, as well as single-author publications. Table I shows the size of the graphs constructed from the full publication data set.

graph	n	m
$G_{\mathbf{PA}}$	2 296 586	3 775 881
$G_{\mathbf{A}}$	852 250	2 785 037

TABLE I
SIZE OF RESULTING GRAPHS

In order to determine whether events have effects detectable in terms of the network (Section IV), we also track groups of authors over the course of time, using a sequence of graphs in which each graph represents a current snapshot of the authorship relations. This *time-resolved* version of $G_{\mathbf{PA}}$ enables us to study the dynamics of the network: Let $t(p)$ denote the publication date of publication p. Then the publications from a time segment $[y, z], z > y$, are

$$\mathbf{P}_{[y,z]} := \{p \in \mathbf{P} : y \leq t(p) \leq z\}$$

The respective authors of these publications are

$$\mathbf{A}_{[y,z]} := \{a \in \mathbf{A} : \exists p \in \mathbf{P}_{[y,z]} : a \smile p\}$$

The graph sequence is constructed on the basis of a sliding time segment, with parameters width w and increment s:

Def. 4. *The* time-resolved authorship graph *is a sequence of graphs* $\mathcal{G}_{\mathbf{PA}}^{w,s}$ *where each graph in the sequence is constructed from the publications in* $\mathbf{P}_{[y,y+w]}$ *and the authors up to* $\mathbf{A}_{[y,y+w]}$ *using a sliding time segment with width* w *and increment* s*.*

Author nodes are aggregated over time, while publications are deleted for each step in the sequence. A time segment and increment of 1 year was chosen for the study in Section IV, the finest time resolution possible with *DBLP* data.

III. NETWORK PROPERTIES AND COMMUNITY STRUCTURE

A. General Network Properties

We briefly review some general properties of the collaboration network:

a) Connectedness: $G_{\mathbf{A}}$ features a *giant connected component* containing about 80% of all authors. (Giant components connecting up to 90% of all authors have previously been detected across scientific fields [3]). Aside from the 6% of the authors without collaborations, about 14 % of author nodes are distributed over a multitude of small components with few publications. We conclude that, in general, authors who have worked on several publications and were part of more than one collaborative team join the large connected component. In terms of average distances between researchers (6.58 for a sample), we confirm the previously reported *small world* property for the field of computer science [4] and *DBLP* in particular [7].

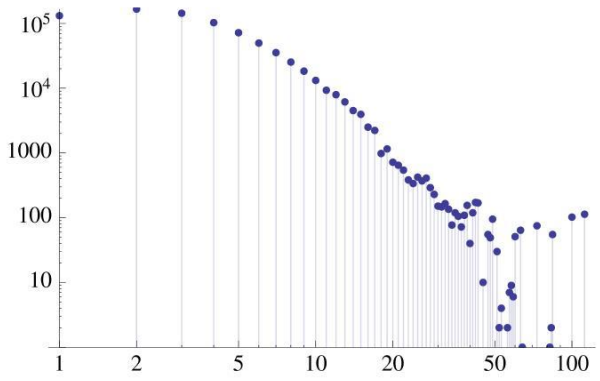


Fig. 1. Histogram of *core numbers* in $G_{\mathbf{A}}$ (x-axis: *core number*, logarithmic y-axis: frequency)

b) k-Core Structure: A k-core is a maximal subgraph in which each node is adjacent to at least k other nodes. k-cores refine the concept of connected components (which form the 1-core); *k-core decomposition* reveals nested, successively more cohesive layers of the graph. We assign each node a *core number*, the highest k for which there is a k-core containing the node. Figure 1 shows a histogram of the resulting *core numbers* in $G_{\mathbf{A}}$ with two logarithmic axes. The rather uniform sequence indicates uniform density and cohesiveness of the graph, showing that the network does not have strongly cohesive groups of authors embedded in shells of weakly connected authors [8]. A more extensive k-core analysis of a *DBLP*-based coauthorship network is presented in [9].

c) Degree Distribution: Node degree in $G_{\mathbf{A}}$ corresponds to the number of coauthors of each author. The degree distribution is highly skewed. It indicates a *scale-free network*, in which the frequency $P(k)$ of nodes with degree k follows a power law, i.e. $P(k) \sim k^{-\gamma}$, with coefficient $\gamma = 2.889$. Newman [3] reports a differing power-law degree distribution

