Social Network Analysis and Mining manuscript No.
(will be inserted by the editor)

Citation Interactions among Computer Science Fields: A Quantitative Route to the Rise and Fall of Scientific Research

Tanmoy Chakraborty[1], Sandipan Sikdar[2], Niloy Ganguly[3], Animesh Mukherjee[4]

Received: 15.12.2013 / Accepted: 24.03.2014

Abstract In this work, we propose for the first time a suite of metrics that can be used to perform post-hoc analysis of the temporal communities of a large scale citation network of the computer science domain. Each community refers to a particular research field in this network and therefore they act as natural sub-groupings of this network (i.e., ground-truths). The interactions between these ground-truth communities through citations over the real time naturally unfolds the evolutionary landscape of the dynamic research trends in computer science. These interactions are quantified in terms of a metric called *inwardness* that captures the effect of local citations to express the degree of *authoritativeness* of a community (research field) at a particular time instance. In particular, we quantify the impact of a field, the influence imparted by one field on the other, the distribution of the "star" papers and authors, the degree of collaboration and seminal publications in order to characterize such research trends. In addition, we tear the data into three subparts representing the continents of North America, Europe and the rest of the world, and analyze how each of them influences one another as well as the global dynamics. We point to how the results of our analysis correlate with the project funding decisions made by agencies like NSF. We believe that this measurement study with a large real-world data is an important initial step towards understanding the dynamics of cluster-interactions in a temporal environment. Note that this paper, for the first time, systematically outlines a new avenue of research that one can practice post community detection.

Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur
India – 721302
E-mail: {[1]its_tanmoy, [2]sandipansikdar, [3]niloy, [4]animeshm}@cse.iitkgp.ernet.in

1 Introduction

Over the last fifty years, the domain of *Computer Science* has moved from its infancy to mature adulthood. The history of this development is peppered with contributions that attempted to increase the computational speed while significantly reducing the physical size of the computers so that they could be used more meaningfully. However, the landscape of the development of the different fields within this domain (e.g., computer hardware, programming languages, compilers, operating systems, databases, artificial intelligence, algorithms and theoretical computer science) is surely not flat [26]; in contrast, it has been throughout guided by the constant shift of focus of the research community causing some of the fields to emerge at the forefront at a particular point in time subsiding the others.

Thousands of scientific papers are being published in this domain every year and it is very crucial to understand which research areas are growing or declining, thereby, unfolding the possibility to rank the popularity and predict the evolutionary trend of various research topics. The utilities of such a ranking scheme are manifold: (a) a new researcher wishing to continue research in computer science can conduct a survey of each field in the light of how it evolved over the years and subsequently decide to choose her topic of research, (b) a funding agency that provides financial support to research projects might be interested in visualizing the landscape of development of the fields to locate bursts of certain research topics as well as to understand how the interactions between different fields change over time, (c) developing nations that are usually also late-beginners can align themselves with the rest of the world at a much quicker pace if they have a worldwide picture of the time-trend of growth and decline of the research fields. Note that precise quantitative estimates of the above factors form a first and a fundamental step toward the design and implementation of a recommendation system capable of predicting research trends in the immediate future. The central objective of this work is to execute this fundamental step by formulating suitable quantitative measures that can accurately unfold the evolutionary trend of the different research fields.

On the other hand, detecting clusters or communities in real-world graphs such as large social networks, web graphs, and biological networks is a problem of considerable practical interest and has, of late, received a great deal of attention [10] [24]. Though several works on detecting and tracking communities in a temporal environment have been conducted [1] [30], the interactive patterns of the detected communities over a temporal scale still remain unexplored mainly due to the lack of standard ground-truth communities. More specifically, one can ask for a metric to understand the dynamics and also rank the importance of various communities over time. This paper stresses on developing ground-truth overlapping communities in the form of the research fields of a large-scale directed citation network of computer science. It then systematically explores the longitudinal (i.e., with the progress of time) intercluster interactive patterns to unfold the latent characteristics of the network

that indeed explains the rise and the fall of the impact of scientific research communities over the last fifty years.

The major contributions of our work are manifold. To start with, we describe for the first time a large-scale paper-paper directed citation network of the computer science domain with the fields annotated thus representing the natural partitioning of the network into ground-truth community structures. Each field represents a community [6] [28], the communities overlap as some papers belong to multiple fields; we rigorously analyze the quality of these community structures as ground-truth using well-known community-centric metrics [33]. Next, we propose a simple edge-centric measurement called "inwardness" of a community (research field in this case) to capture the dynamics of inter-cluster interactions across time points which can explain the varying degree of impact of the scientific research communities. Subsequently, to understand this phenomenon at a more granular level, we postulate several explanations to unveil the possible reasons for such a dynamical behavior of research communities using exhaustive statistical analysis. In particular, we quantify the impact of a scientific community, the influence imparted by one community on the other, the distribution of the "star" papers and authors, the degree of collaboration and seminal publications; all these properties together contribute to quantify the typical dynamics of research communities efficiently. In addition, we classify the entire data into three parts corresponding to the continents of North America, Europe and the rest of the world, and repeat our experiments on these three parts separately to demonstrate how each of them are connected to the worldwide shift in research focus. Finally, we validate our primary predictions by establishing their correlations with the project funding decisions made by NSF (National Science Foundation of the USA). Interestingly, the fields that are presently at the forefront influence the current funding decisions much less than the funding decisions influencing the emergence of a field at the forefront in the immediate future. It is important to remark here that the above observation indicates that predictions of our results are in lines with the intuitions of the expert researchers who are usually involved in such crucial funding decisions. We also believe that this work additionally makes important contributions purely from the perspective of citation network. This is one of the first large scale studies to understand the trends in a research field. A recent work on the computer science knowledge networks [23] has been carried out with the aim to understand its structure and to determine clusters of similar and high-prestige venues. Yang and Leskovec [32] developed ground-truth communities of real-world undirected static networks and detected overlapping communities from these networks [33]. In this experiment, we adopt a longitudinal framework to represent the ground-truth communities of citation network, and understand their evolution using simple statistical analysis. Note that through this work we present for the first time a precise methodology for post-hoc analysis of the community structures obtained from a large scale network.

2 Related work

The analysis of research trend was started with the pioneering study of Kuhn [18]. Scientometric data has been available for several decades and so it was already in the 1960s that de Sola Price [29] first observed power laws in the scientific citation networks and developed models for citation dynamics. Thereafter, a huge number of works on the analysis of citation networks have been conducted either on ranking researchers [9, 14, 31] or on ranking publications [8]. The work by Zhao et al. [34] studies the relationship between authors using community mining techniques. The empirical work by Guimera et al. [11] has shown that new collaborations between experienced authors are more likely to result in a publication in a high impact journal. A related study [27] has shown the information diffusion in citation networks by analyzing the correlations between various citation choices and the subsequent impact of the articles. They tried to establish that citing recent papers and papers within the same scholarly community garners a slightly larger number of citations on average. There has also been interest in visualizing and quantifying the amount of information flow between different areas of science [4] – in effect, mapping the generation of human knowledge through information flow. The development of efficient network algorithms has led not just to discoveries of the overall properties of citation networks, but also the detection of changes in citation patterns where a new trend or paradigm emerges [20]. Recently, Mazloumian et al. [21] consolidated the claim of "rich-gets-richer effect" in predicting the Nobel prize winners through citation network analysis. Redner [25] analyzed the citation statistics of 110 years of Physical Review journal and unleashed few unfamiliar characteristics of citation networks of physics.

On the other hand, there has been some research on modeling evolution of trends that have become popular over the last few years [16, 19]. One of the recent models proposed by Bornholdt et al. [3] considers an interactive agent-based information spreading game with a suitable tuning parameter called "innovation rate". In addition to providing a theoretical understanding of how scientific research trends change over time, the model also provides insights that help explain some related observations in real life. Recently, Pan et al. [22] and Chakraborty et al. [5] demonstrated how, over time, interdisciplinarity is increasingly becoming more dominant thus triggering a shift in the overall trend of physics and computer science research respectively.

Despite such a burgeoning number of research contributions in this area, a systematic approach to analyze how the research focus of the computer science community has evolved over time through a "tug-of-war" amongst the constituent fields remains largely unidentified. This serves as the primary motivation for the current study that attempts to present a full-fledged analysis of the temporal behavior of citation networks. The study attempts to explicitly measure the time-varying importance of various fields along with a rigorous investigation of the factors that regulate the growth and the decline of the popularity of these fields.

