The Role of Research Leaders on the Evolution of Scientific Communities

Bruno Leite Alves
Universidade Federal de Minas Gerais, Brazil
bruno.leite@dcc.ufmg.br

Fabrício Benevenuto
Universidade Federal de Minas Gerais, Brazil
fabricio@dcc.ufmg.br

Alberto H. F. Laender
Universidade Federal de Minas Gerais, Brazil
laender@dcc.ufmg.br

ABSTRACT

There have been considerable efforts in the literature towards understanding and modeling dynamic aspects of scientific communities. Despite the great interest, little is known about the role that different members play in the formation of the underlying network structure of such comminities. In this paper, we provide a wide investigation of the roles that members of the core of scientific communities play in the collaboration network structure formation and evolution. To do that, we define a community core based an individual metric, *core score*, which is an h-index derived measure that captures both, the prolifcness and the involvement of researchers in a community. Our results provide a number of key observations related to community formation and evolving patterns. Particularly, we show that members of the core community work as bridges that connect smaller clustered research groups. Furthermore, members of a scientific community core are responsible for an increase in the average degree of the whole community undrelying network and a decrease on the overall network assortativeness. More important, we note that variations on the members of the community core tend to be strongly correlated with variations on these metrics. We argue that our observations are important for shedding a light on the role of key members on community formation and structure.

Categories and Subject Descriptors

H.4 [**Social Network**]: Temporal Analisys; J.4. [**Computer Applications**]: Social and Behavioral Sciences Miscellaneous

General Terms

Human Factors, Measurement.

Keywords

Scientific Communities, Core Community, Community Evolution.

MSND '13 Rio de Janeiro, Brazil

1. INTRODUCTION

Since its beginning, society has been organizing itself into communities, which are groups of individuals with common interests[1]. Particularly, the proliferation of new communication technologies based on the Internet has facilitated the rapid formation and growth of online communities [13]. Communities exhibit a wide range of characteristics and serve a variety of purposes, from small groups engaged in tightly niche topics such as a very specific scientific community, to millions of users linked by an interest such as a community related to a sport team or fans of a celebrity.

Often, individuals who are socially connected in a tend to share interests and similarities. Although, there are many factors that might determine a community formation and its growth, there are two main driven forces used to explain similarity in a community formation: influence and homophily. On one hand, influence posits that individuals change to become more similar to their friends in the community. On the other hand, homophily postulates that individuals create social connections within a community precisely because they are already similar. Recent efforts have provided quantitative evidences of both forces [1, 3, 4, 6] and existing theories [22, 27], models [11, 12], and approaches [24, 28] rely on identifying a group of influential individuals with the power to affect not only the underlying network structure of a community, but also to interfere on the spread and flow of information within a community.

In this paper, we take a different perspective and study a complementary problem. Here, we focus on studying the roles that influential individuals from a scientific community play on evolving properties of such communities. Intuitively, when prolific research leaders decide to join or leave a community, they take with them resources, experience and students, and they possibly influence other members to do the same, which makes scientific communities very suitable for this kind of study. To construct these communities we used data from DBLP to identify scientific communities, represented by the main ACM SIG conferences. Then, we propose a strategy to infer the community core, the leaders of a given scientific community in a given period of time. Finally, we investigate how aspects of the core impact on the community underlying structure.

The study of the core of scientific communities is of interest from two different perspectives. The first is sociological, coming from the necessity to understand how segments of society evolve as well as to answer longstanding questions related to the interaction among different types of participant.

[1]http://www.merriam-webster.com/dictionary/community

On the other hand, from a technological perspective, understanding these aspects is critical not only for link prediction, but also for the design of better recommendation systems. Such a study, however, has been difficult as essential components like human connections and a proper definition of leadership is hard to be reproduced at a large scale within the confines of a research laboratory.

Among our main observations, our results show that members of the community core work as bridges that connect smaller clustered research groups as well as increase the average degree of the community underlying network, but decrease the overall network assortativeness. More important, we note that variations on the members of the community core tend to be strongly correlated with variations on these metrics.

The rest of this paper is organized as follows. Section 2 surveys related work. Section 3 describes our strategy and dataset used to construct the scientific communities. Section 4 describes our strategy to compute the community core and Section 5 investigates the role that these sets of researchers play within their communities. Finally, Section 6 presents our conclusions and provides directions for future work.

2. RELATED WORK

There has been a number of recent efforts that attempt to analyze community structure and network evolution. Particularly, Kumar *et al.* [14] analyzed two large networks to find a segmentation of these networks into singletons, isolated communities, a giant component. Then, they propose a network growth model able to generate networks with similar characteristics. Ducheneaut *et al.* [7] extracted and characterized explicitly created communities from the World of Warcraft, a massive multiplayer game. Complementarily, Patil *et al.* [21] analyzed and modeled factors that make users to leave or join on-line gaming communities. Viswanath *et al.* [26] studied the evolution of activity between users in Facebook and found that that links in the activity network tend to come and go rapidly over time, and the strength of ties exhibits a general decreasing trend of activity as the social network link ages.

