Automatic Classification and Analysis of Interdisciplinary Fields in Computer Sciences

Tanmoy Chakraborty*, Srijan Kumar†, M Dastagiri Reddy‡, Suhansanu Kumar§, Niloy Ganguly¶, Animesh Mukherjee‖
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur, India – 721302
{*its_tanmoy,†srijan.kumar,‡reddy.dastagiri,§suhansanu.kumar,¶niloy,‖animeshm}@cse.iitkgp.ernet.in

***Abstract*—In the last two decades, there have been studies claiming that science is becoming ever more interdisciplinary. However, the evidence has been anecdotal or partial. Here for the first time, we investigate a large size citation network of computer science domain with the intention to develop an automated unsupervised classification model that can efficiently distinguish the core and the interdisciplinary research fields. For this purpose, we propose four indicative features; three of these are directly related to the topological structure of the citation network, while the fourth is an external indicator based on the attractiveness of a field for the in-coming researchers. The significance of each of these features in characterizing interdisciplinarity is measured independently and then systematically accumulated to build an unsupervised classification model. The result of the classification model shows two distinctive clusters that clearly distinguish core and interdisciplinary fields of computer science domain. Based on this classification, we further study the evolution dynamics at a microscopic level to show how interdisciplinarity emerges through cross-fertilization of ideas between the fields that otherwise have little overlap as they are mostly studied independently. Finally, to understand the overall impact of interdisciplinary research on the entire domain, we analyze selective citation-based measurements of core and interdisciplinary fields, paper submission and acceptance statistics at top-tier conferences and the core-periphery structure of citation network, and observe an increasing impact of the interdisciplinary fields along with their steady integration with the computer science core in recent times.**

I. INTRODUCTION

"Interdisciplinary research is the only way to do research in current times."

– Fritjof Capra, The Turning Point

A field is any comparatively self-contained and isolated domain of human experience which possesses its own community of experts, with distinctive components such as shared goals, concepts, facts, tacit skills and methodologies. Interdisciplinary field, on the other hand, brings in together distinctive components of two or more fields in research or education, leading to new knowledge which would not be possible without this integration. Interdisciplinarity occurs when fields intermesh, integrate and collaborate among themselves [1].

Many believe that the great advances disproportionately take place at the interstices between fields, and that today's research knowledge "knows no field boundaries." The purpose of interdisciplinary research is to advance fundamental understanding or to solve problems whose solutions are beyond the scope of a single field of research practice. Despite a reasonable number of works promoting the increasing trend of cross-field research, the researchers [2][3] still believe that there is a lack of proper quantitative indicator that could efficiently identify interdisciplinary fields (interdisciplinary papers) in a certain domain.

In this paper, we systematically unfold large size citation network of computer science domain, and study the dynamics and emergence of connections across the fields in a longitudinal scale over the last four decades (1975 – 2008). Though the structural and dynamical properties of networks explaining relations among researchers have been an important research subject for the last several decades [4], there is a growing body of literature on the characterization and measurement of interdisciplinarity of scientific journals and researchers [5][6]. Many published attempts try to classify interdisciplinarity into components such as pluridisciplinarity, crossdisciplinarity, and even metadisciplinarity [7][8]. Recently, Pan et al. [9] study the network constructed on the basis of PACS codes (Physics and Astronomy Classification Scheme) of the papers in physics domain to show a clear trend towards increasing interactions between different fields. They conclude that the microscopic observation is missing in their study due to the lack of citation information which naturally bears the inherent interaction pattern among the fields.

We exhaustively analyze the citation network from different perspectives that indeed indicate strong signs of interdisciplinarity of a research field. To begin with, we call computer science including all its sub-branches as a "domain" of research. The different sub-branches like Algorithms, Artificial Intelligence, Operating Systems, Programming Languages, databases etc., constitute the different "fields" of research. The degree of interdisciplinarity of a particular field is measured using various citation indicators that neatly separate out the core from the interdisciplinary fields in a fully unsupervised fashion. In particular, we propose four indicative features that could further help us classifying the interdisciplinary and core fields among the set of 24 research fields identified in the computer science domain (see in section III). Three of the indicators namely *Reference Diversity Index (RDI)*, *Citation Diversity Index (CDI)* and *Membership Diversity Index (MDI)* are directly related to the topological structure of the citation network. The last feature called the *Attraction Index* of a field is based on the propensity of the new researchers to start

The first author of the paper is financially supported by the Google India PhD Fellowship Grant in Social Computing.