3 Preliminary definitions

In this section, we outline the definitions of certain terms which we shall be repeatedly using throughout the rest of the paper.
Domain: We define domain as the broad subject of interest that can be further categorized into multiple sub-classes. Computer science, physics, biology, chemistry are some representative examples of broad research domains.
Field: Fields are more fine-grained research sub-topics of a domain. Algorithm, artificial intelligence, database management system are a few examples of fields within the computer science domain. Sometimes, interdisciplinary research activities across several fields/domains are responsible for the emergence of new fields. For example, computational biology refers to the application of computer science and information technology to the field of biology and medicine.
Citation network: Our method is primarily based on suitable statistical analysis of various properties of citation networks that can be formally defined as a graph $G =< V, E >$ where each node $v_i \in V$ represents a paper and a directed edge e_{ji} pointing from v_j to v_i indicates that the paper corresponding to v_j cites the paper corresponding to v_i in its references. For the purpose of our analysis, we label all the papers in the network with the information about the field to which each of them belong. At a higher tier, each field (i.e., a collection of papers) can be thought of as a single node and two field nodes can be again linked by a directed edge with edge-weights calculated using Equation 1 mentioned below.
Impact of a field: *Impact* defines how important a field is in terms of the research activities going on in that field. We quantify the importance of a paper in terms of the total number of inward citations to the paper (aka *inwardness*). The inwardness $In(f_i)$ of a field f_i (i.e., a collection of papers) can be defined as

$$In(f_i) = \sum_{j \neq i} w_{j \to i} \tag{1}$$

where $w_{j \to i}$ represents the weight of the edge connecting field f_j to f_i. Here, it is worth noting that this metric is only applicable for the network of fields since the weight $w_{j \to i}$ is determined by the ratio of the number of citations ($c_{j \to i}$) from the papers of field f_j to the papers of field f_i to the total number of papers in field f_i (say (p_i)). In other words,

$$w_{j \to i} = \frac{c_{j \to i}}{p_i} \tag{2}$$

Note that this inwardness metric is a measure of the degree of authoritativeness of a research field proposed here for the first time and defined in the lines of what has been already discussed in the context of individual publications in [17].
Lead and Lag: We define $lead(x, y, t)$ to denote that the event x took place t years before the event y. Similarly, we define $lag(x, y, t)$ to denote that the event x took place t years after the event y.

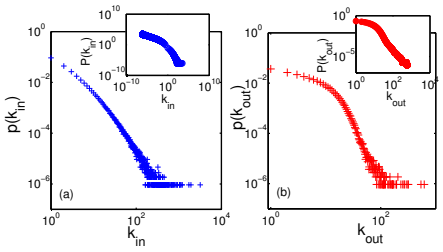

Fig. 1 (Color online) Degree distribution ((a) indegree and (b) outdegree; inset: cumulative degree distribution) of the raw citation dataset.

Table 1 General information of raw and filtered dataset.

	Raw dataset	Filtered dataset
Number of valid indices	1,079,193	702,973
Number of entries with no venue	582	–
Number of entries with no author	5,773	–
Handbook	1,649	–
Archive	86,169	–
Number of papers before 1960	886	–
Number of papers having no in-citation and out-citation	272,325	– –
Partial data of the year 2009	8,836	–
Number of authors	662,324	495,991
Average number of papers by an author	3.82	3.52
Average number of authors per paper	2.615	2.609
Number of unique venues	2,319	1,705

4 Description of the dataset

The traditional information pertaining to citation networks like papers and citation distributions are not adequate in this study to meet all the experimental needs. The analysis needs several other related information about each paper, e.g., publication year, publication venue (journal/conference), research field, authors and their continents. We have used the dataset of the computer science domain developed by Tang et al. [31][1] for our experiments. It has been constructed using the DBLP web repository which contains information about various research papers from different fields of computer science domain published over the years. This information includes the name of the research paper, index of the paper, its author(s), the year of publication, the publication venue, the list of research papers the given paper cites and (in some cases) the abstract of the papers. Certain general information pertaining to the downloaded raw dataset is noted in the second column of the Table 1. Figure 1 shows the degree distribution of the raw citation network.

In order to make the data suitable for our experiments, we extract only those entries which contain the information about the paper index, the title,

[1] http://arnetminer.org/citation, named as *DBLP-Citation-network V4*

the venue of publication, the year of publication and the citations. In general, scientific focus shifts are affected manifold by contributory papers than by reviews, surveys and text books, and therefore we exclude these items from our data. Further, in order to make our data bounded we consider only those papers that cite or are cited by at least one paper. Some of the general information pertaining to the filtered dataset are presented in Table 1. The degree-distribution of the filtered network is in Figure 2.

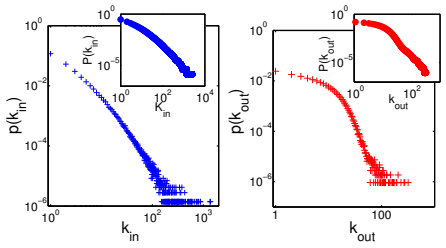

Fig. 2 (Color online) Degree distribution ((a) indegree and (b) outdegree; inset: cumulative degree distribution) of the filtered citation dataset.

4.1 Field tagging

Since the filtered dataset does not have the necessary field information of the papers, we tag them using the Microsoft Academic Search Engine[2]. This website covers more than 38 million publications and over 19 million authors across a wide variety of domains with updates added every week. It categorizes papers of computer science domain into the fields as noted in Table 2. We have crawled the site to find the field(s) of papers present in the filtered dataset using the title of the paper. Approximately, 88.12% of the papers could be tagged with their respective fields when searched with the paper title. Fields of rest 11.88% of the papers have been inserted using the conference/journal name of the paper. About 11.23% of the papers have more than one field. Table 2 notes the percentages (decreasing order) of various fields in the tagged dataset. We also show in the table the average ten-year impact (Equation 3) for each field between the years 1960 and 2008. Note that this value indicates the average number of citations that each individual paper within a field receives from the papers belonging to the other fields.

4.2 Continent tagging

Microsoft Academic Search also provides location of the authors like the name of the university/company he/she is affiliated to and the continent information

[2] http://academic.research.microsoft.com/

Table 2 Percentage of papers in various fields and their average inwardness in each decade (for each decade, top and second ranked inwardness measures are in bold font).

No.	Subject	Abbreviation	% of papers	Average Inwardness				
				60-69	70-79	80-89	90-99	00-08
1.	Artificial Intelligence	AI	15.30	0.02	0.67	**4.94**	**5.14**	**3.29**
2.	Algorithms and Theory	ALGO	14.09	**4.13**	**4.49**	3.39	2.12	0.55
3.	Networking	NW	8.63	0.19	0.53	1.06	**3.42**	1.76
4.	Databases	DB	8.12	**3.75**	**3.67**	1.80	1.14	0.17
5.	Distributed and Parallel Computing	DIST	7.63	0.02	2.02	2.86	1.55	0.56
6.	Hardware & Architecture	ARC	7.29	0.41	2.49	2.29	1.12	1.04
7.	Software Engineering	SE	6.40	1.98	3.21	1.89	1.67	0.52
8.	Machine Learning and Pattern Recognition	ML	6.09	0	0.43	2.51	2.97	**2.62**
9.	Scientific Computing	SC	4.02	0	1.14	2.38	2.91	0.19
10.	Bioinformatics & Computational Biology	BIO	3.88	0	0	0.71	1.27	0.56
11.	Human-Computer Interaction	HCI	3.42	0	0.03	1.65	2.05	1.39
12.	Multimedia	MUL	3.34	0	0.53	2.51	2.22	1.33
13.	Graphics	GRP	3.32	0	0.56	2.58	2.63	1.07
14.	Computer Vision	CV	3.03	0	0.86	1.29	2.73	1.27
15.	Data Mining	DM	3.02	0	0.27	1.80	1.83	1.02
16.	Programming Languages	PL	3.00	0.41	2.49	**3.86**	2.46	1.29
17.	Security and Privacy	SEC	2.94	0	0.86	3.80	2.56	1.59
18.	Information Retrieval	IR	2.26	0	0.42	1.32	2.62	1.79
19.	Natural Language and Speech	NLP	2.11	0	0.13	1.16	2.82	1.92
20.	World Wide Web	WWW	1.76	0	0	1.86	2.10	1.83
21.	Computer Education	EDU	1.67	0	0	0.80	0.83	0.39
22.	Operating Systems	OS	1.07	0.31	1.73	1.39	1.98	1.20
23.	Real Time Embedded Systems	RT	0.90	0	0.67	1.56	2.52	0.54
24.	Simulation	SIM	0.14	0	0.30	1.20	2.70	0.87

