How Connected are the ACM SIG Communities?

Fabrício Benevenuto, Alberto H. F. Laender, Bruno Leite Alves,
Computer Science Department,
Universidade Federal de Minas Gerais, Brazil

Currently, computer scientists publish more in conferences than journals and several conferences are the main venue in many computer science subareas. There has been considerable debate about the role of conferences for computer science research and one of the main arguments in favor of them is that conferences bring researchers together, allowing them to enhance collaborations and establish research communities in a young and fast-evolving discipline. In this work, we investigate if computer science conferences are really able to create collaborative research communities by analyzing the structure of the communities formed by the flagship conferences of several ACM SIGs. Our findings show that most of these flagship conferences are able to connect their main authors in large and well-structured communities. However, we have noted that in a few ACM SIG flagship conferences authors do not collaborate over the years, creating a structure with several small disconnected components.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications—*Data Mining*

General Terms: Measurement

Additional Key Words and Phrases: ACM SIGs, h-index, scientific communities

1. INTRODUCTION

There is a long debate about the role of conference publications in computer science [Fortnow, 2009; Patterson, 2004; Vardi, 2009; Vardi, 2010; Vardi, 2014]. On one hand, some researchers argue that conferences offer a fast and regular venue for publication of research results at the same time that allow researchers to interact with each other. These interactions would be the key for the development of research communities in a relatively young and fast-evolving discipline. On the other hand, there exists some criticism to the conference system due to the short time for referees to review the papers, the limited size of the papers, the review overload of program committee members, and the limited time for authors to revise their papers after receiving reviews.

Despite the existing concerns of this controversial topic, conferences are quite important today as computer scientists give a huge value to them [Franceschet, 2010; Freyne et al., 2010; Laender et al., 2008]. Particularly, the flagship conferences of the ACM Special Interest Groups (SIGs) are often the most prestigious ones, usually being listed among the most important venues of several computer science subareas.

Although the importance of the main ACM SIG conferences to their respective research fields is incontestable, part of the argument in favor of conferences is that they help create and maintain an active research community, by simply offering a place for researchers to meet regularly and promote collaborations. In this work, we aim at investigating two questions related to this context: *(1) How structured are the ACM SIG conference communities?* and *(2) Who are the individuals responsible for connecting each ACM SIG conference community?*

Our effort to answer the first question consists in analyzing the coauthorship graph structure of the communities formed by the flagship conferences of the ACM SIGs. Our findings show that most of the ACM SIG conferences are able to connect their main authors in large and well-structured connected components of a coauthorship network and only very few conferences, such as the ACM Symposium on Applied Computing, flagship conference of SIGAPP, and the ACM Conference on Design of Commu-

nications, flagship conference of SIGDOC, do not form the typical structure of a research community, presenting a set of small and disconnected components.

To approach our second question, we present a tool that allows one to visualize research communities formed by authors from specific ACM SIG conferences, making it possible to identify the most prolific authors with a high level of participation in a given community. To do that, we use data from DBLP[1] and Google Scholar[2] to construct scientific communities and identify their leaders. Our visualization tool also allows a plethora of interesting observations about the authors as we shall see later.

2. ACM SIG COMMUNITIES

In order to construct scientific communities from ACM SIG conferences, we have gathered data from DBLP [Ley, 2002; Ley, 2009], a digital library containing more than 3 million publications from more than 1.5 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 the data and constructing entire scientific communities.

Each publication is accompanied by its title, list of authors, year of publication, and publication venue, i.e., conference or journal. For the purpose of our work, we consider a research network as a coauthorship graph in which nodes represent authors (researchers) and edges link coauthors of papers published in conferences that put together specific research communities [Alves et al., 2013]. In order to define such communities, we focus on the publications from the flagship conferences of major ACM SIGs. Thus, we define a scientific community by linking researchers that have coauthored a paper in a certain conference, making the ACM SIG flagship conferences to act as communities in which coauthorships are formed.

In total, 24 scientific communities have been constructed. Table I lists these communities, including the respective ACM SIG, the conference acronym, the period considered (some conferences had their period reduced to avoid hiatus in the data), the total number of authors, publications and editions as well as ratios extracted from these last three figures.

