Temporal Evolution of Scientific Communities

Bruno Leite Alves
UFMG
Belo Horizonte, Brazil
bruno.leite@dcc.ufmg.br

Fabrício Benevenuto
UFMG
Belo Horizonte, Brazil
fabricio@dcc.ufmg.br

Alberto H. F. Laender
UFMG
Belo Horizonte, Brazil
laender@dcc.ufmg.br

ABSTRACT

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

communities, scientific communities, core community, evolution

1. INTRODUCTION

Since its beginning, society has been organizing itself into communities, which are groups of people with common interests. Particularly, the proliferation of new communication technologies based on the Internet has facilitated the rapid formation and growth of online communities. 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 community 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 [1–4] and existing theories [18,24], models [9,10], and approaches [20,25] 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 scientific community leaders play and how they can impact intrinsic evolving properties of research communities. When prolific research leaders decide to join or leave communities, they take with them resources, experience, students and they possibly influence other authors to do the same, which makes scientific communities very suitable for this kind of study. We used data from DBLP to identify scientific communities, represented by the main ACM SIGs 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 structure.

The study of the community core of scientific communities is of interest from two different perspectives. The first is sociological, steemming 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 participants. On the other hand, from a technological perspective, understanding these aspects is critical not only for link prediction as well as the designing better recommendation systems, but it is also a necessary step for viral marketing strategies and social campaigns. 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.

Our results show that (TO DO).

The rest of this paper is organized as follows. Next section surveys related efforts. Then, Section 3 describes our strategy and dataset used to construct the connections around scientific communities and analyzes the main evolving properties of this communities. Section 4 describes our strategy to compute the community core and Section 5 investigate the main properties of these sets of authors within their communities. Finally, Section 6 concludes the paper and provides directions for future work.

2. RELATED WORK

Studies has been done to understand the structure of social networks, some studies has focus at temporal evolutions. In this terms, [23] shows a studies where they used two Facebook's dataset. The first dataset is about user's profile information, in other words, it is just the user's public information and friendship. The second dataset contains the data of interaction between the user's Facebook. In Facebook, a user's friends can post comment's to the user's wall, these comments appear on the user's wall and can be seen by others who visit the user's profile. In these way, [23] modeled the first dataset as a undirectional graph and the second as a directional graph to analysis. [23] shows that the second dataset contains information about link's disappearance, because user usually stopped to interact, but hardly he removes a friendship tie, it is the cause of the first dataset do not have this information. Other cause of the first dataset do not represent the real world is that some users just accept a friendship invite as courtesy. In this paper also is showed a measures to show the link's evolutions on the time, the resemblance. [23] showed that, in the second dataset, though there is high churn in the user pairs that interact over time, many of the global structural properties remained relatively constant over time.

Is showen in another study [11] the components evolutions of two social networks of a big company. These study analysis the components of three ways. The analysis is done about nodes that have degree one, intermidiate components (components that are not the gigant component and has more than one degree) and the gigant component. In the intermidiate components is showed the concept of star nodes, this nodes are importants to the social networks, because they can make the social network grow. Also proposed is mathematic model that making the prediciton of behavior identified in these study. In this study [11] is done a characterize about the nodes type too where they can be passive, linkers and invites.
[22]
[5]
[2]
[17]
[15]
[19]
[13]
[28]
[26]
[27]
[8]
[6]
[12]

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. In our context, a scientific community is defined in terms of a large and well established scientific conference able to aggregate authors working in similar research topics along a considerable number of years. Next, we describe how we have built a large set of scientific communities and present basic network evolution properties of them.

3.1 Dataset

In order to build a set of scientific communities to study, we have gathered data from DBLP[1] [14], 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 reconstructing 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[2] and the total number of authors, publications and editions as well as ratios extracted from these last three figures.

3.2 Communities Evolution

As an attempt to understand the main structural properties of the scientific communities, next we examine the evolution of the structure of their networks on the light of complex network analysis. To do so, we calculated 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 connected component. Figure 5 shows how each of these four metrics vary over time for a set of six scientific communities selected among those that spam over the longest period in our dataset. Our analysis results are similar for other communities, but we omit them due to lack of space.