(North America, South America, Asia, Europe and Africa) of all the universities. In order to tag the authors with their respective continents, we search for their location through the search engine. Initially, "exact name" of an author is searched to get the location. In case of more than one match, i.e., the case where many authors have exactly the same name, the continents of all the matching authors are checked and the continent of an author is approximated by the continent name that recurs the largest number of times across the search results. Almost 71% of the authors get tagged after this step. For tagging the rest of the authors, we attempt to match an author name with names which have all tokens (ignoring unit length tokens) of the query author name. For instance, the query "Jason A Blake" can be matched with "Jason Blake Audrey". About 9% of the authors get tagged after this step. For tagging the rest of the authors, we find names that have maximum overall token match with the query author name. Around 12.4% of the authors get matched after this step. In both the previous steps, continent of query author is approximated by the one that appears the largest number of times across the search results.

Out of the 7.6% data to be tagged, we could approximate the continent of 6.6% by the most common continent that the collaborators of an author belong to. This is because we find that within the tagged set 73% of times the continent of an author matches with the continent that is most common across his/her collaborators. At the end of the above steps, 99% of the authors

Table 3 Heuristics applied for continent tagging.

Heuristics	Percentage
Exact matching with query name	71%
Matching with all tokens of query name (except unit tokens)	9%
Maximum overall token match	12.4%
Tagging approximated by the most common continent of the collaborators	6.6%
Untagged authors	1%

finally get tagged while the rest 1% of the authors are left untagged and are not used further in our analysis. The above steps are summarized in Table 3. The number of authors from Africa, South America and Asia being relatively low, we merge them together into a new category called "Others" which we use for our future experiments.

5 Characteristics of the citation network

Before proceeding to the main experiments detailing how scientific focus shifts, we analyze the dataset systematically and explore certain interesting results mentioned below.

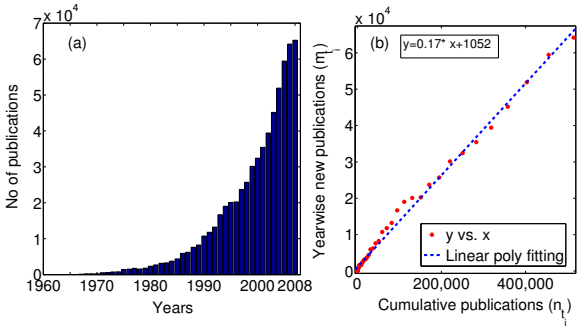

Fig. 3 (Color online) (a) Number of publications over the years; (b) Number of new publications (m_{t_i}) as a function of total number of publications (n_{t_i}).

5.1 Year-wise growth of overall publications

We first plot the number of publications over the years in Figure 3(a). Figure 3(b) shows how the number of new publications (m_{t_i}) correlates with the number of all existing publications (n_{t_i}), i.e., $n_{t_i} = \sum_{t=t_0}^{t_i} m_t$, where t_0 is the starting year and t_i is the year under consideration). The linear correlation between the two quantities indicates that the number of new publications is proportional to the number of already published articles.

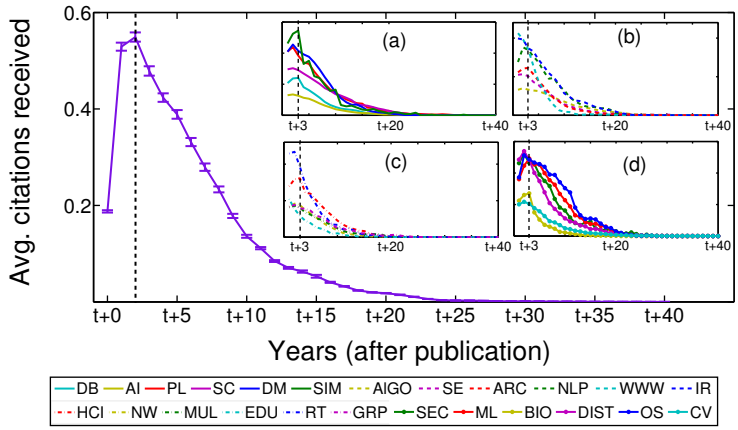

Fig. 4 (Color online) Year-wise average inward citations (inset: same measure for every field; (a) DB, AI, PL, SC, DM, SIM; (b) ALGO, SE, ARC, NLP, WWW, IR; (c) HCI, NW, MUL, EDU, RT, GRP; (d) SEC, ML, BIO, DIST, OS, CV). Y-axis value corresponding to year x indicates the average number of citations received by a paper at the xth year of its publication. Note that, the average inward citation count of a paper in our dataset is 4.36.

5.2 Distribution of average inward citation over the years

Some of the previous experimental results [12,13,15] show that the trend of citations received by a paper after its publication period is not linear in general; rather there is a fast growth of in-citations within the initial few years after the publication, followed by an exponential decay. We notice the same property in our dataset and observe that the average number of inward citations per paper peaks within three years from its publication and then slowly declines over time (see Figure 4). Note that this property is also prevalent across the different fields of the domain (see inset of Figure 4). Therefore, in order to measure the importance of a paper (or a field) around its time of publication, all our analysis throughout the rest of the paper assumes only the citations received by the paper within three years from its publication. This three-year time window helps in capturing the local importance of a paper (say, p) around its publication time and discards those citations coming from the papers published long after the publication of p.

5.3 Year-wise growth of overall inward and outward citations

Here we analyze the total number of citations received (Figure 5(a)) as well as the total number of citations made (Figure 5(c)) by all the papers. It is found that both of the characteristics grow over time. Note that the last two bars in Figure 5(a) have a lower height because it is not possible for the recently published papers to receive a very high number of citations since our dataset is bounded up to the year 2008. In addition, we also study how the number of citations (both inward and outward) received in the current year correlates with

the total number of citations received over all the years. We define $IC(OC)$ and $CIC(COC)$ as the inward (outward) citation and the cumulative inward (outward) citation respectively. Therefore $CIC_{t_i} = \sum_{t=t_0}^{t_i} IC_t$, where t_0 is the starting year and t_i is the current year. A similar definition also holds for COC. Figure 5(b) and Figure 5(d) respectively show the correlation of IC with CIC and the correlation of OC with COC. A linear correlation is observed in both the plots indicating that the number of citations (inward as well as outward) at any point in time is governed by the total number of citations so far (similar to the case of publications illustrated through Figure 3(b)). Note that, there is an order of magnitude difference between the number of inward and outward citations since the number of inward citations is always restricted within the three-year window.

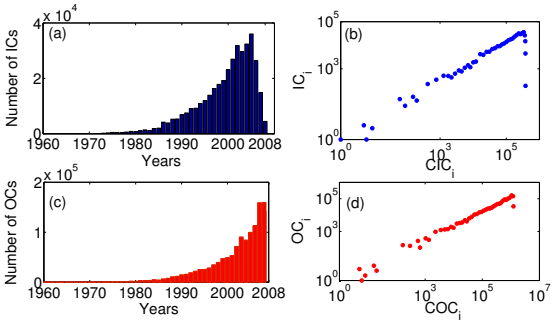

Fig. 5 (Color online) Growth of inward citations (ICs) (Fig.(a)) and outward citations (OCs) (Fig.(c)) over the years; relationship between new inward and outward citations with the existing cumulative inward (CIC) and cumulative outward (COC) citations respectively (Fig.(b) and Fig.(d)). Both axes of (b) and (d) are in logarithmic scale.

6 Community scoring functions

We now discuss various scoring functions defined by Yang and Leskovec [32] that characterize how "community-like" is the connectivity structure of a given set of nodes. The idea is that given a community scoring function, one can find sets of nodes with high/low score (depending upon the function) and consider these sets as communities. All scoring functions are built on the common intuition that communities are sets of nodes with many connections within the members and few connections from the members to the rest of the network. Out of 13 commonly used scoring functions proposed in [32], a few have been proved to be capable enough to capture the effect of all the functions. We will discuss five such effective functions that are again naturally grouped into three coarse-grained categories.