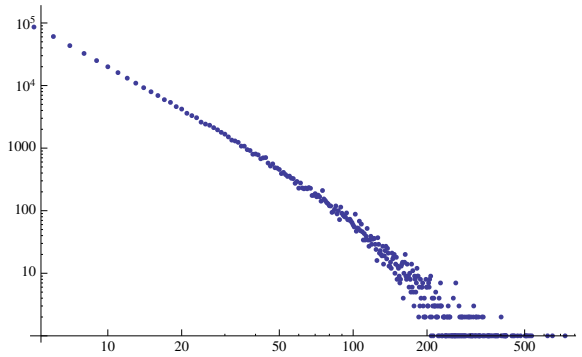


Fig. 2. Degree distribution in G_A (logarithmic x-axis: degree, logarithmic y-axis: frequency)

centrality $\cdot 10^{-5}$	author
9.76232	Diane Crawford
9.45441	Robert L. Glass
9.08697	Chin-Chen Chang
8.30777	Edwin R. Hancock
7.91401	Grzegorz Rozenberg
7.82901	Joseph Y. Halpern
7.75409	Sudhakar M. Reddy
7.69387	Philip S. Yu
7.50894	Moshe Y. Vardi
7.47370	Ronald R. Yager

TABLE II
TOP SEGMENT OF AUTHOR RANKING BY CENTRALITY

in the number of coauthors with $\gamma = 3.41$ for computer science, based on *NCSTRL*.

d) Summary: General properties indicate that the network of collaborations in computer science is in many respects a typical social network: It shows participation inequality (visible as a power-law degree distribution), with a few highly prolific authors and many smaller contributions. It also features a high degree of connectedness, a giant component, and mostly short paths between arbitrary pairs of nodes. Our observations are in agreement with the results of related studies (except for the degree power-law exponent), indicating that these properties are universal features of scientific collaboration networks.

B. Centrality

Centrality measures were formulated to identify nodes which are structurally prominent or influential, due to their position in the center of a network. *Betweenness* and *closeness centrality* have previously been applied to coauthorship graphs with the goal of identifying influential scientists in their respective fields ([4], [10]). Elmacioglu et al. report a ranking of prominent scholars by *closeness* and *betweenness* centrality [7]. As a rationale, it has been stated that authors with high *betweenness* are important intermediates for interactions or information flows, as it allows them to control such flows; high closeness is assumed to be an advantage for accessing or disseminating information [7]. However, it is not clear why academic influence should be understood mainly as the ability to mediate interactions. Furthermore, the network of information flow in academia and the network of coauthorship relations may be quite distinct. We therefore follow a different approach based on *eigenvector centrality* [11] in the bipartite authorship graph: It assumes that an author's influence is first of all proportional to the amount of publications. Additionally, the contribution of a paper to an author's centrality should be weighted depending on the centrality of the coauthors.

Def. 5. Eigenvector centrality*: Given a graph G with adjacency matrix A, we require a centrality score x_i of node v_i to be proportional to the scores of its neighbors:*

$$x_i = c \sum_{j=1}^{n} A(i,j)\ x_j \qquad c \neq 0$$

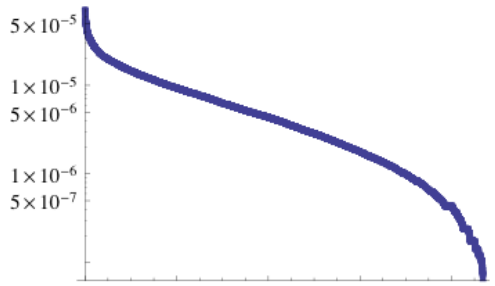
By the Perron-Frobenius theorem, there exists a nonnegative eigenvector x of A (satisfying $Ax = \frac{1}{c}x = \lambda x$) which corresponds to the larges eigenvalue λ. An entry x_i constitutes the desired centrality score for vertex v_i.

Modeling the collaboration network as the bipartite graph $G_{\mathbf{PA}}$ has the benefit that it allows us to assign a centrality score to a publication as a node, rather than just account for a publication as an edge attribute or weight in $G_{\mathbf{A}}$ [12]. Thus, our centrality scores express the concept that authors are central in the collaboration network to the extent that they have collaborated on central publications with other central authors. In this respect, the approach is similar to ranking webpages with the PageRank algorithm, where hyperlinks are treated as votes to the relevance of the target page and are weighted by the relevance of the source page.