978-0-7695-5137-1/13 $26.00 © 2013 IEEE
DOI 10.1109/SocialCom.2013.34

IEEE computer society

research in a particular field. Further, to check the significance of these features in characterizing interdisciplinarity, we rank the fields based on the value of each of the features separately. The results corroborate a consistent ranking of few fields namely Data Mining, WWW, NLP, Computational Biology, Computer Vision at the top position and Algorithm, Programming Languages, Operating System at the bottom of the rank list. Next, we propose an unsupervised classification model (see in section IV) that can efficiently cluster the core and the interdisciplinary fields based on the similarity of the feature sets mentioned above. After that, we perform a two-fold analysis of the results obtained. As a first objective, we compare the evolutionary landscape of a core and an interdisciplinary field while as a second objective we study the overall impact of interdisciplinary research on the computer science domain. To understand the evolutionary landscape, we conduct a case study on one interdisciplinary field (WWW) and one core field (Programming Languages) (see in section V). The results attest to the conclusion that the interdisciplinarity occurs through cross-fertilization of ideas between the fields that otherwise have little overlap as they are studied independently. Finally, to show the impact of the interdisciplinary fields on the computer science domain, we point to two observations in the recent times: citation trends of the core and the interdisciplinary fields separately and the recent publication statistics at top-tier conferences. The conclusion that popularity of the interdisciplinary research now-a-days overshadows the core fields is strengthened on analyzing the core-periphery organization of the citation network in different time periods (see in section VI). We observe that the core region of a domain is gradually dominated by the more applied fields with interdisciplinary fields steadily accelerating towards the core. This measurement study explores a first and fundamental way of quantifying interdisciplinarity of a research field that in turn can again go as a scheme for recommending a field to the new researchers or for recommending grants to funding agencies.

II. Dataset and Network Construction

We have used the DBLP dataset of the computer science domain developed by Tang et al.[1] [10]. The dataset contains 702,973 valid papers and 495,311 authors. The attributes of each paper are as follows: name of 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. Since the filtered dataset of computer science does not have the necessary field information of the papers, we tag them using the Microsoft Academic Search[2] as described by Chakraborty et al. [11]. It categorizes papers of computer science domain into the fields as noted in Table I.

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 with the papers representing nodes and the citations representing directed edges from the citing paper to the cited paper. The citation network of computer science is constructed in longitudinal fashion where the papers in each year are arranged vertically and the unidirectional citation edges point to papers in or before this year.

TABLE I. The fields (with abbreviations) of computer science domain

Fields	Abbrev.	Fields	Abbrev.
Artificial Intelligence	AI	Algorithms and Theory	ALGO
Networking	NETW	Databases	DB
Distributed Systems	DIST	Hardware & Architecture	ARC
Software Engineering	SE	Machine Learning & Pattern Recognition	ML
Scientific Computing	SC	Computational Biology	BIO
Human-Computer Interaction	HCI	Multimedia	MUL
Graphics	GRP	Computer Vision	CV
Data Mining	DM	Programming Languages	PL
Security and Privacy	SEC	Information Retrieval	IR
Natural Language and Speech	NLP	World Wide Web	WWW
Computer Education	EDU	Operating Systems	OS
Real Time & Embedded Systems	RT	Simulation	SIM

III. Indicative features for identifying Interdisciplinarity

The previous literature [5][8] so far coincide to a common conclusion that most of the problems at hand now in Science are beyond the boundaries of any one single field. If one agrees to this fact, then it is natural to be interested to identify proper and permanent indicative features that could efficiently distinguish the core and interdisciplinary fields. Our approach here is to propose such indicative features for each of the fields. The rest of the section elaborately describes the proposed features one by one and their significance in unfolding interdisciplinarity of a field.

A. Reference Diversity Index (RDI)

The National Academy's definition [12] suggests that the key aspect to check the existence of interdisciplinarity is whether the research outputs reflect *knowledge integration*. In citation network, the references of a paper are the indicators of diversity of knowledge sources, i.e., the related subject areas from where the paper has been motivated. Moreover, it is quite intuitive that the more is the breadth of the references of a paper, the more interdisciplinary it should be. Therefore, to formulate the diversity of references, we propose a simple quantitative measure described below.

Definition 1: ***Reference Diversity Index (RDI):*** *The RDI of a paper is the entropy of its reference set in terms of different fields the paper cites. The RDI of a field is the average of the RDIs of all the papers belonging to that field.*

Let X_i be a paper of field f_i, and it refers to papers of k different fields namely $f_1, f_2, ..., f_k$ (f_i may be one of the fields in f_1 to f_k). The *Reference Diversity Index (RDI)* of paper X_i denoted by $RDI(X_i)$ describes the heterogeneity of the bibliometric set of the paper as follows:

$$RDI(X_i) = -\sum_j p_j log(p_j) \tag{1}$$

where p_j is the proportion of references of X_i that are received by the papers of field f_j. In other words, it is the ratio of the number of references made to the field f_j by the paper X_i to the total number of references that the paper X_i makes. The

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

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

average is taken over all the papers in field f_i to get the RDI score of f_i.