In terms of models for network dynamics, Leskovec *et al.* [16] investigated a wide range of real graphs to show that graphs densify over time, with the number of edges growing super linearly in the number of nodes and that the average distance between nodes often shrinks over time. Based on these observations, they develop a graph generation model that incorporate such properties. More recently, Leskovec *et al.* [15] presented a detailed study of network evolution by analyzing four large online social networks. They investigated a wide variety of network formation strategies to show that edge locality plays a critical role in evolution of networks. Based on this observation, they developed a model of network evolution, in which nodes arrive at a pre-specified rate. Differently from the above efforts, our work focuses on community properties and the roles that community leaders play in the underlying network structure.

There are also efforts that attempted to study scientific communities. Particularly, Backstrom *et al.* [3] studied communities in LiveJournal and scientific communities extracted from DBLP to find that the propensity of individuals to join communities, and of communities to grow rapidly, depends in subtle ways on the underlying network structure. Huang *et al.* [10] used DBLP data to construct a network for the Computer Science field covering research collaborations from 1980 to 2005. Among their main observations, they show that the Computer Science field shows a collaboration pattern more similar to Mathematics than to Biology. Different from these efforts, here we focus on studying the properties of the community core, thus, our analyses are complementary to theirs.

Finally, when it comes to identifying the community core, there are many approaches that extract the core based on structural properties of the underlying network [5, 9, 17, 23]. Particular, Seifi *et al.* [25] combined four different approaches to identify a community core and characterized some properties of the obtained cores. Such approach is not applicable to our context, as we are interested in studying network properties of the community core.

3. SCIENTIFIC COMMUNITIES

The notion of community can be understood as a dense group of nodes in a network, with more edges inside than edges linking the rest of the network. There are multiple definitions and strategies of identifying communities and they vary according to the context [13, 17]. In our context, a scientific community is defined in terms of a large and well established scientific conference able to aggregate researchers working in similar research topics along a considerable number of years.

In order to construct a set of scientific communities, we have gathered data from DBLP[2] [18], a digital library containing more than 2.1 million publications from 1.2 million authors that provides bibliographic information on major computer science conference proceedings and journals. DBLP offers its entire database in XML format, which facilitates gathering and constructing entire scientific communities.

Each publication is accompanied by its title, list of authors, year of publication, and publication venue, i.e., the conference or journal. For the purposes of our work, we consider a scientific community as a graph in which nodes represent researchers and edges links coauthors of papers from the same community. In order to define such communities, we focus on the publications from the flagship conferences of major ACM SIGs (Special Interest Groups). Thus, we define a scientific community by linking people that have coauthored a paper in a certain conference, making the flagship conferences of the ACM SIGs to act as communities where coauthorships are formed. We have removed young conferences without enough data for a temporal analysis as well as conferences whose entire history is not registered on DBLP to allow us carrying out temporal analyses.

In total, 24 scientific communities were considered. Table 1 lists these communities, including the respective ACM SIG, the conference acronym, the period considered (some conferences had the period reduced to avoid hiatus in the data), the h-index[3] and the total number of authors, publications and editions as well as ratios extracted from these last three aspects.

4. DEFINING A COMMUNITY CORE

[2]http://dblp.uni-trier.de/

[3]Obtained from the SHINE (Simple H-INdex Estimator) project: http://shine.icomp.ufam.br.

Table 1: The data of DBLP of flagship conferences of ACM SIGs

SIG	Conference	Period	H-Index	Authors	Publications	Editions	Aut/Edi	Pub/Edi	Aut/Pub
SIGACT	STOC	1969-2012	94	2159	2685	44	49.07	61.02	0.80
SIGAPP	SAC	1993-2011	59	9146	4500	19	481.37	236.84	2.03
SIGARCH	ISCA	1976-2011	102	2461	1352	36	68.36	37.56	1.82
SIGBED	HSCC	1998-2012	-	846	617	15	56.40	41.13	1.37
SIGCHI	CHI	1994-2012	144	5095	2819	19	268.16	148.37	1.81
SIGCOMM	SIGCOMM	1988-2011	140	1593	796	24	66.38	33.17	2.00
SIGCSE	SIGCSE	1986-2012	51	3923	2801	27	145.30	103.74	1.40
SIGDA	DAC	1964-2011	98	8876	5693	48	184.92	118.60	1.56
SIGDOC	SIGDOC	1989-2010	23	1071	810	22	48.68	36.82	1.32
SIGGRAPH	SIGGRAPH	1985-2003	119	1920	1108	19	101.05	58.32	1.73
SIGIR	SIGIR	1978-2011	116	3624	2687	34	106.59	79.03	1.35
SIGKDD	KDD	1995-2011	124	3078	1699	17	181.06	99.94	1.81
SIGMETRICS	SIGMETRICS	1981-2011	71	2083	1174	31	67.19	37.87	1.77
SIGMICRO	MICRO	1987-2011	81	1557	855	25	62.28	34.20	1.82
SIGMM	Multimedia	1993-2011	80	5400	2928	19	284.21	154.11	1.84
SIGMOBILE	MobiCom	1995-2011	106	1151	480	17	67.71	28.24	2.40
SIGMOD	SIGMOD	1975-2012	147	4202	2669	38	110.58	70.24	1.57
SIGOPS	PODC	1982-2011	59	1685	1403	30	56.17	46.77	1.20
SIGPLAN	POPL	1975-2012	85	1527	1217	38	40.18	32.03	1.25
SIGSAC	CCS	1996-2011	97	1354	676	16	84.63	42.25	2.00
SIGSAM	ISSAC	1988-2011	-	1100	1177	24	45.83	49.04	0.93
SIGSOFT	ICSE	1987-2011	111	3502	2248	25	140.08	89.92	1.56
SIGUCCS	SIGUCCS	1989-2011	-	1771	1593	23	77.00	69.26	1.11
SIGWEB	CIKM	1992-2011	82	4978	2623	20	248.90	131.15	1.90