3. STRUCTURE OF THE ACM SIG COMMUNITIES

Ideally, it is expected that over the years conferences are able to bring together researchers of mutual interest so that they can collaborate to advance a certain field. Thus, it is expected that with a few decades, the coauthorship graph of a certain community contains a largest connected component (LCC) that puts together a large part (i.e., the majority) of its authors. In other words, one could expect a large LCC in a research community in which authors often interact and collaborate, meaning that there exists at least one path among a large fraction of them.

Table II shows the percentage of the authors of each community that are part of the largest connected component (LCC) of its respective coauthorship graph. Clearly, we can note that most of the research communities formed by SIG conferences have a large connected component that is typically larger than half of the network, suggesting that these conferences have successfully put together their researchers in a collaborative network. Figure 1 depicts the networks of the three conferences with the most representative largest connected components, SIGMOD, STOC and CHI, and the three conferences with the least representative ones, SIGUCCS, SAC and SIGDOC. In these networks, connected components are shown with different colours and the LCC is presented as the most central one. The size of each node represents an estimative of the importance of a researcher to the scientific community, which is discussed in the next section. As we can see, the latter are the only three communities that

[1] http://dblp.uni-trier.de
[2] scholar.google.com

Table I. DBLP statistics for the flagship conferences of the ACM SIGs

SIG	Conference	Period	Authors	Publications	Editions	Aut/Edi	Pub/Edi	Aut/Pub
SIGACT	STOC	1969-2012	2159	2685	44	49.07	61.02	0.80
SIGAPP	SAC	1993-2011	9146	4500	19	481.37	236.84	2.03
SIGARCH	ISCA	1976-2011	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	5095	2819	19	268.16	148.37	1.81
SIGCOMM	SIGCOMM	1988-2011	1593	796	24	66.38	33.17	2.00
SIGCSE	SIGCSE	1986-2012	3923	2801	27	145.30	103.74	1.40
SIGDA	DAC	1964-2011	8876	5693	48	184.92	118.60	1.56
SIGDOC	SIGDOC	1989-2010	1071	810	22	48.68	36.82	1.32
SIGGRAPH	SIGGRAPH	1985-2003	1920	1108	19	101.05	58.32	1.73
SIGIR	SIGIR	1978-2011	3624	2687	34	106.59	79.03	1.35
SIGKDD	KDD	1995-2011	3078	1699	17	181.06	99.94	1.81
SIGMETRICS	SIGMETRICS	1981-2011	2083	1174	31	67.19	37.87	1.77
SIGMICRO	MICRO	1987-2011	1557	855	25	62.28	34.20	1.82
SIGMM	MM	1993-2011	5400	2928	19	284.21	154.11	1.84
SIGMOBILE	MOBICOM	1995-2011	1151	480	17	67.71	28.24	2.40
SIGMOD	SIGMOD	1975-2012	4202	2669	38	110.58	70.24	1.57
SIGOPS	PODC	1982-2011	1685	1403	30	56.17	46.77	1.20
SIGPLAN	POPL	1975-2012	1527	1217	38	40.18	32.03	1.25
SIGSAC	CCS	1996-2011	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	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	4978	2623	20	248.90	131.15	1.90

are formed by a very small largest connected component (i.e., with less than 10% of the researchers in the network) and several other small connected components. Typically, these conferences cover a wide range of topics, making it difficult for their researchers to establish a research community. For example, SAC is an annual conference organized in technical tracks that change at each edition. Although this dynamic format attracts a large number of submissions every year, it does not contribute to the formation of a specific, well-structured research community.

4. LEADERS AND THEIR ROLES IN RESEARCH COMMUNITIES

We now turn our attention to our second research question related to identifying important members of a research community. Our intention here is not to rank researchers within their communities, but to give a sense about which researchers have being engaged in a certain community for consecutive years and mostly helped connecting its final coauthorship graph. Thus, instead of attempting to quantify centrality measures [Freeman, 1977; Girvan and Newman, 2002] of authors and node degree in coauthorship graphs, we have defined a metric that aims at quantifying the involvement of a researcher in a scientific community in terms of publications in its flagship conference over the years. Intuitively, this metric should be able to capture (i) the prolificness of a researcher and (ii) the frequency of her involvement with a certain community. Next we discuss how exactly we have defined this metric.