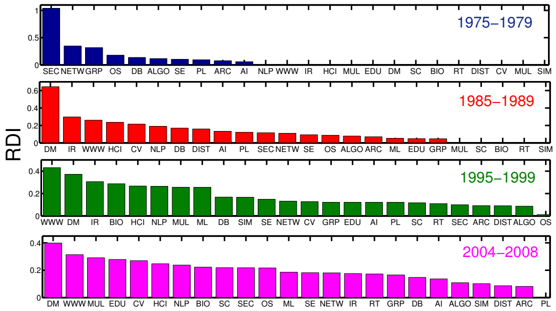

Fig. 1. Reference Diversity Index (RDI) of all the fields in computer science domain in four time-windows. The x-axis is sorted (descending order) by the RDI value.

Figure 1 illustrates the results of the RDI measured for the fields of computer science domain in four different time windows. All the results are sorted in descending order of RDI to get an idea of the rank of the fields in each time window. The more the RDI value of a field the more it is indicative of an interdisciplinary field. After 1975-1979, the interdisciplinary work mainly started emerging and the fields like Data Mining, World Wide Web, Human Computer Interaction, Information Retrieval consistently remain at the top positions in terms of their RDI values (Figure 1). At the same time, the fields like Algorithms, Operating Systems, Hardware and Architecture, Databases, Programming Languages steadily accelerate to the bottom of the rank list. Another important observation is that the degree of interdisciplinarity in terms of RDI for all the fields gradually seems to get uniform over the years (the bars in Figure 1 for all the fields gradually acquire equal height over the years). It is a clear indication of an increasing rate of interdisciplinary activities manifesting across the entire domain in the last few decades.

B. *Citation Diversity Index (CDI)*

When analyzing the inward citation distribution patterns of the fields in our dataset, we land up to an interesting observation that could be another characteristics of interdisciplinary fields distinguishing them from the core fields. We notice that though the skewness of the inward citation pattern (i.e., breadth of the incoming citations of a paper coming from different fields) is reasonably similar for all the fields, there exist few fields exhibiting a sudden sharp rise of citation diversity at certain time points. We quantitatively measure the diversity of the inward citations of a field in the following paragraph.

Definition 3: ***Citation Diversity Index (CDI):*** *The CDI of a paper in a particular time window is the entropy of its incoming citations coming from papers of different fields published in that time window. The CDI of a field is the average of the CDIs of all the papers belonging to that field.*

Let X_i be a paper of field f_i published in the time window t_i[3], and is cited by the papers (also published in t_i) of k different fields namely $f_1, f_2, ..., f_k$ (f_i may be one of the fields in f_1 to f_k). The *Citation Diversity Index (CDI)* of paper X_i in time window t_i denoted by $CDI_{t_i}(X_i)$ is defined to capture the diversity of the inward citations of a paper using the following equation.

$$CDI_{t_i}(X_i) = -\sum_j p_j log(p_j) \tag{2}$$

where p_j is the proportion of citations of paper X_i received from the papers (published in the time window t_i) of field f_j. The average is taken over all the papers in field f_i to get the CDI score of f_i. Similarly, we can find out the CDI of X_i in time window t_{i+1}, i.e., $CDI_{t_{i+1}}(X_i)$ by the diversity of the citations received from the papers published in t_{i+1}. This indicates the diversity of new citations for the same paper in the next time window. Then for a field f_i, the difference in the diversity of inward citations between two successive time windows (t_i and t_{i+1}) which we call *drift* can be expressed as

$$\Delta_{t_i}(f_i) = CDI_{t_{i+1}}(f_i) - CDI_{t_i}(f_i) \tag{3}$$

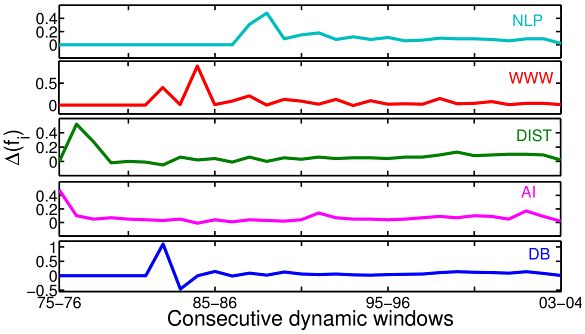

Fig. 2. (Color online) Drift of CDIs in two consecutive time windows for those fields showing sudden fluctuations in their temporal spectrum. The spectra of the other fields are reasonably stable. The value in y-axis corresponding to the label $(t_i - t_{i+1})$ in x-axis indicates the difference of the CDI values obtained from the time windows t_i and t_{i+1}.

