2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Computer Science Fields as Ground-truth Communities: Their Impact, Rise and Fall

Tanmoy Chakraborty*, Sandipan Sikdar†, Vihar Tammana‡, Niloy Ganguly§, Animesh Mukherjee¶
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur, India – 721302
{*its_tanmoy,†sandipansikdar,§niloy,¶animeshm} @cse.iitkgp.ernet.in
‡vihartsk@gmail.com

***Abstract*—Study of community in time-varying graphs has been limited to its detection and identification across time. However, presence of time provides us with the opportunity to analyze the interaction patterns of the communities, understand how each individual community grows/shrinks, becomes important over time. This paper, for the first time, systematically studies the temporal interaction patterns of communities using a large scale citation network (directed and unweighted) of computer science. Each individual community in a citation network is naturally defined by a research field – i.e., acting as ground-truth – and their interactions through citations in real time can unfold the landscape of dynamic research trends in the computer science domain over the last fifty years. 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. Several arguments to unveil the reasons behind the temporal changes of inwardness of different communities are put forward using exhaustive statistical analysis. The measurements (importance of field) are compared with the project funding statistics of NSF and it is found that the two are in sync. 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.**

***Keywords*—community analysis; ground-truth communities; citation network; computer science; temporal network**

I. Introduction

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 [1] [2]. The studies on community formation has gradually moved away from finding exclusive community [3] [4] for each individual node to the domain of *"overlapping communities"* [5] where it is believed that a node may be a member of several communities. Recently, significant research indicating the existence of a consistent partition of nodes across multiple snapshots of an evolving network have been conducted which has shifted the community related study in the direction of *"temporal/time-varying communities"* [6] [7].

Though several works on detecting and tracking communities in a temporal environment have been conducted [8] [9], 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 terms 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) inter-cluster interactive patterns to unfold the latent characteristics of the network that indeed explains the rise and fall of the impact of scientific research communities over the last fifty years.

The major contribution of our work is fourfold. 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 [10], the communities overlap as some papers belong to multiple fields; we rigorously check the goodness of these community structures with well-known community-centric metrics [11]. 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 phenomena in 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 converge to a consensus in quantifying the typical dynamics of research communities efficiently. Finally, we validate our proposed framework with the evidence of another extraneous statistics of the project funding decisions made by NSF (National Science Foundation of the USA). 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 [12] has been carried out with the aim to understand its structure and to determine clusters of similar and high-prestige venues. Yang and Leskovec [13] developed ground-truth communities of real-world undirected static networks and detected overlapping communities from these networks [11]. In this experiment,

The first author is supported by the Google India PhD fellowship Grant in Social Computing.

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.

Besides, the study leads to several interesting findings, some of which are noted below. In almost all cases, the field constituting the current major area of research within the domain is overtaken in the immediate future by its strongest competitor. The density of high impact publications within a field plays a pivotal role in pulling as well as sustaining the field at the forefront. Certain fields produce a huge number of citations (i.e., act as hubs [14]) for a particular field and, thereby, push it to the forefront; an abrupt fall in the number of such received citations, in many cases, triggers the decline of the field currently at the forefront. The inception of seminal papers in a field might trigger the emergence of a field at the forefront. The degree of team work (both within and across continents) in the form of joint publications seem to significantly contribute to the shape of the evolutionary landscape. We also find the fields that are presently at the forefront influence the current funding decisions much less than the funding decisions influence the emergence of a field at the forefront in the immediate future.

The rest of the paper is organized as follows. The process of collecting citation dataset, tagging and the construction of network and ground-truth communities are elaborated in section II. In section III, several community scoring functions are used to judge the goodness of the ground-truth communities. Then in section IV, we outline the time profile of the evolution of scientific communities after analyzing the inter-cluster interactions. Next we present a detailed analysis of the possible causes explaining this temporal dynamics of research communities in section V. In section VI, we point out how our results are correlated to research funding and finally conclude the paper in section VII.

II. Dataset and Construction of the Network

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, research field, authors and their continents. Note that the authors and continent information are required to pose one of the arguments behind the global dynamics of inter-cluster interactions as described in the section V. We have used the dataset of the computer science domain developed by Tang et al. [15][1] for our experiments. It was 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 Table I.