Previous attempts for identifying the community core of a scientific community are based on algorithmic approaches that aim at identifying dense clusters of nodes in the network [25]. However, as we plan to investigate the role of a core in the network structure, any approach that makes use of the network structure to identify such nodes could lead us to a biased set of researchers. Instead, we focus on developing a metric that quantifies the involvement of a researcher in a scientific community during a certain period of time. Intuitively, this metric should be able to capture (i) the prolificness of a researcher in different communities and (ii) the frequency of involvement of that researcher with the community in a certain period of time.

First, in order to capture the prolificness of a researcher, we use the h-index [8], a metric widely adopted for this purpose. This metric consists of an index that attempts to measure both the productivity and the impact of the published work of a researcher. It is based on the set of the researcher's most cited papers and the number of citations that they have received. More specifically, a researcher a has an h-index h_a if she has published h papers which have received at least h citations. Thus, for example, if a researcher has 10 papers with at least 10 citations, her h-index is 10.

Second, as an attempt to capture the importance of a researcher to a specific community in a certain period of time, we multiple her h-index by the number of publications this researcher has in a certain community during a time window. We name this measure *Core Score*. More formally, the Core Score of a researcher r in a community c during a period of time t, $CoreScore_{r,c,t}$, is given by its $h\text{-}index_r$ multiplied by the number of publications r has in c during t ($\#publications_{r,c,t}$), as expressed by Equation 1.

$$CoreScore_{r,c,t} = h\text{-}index_r \times \#publications_{r,c,t} \quad (1)$$

We note that the first part of the equation captures the importance of a researcher to the scientific community as a whole regardless any specific research area or period of time and the second part weights this importance based on the activity of the researcher in a certain community and time. By computing the core score for the members of a community, we define the community core in a certain period of time as the top researchers of that community in terms of their core scores in the given period. Next, in Section 4.1, we detail how we inferred the h-index of researchers. Then, Section 4.2 discusses how we define two important thresholds: the size of the community core and the time window used in our analyses.

4.1 Inferring Researchers' H-index

There are multiple tools that measure the h-index of research authors, out of which Google Citations[4] is the most prominent one. However, to have a profile in this system, a researcher needs to sign up and explicitly create her research profile. In a preliminary collection of part of the profiles of the DBLP authors, we found that less than 30% of these researchers had a profile at Google citations. Thus, this strategy would reduce our dataset and potentially introduce bias when analyzing the communities.

To divert from this limitation, we used data from SHINE, the Simple HINdex Estimation project[5], to infer the researchers' h-index. SHINE provides a website that allows users to check the h-index of almost two thousands computer science conferences. They crawled Google Scholar, searching for the title of papers published in a number of conferences, which allowed them to effectively estimate the h-index of these target conferences based on the citations computed by Google Scholar. Although SHINE only allows one to search for the h-index of a conference, the SHINE developers kindly allowed us to access their database to infer the h-index of researchers based on the conferences they

[4]http://scholar.google.com/citations
[5]http://shine.icomp.ufam.edu.br/

crawled.