We can note from Figure 5 that the largest connected component tend to largely increase as a function of time. This suggests that 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 colabrate with other peers of the same community. Thus, with time, authors from different groups tend to collaborate and increase the size of the largest connected component. As a consequence, the average shortest path, computed only on the largest connected component, tends to increase, becoming stable around typical small-world values (i.e., from 4 to 10 hops) [?,?]. We can also note that the average clustering coefficient tends to values between 0.1 and 0.2, thus suggesting that the coauthors of an author have 10% to 20% of chance to be connected among themselves. These

[1]http://dblp.uni-trier.de/
[2]Obtained from the SHINE (Simple H-INdex Estimator) projet: 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

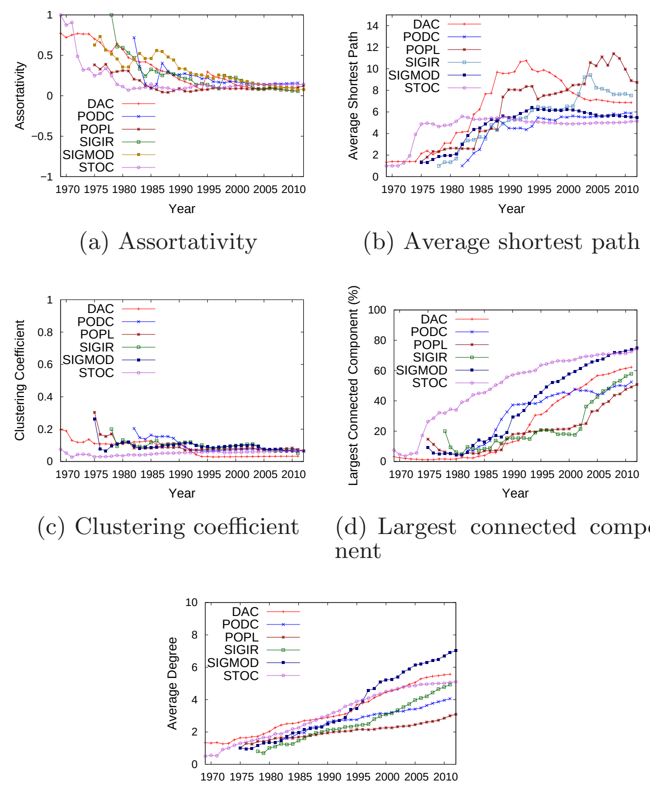
values tend to slightly diminishe over time, as small components tend to connect to form larger components reducing the average clustering coefficient value. When it comes to assortativity, we see that this measure tends to 0, but it is still positive. This means that there is a slight tendency in these communities of nodes to connect with others with similar degree. A positive value for assortativity is a typical characteristic of sociological networks [16].

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 authors are responsible for the social clue that create the paths among smaller and more connected research groups. Next, we seek to further investigate this group of authors. To that end, in the next section we propose an approach to identify the authors' core of scientific communities.

4. DEFINING COMMUNITY CORE

Previous attemps 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 [21]. However, as we plan to investigate the role of a core in the network structure, any approach that uses to network structure to identify such nodes could lead us to a biased set of authors. Instead, we focus on develop a metric that quantify 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 an author in his life as well as in other communities and ii) the frequency of involvement with the community in periods of time.

In order to capture the prolifiness of an author, we use a common metric namely h-index [7]. This metric consists of an index that attempts to measure both the productivity and impact of the published work of a researcher and



(a) Assortativity

(b) Average shortest path

(c) Clustering coefficient

(d) Largest connected component

(e) Average Degree - NAO CITADO NO TEXTO

Figure 1: Metrics accumulated from 1 in 1 year

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 h-index h_a if h of her N papers have at least h citations each, and the other (N âĹŠ h) papers have no more than h citations each. As example, if an author have 10 papers with at least 10 citations, she has h-index 10. Finally, as an attempt to capture the importance of an author to a specific community in a certain period of time, we multiple the h-index by the number of publications this author has in a certain community during a time windown. We name this measure as *Core Score*. More formally, the Core Score of an author a in a certain community c on certain period of time t, $CoreScore_{a,c,t}$, is given by its $h\text{-}index_a$ multiplied by the number of publications a have in c during t.

$$CoreScore_{a,c,t} = h\text{-}index_a * number\ of\ publications_{a,c,t} \quad (1)$$

We note that the first part of the equation captures the importance of an author to the scientific community in different areas and periods and the second part weight this importance based on the activity of the author in a certain community and time. By computing the core score of each author of a community we can determine the community core in a certain period as the top researchers of that community in terms of their core scores in that period. Next, we detail how we inferred the h-index of authors in Section 4.1. Then, Section 4.2 discusses how we define two important thresholds: the size of the community core based on the author's core score and the time windown used in our analyses.