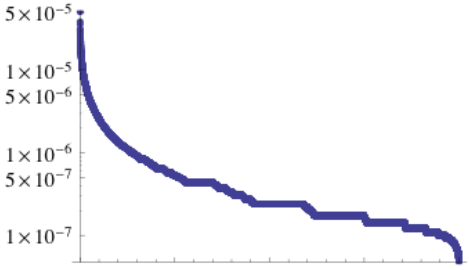
Figure 3 shows the distributions of centrality scores for authors and publications. Extreme values are less frequent, and the distribution does not exhibit a power law. Table II contains the top segment of an author ranking by our approach to centrality. (See [13] for a comparison to a purely productivity-based ranking of *DBLP* authors.) A ranking of publications places papers with unusually high author counts at the top, e.g. work on large supercomputing and database projects, and further study would be needed to interpret publication centrality properly. With respect to the evaluation in Section IV, it should also be noted that *Dagstuhl* seminar invitees have a significantly higher median *eigenvector centrality* score than other authors ($3.8 \cdot 10^{-6}$ versus $2.4 \cdot 10^{-7}$). We therefore propose that *eigenvector centrality* in bipartite author-paper networks is a promising approach for studying the role and impact of collaborating individuals in science, and might serve as an objective measure of influence in scientific publishing.

C. Modularity-driven Clustering

Graph clustering comprises a variety of methods for detecting natural communities in networks. Formally, it is concerned



(a) authors



(b) publications

Fig. 3. Centrality scores (logarithmic y-axis), sorted

with partitioning the node set into disjoint subsets (clusters), the result of which is called a *clustering*. The notion of a cluster is usually based on the *intra-cluster density versus inter-cluster sparsity* paradigm, according to which a clustering should identify groups of nodes which are internally densely connected, while only sparse connections exist between the groups. One of the primary measures of clustering quality based on this paradigm is *modularity* [14].

Def. 6. *For a graph* $G = (V, E)$ *and a clustering* $\zeta = \{C_1, \ldots, C_k\}$ *of* G*,* modularity *is defined as*

$$mod(G, \zeta) := \sum_{C \in \zeta} \frac{|E(C)|}{|E|} - \sum_{C \in \zeta} \frac{\left(\sum_{v \in C} deg(v)\right)^2}{(2 \cdot |E|)^2}$$

The measure considers the clustering's *coverage* (the fraction of edges placed within a cluster) on the actual graph and subtracts the *coverage* it would achieve on a randomly connected version of the graph (preserving degree distribution). *Modularity*-based clusterings often agree with human intuition, although criticism has emerged recently [15]. Since maximizing *modularity* is an $\mathcal{NP}$-hard problem [16], we use a heuristic based on *local greedy agglomeration*. The base algorithm, commonly referred to as the Louvain Method [17], starts with a singleton clustering, considers nodes in turn, moves them to the best neighboring cluster and contracts the graph for the next iteration. This yields a hierarchy of graphs with increasing coarseness where the clustering in the coarsest level induces the resulting clustering in the original graph. Rotta et al. [18] enhance this algorithm by a refinement phase that iteratively projects this clustering to lower levels of the hierarchy and further improves modularity by local node moves. We use this modified algorithm.

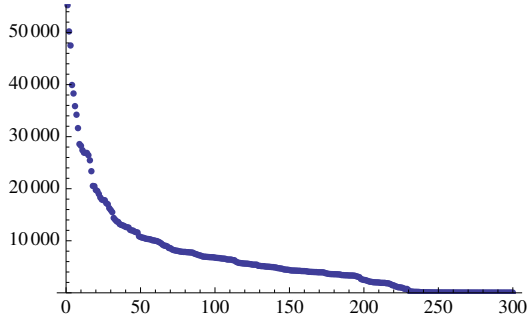


Fig. 4. Size distribution for the 300 largest clusters (x-axis: cluster size, y-axis: frequency)

It is a common approach to apply a clustering method to a real world network and then compare it to a ground-truth partition of the node set in order to interpret the result. For example, Rodriguez et al. [19] study sensor networks research groups and apply clustering techniques like *leading eigenvector*, but not *modularity* maximization; these network-structural communities are then compared to communities defined by socio-academic similarities.

We therefore proceed as follows: Applying local greedy agglomeration to $G_{\mathbf{PA}}$ yields a clustering with 86761 clusters, achieving a *modularity* of 0.896896. The majority of clusters contain only a handful of nodes, and likely correspond to the many tiny components of the graph, while the dominant connected component is divided into several large clusters (see Figure 4). With a clustering of the *authorship graph* at hand, we attempt to interpret such a *modularity*-driven clustering in the context of collaboration networks. The partition found by maximizing *modularity* locally identifies groups of authors who are densely connected through collaborative ties. Our hypothesis is that we can infer a topical similarity from these connections. More precisely, we conjecture that researchers form collaborative ties around distinct areas of research, which is reflected in the clustering structure of the graph. To put this hypothesis to the test, we compare the *modularity clustering* of $G_{\mathbf{PA}}$ to a ground-truth subdivision of the author set based on conferences: Assuming that distinct areas of computer science generally have dedicated conferences, we assign all authors who have published at a particular conference to an author-cluster. (Unlike the *modularity clustering*, this does not yield a proper, complete and disjoint partition of the author set, but is nonetheless informative.) Thereby we arrive at *topical clusters* of authors, which are suited as a ground-truth to compare the *modularity clustering* to.