4.1 Quantifying a Researcher's Engagement in a Community

First, in order to capture the prolificness of a researcher, we use the h-index [Hirsch, 2005], 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 publications and the number of citations that they have received. For example, a

Table II. Structure of Scientific Communities

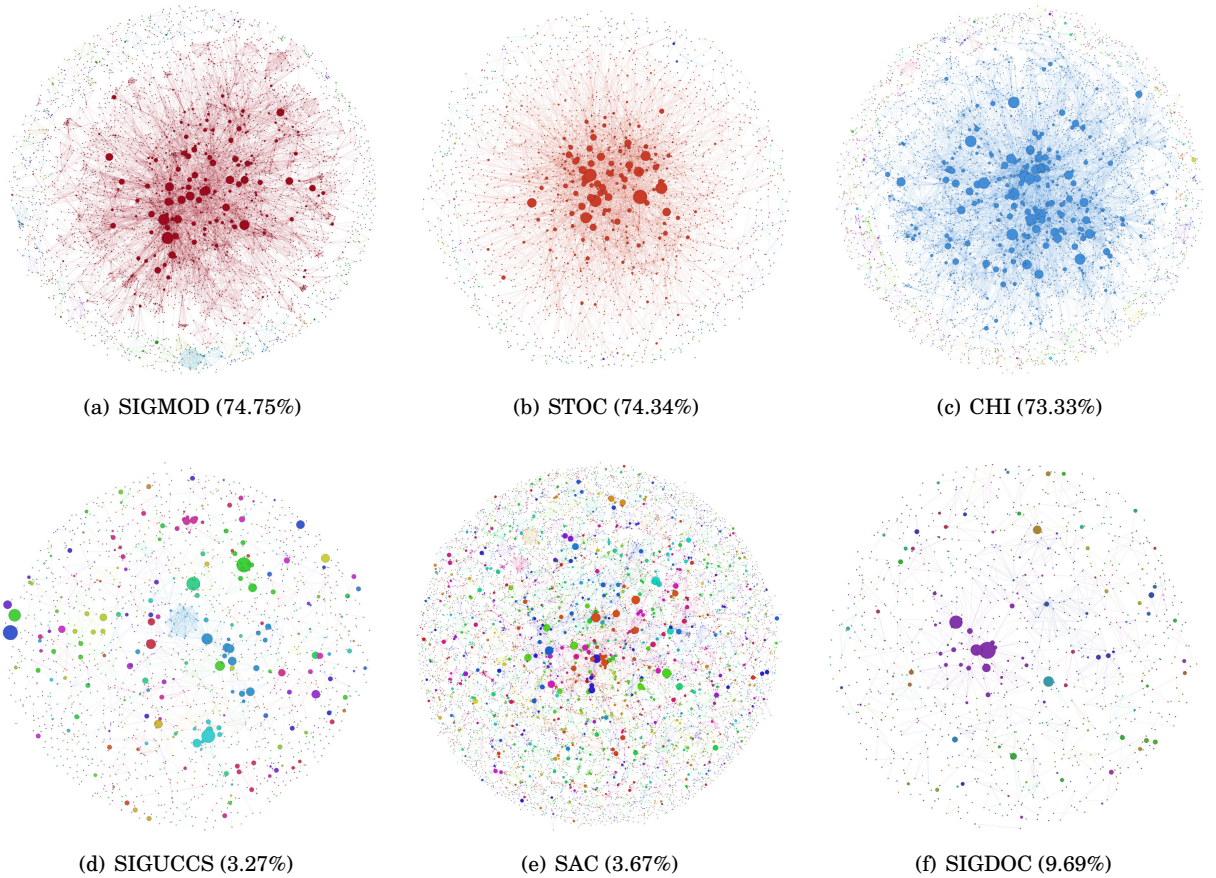
Conference	Largest Connected Component
SIGMOD	74.75%
STOC	74.34%
CHI	73.33%
MICRO	65.13%
HSCC	62.53%
DAC	62.21%
KDD	61.24%
ISCA	58.72%
SIGCOMM	57.88%
SIGIR	57.86%
SIGCSE	55.31%
ICSE	52.68%
PODC	52.46%
CIKM	51.81%
CCS	51.70%
SIGMETRICS	50.89%
POPL	50.82%
MM	50.06%
SIGGRAPH	46.72%
ISSAC	44.09%
MOBICOM	37.88%
SIGDOC	9.69%
SAC	3.67%
SIGUCCS	3.27%