4.1 Inferring Authors H-index

There are multiple tools that measure the h-index of research authors, out of which Google Citations[3] is the most prominent one. However, to have a profile in this system, a researcher needs to sig up and explicitly create her research profile. In a prelimianry data collection of part of the profiles of the DBLP SIGMOD authors we found that less than 30% of the authors of these communities had a profile in the Google citations. Thus, this strategy would reduce our dataset and potentiallly introduce bias for network analysis.

To divert from this limitation, we used data from SHINE, the Simple HINdex Estimation project[4], to infer community authors' h-index . SHINE is a search system that allow users to check the h-index of thousands of 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 does not allow one to search for an author instead of a conference, the shine builders kindly allowed us to access their database to infer the h-index of authors based on the conferences they crawled.

There is a limitation with this strategy. As SHINE does not track all the existent computer conferences, authors' h-index might be underestimated when computed with this data. To investigate this issue, we compared the h-index of a set of authors with account on Google citations with their estimated h-index based on the SHINE dataset. We randomly selected 10 researchers for each conference from

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

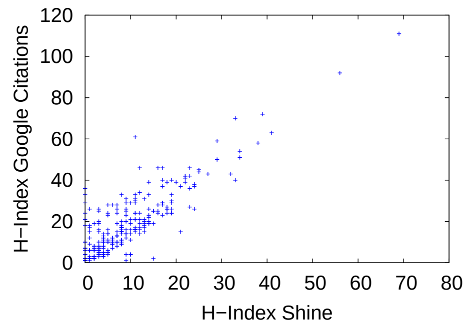


Figure 2: Correlation between the inferred H-Index and Google Scholar Citations

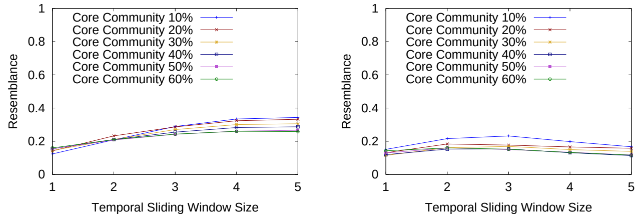
Table 1 and we extracted their h-index from their Google scholar profiles. In comparison with the H-index we estimated from SHINE, the Google scholar values are in average 51% higher. Figure 2 shows the scatter plot for the two h-index. We can note that although the SHINE h-index is smaller, the two measures are highly correlated. Indeed, the pearson correlation coefficient is 0.85, which indicates that authors might have proportional h-index estimations in the two systems.

4.2 Setting the thresholds

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 windown in which the core community is computed. In other words, should we compute the community core at each year, at each two years, or for a larger time windown? The second thereshold 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 windown, it is important to define the threshold for choosing the top ones.

Our strategy to define these thresholds consists of varying each of them and quantify how they impact on the changes on the members of the community core. To measure these changes, we compute a metric namely ressemblance, as used in [23], which measures the fraction of members in the core at time t_0 and remain in the core at the time t_1. For each conference, we vary the window size from 1 to 5 years and the size of community core from 10% to 60% of the entire community.

Intuitively, high ressamblance variations indicate bad threshold choices, and thus, we should seek for values in which threshold changes causes slightly changes on ressamblance. Figure 3 shows the ressemblance values as a function of the window size, providing different curves for the community core size. We choose the SIGMOD and SIGDOC for these analysis. The rest of the communities are ommited due to lack of space, but the same observations hold for them. By 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 ressamblence after these values. To help us decide we compute the angular coeficient for the 10% core size curves of each community and obtained the average angular coefficient for them. Based on this value, we choose the window size for our experiments as 3.