Let $G(V, E)$ be a graph with $n = |V|$ nodes and $m = |E|$ edges. Given a set of nodes S with $n_S = |S|$, $m_S = |(u, v) \in E : u \in S, v \in S|$, $c_S =$

$|(u,v) \in E : u \in S, v \notin S|$ and $d(u)$ the degree of node u, we consider a function $f(S)$ that characterizes how community-like is the connectivity of nodes in S. The symbols used in the following definitions are described in Table 4.

(A) Based on external connectivity:
1. Expansion (EXPN): It measures the number of edges per node that point outside the cluster, i.e., $f(S) = \frac{c_S}{n_S}$.
2. Cut Ratio (CUT): It is the fraction of edges (out of all possible edges) leaving the cluster, i.e., $f(S) = \frac{c_S}{n_S \times (n - n_S)}$.

(B) Based on internal connectivity:
3. Fraction over median degree (FOMD): It is the fraction of nodes of S that have internal degree higher than the median degree of a vertex in the entire network, i.e., $f(S) = \frac{|u:u \in S, |(u,v):v \in S| > d_m|}{n_s}$ where d_m is the median value of $d(v)$ for all $v \in V$.

(C) Combining internal and external connectivity:
4. Conductance (COND): It measures the fraction of total edge volume that points outside the cluster, i.e., $f(S) = \frac{c_S}{m_S + c_S}$.
5. Flake-ODF (ODF): It is the fraction of nodes in S that have fewer edges pointing inside than to outside of the cluster, i.e., $f(S) = \frac{|u:u \in S, |(u,v) \in E:v \in S| < d(u)/2|}{n_S}$.

Table 4 Used symbols to describe the community scoring functions.

Notation	Description
V	Set of nodes in the graph G
E	Set of edges in the graph G
n	Number of nodes in the graph G
m	Number of edges in the graph G
S	A selected subset of nodes drawn from V (i.e., $S \subseteq V$)
n_S	Number of nodes in S
m_S	Number of edges whose both end points are in S
c_S	Number of edges whose one end point is in S and another is outside S
$d(u)$	Degree of vertex u
d_m	Median value of the degree of vertices present in V

Note that, the less the values of EXPN, CUT, COND, and ODF, the better is the community structure of the network. But for FOMD, the reverse argument is true. However, the above mentioned functions have been proposed for the undirected graphs [32]. In the present experiment, we calculate each of the functions separately for incoming and outgoing edges and report the value after averaging them. These scoring functions are used to obtain individual scores for each community, and by averaging them we get the scores for the entire network. For the purpose of comparison, all the scores reported are rescaled within the range of 0 and 1. Since the present work deals with the time-varying communities, we report the above functions for the network in five

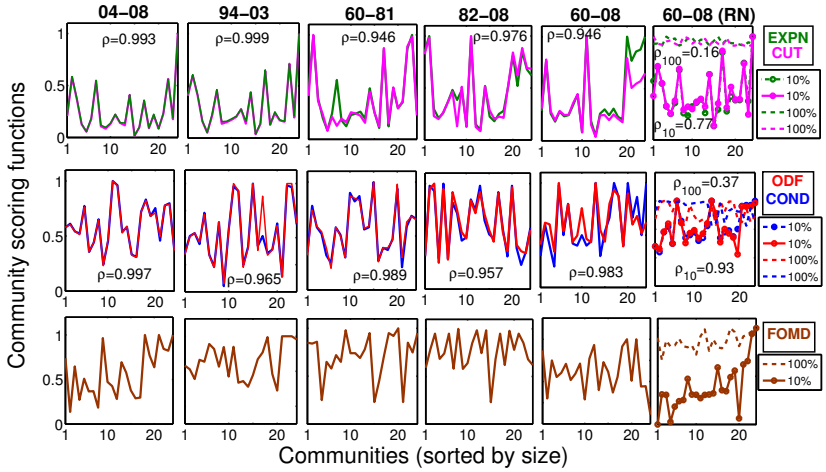

Fig. 6 (Color online) Community scoring functions for real-world ground-truth communities (sold lines) in different time windows (2004-2008, 1994-2003, 1960-1981, 1982-2008 and 1960-2008). Results from the randomized versions (10% and 100%) of the ground-truth communities are presented in the right-most panel (indicated by RN). For better visualization, all the functional values are rescaled between 0 and 1. In each slice of the figure, Pearson's correlation coefficient (ρ) between two similar functions is reported.

time-windows (2004-2008, 1994-2003, 1960-1981, 1982-2008 and 1960-2008)[3] to demonstrate the robustness of these natural groupings to different sample sizes of data (ranging from 5-year aggregate to 49-year aggregate) (see Figure 6). For each time-window, we calculate Pearson's correlation coefficient (ρ) [7] between the functions in each category (except FOMD). Across all different time points and for all different data sets we observe that the correlation between the scoring functions from within a group of measures is always almost close to one. In order to further show that the ground-truth communities are not arbitrarily formed and are actually tightly knit, we randomly swap members between communities (10% and 100% of all the nodes) keeping the community sizes intact and show that the scores as well as the correlations heavily degrade as one increases the degree of random swaps (see the last column of Figure 6). Note that, we perform the randomization experiment only on the entire network (containing all the papers published in 1960 to 2008). The intuition is that if the community scoring functions and their correlations obtained from the actual ground-truth communities outperform those values obtained from the randomized communities in the entire graph, then a similar behavior should also be observed for the subgraphs constructed at different time stamps as well [2]. We report further the actual value of the functions

[3] Note that, these results are representative and therefore hold for any reasonable size sampling of the data. The first set represents a period of the most recent 5 years; the second set corresponds to a period of 10 years from the immediate past; the third and fourth sets represent the full data partitioned into two chunks, and the last set presents the results on the entire dataset.

Table 5 Community scoring functions of the network in different time-windows with the ground-truth (GT) and random (RN) communities.

Time-windows	EXPN	CUT	COND	ODF	FOMD
GT (04-08)	0.411	0.84e(-6)	0.251	0.003	0.542
GT (94-03)	0.437	1.40e(-6)	0.332	0.004	0.522
GT (60-81)	0.710	1.02e(-6)	0.381	0.006	0.538
GT (82-08)	0.701	9.06e(-6)	0.283	0.002	0.559
GT (60-08)	0.610	1.02e(-6)	0.270	0.002	0.593
RN-10% (60-08)	0.768	1.18e(-6)	0.328	0.006	0.465
RN-100%(60-08)	0.985	2.15e(-6)	0.485	0.008	0.216

for the entire network in Table 5. Once again, note that for all different time points and sample sizes, the ground-truth data have significantly better scores as compared to their randomized counterparts. We also notice in Table 5 that the community scores obtained in the latest time-window (graph in 2004-2008) are better in comparison to the earlier time-window (graph in 1994-2003). This observation is also true for the two twenty-year time windows (i.e., community scores in 1982-2008 time-window outperform the scores in 1960-1981). This indicates that as time progresses, due to the maturity of the fields, the communities tend to become more well-formed and well-knit thus serving as better ground-truth structures. These time-varying community structures might also be very interesting in order to explain the evolutionary landscape of different fields in a particular domain [5].

7 Time transition of scientific communities

In this section, we analyze the time transition of the scientific focus showing how one field has taken over another during the time evolution of the computer sciences. In particular, we measure the inwardness of a field so as to construct the time transition diagram reflecting the focus shifts. However, here again we restrict our analysis to citations that are received within the three-year window. Consequently, we redefine the inwardness metric as follows:

$$In(f_i^t) = \sum_{j \neq i} w_{j \to i}^t \tag{3}$$

where $w_{j \to i}^t = \frac{c_{j \to i}^t}{p_i^t}$ with $c_{j \to i}^t$ corresponding to the number of citations received by the papers of field f_i from the papers of field f_j, p_i^t corresponding to the total number of papers in field f_i and $1 \leq t \leq 3$. Note that for all our estimates, in addition to this three-year window we also include the year of publication of the paper.