	random	topical
O	0.04404	0.22832
J	0.00372	0.01390

TABLE III
MEAN MAXIMUM OVERLAP FOR MODULARITY CLUSTERING AND RANDOM VS TOPICAL CLUSTERING

In order to evaluate the similarity between the two community structures, one being the *modularity clustering*, the other the *topical clustering* defined by conferences, we apply overlap measures to each pair of clusters: The *Jaccard index* $J(A,B) := \frac{|A \cap B|}{|A \cup B|}$ favors exact match of the two sets; the *overlap coefficient* $O(A,B) := \frac{|A \cap B|}{\min(|A|,|B|)}$ treats containment of one set in the other set as a strong match, which is more equitable when dealing with clusters of uneven sizes. Applying these measures yields matrices of overlap values between *modularity clusters* and *topical clusters*. Additionally, we arrive at a baseline for the overlap values by calculating the overlap matrix of *modularity* clustering and a random clustering. The random clustering is constructed by copying the size distribution of the 250 largest modularity clusters, but randomly assigning authors to the clusters.

In these overlap matrices, we are interested in the maximum entry for each row, pointing to pairs of clusters that are most similar. Table III shows the means of these maximum overlap values. It is evident that the maximum J and O overlap is significantly better for *modularity clusters* than for random clusters. This shows that a more than coincidental relation between *modularity clusters* and *topical clusters* exists. However, the values are not close to 1.0 and indicate that the correspondence is not very strong. Thus, factors in addition to joint conferences are influential in shaping the community structure of the network. In the following section, we take an in-depth look at one possible factor of this kind, namely participation in research seminars.

IV. Impact of Seminars on Network Evolution

After describing static aspects of the network in the previous section, this section is concerned with its dynamics: We examine whether the *Dagstuhl Seminars*, as academic and social events, leave a track in the structure of the network, preferably in the form of increased collaboration between the participants. In the authors' subjective experience, the seminars present valuable opportunities for networking. Our approach to this question can be summarized as follows: Track groups of researchers (seminar participants and others selected as reference groups) in the time-resolved graph $\mathcal{G}^{1,1}_{\mathbf{PA}}$ and observe their publication output as well as their collaborative links; take into account the date of a seminar in order to observe immediate or long-term effects. The preparations necessary for this approach are described in the following:

A. Preparations

a) Aligning Data Sets: Our data sets record a total of 11 625 seminar guests in the *Dagstuhl* database and 852 250 authors in *DBLP*. All seminars took place in the 2000s. We align the tests by author name, whereby some false (mis)matches cannot be avoided. Still, a matching author in the publication database was found for 72 percent of the seminar invitees.

b) Area Launchers: In order to detect increased collaboration which can be clearly attributed to the seminars, we first try to identify *area launchers*. These are seminars intended to bring together a group of researchers who have not collaborated much before. A stated goal of the *Dagstuhl Seminars* is that some of them are intended to launch new areas of research by fostering collaboration between previously unaffiliated researchers, thereby contributing to emerging fields. *Area launchers* are relevant to us due to the following argument: If participants develop collaborative ties in the aftermath of an *area launcher* seminar, it is possible to attribute this more clearly to the seminar rather than existing relationships, developed, for instance, in the course of a common conference.

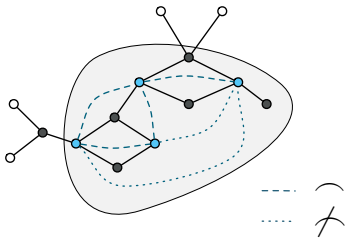


Fig. 5. Illustrating collaboration measure cad: $cad(A) = 2/3$

We classify a set of seminars as *area launchers* without special knowledge about the intent or content of the seminar, but solely from participation data: It is assumed that well-established areas of research generally spawn their own dedicated conference, and that the participants of such a conference represent the researchers active in this area. By this logic, a seminar corresponds to an established area of research if the invitees have a strong overlap with the participants of the respective conference. Furthermore, if researchers attend the same conference, it is likely that they are already familiar with each other as well as each other's work. We therefore reason that a seminar is an *area launcher* if its invitees do not overlap strongly and clearly with the participants of any particular conference. From this calculated set of seminars, 10 seminars are selected by hand and classified as *area launchers*.