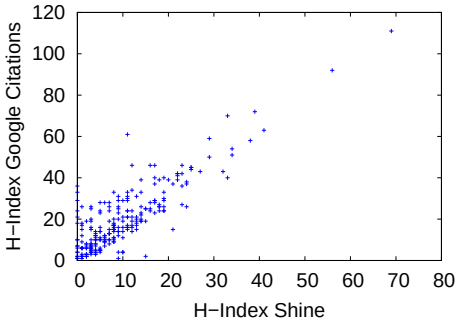


Figure 1: Correlation between the inferred h-index and Google Citations one

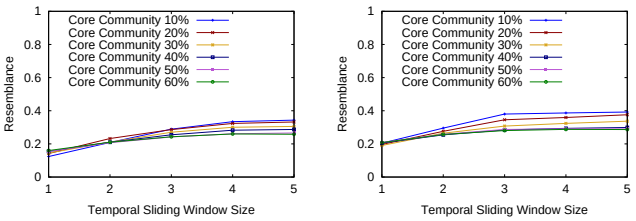
However, there is one important limitation with this strategy. As SHINE does not track all the existent computer science conferences, researchers' h-index might be underestimated when computed with this data. To investigate this issue, we compared the h-index of a set of researchers with a profile on Google Scholar with their estimated h-index based on the SHINE data. For this, we randomly selected 10 researchers for each conference from Table 1 and extracted their h-indexes from their Google Scholar profiles. In comparison with the h-index we estimated from SHINE, the Google scholar values are, on average, 50% higher. Figure 1 shows the scatter plot for the two h-index measures. We can note that although SHINE h-index is smaller, the two measures are highly correlated. The pearson correlation coefficient is 0.85, which indicates that researchers might have proportional h-index estimations in both systems.

4.2 Setting the Thresholds

Our strategy to define these two thresholds consists of varying each of them and quantifying how they impact on the changes on the members of the community core. To measure these changes, we compute the resemblance metric, as used in [26], which measures the fraction of members in the core at time t_0 that remains in the core at the time t_1. For each community, we varied the window size from 1 to 5 years and the size of the community core from 10% to 60% of the entire community.

There are two important thresholds in our approach we need to define to determine the core of a scientific community. The first is related to the time window in which the community core is computed. In other words, should we compute the community core at each year, at each two years, or for a larger time window? The second threshold is related to the size of the community core. As we define the core of a community as the top researchers in terms of their core score during a certain time window, it is important to define the threshold for choosing the top ones.

Intuitively, high resemblance variations indicate bad threshold choices and, thus, we should seek for values in which threshold changes cause slight changes on resemblance. Figure 2 shows the resemblance values as a function of the window size, providing different curves for the community core size. We chose the SIGMOD and CHI communities for this analysis. The rest of the communities are omitted due to lack of space, but the same observations hold for them. By



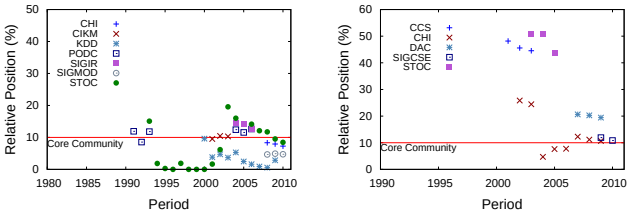
(a) SIGMOD (b) CHI

Figure 2: Average of the values of resemblance

visual inspection we would set the core size as 10% due to the proximity of the curves, and the window size as 2 or 3, as most of the communities showed a more stable resemblance after these values. To help us decide, we computed the angular coefficient for the 10% core size curves of each community and obtained the average angular coefficient for them. Based on this value, we chose the window size for our experiments as 3 years.

4.3 Validation

Based on the core score value, we expect that the members of the community core would be standing researchers that actively contribute with publications to a certain community. The validation of this assumption is, by nature, subjective. Thus, we provide next evidence that our approach correctly captures this expected characteristic.



(a) Jon Kleinberg (b) Luis von Ahn

Figure 3: Core score of two WWW 2013 keynote speakers

First, we analyzed the core score of two WWW 2013 keynote speakers: Jon Kleinberg and Luis von Ahn. Figure 3 shows the ranking position in terms of percentage (e.g., position 5% of that community) of these two researchers in the communities they have published. The botton line divides the members of the community core from the others. We can note that Jon Kleinberg was a member of the community core of STOC, a theoretical conference, for years. More precisely, he was part of the STOC core for twelve years, publishing seven STOC papers in a single period of three years. With Kleinberg's involvement on KDD, he became less active in STOC and left the core of that community for some time. During this period, he published several KDD papers, while his STOC publications were drastically reduced. When it comes to Luis von Ahn, we can note that he is more active in the CHI community, a community in which he published six papers along his academic life. He

reached the core of the CHI community along three consecutive time windows, publishing four CHI papers in a single period.

Next, we computed a ranking of researchers that appear most often in the community core of each scientific community. We chose the KDD, SIGCOMM, SIGIR, and SIGMOD communities to show their top 20 researchers in Table 2. As we can note, several big names appear in this top list, including past keynote speakers of these conferences as well as awarded researchers by their life time contributions in that community. Indeed, by analyzing the awarded researchers from each community we found that a large fraction of them appeared in the community core at least one time in the conference history. More specifically, these fractions are 75% of the awarded KDD[6] members, 35% for SIGCOMM[7], 60% for SIGIR[8], and 80% for SIGMOD[9]. Except for SIGCOMM, a community with many sponsored event that were not considered in our datasets, the other three communities presented very high numbers of awarded members that appear at least one time a community core. These observations provide evidence that our approach correctly captures the notion of a scientific community core.