In order to investigate the global time transition pattern (i.e., the world-wide behavior) we compute the inwardness of each field (Equation 3), rank them and plot the top two values (see the solid and broken lines respectively in Figure 7(a)) as a function of time. Each field is uniquely color coded and

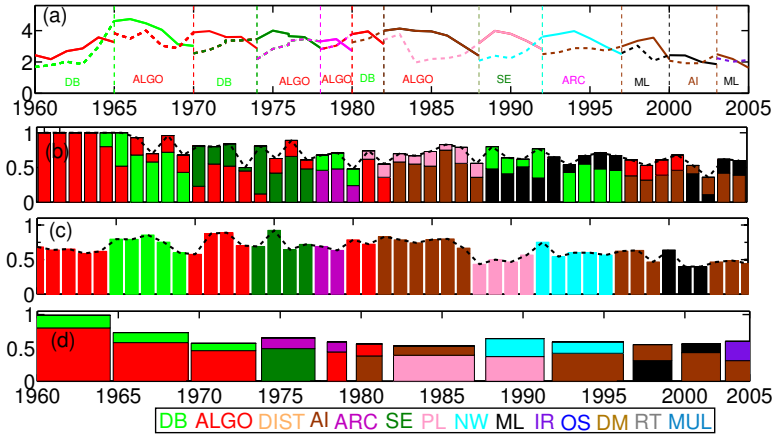

Fig. 7 (Color online) (a) Top two scientific communities (based on inwardness) at the forefront of scientific research trend (names of topmost backup communities for the communities in the forefront of every trend-window are mentioned). Cause analysis: Fig.(b) fraction of papers for the top and second ranked communities among the 10% high impact papers in each year; Fig.(c) change of citations from the topmost backup communities; Fig.(d) fraction of papers for the top and second ranked communities among the 10% highly influential papers in each trend-window. To smoothen the curves, the best sliding window size of five years has been used.

the relative height of the y-axis shows the inwardness of the field for a particular year. In each focus-window, we also mention the name of the top hub (backup) field that on an average produces the largest number of citations for the top ranking field. This information, as we shall see in section 8, forms one of the major reasons for focus shifts. The total number of transitions of research focus during 1960 to 2005 is 11 (i.e., there are 12 trend-windows in the global time transition diagram). A careful inspection of the behavior of the curves shows that in every focus-window, a similar pattern is followed with the inwardness first rising and then gradually declining near the transition. Simultaneously, the second rank field which comes to the top position in the next focus-window in every case starts reflecting a relative growth of inwardness at the middle of the current focus-window. Bornholdt et al. [3] mention a similar observation that the competing communities are as if running in a continuous race to dominate others and when the magnitudes of dominance (in this case, it is In) are nearly equal between the top and second top ranked communities, a sudden chaos among the research communities suppresses one of them and makes the other popular. However, in their model once a field declines it never rises again; in contrast, real data analysis here shows that there are at least two cases (Algorithm: 3 times, AI: 3 times) where a field can decline and then rise again at a later time. Another important issue is that the differences of inwardness between the top and the second top ranked fields in the long-ranged and short-ranged focus-windows are largely different. We investigate this property in further detail in the next section.

Table 6 Ranking of top fields in each trend-window in terms of collaborative papers, multi-continent papers and diversity (average ranks of top fields in two segments of 6 trend-windows are shown in third, fifth, seventh, tenth, twelfth and fourteenth rows).

		60-64	65-69	70-73	74-77	78-79	80-81
Collaborative	Rank	13	8	13	11	3	13
	Avg.	10.16					
Multi-continent	Rank	12	8	12	10	1	12
	Avg.	9.87					
Diversity	Rank	11	8	11	13	12	11
	Avg.	11					
		82-87	88-91	92-96	97-99	2000-2002	2003-2005
Collaborative	Rank	6	12	2	6	1	6
	Avg.	5.5					
Multi-continent	Rank	7	11	3	7	2	7
	Avg.	6.17					
Diversity	Rank	3	9	10	3	4	3
	Avg.	5.33					

8 Reasons for transitions

In this section, we conduct a diverse set of experiments to investigate the reasons behind the typical dynamics of scientific communities in the longitudinal scale observed in the previous section. We focus on different orthogonal characteristics all of which converge to reasons for the transitions observed. While the first cause that we propose is from an overall estimate of the data, the following three are time-varying estimates of the data.

Cause I: Impact of collaborations

In this section, we show that, in the current years, the expansion of collaborative work within and across continents as well as the diversity in research interest can have direct influence on the emergence of a scientific community at the forefront. To this purpose, we measure the impact of collaborative research by ranking all fields globally based on (i) the number of papers in that field having multiple authors (collaborative papers), (ii) the number of papers involving authors from multiple continents (multi-continent papers) and (iii) the diversity of a field (say, f) measured by the average number of fields that the authors of f have worked. These three ranks act as three different indicators of collaboration. Note that in case (iii), the more the diversity the higher is the rank of the field. Moreover, we suitably normalize each of the above three factors for any particular field by the total number of papers in that field. Thus, each factor indicates the average collaborative nature of a field. We then rank the fields based on each of the three normalized scores. Table 6 notes the ranks in cases (i), (ii) and (iii) for those fields that are at the forefront in terms of inwardness in each trend-window and the average rank of these fields in two segments each composed of six trend-windows. We observe that in all the three cases the average rank in the second segment is

much higher[4] than that in the first segment. This indicates that in the current years, those fields that enjoy a higher number of collaborations and a higher overall diversity in the research interests of its constituent authors have an increased chance of emerging at the forefront. The collaborative ranks of the top fields in the earlier time periods are lower mainly because of the less proportion of the collaborative/multi-continent/diverse papers in those fields. We also observe that during the earlier time periods the high ranked collaborative fields are mostly the newly emerging fields such as AI, ML, NLP. Earlier, these emerging fields contained very few papers compared to the papers in the core fields. It seems that in the early years, the top ranking fields like Algorithms and Databases (the so-called core-fields of computer science) acted as the only and therefore indispensable sources of citation for any other field. Therefore, they were able to maintain their high ranks at least in the initial years even without having much collaborations. This is precisely the reason for their low collaboration score in spite of a high inwardness score.

Cause II: High impact papers
We extract the top 10% of the papers that have the highest number of incitations (considering the last three years and the current year) from among all the papers published in a year. We call them as high-impact papers. Next we measure the fraction of papers out of this 10% that belong to a particular field. The fields are then ranked by this fraction and the fractional values are plotted in Figure 7(b) for the top and the second ranked fields. We observe that in 9 out of 11 cases a decline in the fraction of high-impact papers of the top ranked field and the simultaneous increase of high-impact papers in the second ranked field trigger a transition in Figure 7(a). Another important point to note is that in the later years, out of the 10% high impact papers, the fractions from the top and the second ranked fields diminish rapidly. While in the initial years this fraction is found to be close to 1, in the later years it drops to around 0.5. This partially indicates the maturity of the computer science domain as a whole, whereby several fields become effective and now have a place in the list of 10% high-impact papers unlike in the earlier years.

Cause III: Citation patterns of backup communities
The impact of a paper in our experiment is determined by the citations received from other papers. Therefore, one of the important factors that helps a particular scientific community to rise up to the top is the contribution of its backup communities that direct most of their outward citations to push this community to the top. In Figure 7(c), we plot bars for each year indicating the fraction of citations that the top ranked community (according to Figure 7(a)) received from its primary backup community (i.e., the backup community that brings in the largest number of citations). Note that, in 75% of the cases, the citation received from the primary backup community falls abruptly close to

[4] Note that, in this case, the rank x is higher than rank y if $x < y$ conforming to the usual notion of any ranking system.

the transition indicating that they play a pivotal role in keeping the dominant field "dominant". This abrupt fall could be possibly caused because the citations coming from the backup communities start getting shared by other competing communities and the current community at the forefront start losing its charm owing to its member topics slowly becoming dated, thereby, losing the "timeliness" advantage. We observe that the backup fields for a particular top field are not same in all time windows. Moreover, the citations from the backup fields which are mostly focused towards the top field in the initial time periods, split among multiple fields at the time of transition. In [5], Chakraborty et al. pointed out that as a field becomes more and more mature, the citations emitted to the other fields get more dispersed with a significant proportion being retained within the field (unlike in the initial years). In our experiment, we observe that the increase in the diversity of citations from the backup field is one of the main triggering factors behind the time transition, and this might be possibly tied to the overall maturity of the backup field itself to emerge as an altogether new scientific paradigm (see the example of the emergence of WWW as a new field in [5]).