The interpretation of this difference Δ is as follows. If the temporal profile of Δ is roughly stable for a field then it would mean that the diversity of inward citations does not change over time. However, there are certain fields where at some point Δ rises abruptly indicating a sudden huge difference in the diversity between t_i and t_{i+1}. Following this point, the diversity remains high at all time points thus keeping the difference Δ stable once again for the rest of the time span. In Figure 2, we plot the Δ values of those fields for which we are able to detect such a large fluctuation at some time point in the entire profile. As shown in Figure 2, WWW shows a sudden peak between the time windows 1984-1988 and 1985-1989 and then gets stabilized. Similar behavior is observed for NLP between the time windows 1988-1992 and 1989-1993. Other fields mentioned in Figure 2 indicate similar characteristics. If the fields mentioned in this figure are proved to be interdisciplinary, then the sharp rise in the value of Δ followed by a stability can be a promising signature to show the emergence of an interdisciplinary field. We will look into this issue in the next section. However, the only exception in Figure 2 is the Databases field which although seems to

[3] Note that, by the term "time window t_i" we refer to the five year time period from t_i to $t_i + 4$.

be a relatively core area of research shows a peak in Δ at around 1982-1986. Within a very short period, the Δ falls abruptly again (1983-1987) which is unlike the case of other fields discussed earlier. A closer inspection of our data shows that during the years 1982-1986, Databases received a variety of citations from fields like Computer Vision, Security and Privacy and Operating Systems. However, in the later years such citations to the Databases field are not found any more. A possible reason could be that in the later years Data Mining that had its birth from Databases (see Figure 6 later) started enjoying the cross-field citations rather than the Databases itself.

C. Membership Diversity Index (MDI)

The communities in citation network of a domain generally indicate different areas of research [11] where the intra-community citation density is higher than across communities [13]. We hypothesize that the diverse range of membership of a paper in different communities could be an indicator of its degree of interdisciplinarity. To verify our hypothesis, we conduct a community-centric measurement on the networks of four dynamic-windows (1975-1979, 1985-1989, 1995-1999 and 2004-2008). We use SLPA (Speaker listener Label Propagation Algorithm) [14] to detect overlapping communities in each dynamic-window. Then based on the membership of the overlapping nodes (papers) in each field, we define another metric called *Membership Diversity Index (MDI)* for each field as a measure of its interdisciplinarity.

Definition 2: ***Membership Diversity Index (MDI):*** *The MDI of a paper is the entropy of its membership in terms of different communities it belongs to. The MDI of a field is the average of the MDIs of all the papers belonging to that field.*

We run SLPA on the network of each dynamic-window that extracts the overlapping communities (say, $c_1, c_2, ..., c_n$). Since we know the actual field information of the papers, for each community c_j we can then find out the major field f_i such that c_j contains most of the papers from f_i. In this way, we can mark each community with a field tag that roughly signifies the research area indicated by this community. Note that it might be possible that more than one communities are marked by the same field tag since we have very few field categories (24 fields in the computer science domain) compared to the number of communities in each dynamic-window. Now for the field f_i, we extract only those papers that are part of overlapping communities in that time-window. These papers 'flagged' as overlapping papers within the field f_i form the basic constituent of the MDI measure.

We find out the membership of each such overlapping paper in the different field-tagged communities. Now, the MDI of the field f_i in a particular time-window is defined by the following equation:

$$MDI(f_i) = -\sum_{j=1}^{m} p_j log(p_j) \tag{4}$$

where p_j is the fraction of papers flagged as overlapping in f_i and is a member of the community tagged as f_j, while m is the number of fields (i.e., $m = 24$). The more the MDI value of a field the more it shows its interdisciplinarity.

Note that, since the overlaps are measured in different dynamic sliding windows, a node that belongs to a specific community in one dynamic window may move to a different community (communities) in the subsequent dynamic window because its surrounding connectivity might change in the next time window. Figure 3 shows the fields of computer science domain in four different time windows in decreasing order of MDI. Here, while in the time windows (1975-1979) and (1985-1989), Data Mining is consistently found to be at the top, in the later years the fields like NLP and Computational Biology seem to acquire the top positions. Therefore, this feature can be another indicative feature in characterizing interdisciplinarity.

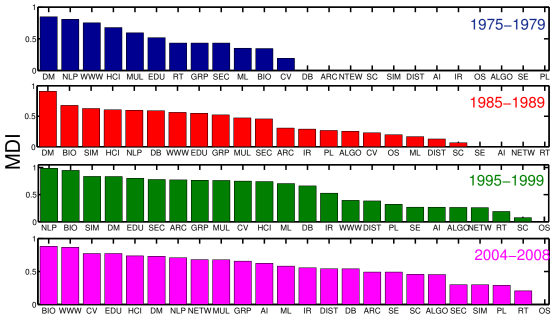

Fig. 3. Membership Diversity Index (MDI) of all the fields in computer science domain in four time-windows. The x-axis is sorted (descending order) by the MDI value.