TABLE I. 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	495,311
Average number of papers by an author	3.82	3.52
Average number of authors per paper	2.615	2.609
Number of unique venue name	2,319	1,705

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, publication venue (conference/journal) of the paper (required for field tagging), the year of publication and the citations. In general, the trend shifts of scientific communities 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 I.

A. Field Tagging

The natural intuition behind considering each scientific field as a separate community is that the intra-field citation density is generally much higher than the cross-field citation density which concurs with the traditional definition of a community (higher edge-density within a community than across communities) in a network [16]. People tend to inherently build these natural groupings due to their common research interest. Moreover, the increasing rate of interactions across multiple research communities now-a-days enhances the possibility of overlapping communities resulting in the emerging trend of interdisciplinary research (e.g., computational biology). 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 II. 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 II notes the percentages (decreasing order) of papers in various fields in the tagged dataset. We also show in the table the average ten-year impact (Equation 1) for each field between the years 1960 and 2008. Note that this value indicates the average impact of a research community due to the incoming citations emanating from the papers of the other communities.

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

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

TABLE II. 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

B. Continent Tagging

As mentioned earlier, the continent information of an author is used for analyzing probable reasons behind the rise and fall of scientific research communities described in section V. Microsoft Academic Search also provides location of the authors like the name of the university/company they are affiliated to and the continent information (North America, South America, Asia-Oceania, 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 in the site 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 with 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 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 III. The number of authors from Africa, South America and Asia-Oceania being relatively low, we merge them together into a new category called "Others" which we use for our future experiments.

TABLE III. 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%

Since our method is primarily based on suitable statistical analysis of various properties of paper-paper citation network that in turn characterizes the inter-cluster interactions, the next task is to construct the citation network from the tagged dataset. Formally, a citation network is 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. From our tagged dataset, a citation network was constructed by the papers representing nodes and the citations representing directed edges from the citing paper to the cited paper. At a higher tier, each field (i.e., a collection of papers) can be thought of as a single community and two communities can again be linked by a directed edge with edge-weight calculated using Equation 1 mentioned in section IV. Following this strategy, we essentially obtain a field-field directed and weighted network on top of the paper-paper citation network which attempts to capture the interaction patterns of the scientific communities. Note that in each year, there are at most 24 communities (if there exists

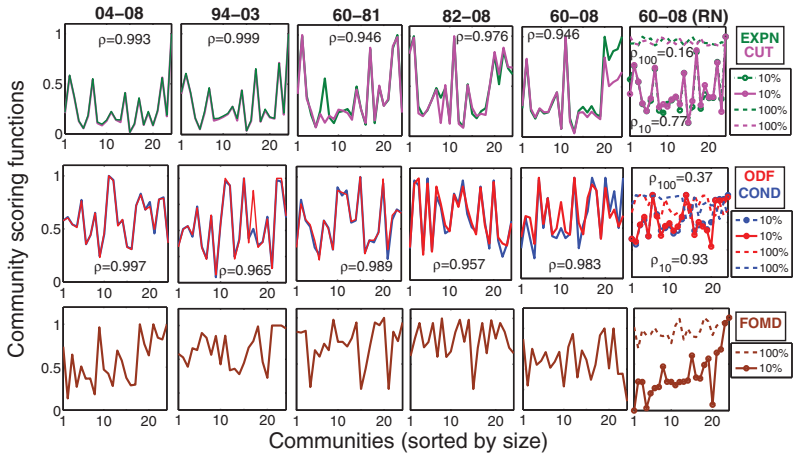

Fig. 1. (Color online) Community scoring functions for real-world ground-truth communities (sold lines) in different time windows (2004-2008, 1994-2003, 1960-1981, 1982-2008 and 1960-2008). Results from the randomized versions (10% and 100%) of the ground-truth communities are presented in the right-most 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.

at least one paper from each of the fields) and the size of each community changes over the years depending upon the number of publications in that field. A community at time t can interact with any other communities at or before t.