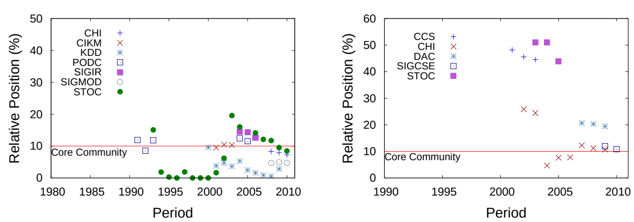
(a) SIGMOD (b) SIGDOC

Figure 3: Average of the values of resemblance

4.3 Validation

Based on the measure of core score, we expect that the members of the community core would be members of standing that actively contribute with publications to a certain community. The validation of these matters is, by nature, subjective. Next, we provide evidences that our approach correctly captures these expected properties.

First, we analyze the core score of two WWW 2013 keynote speackers: Jon Kleinberg and Luis von Ahn. Figure 4 shows the ranking position in terms of percentage (i.e. position 5% of that community) of these two authors in the communities they have published. The red line divides the members of the community core from the other members. We can note that Jon Kleinberg was a member of the community core of STOC for years, a theorical conference. Indeed, he was part of the STOC core for twelve years, publishing more than min(one), max(seven) papers at each period of three years. With Kleinberg's involvements on KDD, he became less active in STOC and left the that community core. During this period he published more than min(one), max(five) KDD papers at each period of three years whereas his publications in STOC were reduced to 1 or 2 in nine - one is 0 consecutive periods of three years. 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 XX papers along his acadamic life. He reached the core of the CHI community along three consecutive windows, with more than min(one), max(four) CHI papers per period.



(a) Jon Kleinberg (b) Luis von Ahn

Figure 4: Core score of two WWW 2013 keynote speakers

Next, we compute a ranking of researchers that appear most often in the community core of each scientific community. We choose the SIGCHI, SIGCOMM, SIGIR, and SIGMOD to show the top 20 researchers in Table 2. We can note that several big names appear in this ranking, including past keynote speakers of these events as well as awarded authors 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 appear in the top 100 list of researchers that appear in the community core. More specifically, these fractions are XX% of the awarded SIGCHI [5] members, XX% for SIGCOMM[6], XX% for SIGIR[7], and XX% for SIGMOD[8]. These observations provides evidences that our approach correctly captures the notion of a scientific community core.

5. PROPERTIES OF THE COMMUNITIES CORE

Next, we present a series of analyses about the scientific communities core. First, we analyze the network properties of scientific communities evolution in Section 5.1. Then, Section 5.2 contrasts the properties of the communities core over time against the properies of the remaining communities authors. Then, in Section **??** we compute the average core score of a community to investigate variations properties of the members of the core of each community. Finally, Section 5.3 correlates these variations with the network properties of the communities.

5.1 The Core Over Time

As an attempt to understand the main structural properties of scientific communities, next we examine the evolution of the network structure of scientific communities on the light of complex network analysis. To do so, we calculated various network metrics for each of the scientific community. We present four popular metrics here: assortativity, average clustering coefficient, average path length, and the size of the largest connected component. Figure 5 shows how each of the four metrics vary over time for a set of selected scientific communities. Our analysis results are similar for other communities, but we omit these due to lack of space.

We can note from Figure 5 that the largest connected component tend to largely increase as a function of time. This suggests that at early stages, scientific communities are formed by several small and segregated research groups. With time some students become professors, leave an institute and begin collaborations with other research groups. Additionally, as the community evolves, head of research labs tend to colabrate with peers of the same community. Thus, with time, authors from different groups tend to collaborate and increase the size of the largest connected component. As a consequence the average shortest path, computed only on the largest connected component, tend to increase and then become stable around typical small-world values (i.e. from 4 to 10 hops). We can also note that the average clustering coefficient tend to values between 0.1 and 0.2, suggesting that the co-authors of an author have 10% to 20% of chance to be connect among themselves. This value tend to slightly diminishes over time, as small components tend connect to form larger components reducing the average clustering coefficient value. When, it comes to assortativity, we see that this measure tend to 0, but is still positive. This means that there is a slight tendency in these

[5]http://www.sigchi.org/about/awards/
[6]http://www.sigcomm.org/awards/sigcomm-awards
[7]http://www.sigir.org/awards/awards.html
[8]http://www.sigmod.org/sigmod-awards

Table 2: The researchers who appear most often in the Core Community over the years.