c) Measures: We quantify the publication output and intensity of collaboration among researchers using several measures which map sets of authors to real numbers. For example, Figure 5 shows a small number of authors (light nodes) and their publications (dark nodes) in the *authorship graph*. Authors belonging to A are colored blue. Blue lines show existing (dashed line) and nonexisting (dotted line) coauthorship relations between pairs of authors in A. This illustrates the measure $cad(A)$, which is the fraction of actually existing coauthorship relations within an author set. Before introducing all measures, it is helpful to define sets of (co)publications, copublications internal to a group, and coauthors first: Given a set of authors $A \subseteq \mathbf{A}$, the set of their publications $P(A)$ is equal to

$$P(A) := \bigcup_{a \in A} P(a) = \bigcup_{a \in A} \{p \in \mathbf{P} : a \smile p\}$$

The set of *copublications* for an author a consists of publications which were written as collaborations with another

author:

$$CP(a) := \{p \in P(a) : \exists b \in \mathbf{A} : b \smile p\}$$

For an author set $A \subseteq \mathbf{A}$, the *aggregated copublications* are

$$CP(A) := \bigcup_{a \in A} CP(a)$$

The set of *intra-copublications* of a set of authors is defined as

$$CP_{\text{intra}}(A) := \{p \in CP(A) : \exists a, b \in A : a \smile p, b \smile p\}$$

The set of coauthors for a given author $a \in \mathbf{A}$ are those authors with whom a has authored a collaboration.

$$CA(a) := \{b \in \mathbf{A} : b \frown a\}$$

This can be generalized for a set of authors A:

$$CA(A) := \bigcup_{a \in A} CA(a)$$

Based on these sets, we formulate five measures, listed and defined in Table IV. These measures are intended to answer the following questions:

- $ap(A)$: What is the general productivity of an average author from the group?
- $acp(A)$: What is the productivity of such an author in terms of collaborations?
- $aca(A)$: With how many other authors does an average author from the group collaborate?
- $cpr_{\text{intra}}(A)$: Do the authors collaborate more often within the group or outside of the group?
- $cad(A)$: How close is the group to a collaborative clique, i.e. a group in which all authors have collaborated with each other?

measure	definition
$ap(A)$	$\frac{\lvert P(A)\rvert}{\lvert A\rvert}$
$acp(A)$	$\frac{\lvert CP(A)\rvert}{\lvert A\rvert}$
$aca(A)$	$\frac{\lvert CA(A)\rvert}{\lvert A\rvert}$
$cpr_{\text{intra}}(A)$	$\frac{\lvert CP_{\text{intra}}(A)\rvert}{\lvert CP(A)\rvert}$
$cad(A)$	$\lvert\{\{a,b\} \in \binom{A}{2} : a \frown b\}\rvert / \lvert\binom{A}{2}\rvert$

TABLE IV
OVERVIEW OF COLLABORATION MEASURES AND THEIR DEFINITIONS

d) Author Classes: The classes of author groups which we track are the seminar participants on the one hand and several reference classes on the other:

- *seminar attendees* ($\mathcal{A}t_s$): For each seminar s, the set of researchers who attended the seminar.
- *seminar absentees* ($\mathcal{A}b_s$): For each seminar s, the set of researchers who were invited to the seminar but did not attend. (For some seminars, the set was empty or very small, so these are only included if they have a sufficient size.)
- *random samples* ($\mathcal{RS}_i$) Contains randomly assembled sets of authors with the size of a typical seminar.
- *connected samples* ($\mathcal{CS}_i$) Contains sets of authors found by collecting nodes from $G_{\mathbf{PA}}$ in a breadth-first search from a random initial node until the typical size of a seminar is reached.
- *all authors* ($\mathbf{A}$) A single set containing all authors.

B. Evaluation and Results

We speculate that joint participation in a seminar leads to increased collaboration between the participants. This would be measurable as higher values for the collaboration measures (cad, cpr_{intra}) on the respective subgraph. Additionally, we measure whether seminar participation leads to a higher publication output for the participants (ap, acp, aca). In order to test this, seminar-related groups as well as reference groups are tracked within the graph $\mathcal{G}^{1,1}_{\mathbf{PA}}$: For any author set A, a subset $A' \subseteq A$ has corresponding nodes $V_{A'}$ in the graph G_y. For all measures M, we evaluate $M(A')$, yielding a sequence of values for each group. The evaluation yields one value sequence per author group, and thus several data points per year. All seminar-related sequences are aligned according to the time of the seminar, in order to compare values before and after seminar participation. We present these data points in boxplot form (e.g. Figure 6), with the horizontal axis denoting time relative to the seminar date and the vertical axis values of the respective measure. By following the plotted median and quantiles along the time axis, one can identify trends for the author class as a whole. The point in time where a seminar occurs is marked by an arrow.