III. Community Scoring Functions

We now discuss various scoring functions defined by Yang and Leskovec [13] 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 between the members and few connections from the members to the rest of the network. Out of 13 commonly used scoring functions proposed in [13], a few have been proved to be necessary 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.

(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., $f(S) = \frac{|u:u \in S, |(u,v) \in E:v \in S| < d(u)/2|}{n_S}$

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 [13]. 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 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 1). For each time-window, we calculate Pearson's correlation coefficient (ρ) [17] 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

[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 IV. 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

of random swaps (see the last column of Figure 1). We report further the actual value of the functions for the entire network in Table IV. 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.

IV. TIME TRANSITION OF SCIENTIFIC COMMUNITIES

In this section, we analyze the time profile of scientific research communities showing how one community has taken over another during the temporal evolution of the computer sciences. In particular, we measure the impact of a field so as to construct the time transition diagram reflecting the trend shifts. Some of the previous experimental results [18][19] show that the pattern 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 the publication and then slowly declines over time (see Figure 2). Note that this property is also prevalent across the different fields of the domain (see inset of Figure 2). Therefore, all our analysis throughout the rest of the paper assume only the citations received by a paper within three years from its publication since it is more appropriate to predict the emergence of a field in the forefront based only on the recently received citations by its constituent papers. We quantify the importance of a paper (aka inwardness) in terms of the total number of inward citations to the paper. Consequently, the temporal inwardness of a field f_i at time t denoted by $In(f_i^t)$ that captures the local citation count (within three-year window) suitably normalized by the number of papers in that field can be defined as

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

where $w_{j \to i}^t = \frac{c_{j \to i}^t}{n_i^t}$ 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, n_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. This inwardness metric is a measure of the degree of authoritativeness of a research community proposed here for the first time and defined in the line of what has been already discussed in the context of individual publications [14].

In order to investigate the global time transition pattern (i.e., the worldwide behavior) we compute the inwardness of each scientific community (Equation 1) and rank them accordingly. For better visualization, we plot the top two values (see the solid and broken lines respectively in Figure 3(a)) as a function of time. Each field is uniquely color coded and the relative height of the y-axis shows the inwardness of the field for a particular year. In each trend-window, we also mention the name of the top hub (backup) field that on an average produces the largest number of the citations for the top ranking field. This information, as we shall see in the next section, forms one of the major arguments explaining the dynamics of scientific communities. The total number of transitions of research trend during 1960 to 2005 is 11. A careful inspection of the behavior of the curves shows that in every trend-window, a similar pattern is followed with the inwardness of the top field first rising and then gradually declining near the transition. Simultaneously, the second rank field which comes to the top position in the next trend-window in every case starts reflecting a relative growth of inwardness at the middle of the current trend-window. Bornholdt et al. [20] 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 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 trend-windows are largely different. We plan to investigate this property in more detail in the next section. The average values of inwardness of all the fields in each decade are mentioned in Table II that precisely illustrate the "trending" of the research communities in the last fifty years.

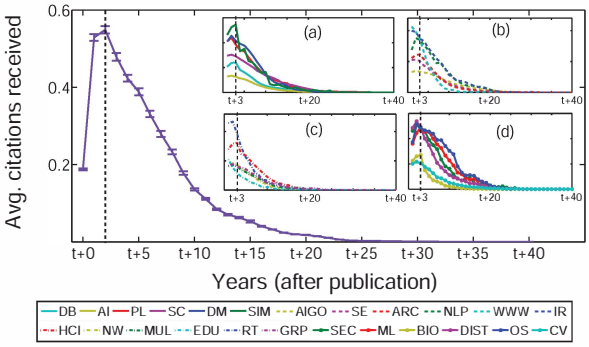

Fig. 2. (Color online) Average distribution pattern of inward citations (with variances) for a paper after publication (inset: same measure for every field).

V. REASONS FOR TRANSITIONS

In this section, we conduct an exhaustive 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.