	CHI	SIGCOMM	SIGIR	SIGMOD
1st	Scott E. Hudson	Scott Shenker	W. Bruce Croft	David J. DeWitt
2nd	Elizabeth D. Mynatt	George Varghese	Clement T. Yu	Michael Stonebraker
3rd	Hiroshi Ishii	Hui Zhang	Susan T. Dumais	H. V. Jagadish
4th	Steve Benford	Donald F. Towsley	James Allan	Rakesh Agrawal
5th	Shumin Zhai	Hari Balakrishnan	Justin Zobel	Christos Faloutsos
6th	Brad A. Myers	Ion Stoica	Alistair Moffat	Raghu Ramakrishnan
7th	Ravin Balakrishnan	Srinivasan Seshan	Norbert Fuhr	Jiawei Han
8th	James A. Landay	Deborah Estrin	James P. Callan	Gerhard Weikum
9th	George G. Robertson	David Wetherall	Yiming Yang	Philip A. Bernstein
10th	Michael J. Muller	Thomas E. Anderson	Edward A. Fox	Jeffrey F. Naughton
11th	Mary Czerwinski	Jennifer Rexford	Gerard Salton	Hector Garcia-Molina
12th	Robert E. Kraut	Jia Wang	Ricardo A. Baeza-Yates	Michael J. Carey
13th	Loren G. Terveen	Ratul Mahajan	Jian-Yun Nie	Joseph M. Hellerstein
14th	Carl Gutwin	Vern Paxson	Mark Sanderson	Philip S. Yu
15th	Ken Hinckley	Mark Handley	Charles L. A. Clarke	Divesh Srivastava
16th	W. Keith Edwards	Yin Zhang	Chris Buckley	Michael J. Franklin
17th	Gregory D. Abowd	Peter Steenkiste	Chengxiang Zhai	Jennifer Widom
18th	Anind K. Dey	Walter Willinger	Alan F. Smeaton	Hans-Peter Kriegel
19th	Saul Greenberg	Ramesh Govindan	Zheng Chen	Hamid Pirahesh
20th	Susan T. Dumais	Jon Crowcroft	Ophir Frieder	Surajit Chaudhuri

communities of nodes to connect with others with similar degree. A positive value for assortativity is a typical characteristic of sociological networks [16].

All in all, 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 authors are responsible for the social clue that create the paths among smaller and more connected research groups. Next, we seek to further investigate this group of authors. To that end, in the next section we propose an approach to identify the author's core of scientific communities.

5.2 Core vs. Other Members

So, to what extend the properties of the core community differ from the rest of the community? Next, we compute node network properties for members and non-members of the core community. We consider the time window analysis to understand the variations that these two classes might have in the global measure. Figure 6 shows the average degree and the average clustering coefficient computed by the members and non-members of the SIGMOD core community. Additionally, we also measure the fraction of community core members as well as non-members that are in the largest connected component. We can make key observations from theses analysis. First, we can note that the average degree of the members of the core considerably higher in comparison with non-members, as they tend to stablish more and more connections as a function of time. However, the clustering coefficient of the members of the core tend to be slighly smaller in comparison with non-members meaning that they might act like hubs, by connecting different groups with small intersection. Indeed, by analyzing the fraction of members of the core community that are part of the largest connected component, we can note that it is much larger than the fraction of non-members, suggesting that they might be connecting smaller components. 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

reproduzir trabalho do rich club [28]

We now examine to what extent the community core fluctuations affect the network structure.