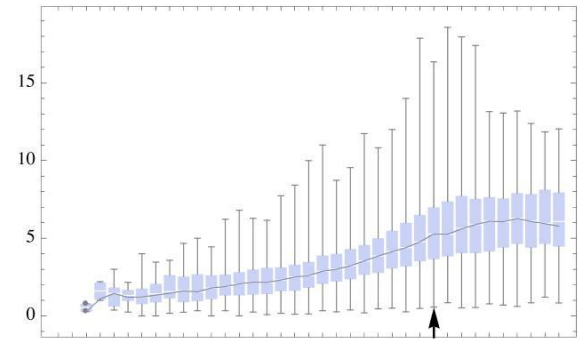
In the following section, we describe a selection of notable observations:

a) Average publication output remains rather constant: For the authors as a whole ($\mathbf{A}$), average publication output and number of coauthors remain stable over time, even as the graph grows at an increasing rate and author nodes accumulate.

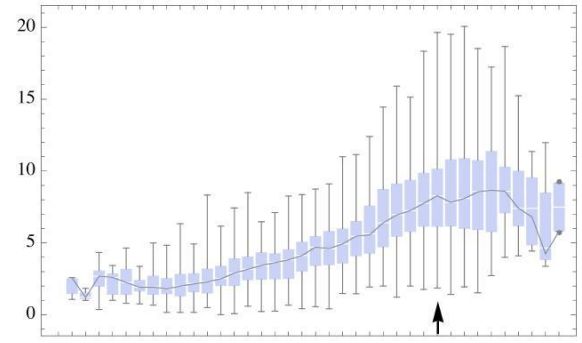
b) Randomly grouped authors as a baseline for publication output: As a reference class, we evaluate the randomly compiled author groups $\mathcal{RS}$. Both ap and aca are, on average, in the range of 0.6-0.8, showing that there are typically inactive authors in any given time frame. As expected, there is no collaboration between authors in the random samples.

c) Connected Sample Groups: Authors from the $\mathcal{CS}$ have a significantly higher productivity than randomly selected authors, since breadth-first search finds high-degree nodes with a higher probability. There is also an upward trend over time for all measures. A possible explanation for this is that nodes gain connections over time according to degree, if there is an underlying *preferential-attachment* process at work (as suggested by the power-law degree distribution). Overall cpr_{intra} remains clearly below 0.5, showing that these sample groups are just sections from greater collaborative clusters.

d) Attendees and absentees are equally productive: The effect of seminar participation is best judged by contrasting attendees with absentees. With respect to productivity, measured by the number of coauthors and the number of publications,

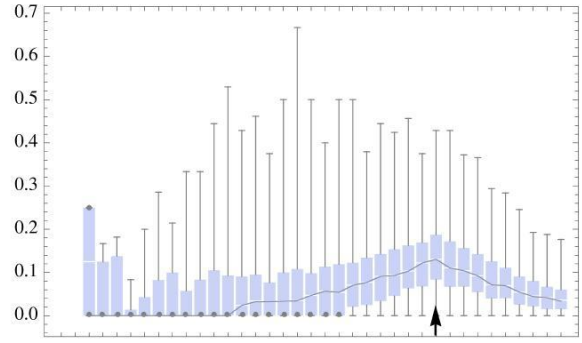


(a) aca: At

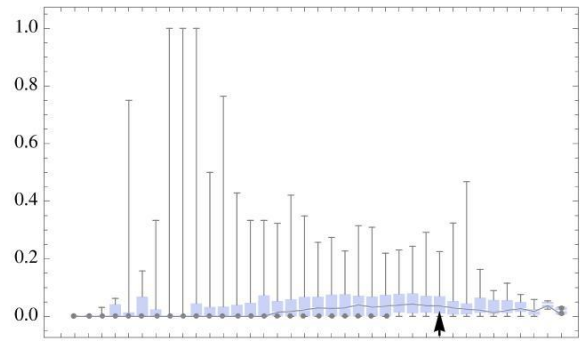


(b) aca: Ab

Fig. 6. aca (y-axis) for seminar attendees and absentees (x-axis: time relative to seminar, arrow: seminar date)



(a) cpr_{intra}: At



(b) cpr_{intra}: Ab

Fig. 7. cpr_{intra} (y-axis) for seminar attendees and absentees (x-axis: time relative to seminar, arrow: seminar date)