Cause IV: Effect of seminal papers

The two causes discussed above have a direct bearing with the time transition of the research trend. However, there can be indirect factors affecting the rank of a community – one such factor could be the inception of seminal papers that have potential to completely mould the direction of research in the immediate future. In this section, we attempt to quantify the impact of such papers by introducing a metric called *Influence*. In particular, we consider only those citations that a paper receives from the papers belonging to its own field published within the three-year window, however, ensuring that the paper being cited does not have any author in common with the paper citing it. This expresses how important a particular paper is within its own scientific community. The influence ($Influence(p_i^t)$) of paper p_i at time t is defined as follows:

$$Influence(p_i^t) = \sum_{p_j \in P^t} \frac{1}{d_{p_j}} \tag{4}$$

where P^t is the set of all papers that cite p_i within the three year window ($1 \leq t \leq 3$) and belong to the same field as of p_i, and d_{p_j} corresponds to the total number of outward citations from the paper p_j - the fraction is used to suitably normalize the impact of citation.

We extract the top 10% influential papers in each trend-window and find out from among them the fraction of influential papers for each field. We then rank the fields based on this fraction and plot once again the top and second ranked influential fields in each trend-window in Figure 7(d). The results corroborate our hypothesis that the top rank field (inwardness based) in a certain trend-window has the highest number of influential papers in the previous window (almost in 65% cases). In the earlier years (1960 to 1975), the two fields,

namely Algorithms and Databases completely shadow all other fields in terms of papers and citations. The competitive pressure starts to appear mainly after 1975. If we measure this fraction from after 1975, we observe that in six out of seven cases (excluding the last window) the field that sees the birth of the largest number of influential papers in a trend-window emerges in the forefront in the immediate next trend-window. This observation points to the fact that the influential papers can play a very crucial role in determining the shape of the future research trend.

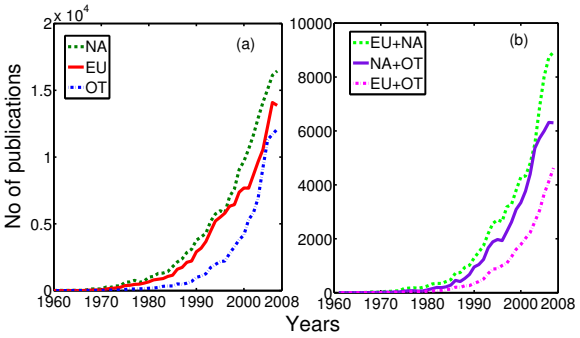

Fig. 8 (Color online) (a) Number of publications over the years where all authors are from the same continent; (b) Number of publications over the years where either of the authors belong to one of the considered continents.

9 Comparison of the continents

The citation patterns within and across different continents have an important role in the transition of worldwide research trend. Therefore, in order to complete our empirical study, in this section, we inspect the continents separately and analyze their time transition patterns in comparison to the global pattern. As a first step, we note in Table 7 the number of authors as well as the percentage authors (among the total number of authors) who belong to a particular continent. Table 8 shows the number of publications as well as the percentage of publications (among the total number of publications) that have at least one author of a given continent. As a following step, we investigate the distribution of the number of publications over the years for the three continents separately. From Figure 8(a), it is evident that this number is always higher for North America. While considering number of pairwise collaborative publications across the three continents, the collaboration between North America and Europe emerges as the strongest (Figure 8(b)).

Table 7 Number of authors from different continents.

Continent	Number of Authors	Percentage
Europe (EU)	181,978	36.69%
North America (NA)	162,104	32.68%
Others (OT)	151,909	30.63%

Table 8 Number of papers having at least one author from the corresponding continent.

Continent(C)	# of publications with at least 1 author of C	% with total publications
Europe	340,191	47.79%
North America	355,143	49.89%
Others	253,743	35.65%

9.1 Effect of cross-continent collaborations

In this section, we analyze the effects of the cross-continent collaborations. To this purpose, we first illustrate through Figure 9(a) that there has been a continuous increase in team size over the years. Figure 9(b) makes this picture more clear by showing that not only team size but also multi-continent collaboration has increased largely over the years.

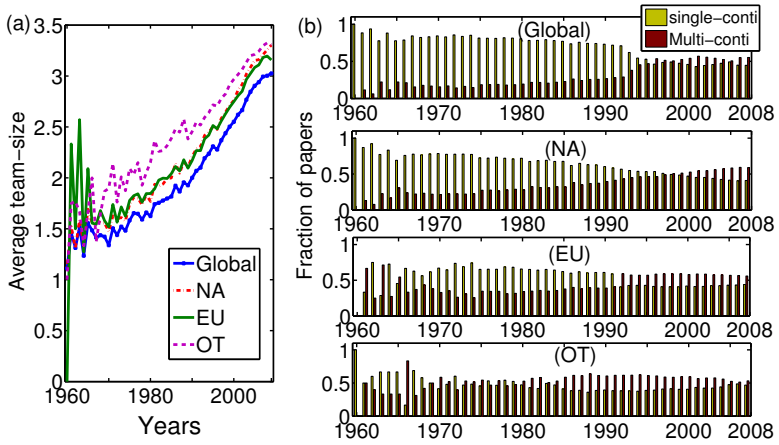

Fig. 9 (Color online) Increase in (a) average team-size and (b) multi-continent collaborative papers over the years.

9.2 Continent-wise impact of papers

The Venn diagram in Figure 10(a) depicts the distribution of publications by author(s) within and across continents. For instance, 11.62% of the publications have authors from Europe and North America only. In addition, the Venn

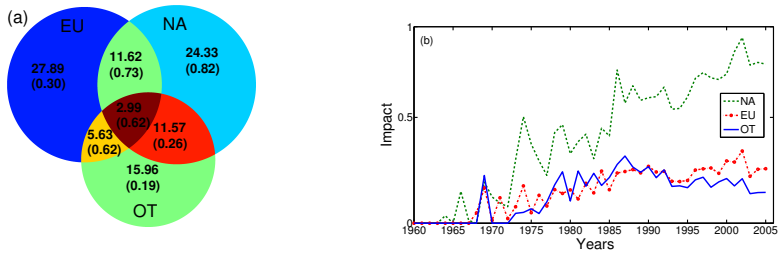

Fig. 10 (Color online) (a) Percentage of intra and inter-continent publications (terms within the parenthesis in each segment denote the average impact of the papers in that segment; (b) average impact of papers written by single-continent authors.

diagram also indicates the average inwardness (equation 3) of the publications within and across the continents. Note that North American papers by themselves have the largest inwardness. Next we compare the average impact of the papers written by authors all of whom belong to a single continent. Figure 10(b) indicates this value over the years for the three continents. Clearly, once again North American papers on an average seem to have a much more pronounced inwardness in comparison to the other continents.

9.3 Time transition patterns in the continents

We plot the time transition pattern for the three continents in Figure 11 with the continents marked as "NA" (North America), "EU" (Europe), "OT" (Others). One of the first important points to note is that the transitions take place more frequently here. The total number of transitions for NA, EU and OT are 19, 15 and 17 respectively (as opposed to 11 for the global case). Note that, in this experiment we have considered that a paper belongs to continent C if one of the authors belongs to C. Surprisingly, in 60% of the cases in NA, the top ranking field of a certain trend-window is not the second rank field in the previous trend-window which is in contrast with the global behavior. For EU and OT, this value is 62% and 60% respectively. Furthermore, for all the above cases (60% (NA), 62% (EU) and 60%(OT)), we have found that the third rank field in the previous trend-window plays the lead role in the immediate future. This is possibly because the behavioral patterns of the continents have a correspondence with the global behavior that either lags or leads in time (we shall revisit this issue in the following section). Another observation is that in most of the cases, the second rank fields continuously remain second for significantly long times. For instance, Operating System (OS) consistently holds the second position from 1978 to 1985 constituting four consecutive trend-windows in NA, however never coming to the top in any of the subsequent trend-windows. This implies that the worldwide behavior is not fully governed by any one single continent, rather it could be a combined effect manifesting itself in the form of the observed global patterns.