researcher r has an h-index h_r if she has at least h publications that have received at least h citations. Thus, for instance, if a researcher has 10 publications with at least 10 citations, her h-index is 10.

Then, as an attempt to capture the importance of a researcher to a specific community in a certain period of time, we multiply her h-index by the number of publications this researcher has in a certain community (conference) during a time window. We name this metric *CoScore*, as it aims to measure the importance of a researcher as a member of the community. More formally, the *CoScore* of a researcher r in a community c during a period of time t, $CoreScore_{r,c,t}$, is given by her h-index h_r multiplied by the number of publications r has in c during t (#$publications_{r,c,t}$), as expressed by the following equation:

$$CoScore_{r,c,t} = h_r \times \#publications_{r,c,t} \tag{1}$$

We note that the first part of the above equation captures the importance of a researcher to the scientific community as a whole regardless of 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 over a period of time. The idea is to compute the amount of time a certain research appeared among the top researchers in terms of this metric over periods of a few consecutive years. For example, if a researcher that today has a high h-index has published four papers at KDD in a period of three years, it means she is engaged with that community at least for that short period of time. If a researcher appears among the top ones within a community for several of these periods, it suggests that she has a life of contributions dedicated to that community. Next, we briefly describe how we have inferred the h-index of the researchers.



(a) SIGMOD (74.75%) (b) STOC (74.34%) (c) CHI (73.33%)

(d) SIGUCCS (3.27%) (e) SAC (3.67%) (f) SIGDOC (9.69%)

Fig. 1. Scientific communities and the size of their LCC

4.2 Inferring Researchers' H-index

There are multiple tools that measure the h-index of researchers, out of which Google Scholar Citations[3] 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 authors 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 the SHINE (Simple HINdex Estimator) project[4] to infer the researchers' h-index. SHINE provides a website that allows users to check the h-index of almost eighteen hundred computer science conferences. The SHINE developers crawled Google Scholar, searching for the title of papers published in these conferences, which allowed them to effectively estimate the h-index of the target conferences based on the citations computed by Google Scholar. Although SHINE only allows one to search for the h-index of conferences, the SHINE developers kindly allowed us to access their dataset to infer the h-index of researchers based on the conferences they crawled.

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

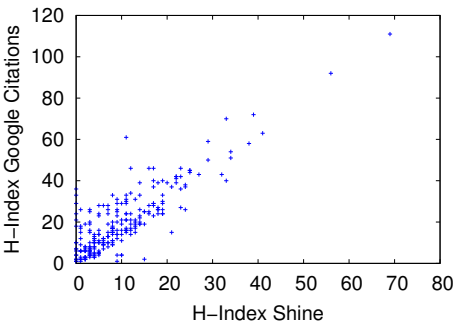


Fig. 2. Correlation between the inferred h-index and Google Scholar Citations one

However, there is a limitation with this strategy. As SHINE does not track all existing 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 from each conference in Table I 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 2 shows the scatter plot for the two h-index measures. We can note that although the SHINE-based h-index is smaller, the two measures are highly correlated. The Pearson's correlation coefficient is 0.85, which indicates that researchers might have proportional h-index estimations in both systems.