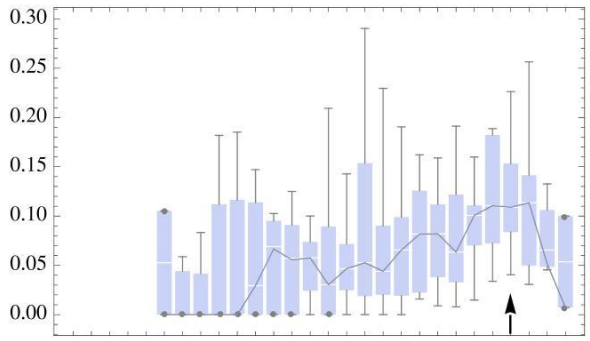
attendees and absentees are quite similar, with some outliers among the absentees surpassing the attendees (see Figure 6). For the productivity measures, an upward trend before the seminar continues for a few years but then tends to reverse.

e) Attendees form a more cohesive group: For seminar attendees, a larger fraction of their collaborations are internal to the seminar group, both before and after the seminar (Figure 7). This indicates that attendees already come from a more cohesive group. Values for cad agree with this interpretation: Clearly, those who choose to attend the seminar form a denser subgraph in the collaboration network. There seems to be no lasting increase in collaboration after the seminar, but a downward trend for both attendees and absentees.

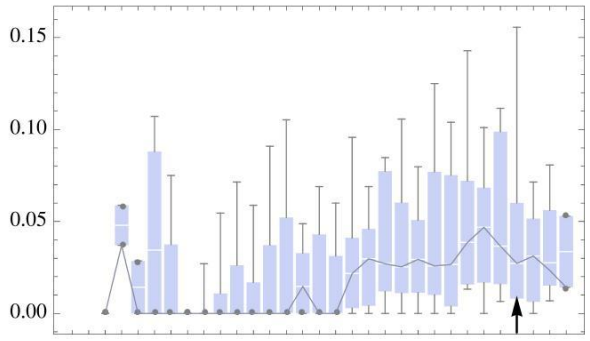
f) Area launchers are not exceptional: For the subset of seminars classified as *area launchers*, we expect comparatively less collaboration before the seminar, and a stronger increase after. This effect would be most clearly captured by the measures cpr_{intra} (Figure 8) and cad. The plots in Figure 8 support our reasoning about area launchers, namely that the authors invited have a comparatively low probability of collaboration in the time prior to the seminar: Values for cpr_{intra} are generally in the lower range compared to all seminars. Still, a visible change after the time of the seminar is missing. The influence of an *area launcher* seminar does not seem to differ from the other seminars.

g) Subdivision by career stage: Suspecting that seminar participation affects researchers in early stages of their career more strongly, we repeat a part of the evaluation with the authors classified by career length (≤ 5, ≤ 15, > 15 years of publication history). However, the results do not modify our conclusions: A seminar effect for academic newcomers is no more observable than for all other authors.

h) Summary and Interpretation: Seminar invitees are more productive and more collaborative than randomly selected authors. Yet there is little difference between attendees and absentees in terms of their productivity. Invited researchers are already actively publishing, with an upward trend, prior to the time of the seminar. For cpr_{intra} and cad, attendees are consistently better than absentees. This indicates that those who attend are already a tightly connected collaborative group before the seminar, possibly influencing their decision to participate. The general trend over time is an increase up to the seminar and a slight decrease afterwards for both classes of researchers. A possible explanation for the increase and decrease over time is that invitations are biased towards researchers who are currently most active: Invitations to seminars occur at a period of peak activity. There is, however, no significant change of structure connected to seminars (either significant short-term increase in collaboration directly after the seminar or long-term increase). Most importantly, attendees and absentees do not differ in this respect. While



(a) cpr_{intra}: *At*



(b) cpr_{intra}: *Ab*

Fig. 8. cpr_{intra} (y-axis) for attendees and absentees of *area launchers* (x-axis: time relative to seminar, arrow: seminar date)

the focus on *area launcher* seminars supports our assumption that the invited researchers had collaborated less, a significant structural change after the seminar is not visible. These results suggest that a single event like a seminar is not influential enough to alter the network structure of collaboration for the group of participants in ways observable with our measures. Clearly, other factors have additional and apparently more influence on the structure. Rather in the opposite direction, the network structure might be employed to predict who will attend the seminar and who will decline, since the participants evidently come from a more cohesive group.