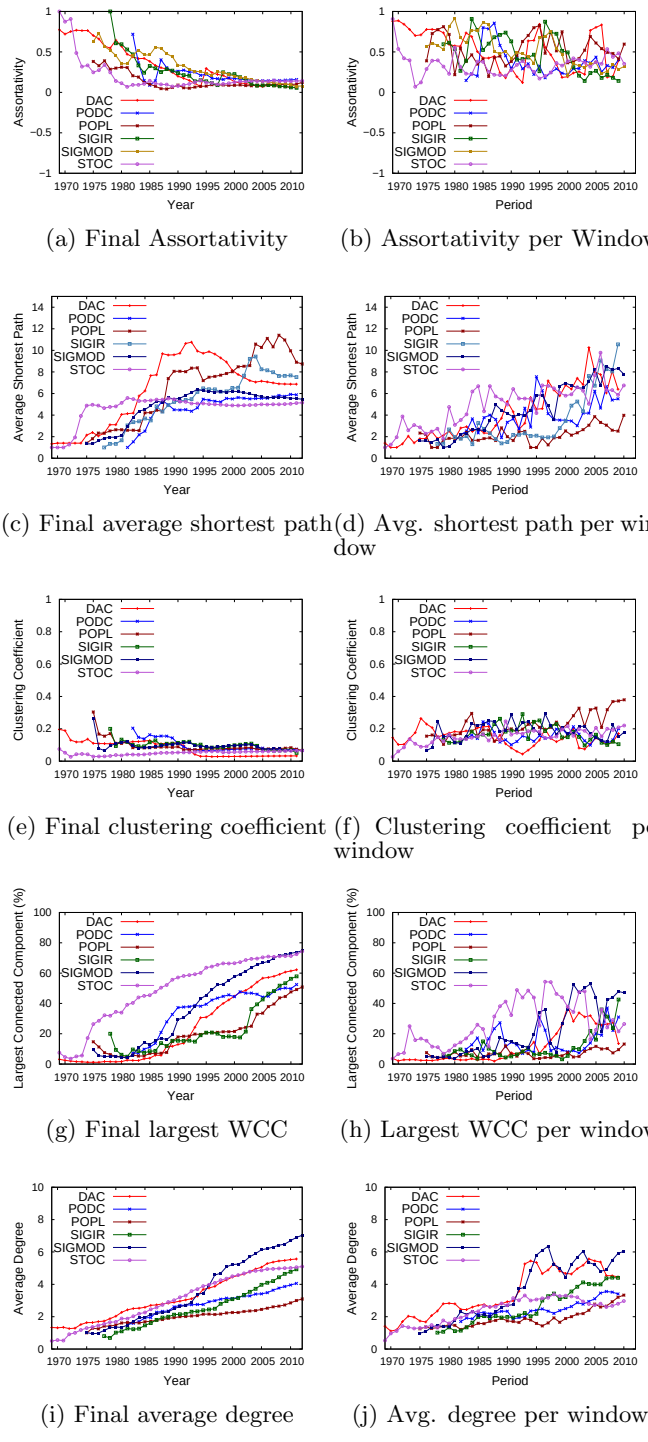
5. PROPERTIES OF COMMUNITY CORES

In this section, we present a series of analyses about the scientific community cores. First, we analyze how the network properties of the scientific communities have evolved. Then, we contrast the properties of the community cores over time against the properties of the remaining members of the respective communities. Finally, we compute the average core score of a community to investigate fluctuations in the properties of the members of the community cores and correlate these fluctuations with the network properties of the communities.

5.1 Evolution of the Scientific Communities

In order to study the evolution of the main structural properties of the scientific communities, we examine various network metrics for each of the scientific communities. We present four popular metrics here: assortativity, average clustering coefficient, average path length, and the size of the largest weakly connected component (WCC). Figure 4 shows how each of these four metrics vary over time for a set of six scientific communities selected among those that span over the longest period in our dataset. Our analysis are performed under two perspectives. The first consists on analyzing the network evolution year by year by accumulating nodes and edges to a single final snapshot of the graph. This perspective allow us to observe the final network structure of a community as a function of time. The second perspective consists of analyzing snapshots constructed based on nodes and edges created on a predefined time window (three years, as discussed in Section 4.2). This analysis allows us to investigate network variations with potential to impact the final network structure. Our analysis results are similar for the other communities, but we omit them due to lack of space.