4.3 Visualizing Community Members and their Roles within the Communities

In order to make our results public, we have developed an interactive tool[5] that allows one to browse the scientific communities, visualizing their structures and the contribution of each specific researcher to connect their coauthorship graph. Our effort consists in allowing users to search for researchers based on the metric presented in the previous section. The size of each author's node is proportional to the number of times she appears within the top 10% researchers with highest *CoScore* values in a time windows of three years. Figure 3 shows, for example, the coauthorship graph of Michael Stonebracker, the winner of the 2014 A.M. Turing Award[6], and his connections within the SIGMOD community. These connections are highlighted when one passes the mouse over the researcher's name. In addition, our tool allows one not only to search for authors but also to visualize statistics about them within the communities.

To check if our approach really identifies those who are prolific and engaged in a specific community, we notice that several research communities have established different awards to recognize those who were important to a certain field and helped to advance or even build a certain community. Thus, we use some of these awards to corroborate the effectiveness of our metric in establishing the importance of a researcher within a specific community. We have computed a ranking of the researchers that appear most often in the top 10% of the *CoScore* ranking over the years for each community. We have chosen the CHI, ICSE, KDD, POPL, SIGCOMM, SIGGRAPH, SIGIR, and SIGMOD communities to show their top 20 researchers in Tables III and IV. As we can see, several well known names appear in these top lists, including past keynote speakers of those conferences and awardees for their life time

[5]Available at www.acmsig-communities.dcc.ufmg.br
[6]http://amturing.acm.org/stonebraker_1172121.pdf