6. CONCLUSIONS

7. FUTURE WORK

-Ajustes na formula de Core Score:
-Computar h-index a todo momento, considerando o tempo t em analise
-Mudar a formula para considerar publicacoes feitas no passado, no entanto, colocar uma penalidade. (vida trabalho do peterson)

-Realizar estudo semelhante em outra rede, tipo twitter, o h-index poderia ser alterado para um calculo de infuencia

8. REFERENCES

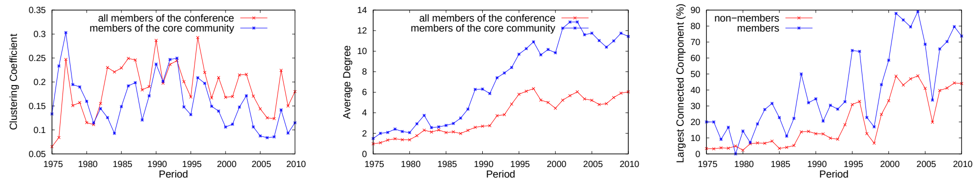
[1] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *Proc. of KDD*, 2008.

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

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

[4] 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.

[5] N. Ducheneaut, N. Yee, E. Nickell, and R. J. Moore. The life and death of online gaming communities: a



(a) Clustering coefficient (b) Avg. Degree (c) Largest connected component

Figure 6: SIGMOD network properties for members and non-members of the core

Table 3: Corelation between average core score of the core community and the metrics of complex networks

Conference	**Diameter**	**Avg. Short P.**	**Clus. Coef.**	**Assort.**	**Larg. WCC**	**Avg. Deg.**	**Num. Nodes**
CCS	0.34	0.2	0.23	-0.2	0.45	0.14	-0.13
CHI	0.75	0.79	-0.62	-0.74	0.76	0.77	0.58
CIKM	0.56	0.56	-0.52	-0.67	0.39	0.87	0.64
DAC	0.8	0.85	-0.49	-0.63	0.76	0.92	0.84
HSCC	0.17	0.45	-0.62	-0.71	0.87	0.55	-0.55
ICSE	0.81	0.83	-0.52	-0.84	0.68	0.8	0.78
ISCA	0.63	0.55	0.54	-0.32	0.63	0.81	0.41
ISSAC	0.05	0.01	-0.25	-0.43	-0.07	0.21	0.78
KDD	0.1	0.17	-0.33	-0.67	0.2	0.14	0.2
MICRO	0.35	0.35	0.28	-0.36	0.52	0.51	0.36
MOBICOM	-0.04	0.11	0.13	-0.65	0.23	-0.09	0.02
Multimedia	0.67	0.68	-0.91	-0.95	0.67	0.69	0.75
PODC	0.4	0.42	-0.23	-0.2	0.13	0.68	0.57
POPL	0.21	0.2	0.23	-0.43	0.25	0.19	0.05
SAC	0.48	0.59	0.16	-0.39	-0.55	0.16	0.23
SIGCOMM	0.18	0.19	0.05	-0.81	0.49	0.41	-0.03
SIGCSE	0.88	0.84	-0.22	-0.5	0.93	0.87	0.8
SIGDOC	0.73	0.78	-0.36	-0.89	0.66	0.76	0.05
SIGGRAPH	0.79	0.85	-0.45	-0.75	0.94	0.88	0.55
SIGIR	0.83	0.85	-0.42	-0.77	0.7	0.89	0.88
SIGMETRICS	0.31	0.24	0.3	-0.44	0.37	0.64	0.43
SIGMOD	0.78	0.81	0.27	-0.61	0.77	0.87	0.68
SIGUCCS	0.38	-0.22	0.53	-0.13	0.51	0.7	0.57
STOC	0.61	0.63	0.54	-0.37	0.82	0.88	0.68
Average	**0.49**	**0.49**	**-0.11**	**-0.56**	**0.5**	**0.59**	**0.42**

look at guilds in world of warcraft. In *Proc. of CHI*, 2007.

[6] S. Garg, T. Gupta, N. Carlsson, and A. Mahanti. Evolution of an online social aggregation network: an empirical study. In *Proc. of IMC*, 2009.

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

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

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

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

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

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

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

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

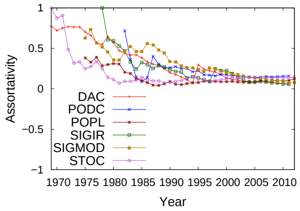
[15] G. R. Lopes, M. M. Moro, and J. P. M. de Oliveira. Temporal influence in collaborators recommendation in social networks. In *Proc. of IADIS*, 2011.