D. Attraction Index

The selection of the new research field for both the budding and experienced researchers mostly depends on the impact and popularity of the existing fields in that time period. Therefore, the study of inclination of the authors to adopt a new field can be one of the real and relevant evidences supporting the popularity of the fields in that time period. To quantify the attractiveness of a field, we use a simple measurement called *Attraction Index* (χ) discussed below.

Definition 4: ***Attraction Index*** (χ)*: The Attraction Index of a field in a time window is defined by the number of new authors (normalized by the number of papers in that time window) who start research in that field at that time period.*

Let us assume that the number of unique authors from the beginning to the year t_i and to the year t_{i+4} who published papers in field f are n_i and n_{i+4} respectively. The number of papers of field f published in time window $(t_i - t_{i+4})$ is c_i. Therefore, the *Attraction Index* of a field f at that time window denoted by χ_f is measured by the following equation.

$$\chi_f = \frac{n_{i+4} - n_i}{c_i} \tag{5}$$

In Figure 4, we plot the value of χ for all the fields (in decreasing order of χ) in four different time windows. We can observe that though the fields like OS, Networking hold the top few positions in terms of χ in the earlier two time windows (1975-1979 and 1985-1989), in the recent years, these positions are gradually occupied by the fields like Computational Biology, WWW, Data Mining. This observation can be a distinctive factor to categorize core and interdisciplinary fields.

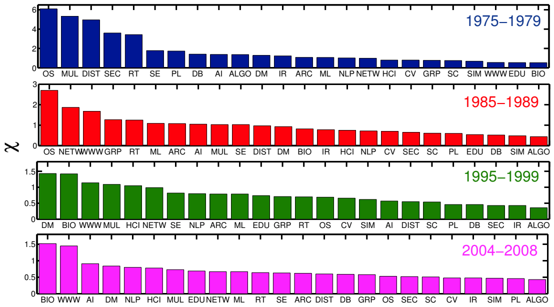

Fig. 4. Attraction Index of all the fields in computer science domain in four time-windows. The x-axis is sorted (descending order) by the χ value.

IV. Unsupervised Classification model

In the previous section, we have proposed four features with the intention that they would be indicative to explore the degree of interdisciplinarity of a field as well as help classifying the core and interdisciplinary fields. In this section, we propose an unsupervised classification model that can effectively cluster the fields based on the similarity of these features. Note that, we only consider the most recent time period of 1995-2008 for this classification[4]. In this model, each field f is represented by a feature vector of size four. The entries of the vector correspond to the value of four features namely $RDI(f)$, Δ_f, $MDI(f)$ and χ_f. Then we create a symmetric adjacency matrix $A_{24\times 24}$ whose (i, j) cell, $A(i, j)$, denotes the cosine similarity of the feature vectors corresponding to the fields f_i and f_j. For instance, let us assume that V_i and V_j represent the feature vectors corresponding to the fields f_i and f_j respectively. Then $A(i, j)$ represents the cosine-similarity between the feature vectors V_i and V_j as indicated by the following equation:

$$A(i,j) = cos(V_i, V_j) = \frac{\sum_{k=1}^{4} V_{ik} \times V_{jk}}{\sqrt{\sum_{k=1}^{4} V_{ik}^2} \times \sqrt{\sum_{k=1}^{4} V_{jk}^2}} \qquad (6)$$

An undirected and weighted network is created based on the adjacency matrix A, and the network is fed into the classification module. We use the algorithm proposed by Waltman et al. [15] for the unsupervised clustering.

The result of the clustering algorithm is pictorially depicted in Figure 5. It can clearly observed from the figure that the fields get divided into two distinct clusters. The cluster represented by the green color comprises of eight fields; all of them seem to be interdisciplinary fields except Databases. The reason could be that the fields like WWW, NLP, Data Mining got the major motivation and ideas from Databases when emerging as separate fields (see Figure 6 for further details). Therefore, though individual features could not reflect this similarity properly, their combination efficiently unveils the latent similarity in the clustering results. On the other hand, the cluster represented by the red color consists mainly of the fields which show their consistent existence from the very beginning. Therefore, this cluster seems to be representing the core fields of computer science. To the best of our knowledge, this is the first attempt to present a quantitative definition of interdisciplinarity in terms of a set of distinctive features that neatly separates out the core from the interdisciplinary research areas.