We note from Figure 4 that the largest WCC tend to largely increase as a function of time. This suggests that



(a) Final Assortativity (b) Assortativity per Window

(c) Final average shortest path (d) Avg. shortest path per window

(e) Final clustering coefficient (f) Clustering coefficient per window

(g) Final largest WCC (h) Largest WCC per window

(i) Final average degree (j) Avg. degree per window

Figure 4: Network evolution metrics for scientific communities

[6]http://www.sigkdd.org/awards_innovation.php
[7]http://www.sigcomm.org/awards/sigcomm-awards
[8]http://www.sigir.org/awards/awards.html
[9]http://www.sigmod.org/sigmod-awards

Table 2: Researchers who appear most often in the community core over the years

KDD	SIGCOMM	SIGIR	SIGMOD
Heikki Mannila⋆	Scott Shenker⋆	W. Bruce Croft⋆	David J. DeWitt⋆
Jiawei Han⋆	George Varghese	Clement T. Yu	Michael Stonebraker⋆
Eamonn J. Keogh	Hui Zhang	Susan T. Dumais⋆	H. V. Jagadish
Martin Ester	Donald F. Towsley⋆	James Allan	Rakesh Agrawal⋆
Bing Liu	Hari Balakrishnan	Justin Zobel	Christos Faloutsos
Padhraic Smyth⋆	Ion Stoica	Alistair Moffat	Raghu Ramakrishnan
Charu C. Aggarwal	Srinivasan Seshan	Norbert Fuhr⋆	Jiawei Han
Philip S. Yu	Deborah Estrin	James P. Callan	Gerhard Weikum
Ke Wang	David Wetherall	Yiming Yang	Philip A. Bernstein⋆
Hans-Peter Kriegel	Thomas E. Anderson	Edward A. Fox	Jeffrey F. Naughton
Rakesh Agrawal⋆	Jennifer Rexford	Gerard Salton⋆	Hector Garcia-Molina⋆
Jian Pei	Jia Wang	Ricardo A. Baeza-Yates	Michael J. Carey⋆
Wynne Hsu	Ratul Mahajan	Jian-Yun Nie	Joseph M. Hellerstein
Qiang Yang	Vern Paxson⋆	Mark Sanderson	Philip S. Yu
Christos Faloutsos⋆	Mark Handley	Charles L. A. Clarke	Divesh Srivastava
Huan Liu	Yin Zhang	Chris Buckley	Michael J. Franklin
Mohammed Javeed Zaki	Peter Steenkiste	Chengxiang Zhai	Jennifer Widom⋆
Pedro Domingos	Walter Willinger	Alan F. Smeaton	Hans-Peter Kriegel
Jon M. Kleinberg	Ramesh Govindan	Zheng Chen	Hamid Pirahesh
Vipin Kumar⋆	Jon Crowcroft⋆	Ophir Frieder	Surajit Chaudhuri⋆

⋆ Researchers awarded by a lifetime of innovation and leadership inside that community.

at early stages, scientific communities are formed by several small and segregated research groups. With time, some reserachers (e.g., students) leave an institute and begin collaborations with other research groups. Additionally, as the community evolves, heads of research groups tend to colaborate with other peers of the same community. Thus, with time, researchers from different groups tend to collaborate and increase the size of the largest WCC. As a consequence, the average shortest path, computed only on the largest WCC, tends to increase, becoming stable around typical small-world values (i.e., from 4 to 10 hops) [2, 19]. We can also note that the average clustering coefficient tends to values between 0.1 and 0.2, thus suggesting that the coauthors of a researcher have 10% to 20% of chance to be connected among themselves. These values tend to slightly diminishe over time, as small components tend to connect to form larger components reducing the average clustering coefficient value. When it comes to assortativity, we see that this measure tends to 0, but it is still positive. This means that there is a slight tendency in these communities of nodes to connect with others with similar degree. A positive value for assortativity is a typical characteristic of sociological networks [20].

In general, we can note that scientific communities have similar evolving characteristics and these properties are dynamic as they change over time. More important, our observations suggest that a small set of core researchers are responsible for the social clue that creates the paths among smaller and more connected research groups. In order to further investigate these core researchers in the next subsection we contrast members and non-members of the communities core.

5.2 Core vs. Other Members

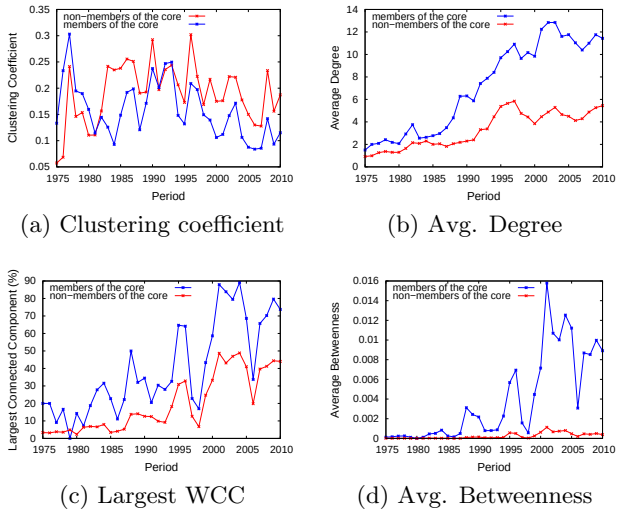
To what extend the properties of the core community differ from the rest of the community? To answer this question, we compute node network properties for members and non-members of the community core. We consider the time window analysis to understand the variations that these two classes might have in the global measure. Figure 5 shows the average degree and the average clustering coefficient computed by the members and non-members of the SIGMOD community core. Additionally, we also measure the fraction of community core members as well as non-members that are in the largest WCC and we compute the average betweenness of each of these group of members.