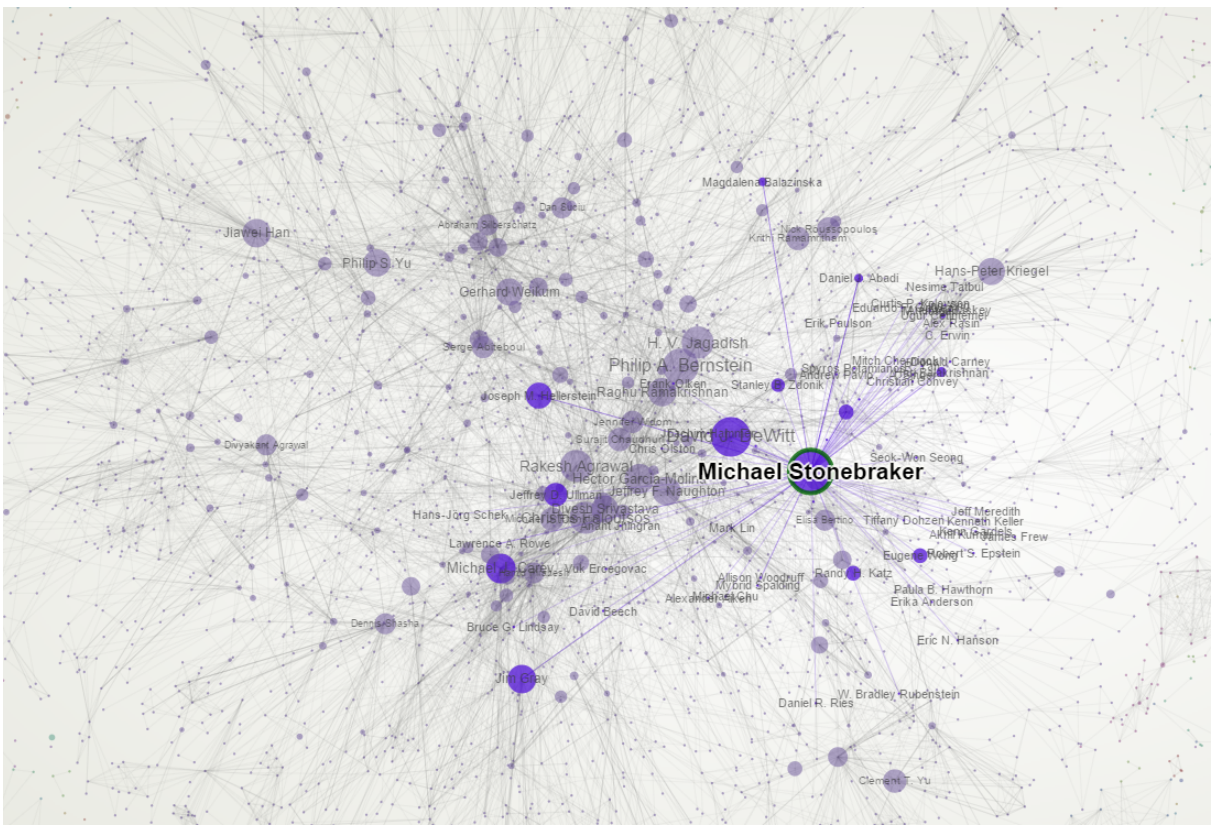


Fig. 3. Michael Stonebraker and his connections within the SIGMOD community

contributions in the respective community (names in bold). In addition, besides Michael Stonebraker, these top lists include four other winners of the A.M. Turing Award (indicated by asterisks): Amir Pnueli (1996), Barbara Liskov (2008), Edmund M. Clarke (2007) and Jim Gray (1998). Indeed, by analyzing all these awardees from each community, we found that a large fraction of them appeared in the top 10% of the *CoScore* ranking at least once in the conference history. For example, according to the respective ACM SIG websites, these fractions are 75% for KDD[7], 35% for SIGCOMM[8], 60% for SIGIR[9], and 80% for SIGMOD[10]. Except for SIGCOMM, a community with many sponsored conferences that were not considered in our dataset, the other three communities presented very high numbers of awardee members that appear at least once in the top 10% of the *CoScore* ranking over the years. These observations provide evidence that our approach correctly captures the notion we wanted to.

5. CONCLUSIONS

This work analyzes the structure of the communities formed by the flagship conferences of ACM SIGs. Our findings show that most of the ACM SIGs are able to connect their main authors in connected and

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

Table III. Researchers that appear most often in the top 10% of the CoScore ranking over the years

CHI	ICSE	KDD	POPL
Scott E. Hudson	**Victor R. Basili**	**Heikki Mannila**	Thomas W. Reps
Hiroshi Ishii	**Barry W. Boehm**	Hans-Peter Kriegel	Martn Abadi
Steve Benford	**Jeff Kramer**	**Jiawei Han**	John C. Mitchell
George G. Robertson	**Mary Shaw**	Martin Ester	Robert Harper
Shumin Zhai	Dewayne E. Perry	**Rakesh Agrawal**	Zohar Manna
Brad A. Myers	Don S. Batory	Bing Liu	Benjamin C. Pierce
Robert E. Kraut	Mary Jean Harrold	Ke Wang	Amir Pnueli*
Elizabeth D. Mynatt	**Lori A. Clarke**	**Padhraic Smyth**	**Barbara Liskov***
Ravin Balakrishnan	Gruia-Catalin Roman	Philip S. Yu	Martin C. Rinard
James A. Landay	Premkumar T. Devanbu	Charu C. Aggarwal	Luca Cardelli
Ken Hinckley	Gail C. Murphy	**Vipin Kumar**	Thomas A. Henzinger
Mary Czerwinski	**Richard N. Taylor**	Wynne Hsu	**Ken Kennedy**
Carl Gutwin	**David Garlan**	Qiang Yang	**Matthias Felleisen**
Gregory D. Abowd	Michael D. Ernst	**Christos Faloutsos**	Edmund M. Clarke*
Michael J. Muller	James D. Herbsleb	William W. Cohen	Mitchell Wand
Susan T. Dumais	Lionel C. Briand	Pedro Domingos	David Walker
Loren G. Terveen	Gregg Rothermel	Eamonn J. Keogh	Simon L. Peyton Jones
Steve Whittaker	Kevin J. Sullivan	Alexander Tuzhilin	Shmuel Sagiv
W. Keith Edwards	**David Notkin**	Mohammed Javeed Zaki	Barbara G. Ryder
John M. Carroll	Douglas C. Schmidt	Mong-Li Lee	Alexander Aiken

Table IV. Researchers that appear most often in the top 10% of the CoScore ranking over the years