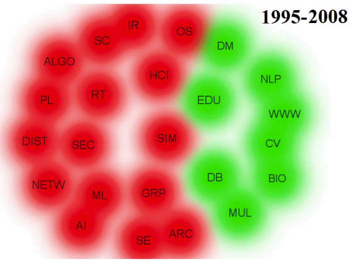

Fig. 5. (Color online) The result of the unsupervised classification model. We only consider the recent time window of 1995-2008 for this classification. Two clusters are represented by two different colors (red and green).

V. Evolutionary landscape of interdisciplinary fields

Since from the previous section, we obtain two distinct clusters of core and interdisciplinary fields in the computer science domain, the immediate question we ask is that how such an interdisciplinary field could have evolved from the cross-fertilization of the various core fields. Are the citation-based evidences capable of unfolding the evolutionary landscape of an interdisciplinary field? To answer this question, we concentrate on the temporal interaction patterns among the fields through citations over the last four decades. We hypothesize that if an interdisciplinary field has evolved from two or more fields (say, $f_1, f_2, ..., f_n$), the interactions among the fields $f_1, f_2, ..., f_n$ over the years should show a steady growth due to the sharing of knowledge and principles through cross-citations.

For this purpose, we construct a field-field citation network $G_f =< V_f, E_f >$ on top of the paper-paper citation network in each time window, where each node $f_i \in V_f$ indicates a field f_i (a collection the papers related to f_i), and a directed and weighted edge $e_{ij} \in E_f$ from f_i to f_j denotes the number of citations from the papers of field f_i to the papers of field f_j. Thus, in our experiment, we have maximum 24 vertices (if there exists at least one paper in a field, it qualifies as a vertex) in G_f at any time point. Then, we study the temporal interactions of the vertices in each time window. For the sake of conciseness, here we present the evolutionary landscape of one interdisciplinary field (WWW) and one core field (Programming Languages) which exhibit a consistent ranking for all the metrics discussed in section III. In Figure 6, we draw the contour heat maps showing the evolution pattern of WWW (top panel) and PL (bottom panel) over the last four decades. This figure has following two utilities. First of all, it takes into account the distance of two vertices as the inverse of the edge weight connecting them and groups them accordingly (green regions). In addition, the size of the font and the red circle around each vertex (field) indicates the relative importance of the vertex. Here the size of each vertex in Figure 6 (a) indicates the amount of citation received by the field (corresponding

[4] The features are most discriminative in this time-window.

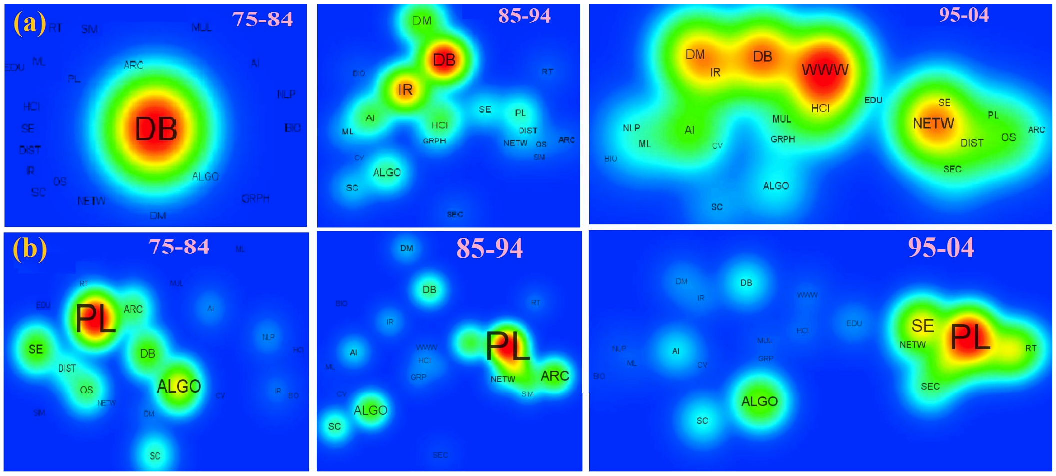

Fig. 6. (Color online) Evolutionary landscape of (a) WWW and (b) Programming Languages (PL) based on the references. Top panel shows a constant level of interaction among Databases, Data Mining, IR resulting in a new field - WWW; whereby core field like PL remains same over the years.