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

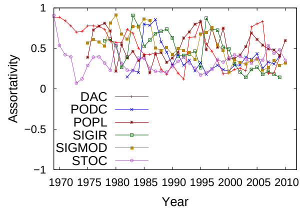
[17] A. Patil, J. Liu, B. Price, H. Sharara, and O. Brdiczka. Modeling destructive group dynamics in on-line gaming communities. In *Proc. of ICWSM*, 2012.

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

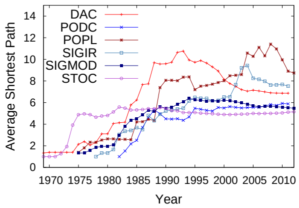
[19] M. Sachan, D. Contractor, T. A. Faruquie, and L. V. Subramaniam. Using content and interactions for discovering communities in social networks. In *Proc. of WWW*, 2012.



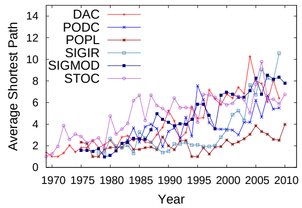
(a) Final Assortativity



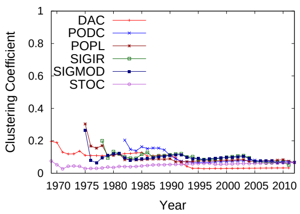
(b) Assortativity per Window



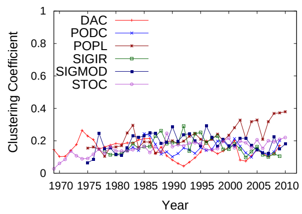
(c) Final average shortest path



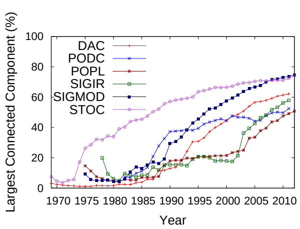
(d) Avg. shortest path per window



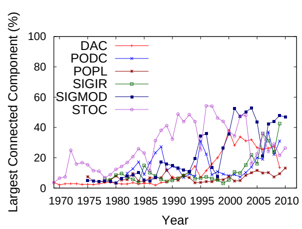
(e) Final clustering coefficient



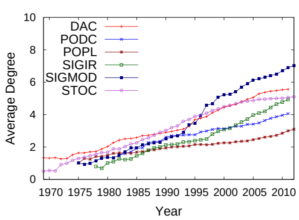
(f) Clustering coefficient per window



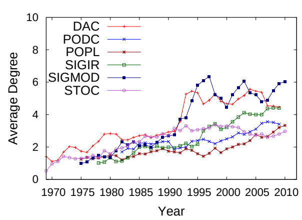
(g) Final largest connected component



(h) Largest connected component per window



(i) Final average degree



(j) Avg. degree per window

Figure 5: Metrics accumulated from 1 in 1 year

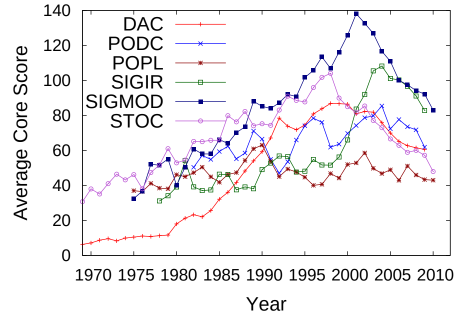


Figure 7: Average Core Score

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

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

[22] X. Sun, J. Kaur, S. Milojević, A. Flammini, and F. Menczer. Social dynamics of science, 2012.

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

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

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

[26] W. Willinger, R. Rejaie, M. Torkjazi, M. Valafar, and M. Maggioni. Research on online social networks: time to face the real challenges. *SIGMETRICS Perform. Eval. Rev.*, 37(3):49–54, 2010.

[27] B. Wu, F. Zhao, S. Yang, L. Suo, and H. Tian. Characterizing the evolution of collaboration network. In *Proc. of SWSM*, 2009.

[28] X.-K. Xu, J. Zhang, and M. Small. Rich-club connectivity dominates assortativity and transitivity of complex networks. *Phys. Rev. E*, 82:046117, 2010.