SIGCOMM	SIGGRAPH	SIGIR	SIGMOD
Scott Shenker	**Donald P. Greenberg**	**W. Bruce Croft**	**Michael Stonebraker***
George Varghese	**Pat Hanrahan**	Clement T. Yu	**David J. DeWitt**
Donald F. Towsley	Demetri Terzopoulos	**Gerard Salton**	**Philip A. Bernstein**
Ion Stoica	**David Salesin**	Alistair Moffat	H. V. Jagadish
Hui Zhang	**Michael F. Cohen**	**Susan T. Dumais**	Christos Faloutsos
Deborah Estrin	**Richard Szeliski**	James Allan	**Rakesh Agrawal**
Hari Balakrishnan	John F. Hughes	Yiming Yang	**Michael J. Carey**
Robert Morris	N. Magnenat-Thalmann	Edward A. Fox	**H. Garcia-Molina**
Thomas E. Anderson	**Tomoyuki Nishita**	James P. Callan	Jiawei Han
Ramesh Govindan	**Andrew P. Witkin**	Chris Buckley	Raghu Ramakrishnan
Srinivasan Seshan	Norman I. Badler	**C. J. van Rijsbergen**	Jeffrey F. Naughton
David Wetherall	**Peter Schrder**	Justin Zobel	**Jim Gray***
Yin Zhang	Steven Feiner	Ellen M. Voorhees	Hans-Peter Kriegel
Jennifer Rexford	**Hugues Hoppe**	Mark Sanderson	Gerhard Weikum
Jia Wang	**Jessica K. Hodgins**	**Norbert Fuhr**	Philip S. Yu
J. J. Garcia-Luna-Aceves	**Greg Turk**	Nicholas J. Belkin	Divesh Srivastava
Randy H. Katz	**Marc Levoy**	Chengxiang Zhai	Joseph M. Hellerstein
Albert G. Greenberg	**P. Prusinkiewicz**	Charles L. A. Clarke	Krithi Ramamritham
Mark Handley	Eihachiro Nakamae	Alan F. Smeaton	Nick Roussopoulos
Simon S. Lam	Dimitris N. Metaxas	Gordon V. Cormack	**Surajit Chaudhuri**

visually well-formed communities. However, we note that a few conferences, such as the ACM Symposium on Applied Computing, flagship conference of SIGAPP, and the ACM Conference on Design of Communications, flagship conference of SIGDOC, do not form a strong research community, presenting a structure with several disconnected components. We have opened our results to the research community as an interactive visualization tool that allows one to browse the scientific communities, visualizing their structures and the contribution of each specific researcher to connect its coauthorship graph.

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.

REFERENCES

Alves, B. L., Benevenuto, F., and Laender, A. H. F. (2013). The Role of Research Leaders on the Evolution of Scientific Communities. In *Proceedings of the 22nd International Conference on World Wide Web (Companion Volume)*, pages 649–656.

Fortnow, L. (2009). Time for Computer Science to Grow Up. *Commun. ACM*, 52(8):33–35.

Franceschet, M. (2010). The Role of Conference Publications in CS. *Commun. ACM*, 53(12):129–132.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

Freyne, J., Coyle, L., Smyth, B., and Cunningham, P. (2010). Relative Status of Journal and Conference Publications in Computer Science. *Commun. ACM*, 53(11):124–132.

Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–16572.

Laender, A. H. F., de Lucena, C. J. P., Maldonado, J. C., de Souza e Silva, E., and Ziviani, N. (2008). Assessing the Research and Education Quality of the Top Brazilian Computer Science Graduate Programs. *SIGCSE Bulletin*, 40(2):135–145.

Ley, M. (2002). The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. In *Proceedings of the 9th International Symposium on String Processing and Information Retrieval*, pages 1–10.

Ley, M. (2009). DBLP: Some Lessons Learned. *Proc. of VLDB Endow.*, 2(2):1493–1500.

Patterson, D. A. (2004). The Health of Research Conferences and the Dearth of Big Idea Papers. *Commun. ACM*, 47(12):23–24.

Vardi, M. Y. (2009). Conferences vs. Journals in Computing Research. *Commun. ACM*, 52(5):5–5.

Vardi, M. Y. (2010). Revisiting the Publication Culture in Computing Research. *Commun. ACM*, 53(3):5–5.

Vardi, M. Y. (2014). Scalable Conferences. *Commun. ACM*, 57(1):5–5.