to the vertex) from the papers of WWW (similarly from the papers of PL in Figure 6 (b)). Furthermore in each time step, an automated threshold is defined to exclude the fields which have not received sufficient amount of citations compared to the others. As shown in Figure 6 (a), the papers of WWW have cited only the papers of Databases in early times (1975-1984); but in the later time window (1985-1994), the citations get divided among Databases, IR and Data Mining. Moreover, a distinct group comprising Databases and Information Retrieval starts evolving with small contributions from Data Mining, AI and Human Computer Interaction. Till this point, WWW is missing from the frame due to the small number of inward citations. In the latest time stamp (1995-2004), WWW is found to receive huge self citations and the previous group is enlarged with the pronounced involvement from WWW, DM, DB and IR. It clearly explains the evolution dynamics of WWW. However, another group is noticed in the last time window consisting mainly of Networking, Software Engineering, Distributed Systems and Security & Privacy. It probably indicates another line of interdisciplinary research manifesting in the form of secured distributed networking. On the other hand, if we look at Figure 6 (b) demonstrating the evolution of Programming Languages (PL), a constant appearance of PL is noticed from the very beginning. This indicates that Programming Languages was one of the contributory fields in computer science domain earlier and remains significant afterwards. This could be the first and fundamental study to understand the basic ingredients responsible for generating a new field of research and helps develop a prediction system capable of recommending the probable fields whose cross-fertilization can produce another field of research in the near future.

VI. Impact of Interdisciplinary research on the entire domain

The results that we have obtained so far lead us to advocate that interdisciplinary research has started emerging from the cross-hybridization of the ideas between the fields that otherwise have little overlap as they are studied independently. But how does it affect the entire domain as a whole? Where is the exact position of interdisciplinary research in the computer science research communities? In this section, we demonstrate three evidences that can systematically explain the raised questions.

TABLE II. Citation-based measurements extracted from the time window 1995-2008.

Statistics	Core fields (CR)	Interdisciplinary fields (INT)
Avg. number of papers per field	21,846	28,091
Avg. number of citations received by a paper	1.343	2.433
Avg. number of citations received by an author	1.349	1.813
Avg. number of citations received by a venue (in terms of the papers published in that venue)	7,276	8,382
Fraction of papers among the top 10% high impact papers	0.338	0.662
Fraction of authors among the top 10% high impact authors	0.257	0.743
Fraction of cross-citations	0.081 (CR → INT)	0.067 (INT→CR)

A. Citation-based measurements

The classification result obtained in section IV provides the scope of further analyzing the citation pattern of core and interdisciplinary fields separately. For this, we measure citation-based statistics listed in Table II to show individual impact of core and interdisciplinary fields. Note that, since the classification in section IV is based on the network constructed within 1995 to 2008, here also all the experiments are conducted in this time window. The first result in Table II shows that the number of papers per interdisciplinary field is higher than the core field in this time period. Next, three analogous measures based on the average number of incoming citations received by (i) a paper, (ii) an author and (iii) a venue (journal/conference) show that all values are higher for interdisciplinary fields. Subsequently, we take a step further, and show two more fine-grained properties related to each field. The first one extracts top 10% highly cited papers in 1995-2008 and finds how many of them belong to the core and the interdisciplinary fields separately. Similar experiment is conduced based on top 10% highly cited authors [11]. These two measures once again corroborate the same conclusion that the majority of the top cited papers and authors present in 1995-2008 belong to the interdisciplinary fields. Finally, we measure the fraction of citations emitting from the papers of the core fields and pointing to the papers of the interdisciplinary fields and vise-versa. These cross-citation measures in the last row of Table II show the increasing importance of interdisciplinary research even among the researchers of the core fields.

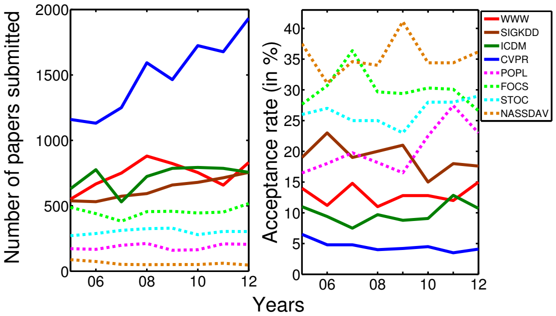

Fig. 7. (Color online) (a) Number of submissions and (b) acceptance rate of papers at top-tier conferences over the last eight years.

B. *Submission and acceptance history of top-tier conferences*

The most practical and real-world observations conforming to the popularity of different research fields would be the history of submission and acceptance statistics of papers in top-tier conferences. For this purpose, we collect the recent eight years (2005-2012) statistics of four top-tier annual conferences identified by Microsoft Academic Search both for the core and the interdisciplinary fields. The selected conferences representing interdisciplinary fields are: WWW (World Wide Web), SIGKDD (Data Mining), ICDM (Data Mining) and CVPR (Computer Vision). The selected conferences representing the core fields are: POPL (Programming Languages), FOCS (Algorithms & Theory), STOC (Algorithms & Theory) and NSSDAV (Operating Systems). We mainly focus on the submission statistics of the conferences indicating the productivity of various research fields in terms of the number of papers, and the acceptance rate indicating the competitiveness across the different fields. We observe that while the growth of the number of submissions (Figure 7 (a)) is much higher in the conferences related to the interdisciplinary fields over the last eight years, their paper acceptance rates (Figure 7 (b)) consistently remain lower compared to the conferences related to the core fields. It clearly indicates that with the increasing trend of interdisciplinary research in the recent years, the competition is mounting in leaps and bounds thus making such interdisciplinary venues more tough than ever.