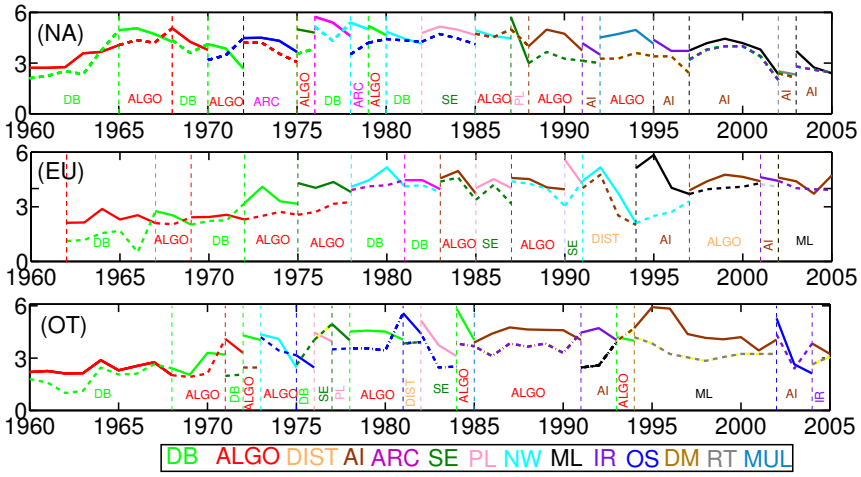

Fig. 11 (Color online) Continent-wise transition of research trend (North America (NA), Europe (EU) and others (OT)). Names of the topmost backup fields for the field at the forefront in each trend-window are also mentioned. To smooth the curves the best sliding window size of five years has been taken. Note that, the time transition for EU starts from 1962 since before that year there was no paper with at least one author from EU in our dataset.

9.4 Cross correlation in time transition patterns

An important question that has so far remained unanswered is whether the individual time-transition patterns of the continents affect the global behavior. For this purpose, we find the extent of similarity between the field that is most dominant at the global level and the field that is most dominant within a continent for each individual year. Since, there are data for $n = 45$ years in our dataset, we calculate a similarity metric τ that is defined as

$$\tau = \frac{s}{n} \tag{5}$$

where s is the number of similar pairs. As the number of data points are not many, exact similarity might be a very strict assumption in this case. Therefore, we relax τ by calling a pair similar if there is any match between the top two pairs (instead of top one). The pairwise similarity measures are reported in Table 9. The global behavior at any point in time seems to be most similar to that of Europe. However, what is possibly more important is to investigate if a field that is at the peak in one continent at any point in time emerges as the globally dominant one in the near future. This would then indicate that what one continent finds important today is adopted by the rest of the world within a few years. In order to quantify this delayed similarity, we redefine τ where we compare (a) the current global figures (GLO) with that of any continent t years earlier (i.e., $lead(\text{continent}, \text{GLO}, t)$) as well as (b) the current global figures (GLO) with that of any continent t years later

Table 9 Correlations between transitions in research trend (GLO: global transition, NA: transition in North America, EU: transition in Europe and OT: transition in the other continents). $lead(X, Y, t)$ denote that the event x took place t years before the event y. Similarly, $lag(x, y, t)$ denote that the event x took place t years after the event y.

Pairs	τ			
	Time (t)			
	0 yr	1 yr	2 yrs	3 yrs
$lead(GLO, NA, t)$	0.76	0.72	0.68	0.58
$lag(GLO, NA, t)$	0.76	**0.86**	0.68	0.63
$lead(GLO, EU, t)$	**0.91**	0.84	0.79	0.68
$lag(GLO, EU, t)$	0.91	0.77	0.72	0.51
$lead(GLO, OT, t)$	0.65	0.60	0.57	0.49
$lag(GLO, OT, t)$	0.65	0.60	0.50	0.55
$lead(NA, EU, t)$	0.09	0.73	0.72	0.69
$lag(NA, EU, t)$	0.09	0.71	0.63	0.53
$lead(NA, OT, t)$	0.40	0.66	0.59	0.65
$lag(NA, OT, t)$	0.40	0.62	0.54	0.46
$lead(EU, OT, t)$	0.13	0.57	0.50	0.37
$lag(EU, OT, t)$	0.13	0.55	0.61	0.51

(i.e., $lag(\text{continent}, \text{GLO}, t)$) as defined earlier in Section 3. In similar line, we compute pairwise this measure for all the continents. Note that, $lead(\text{X}, \text{Y}, t = 0)$ is equivalent to $lag(\text{Y}, \text{X}, t = 0)$. The results are presented in Table 9. While the European time transition pattern is almost similar to the global pattern at a particular year ($lead(\text{GLO}, \text{EU}, t = 0)$=0.91), we observe the maximum concordance of North American transition one year before with the transition of the global pattern at the current time (i.e., $lag(GLO, NA, t = 1)$=0.86). It indicates whatever is popular in North America today would become popular in the rest of the world within one year. In fact, this trend is also observed in the similarity values across continent pairs – North America seems to be a "torch-bearer" for both the other continents.

10 Correlation with research funding

It could be interesting as well as important to validate our measurements with other extraneous real-world statistics directly or indirectly reflecting the evolution of scientific research in the computer science domain. To this purpose, we collect the fund disbursal data of one of the major funding agencies of the United States – the National Science Foundation (NSF)[5]. Although this agency has a long funding history, the publicly available data that we could gather is from 2003 to 2009. In Table 10, we compare the top three fields ranked by our inwardness metric with the top three fields ranked by (i) the number of NSF proposals submitted and (ii) the number of proposals accepted in that field. The high-impact fields predicted by our method match accurately with the trend of proposal submission. To compare the two statistics, we propose a

[5] http://www.nsf.gov/

similarity metric τ that is defined as

$$\tau = \frac{s}{n} \tag{6}$$

where s is the number of similar pairs and n is the number of data points. As the number of data points are not many, exact similarity might again be a very strict assumption in this case. Therefore, we relax τ by calling a pair similar if there is any match between the top two pairs (instead of top one). In Table 11, we report the pairwise similarity (τ) between the fields ranked by our method and fields ranked by (a) the number of proposals submitted and (b) the number of proposals granted in those fields. While measuring the similarity using equation 6, we increment the value of s when (i) at least one field is matching, and (ii) at least two fields are matching with 50% weight for each matching. We report the similarity values in the first row (REC vs. SUBMIT) and fourth row (REC vs. AWARD) of Table 11 for the same year where REC refers to what is recommended by our method based on inwardness. The results clearly show that our predictions are very well aligned with proposal submission while it is moderately aligned with the fund disbursal patterns.

Table 10 Funding statistics compared with the inwardness results (top three ranked fields are tabulated from left to right).

Years	Inwardness results	NSF	
		Proposal submitted	Proposal awarded
2003	AI/IR/NW	NW/AI/HCI	NW/ALGO/SE
2004	AI/IR/NW	AI/HCI/RT	RT/ARC/DIST
2005	AI/IR/NW	AI/ML/HCI	GRP/SE/ALGO
2006	IR/ML/AI	ML/ALGO/SEC	ALGO/SEC/ML
2007	ML/AI/ALGO	ALGO/ML/HCL	ALGO/HCI/SEC
2008	ML/AI/ALGO	ML/ALGO/SE	ALGO/ML/SE

It is often observed that the current funding patterns significantly affect the research directions of the future. Further, at times, the current research trend seems to strongly influence the funding decisions of the immediate future. The above observations can be illustrated quantitatively here. In order to do so, we introduce lagging and leading similarities between fields ranked by the inwardness metric (REC) and those ranked by the number of proposals submitted/awarded. We measure two different similarity values – $lead(\text{fund}, \text{REC}, t = 1)$ and $lag(\text{fund}, \text{REC}, t = 1)$. From the results depicted in Table 11, we observe that the influence of funding decisions on the future research trend is much more (lead) than the influence of the current research trend on the future funding decisions (lag). This shows that our results are remarkably in line with the decisions made by the expert researchers involved in such important proposal selection committees. However, we remark that all our results are based on only a small number of data points and should therefore be considered indicative.

Table 11 Correlations between our recommendations (REC) with the submit (SUBMIT) and award (AWARD) patterns of grants.