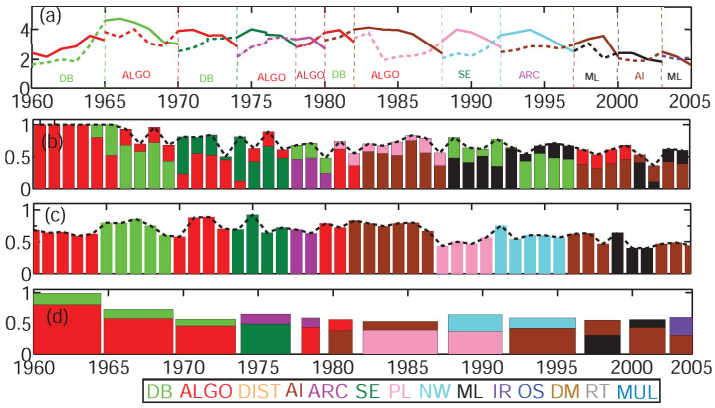

Fig. 3. (Color online) (a) Top two scientific communities (based on inwardness) at the forefront of scientific research trend (names of topmost backup communities for the communities in the forefront of every trend-window are mentioned). Cause analysis: Fig.(b) fraction of papers for top and second ranked 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 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.

TABLE V. 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 AND SEVENTH ROWS).

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

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 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 measured by the average number of fields that the authors of that particular field have worked. Note that in case (iii), the more the diversity the higher is the rank of the field. Table V 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.

Cause II: High impact papers

We extract the top 10% of the papers that have the highest number of in-citations (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 3(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 3(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 0.5. This could indicate the presence of a tremendous competitive pressure from the other fields many of which 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 3(c), we plot bars for each year indicating the fraction of citations that the top ranked community (according to Figure 3(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 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 get-

[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.

ting 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.

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:

$$Influence(p_i^t) = \sum_{p_j \in P^t} \frac{1}{d_{p_j}} \quad (2)$$

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.

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 3(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 Algorithm 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.

VI. 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 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 VI, 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 similarity metric τ that is defined as

$$\tau = \frac{s}{n} \quad (3)$$

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 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 VII, 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 3, 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 (OUR vs. SUBMIT) and fourth row (OUR vs. AWARD) of Table VII for the same year. The results clearly show that our predictions are very well aligned with proposal submission while it is moderately aligned with the fund disbursal patterns.

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[6] and leading[7] similarities between fields ranked by the inwardness metric and those ranked by the number of proposals submitted/awarded. We measure two different similarity values – $lead(\text{fund}, \text{inwardness}, 1)$ and $lag(\text{fund}, \text{inwardness}, 1)$. From the results depicted in Table VII, 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 VI. Funding statistics compared with the inwardness results (top three ranked fields are tabulated from left to right).

Yrs	Inwardness results	NSF	
		Proposal submitted	Proposal awarded
03	AI/IR/NW	NW/AI/HCI	NW/ALGO/SE
04	AI/IR/NW	AI/HCI/RT	RT/ARC/DIST
05	AI/IR/NW	AI/ML/HCI	GRP/SE/ALGO
06	IR/ML/AI	ML/ALGO/SEC	ALGO/SEC/ML
07	ML/AI/ALGO	ALGO/ML/HCL	ALGO/HCI/SEC
08	ML/AI/ALGO	ML/ALGO/SE	ALGO/ML/SE

VII. Conclusion

The lack of reliable ground-truth communities has made network community detection a very challenging task. In this

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

[6] $lag(x, y, t)$ means the event x took place t years after the event y.

[7] $lead(x, y, t)$ means the event x took place t years before the event y.

TABLE VII. CORRELATIONS BETWEEN OUR RECOMMENDATIONS (OUR) WITH THE SUBMIT (SUBMIT) AND AWARD (AWARD) PATTERNS OF GRANTS.