V. Conclusion

This paper ties in with the existing work on scientific collaboration networks and explores several new variations of network analysis methods. The coauthorship graph in the field of computer science constitutes in many respects a typical social network, as observed before in similar studies: We encounter properties such as low average distances between researchers, a *giant connected component*, a power-law distribution with regard to publications and coauthors (making it a *scale-free network*), and a regular k-*core* structure. We detect dense communities of researchers through *modularity* maximization, and compare the resulting clustering to ground-truth communities defined by conferences, from which topical similarity is inferred. The overlap between the two partitions is clearly not coincidental, although other factors seem to be at work in shaping the community structure. In order to identify influential researchers by their network centrality, we test a novel combination of bipartite author-paper graph and *eigenvector centrality*. We are the first to incorporate data on participants of the *Schloss Dagstuhl* research seminars and use it to evaluate the impact of such seminars on the evolution of collaborative ties. Since the seminars are designed to foster collaboration on cutting-edge research topics, and many participants experience the seminars as a valuable opportunity for networking, we investigate whether such effects can be observed as structural changes in the collaboration network. Seminar invitees are more productive, more collaborative and structurally prominent compared to the average researcher. However, our methods suggest that seminar participation does not directly affect the structure of the collaboration network. An interesting finding of this analysis was that researchers who choose to attend the seminar form a distinctly more cohesive subgraph than those who decline.

Acknowledgment

We thank Ulrik Brandes for helpful discussions during the preparation of this work. We also thank the *Schloss Dagstuhl* conference center for providing us with the necessary data on their seminars.

References

[1] M. E. Newman, "The structure of scientific collaboration networks." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 2, pp. 404–9, Jan. 2001. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=14598

[2] M. E. J. Newman, "Coauthorship networks and patterns of scientific collaboration." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101 Suppl, pp. 5200–5, Apr. 2004. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=387296

[3] M. Newman, "Scientific collaboration networks. I. Network construction and fundamental results," *Physical Review E*, vol. 64, no. 1, pp. 1–8, Jun. 2001. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevE.64.016131

[4] M. E. J. Newman, "Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality," *Physical Review E*, vol. 64, no. 1, pp. 1–7, Jun. 2001. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevE.64.016132

[5] "DBLP - Digital Bibliography and Library Project," 2007, http://dblp.uni-trier.de/.

[6] "Schloss Dagstuhl," http://www.dagstuhl.de.

[7] E. Elmacioglu, "On Six Degrees of Separation in DBLP-DB and More," *Distribution*, vol. 34, no. 2, pp. 33–40, 2005.

[8] J. Scott, *Social Network Analysis - a Handbook*, 2nd ed. SAGE Publications, 2000.

[9] C. Giatsidis, D. Thilikos, and M. Vazirgiannis, "Evaluating Cooperation in Communities with the k-Core Structure," *Social Networks*. [Online]. Available: http://graphdegeneracy.org/k-cores.pdf

[10] K. Boerner, L. Dall'Asta, W. Ke, and A. Vespignani, "Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams," *Complexity*, vol. 10, no. 4, pp. 57–67, Mar. 2005. [Online]. Available: http://doi.wiley.com/10.1002/cplx.20078

[11] P. Bonacich, "Factoring and Weighting Approaches to Status Scores and Clique Identification," *Journal of Mathematical Sociology*, vol. 2, pp. 113–120, 1972.

[12] P. Bonacich, A. Cody Holdren, and M. Johnston, "Hyper-edges and multidimensional centrality," *Social networks*, vol. 26, no. 3, pp. 189–203, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378873304000024

[13] "Most Prolific DBLP Authors," http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/prolific/.
[14] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 026113, pp. 1–16, 2004. [Online]. Available: http://link.aps.org/abstract/PRE/v69/e026113
[15] A. Lancichinetti and S. Fortunato, "Limits of modularity maximization in community detection," *Phys. Rev. E*, vol. 84, p. 066122, Dec 2011. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevE.84.066122
[16] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Höfer, Z. Nikoloski, and D. Wagner, "On Modularity Clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 2, pp. 172–188, February 2008. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TKDE.2007.190689
[17] V. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, 2008. [Online]. Available: http://dx.doi.org/10.1088/1742-5468/2008/10/P10008
[18] R. Rotta and A. Noack, "Multilevel local search algorithms for modularity clustering," *ACM Journal of Experimental Algorithmics*, vol. 16, pp. 2.3:2.1–2.3:2., July 2011. [Online]. Available: http://doi.acm.org/10.1145/1963190.1970376
[19] M. a. Rodriguez and A. Pepe, "On the relationship between the structural and socioacademic communities of a coauthorship network," *Journal of Informetrics*, vol. 2, no. 3, pp. 195–201, Jul. 2008. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1751157708000230