We can make key observations from theses analysis. First, we can note that the average degree of the members of the core are considerably higher in comparison with non-members, as they tend to establish more and more connections as a function of time. However, the clustering coefficient of the members of the core tend to be slightly smaller in comparison with non-members meaning that they might act like hubs, by connecting different groups with small intersection. By analyzing the fraction of members of the core community that are part of the largest WCC, we can note that it is much larger than the fraction of non-members, suggesting that they might be connecting smaller components. We confirm these observations by analyzing the betweenness centrality of these groups of researchers. We can note that the average betweenness of the core community is much higher, meaning that a higher number of shortest paths include these nodes.

Next, we investigate how aspects of the members of a core community can impact in the overall structure of the community.

5.3 Core Communities and Network Structure

We now examine to what extent the community core fluctuations affect the network structure. To that end, we compute the average core score of the members of each community over time. Intuitively, this measure captures the overall proliftness and involvement of the core members of a scientific community. Figure 6 shows this value for a set of communities as a function of time. We plot it in two separate figures to facilitate visualization. We can note that all communities experimented rises and falls along its life time, indicating strong variations in the core score values of the



(a) Clustering coefficient (b) Avg. Degree

(c) Largest WCC (d) Avg. Betweenness

Figure 5: SIGMOD network properties for members and non-members of the core

members of the core. We can speculate innumerous factors that are able to explain such variations, including expansion or reduction in the number of published papers, raise and fall of hot topics with ability to attract or loose important core members, members involved in the conference organization, etc. However, disregarding what caused these variations, we want to investigate if such variations can directly impact the network structure.

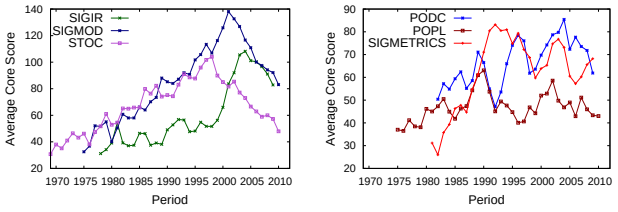


Figure 6: Avg. core score of scientific communities

Our approach to investigate this issue, consists of computing the pearson correlation coefficient between the average core score of each scientific community and a number of network metrics for that community. Table 3 presents these values.

We make key observations from these analysis. First, we can note that the diameter of a conference is positive correlated with the average core score. Although for some communities we cannot see values close to 0 or even negative values (e.g. Mobicom, with -0.04), the average correlation coefficient for all communities is 0.49, which indicates an overall positive tendency. This means that when the average core score of a community increases or decreases, the diameter tends to follow the same tendency. This suggests that core score members might connect smaller components, creating bridges among them, which contributes to increase the overall diameter. This conjecture is also supported by the high coefficient correlation for the average shortest path (in average 0.49) and the size of the largest WCC (in average 0.5).

Second, on one hand, we can note that a highly positive correlation coefficient between the average core score of communities and the average degree of the network and, in the other hand, we can observe a strong negative correlation with the assortativeness of the network. This suggests that an increase in the average community core increases the set of highly connected nodes in the network. But, although they create paths among components, they tend to connect themselves mostly with nodes of small degree values, decreasing the assortativeness of the network. Indeed, a senior researcher might tend to be coauthor of a high number of students and young researchers, but also keep collaborations with other senior researchers from other groups.

Finally, despite the expected variations, we note a clear pattern for most of the conferences on each of the analyzed metrics (i.e. clear positive or negative correlations for most of the communities). This reinforces that our observations holds for a significant number of scientific communities.

6. CONCLUSIONS

In this work we provide a deep investigation of the roles that members of the core of scientific communities play in the coauthorship network structure formation and evolution. Our effort builds upon previous existent studies as it focuses on the core community instead of analyzing the evolutionary aspects of entire communities. To do that, we define a community core based on a metric namely *core score*, an h-index derived measure that capture both, the prolifcness and involvement of authors in a community. Our analysis suggests that the members of the core community work as bridges that connect smaller clustered research groups. Additionally, we note that the members of the core community tend to increase the average degree of the network and decrease the assortativeness. More important, we note that variations on the members of the community core are strongly correlated with variations on network properties. Our study also highlights the importance to study the members of the community core and we hope that our observations might inspire future community formation models.

As future work, we would like to extend and apply our analysis of the community core to other contexts such as massive multiplayer games and online social networks.

Acknowledgments

This work was partially funded by InWeb - The Brazilian National Institute of Science and Technology for the Web (grant MCT/CNPq 573871/2008-6), and by the authors' individual grants from CNPq, CAPES e FAPEMIG.

7. REFERENCES

[1] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *Proc. of KDD*, 2008.

[2] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. Four degrees of separation. In *Proc. of ACM Web Science*, 2012.

[3] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proc. of KDD*, 2006.