Pairs		τ	
		At least 1 matching	At least 2 matching
OUR vs. SUBMIT	Same year	1	0.78
	$lead$(SUBMIT, OUR, 1)	1	0.83
	lag(SUBMIT, OUR, 1)	0.83	0.50
OUR vs. AWARD	Same year	0.71	0.50
	$lead$(AWARD, OUR, 1)	0.75	0.42
	lag(AWARD, OUR, 1)	0.33	0.25

paper, we developed ground-truth overlapping communities of a directed paper-paper citation network that emerge from the natural grouping of research papers into the 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 and outlining some of the possible future directions. We observe that
(i) the ground-truth communities effectively capture the notion of natural groupings in scientific research world,
(ii) the shift in the trend of research communities is, quite strikingly, similar across past fifty years, i.e., the community that is the strongest competitor of the community currently at the forefront, emerges as the top ranker in the next trend-window,
(iii) a research community that declines after remaining at the top for sometime, can again emerge as the top ranker in future,
(iv) the citation support received by a field from its backup fields plays an important role in keeping the field in the forefront,
(v) presence of a significant number of high-impact papers and the inception of seminal papers in a field accelerate it to the forefront,
(vi) collaborative research, in general, seems very effective in producing high impact publications,
(vii) 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 shall be 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] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," PNAS, vol. 99, no. 12, pp. 7821–7826, Jun. 2002.

[2] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," PNAS, vol. 101, no. 9, p. 2658, 2004.

[3] G. W. Flake, S. Lawrence, and C. L. Giles, "Efficient identification of web communities," in Proceedings of the sixth ACM SIGKDD, New York, USA, 2000, pp. 150–160.

[4] A. Clauset, "Finding local community structure in networks," Phys. Rev. E, vol. 72, no. 2, p. 026132, Aug. 2005.

[5] J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping community detection in networks: the state of the art and comparative study," CoRR, vol. abs/1110.5813, 2011.

[6] V. Kawadia and S. Sreenivasan, "Sequential detection of temporal communities by estrangement confinement," Scientific Reports, vol. 2, Nov. 2012.

[7] D. Greene, D. Doyle, and P. Cunningham, "Tracking the evolution of communities in dynamic social networks," in Proceedings of ASONAM '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 176–183.

[8] M. Spiliopoulou, I. Ntoutsi, Y. Theodoridis, and R. Schult, "Monic: modeling and monitoring cluster transitions," in Proceedings of the 12th ACM SIGKDD, New York, USA, 2006, pp. 706–711.

[9] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in Proceedings of the 23rd international conference on Machine learning, ser. ICML '06. New York, USA: ACM, 2006, pp. 113–120.

[10] N. Shibata, Y. Kajikawa, Y. Takeda, and K. Matsushima, "Detecting emerging research fronts based on topological measures in citation networks of scientific publications," TECHNOVATION, vol. 28, no. 11, 2008.

[11] J. Yang and J. Leskovec, "Overlapping community detection at scale: a nonnegative matrix factorization approach," in WSDM, 2013, pp. 587–596.

[12] M. C. Pham and R. Klamma, "The structure of the computer science knowledge network," in Proceedings of ASONAM '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 17–24.

[13] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," in Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, ser. MDS '12, New York, USA, 2012, pp. 3:1–3:8.

[14] J. Kleinberg, "Authoritative sources in a hyperlinked environment," J. of the ACM, vol. 46, no. 5, pp. 604–632, 1999.

[15] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: extraction and mining of academic social networks," in ACM SIGKDD, 2008, pp. 990–998.

[16] N. Shibata, Y. Kajikawa, Y. Takeda, and K. Matsushima, "Detecting emerging research fronts based on topological measures in citation networks of scientific publications," TECHNOVATION, vol. 28, no. 11, 2008.

[17] L. Egghe and L. Leydesdorff, "The relation between Pearson's correlation coefficient r and Salton's cosine measure," Journal of the American Society for Information Science and Technology, vol. 60, no. 5, pp. 1027–1036, 2009.

[18] R. Guns and R. Rousseau, "Real and rational variants of the h-index and the g-index," J. of Informetrics, vol. 3, no. 1, pp. 64–71, 2009.

[19] B. Jin, L. Liang, R. Rousseau, and L. Egghe, "The R- and AR-indices: Complementing the h-index," Chin. Sci. Bull., vol. 52, no. 6, pp. 855–863, Mar. 2007.

[20] S. Bornholdt, M. H. Jensen, and K. Sneppen, "Emergence and Decline of Scientific Paradigms," Phys. Rev. Lett., vol. 106, no. 5, p. 058701, 2011.