Pairs		τ	
		At least 1 matching	At least 2 matching
REC vs. SUBMIT	Same year	1	0.78
	$lead(\text{SUBMIT}, \text{REC}, t = 1)$	1	0.83
	$lag(\text{SUBMIT}, \text{REC}, t = 1)$	0.83	0.50
REC vs. AWARD	Same year	0.71	0.50
	$lead(\text{AWARD}, \text{REC}, t = 1)$	0.75	0.42
	$lag(\text{AWARD}, \text{REC}, t = 1)$	0.33	0.25

11 Conclusion and future work

The lack of reliable ground-truth communities has made network community detection a very challenging task. In this paper, we developed ground-truth overlapping communities of a directed paper-paper citation network that emerge from the natural grouping of research papers in various fields of the computer science domain. Subsequently, we validated the existence of such tightly knit ground-truth communities through well-established scoring functions proposed in the literature. We demonstrated the dynamics of inter-community interactions across a longitudinal timescale that in turn unfolds the research trend in the computer sciences for the last fifty years. We conclude by summarizing our main observations as follows:

- the ground-truth communities indeed represent natural groupings of any scientific research discipline,
- quite remarkably, for the last fifty years one observes a very robust behavior of the dynamics – the field that is the strongest contender of the field currently at the forefront almost surely emerges as the top ranked field after the transition,
- in contrast to what is predicted by the model described in [3], our empirical analysis shows that a field can once again emerge as a top ranker after undergoing a decline,
- the key factors that keep a field at the forefront include the citations from the backup field, the inception of the seminal papers and the existence of the highly cited papers,
- North American papers seem to have the largest overall impact; in addition, North America seems to regulate the research focus of the rest of the world. However, North America enjoys very less citation support from the papers of the other continents in comparison to what it receives from within,
- finally, funding statistics obtained from NSF is in very good agreement with the results predicted by our method.

The availability of ground-truth communities allows for a range of interesting future investigations. For example, further examining the connectivity structure in and across ground-truth communities could lead to novel community detection methods especially in citation network. Moreover, the present

empirical study marks the foundation for the design and implementation of a specialized recommendation engine that would be capable of answering search queries pertaining to the (a) impact of papers/authors, (b) fields at the forefront (currently and in the near future), (c) seminal papers within a field and many such other factors. These results can be useful for (i) the funding agencies to make appropriate decisions as to how to distribute project funds, (ii) the universities in their faculty recruitment procedure. The dataset is available at `http://cnerg.org` for the research community to facilitate further investigations. In summary, this paper shows that the usual consensus on the fact that suggesting an efficient community detection technique usually marks the "endpoint" in research in this area might not be true; in contrast, it possibly triggers the beginning of a new dimension of research, whereby, the temporal interaction, influence, shape and size of the communities so obtained can be suitably analyzed thus allowing for newer insights into the complex system under investigation.

References

1. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning, International Conference on Machine Learning (ICML), pp. 113–120. ACM, New York, USA (2006)
2. Booth, K.S., Lueker, G.S.: Testing for the consecutive ones property, interval graphs, and graph planarity using pq-tree algorithms. Journal of Computer and System Sciences **13**(3), 335–379 (1976)
3. Bornholdt, S., Jensen, M.H., Sneppen, K.: Emergence and Decline of Scientific Paradigms. Phys. Rev. Lett. **106**(5), 058,701 (2011)
4. Boyack, K.W., Klavans, R., Börner, K.: Mapping the backbone of science. Scientometrics **64**(3), 351–374 (2005)
5. Chakraborty, T., Kumar, S., Reddy, M.D., Kumar, S., Ganguly, N., Mukherjee, A.: Automatic classification and analysis of interdisciplinary fields in computer sciences. In: 2013 ASE/IEEE International Conference on Social Computing (SocialCom), pp. 180 – 187. Washington DC, USA (2013)
6. Chakraborty, T., Sikdar, S., Tammana, V., Ganguly, N., Mukherjee, A.: Computer science fields as ground-truth communities: Their impact, rise and fall. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 426 – 433. Nigara Falls, Canada (2013)
7. Egghe, L., Leydesdorff, L.: The relation between Pearson's correlation coefficient r and Salton's cosine measure. Journal of the American Society for Information Science and Technology **60**(5), 1027–1036 (2009)
8. Franceschet, M.: A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar. Scientometrics **83**(1), 243–258 (2010)
9. Garfield, E., Sher, I.H., Torpie, R.J.: The Use of Citation Data in Writing the History of Science. Institute for Scientific Information Inc. (1984)
10. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States (PNAS) **99**(12), 7821–7826 (2002)
11. Guimerà, R., Uzzi, B., Spiro, J., Nunes Amaral, L.A.: Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. Science **308**(5722), 697–702 (2005)
12. Guns, R., Rousseau, R.: Real and rational variants of the h-index and the g-index. Journal of Informetrics **3**(1), 64–71 (2009)

13. Hirsch, J.E.: Does the h index have predictive power? Proceedings of the National Academy of Sciences of the United States (PNAS) **104**(49), 19,193–19,198 (2007)
14. Hirsch, J.E.: An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. Scientometrics **85**(3), 741–754 (2010)
15. Jin, B., Liang, L., Rousseau, R., Egghe, L.: The R- and AR-indices: Complementing the h-index. Chinese Science Bulletin **52**(6), 855–863 (2007)
16. Kawamae, N., Higashinaka, R.: Trend detection model. In: Proceedings of WWW '10, pp. 1129–1130. ACM, New York, USA (2010)
17. Kleinberg, J.: Authoritative sources in a hyperlinked environment. Journal of the ACM **46**(5), 604–632 (1999)
18. Kuhn, T.S.: The structure of scientific revolutions. University of Chicago Press (1970)
19. Lee, K., Palsetia, D., Narayanan, R., Patwary, M.M.A., Agrawal, A., Choudhary, A.N.: Twitter trending topic classification. In: International Conference on Data Mining (ICDM) Workshops, pp. 251–258 (2011)
20. Leicht, E.A., Clarkson, G., Shedden, K., Newman: Large-scale structure of time evolving citation networks. The European Physical Journal B **59**(1), 75–83 (2007)
21. Mazloumian, A., Eom, Y.H., Helbing, D., Lozano, S., Fortunato, S.: How Citation Boosts Promote Scientific Paradigm Shifts and Nobel Prizes. PLoS ONE **6**(5), e18,975 (2011)
22. Pan, R.K., Sinha, S., Kaski, K., Saramäki, J.: The evolution of interdisciplinarity in physics research. Nature Scientific Reports **2**(551) (2012)
23. Pham, M.C., Klamma, R.: The structure of the computer science knowledge network. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 17–24. IEEE Computer Society, Washington, DC, USA (2010)
24. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. Proceedings of the National Academy of Sciences of the United States (PNAS) **101**(9), 2658 (2004)
25. Redner, S.: Citation Statistics from 110 Years of Physical Review. Physics Today **58**(6), 49–54 (2005)
26. Shaparenko, B., Caruana, R., Gehrke, J., Joachims, T.: Identifying temporal patterns and key players in document collections. In: IEEE International Conference on Data Mining (ICDM), pp. 165–174 (2005)
27. Shi, X., Tseng, B.L., Adamic, L.A.: Information diffusion in computer science citation networks. In: International AAAI Conference on Weblogs and Social Media (ICWSM), pp. 319–322 (2009)
28. Shibata, N., Kajikawa, Y., Takeda, Y., Matsushima, K.: Detecting emerging research fronts based on topological measures in citation networks of scientific publications. TECHNOVATION **28**(11) (2008)
29. de Solla Price, D.J.: Networks of scientific papers. Science **149**(3683), 510–515 (1965)
30. Spiliopoulou, M., Ntoutsi, I., Theodoridis, Y., Schult, R.: Monic: modeling and monitoring cluster transitions. In: Proceedings of the 12th ACM SIGKDD, pp. 706–711. New York, USA (2006)
31. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: extraction and mining of academic social networks. In: ACM SIGKDD, pp. 990–998 (2008)
32. Yang, J., Leskovec, J.: Defining and evaluating network communities based on ground-truth. In: Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, MDS '12, pp. 3:1–3:8. New York, USA (2012)
33. Yang, J., Leskovec, J.: Overlapping community detection at scale: a nonnegative matrix factorization approach. In: International conference on web search and data mining (WSDM), pp. 587–596 (2013)
34. Zhao, Q., Bhowmick, S.S., Zheng, X., Yi, K.: Characterizing and predicting community members from evolutionary and heterogeneous networks. In: ACM International Conference on Information and Knowledge Management (CIKM), pp. 309–318 (2008)