[4] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proc. of ICWSM*, 2010.

Table 3: Corelation between average core score of the communities core and network metrics

	Diameter	Avg. Short P.	Clus. Coef.	Assort.	Larg. WCC	Avg. Deg.
CCS	0.34	0.2	0.23	-0.2	0.45	0.14
CHI	0.75	0.79	-0.62	-0.74	0.76	0.77
CIKM	0.56	0.56	-0.52	-0.67	0.39	0.87
DAC	0.8	0.85	-0.49	-0.63	0.76	0.92
HSCC	0.17	0.45	-0.62	-0.71	0.87	0.55
ICSE	0.81	0.83	-0.52	-0.84	0.68	0.8
ISCA	0.63	0.55	0.54	-0.32	0.63	0.81
ISSAC	0.05	0.01	-0.25	-0.43	-0.07	0.21
KDD	0.1	0.17	-0.33	-0.67	0.2	0.14
MICRO	0.35	0.35	0.28	-0.36	0.52	0.51
MOBICOM	-0.04	0.11	0.13	-0.65	0.23	-0.09
Multimedia	0.67	0.68	-0.91	-0.95	0.67	0.69
PODC	0.4	0.42	-0.23	-0.2	0.13	0.68
POPL	0.21	0.2	0.23	-0.43	0.25	0.19
SAC	0.48	0.59	0.16	-0.39	-0.55	0.16
SIGCOMM	0.18	0.19	0.05	-0.81	0.49	0.41
SIGCSE	0.88	0.84	-0.22	-0.5	0.93	0.87
SIGDOC	0.73	0.78	-0.36	-0.89	0.66	0.76
SIGGRAPH	0.79	0.85	-0.45	-0.75	0.94	0.88
SIGIR	0.83	0.85	-0.42	-0.77	0.7	0.89
SIGMETRICS	0.31	0.24	0.3	-0.44	0.37	0.64
SIGMOD	0.78	0.81	0.27	-0.61	0.77	0.87
SIGUCCS	0.38	-0.22	0.53	-0.13	0.51	0.7
STOC	0.61	0.63	0.54	-0.37	0.82	0.88
Average	**0.49**	**0.49**	**-0.11**	**-0.56**	**0.5**	**0.59**

[5] D. Chakrabarti, R. Kumar, and A. Tomkins. Evolutionary clustering. In *Proc. of KDD*, 2006.

[6] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *Proc. of KDD*, 2008.

[7] N. Ducheneaut, N. Yee, E. Nickell, and R. J. Moore. The life and death of online gaming communities: a look at guilds in world of warcraft. In *Proc. of CHI*, 2007.

[8] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proc. of PNAS*, 102(46):16569–16572, 2005.

[9] J. Hopcroft, O. Khan, B. Kulis, and B. Selman. Tracking evolving communities in large linked networks. *Proceedings of the National Academy of Sciences*, 101:5249–5253, April 2004.

[10] J. Huang, Z. Zhuang, J. Li, and C. L. Giles. Collaboration over time: characterizing and modeling network evolution. In *Proc. of WSDM*, 2008.

[11] D. Kempe, J. Kleinberg, and Éva Tardos. Influential nodes in a diffusion model for social networks. In *Proc. of ICALP*, 2005.

[12] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proc. of KDD*, 2003.

[13] J. Kleinberg. The convergence of social and technological networks. *Commun. ACM*, 51(11):66–72, Nov. 2008.

[14] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Proc. of KDD*, 2006.

[15] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *Proc. of KDD*, 2008.

[16] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proc. of KDD*, 2005.

[17] J. Leskovec, K. J. Lang, and M. Mahoney. Empirical comparison of algorithms for network community detection. In *Proc. of WWW*, 2010.

[18] M. Ley. Dblp: some lessons learned. *VLDB*, 2(2):1493–1500, 2009.

[19] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proc. of IMC*, 2007.

[20] M. Newman and J. Park. Why social networks are different from other types of networks. *Phys. Rev. E*, 68, 2003.

[21] A. Patil, J. Liu, B. Price, H. Sharara, and O. Brdiczka. Modeling destructive group dynamics in on-line gaming communities. In *Proc. of ICWSM*, 2012.

[22] E. M. Rogers. *Diffusion of Innovations*. 1962.

[23] M. Sachan, D. Contractor, T. A. Faruquie, and L. V. Subramaniam. Using content and interactions for discovering communities in social networks. In *Proc. of WWW*, 2012.

[24] D. Saez-Trumper, G. Comarela, V. Almeida, R. Baeza-Yates, and F. Benevenuto. Finding trendsetters in information networks. In *Proc. of KDD*, 2012.

[25] M. Seifi and J.-L. Guillaume. Community cores in evolving networks. In *Proc. of MSND*, 2012.

[26] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the evolution of user interaction in facebook. In *Proc. of WOSN*, 2009.

[27] D. Watts and P. Dodds. Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 34(4):441–458, 2007.

[28] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proc. of WSDM*, 2010.