C. *Core-periphery analysis*

A more systematic way of understanding the impact of the research fields on the entire domain is the study of core-periphery analysis [16] of the citation network. The idea is to decompose the fields into various shells in a particular year (or in a dynamic time window) such that a high $k^s - shell$ index of a field reflects a central position in the core of the network. As mentioned earlier, both the inward and outward citations play pivotal roles in determining the impact of a field in its domain. Therefore, we take into account both of them separately to perform the k-core decomposition in four different dynamic windows (i.e., 1975-1979, 1985-1989, 1995-1999, 2004-2008).

We start by recursively removing nodes that have single link until no such nodes remain in the network. These nodes form the 1-shell of the network ($k^s - shell$ index $k^s = 1$). Similarly, by recursively removing all nodes with degree 2, we get the 2-shell. We continue increasing k until all nodes in the network have been assigned to one of the shells. The union of all the shells with index greater than or equal to k^s is called the k^s-core of the network. We repeat the experiment both for in-citation and out-citation of a node separately. Since the shell index is assigned to each paper, we calculate the fraction of papers of a field in each k^s-core of the network in each dynamic window to identify the core fields of a domain.

The multi-level pie charts in Figure 8 (a) in four dynamic time-windows show how the different branches of computer science are positioned with respect to the core-periphery organization of the citation network (considering inward citations). Each level of the pie-chart represents one of the k^s-shell regions, i.e., the innermost layer represents Region I (largest k^s-shell index), followed by Region II, Region III, and finally the outermost layer represents the peripheral Region IV. In each layer, we show the fraction of papers belonging to a field. The pie charts for the time windows 1975-1979 and 1985-1989 show that the Region I consists mostly of core fields like Databases, Programming Languages and Software Engineering; while after that it is dominated by the more applied fields like Networking, Distributed Systems, Data Mining with a small contribution from Hardware & Architecture and Databases. In all other regions, all branches of computer science are present. From these results, we can infer that the core of the computer science is gradually being shaped by the more applied fields.

As mentioned earlier, while inward citation represents the *authoritativeness* of a field, the outward citation shows the *hubness* of a field, i.e., the propensity of a field to cite others. The degree of hubness of a field is equally important to measure its impact since the high degree hub papers (fields) usually act as the connectivity backbone of the network, sometimes creating paths between distant fields thereby, unfolding a scope for the emergence of new transdisciplinary fields. Therefore, we extract the core-periphery organization of citation network with respect to the outward citations as shown in Figure 8 (b). Surprisingly, while Algorithms and Theory has been consistently appearing at the periphery region in Figure 8 (a), the core regions are heavily dominated by Algorithms in Figure 8 (b) along with an additional contribution from Databases. Recently, the core region is covered by the emerging fields like Computer Vision, Multimedia and Distributed Systems. In short, Figure 8 presents a clear indication of the position of different fields within the domain and that the interdisciplinary fields are accelerating steadily toward the core of computer science domain.

VII. Discussion and Conclusion

This paper attempts to present a measurement study in categorizing core and interdisciplinary fields in computer science

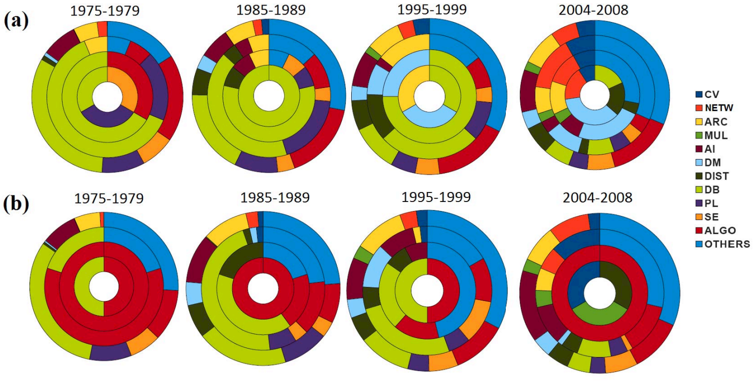

Fig. 8. (Color online) Multilevel pie-chart for the dynamic windows 1975-1979, 1985-1989, 1995-1999 and 2004-2008 showing the core-periphery organization of the citation network of computer science with respect to (a) inward citations and (b) outward citations.

domain which is so far missing in the existing literature. In doing so, we propose several indicative features to unfold the degree of interdisciplinarity of a field that can indeed help to build the classification model. The perspectives provided by the citation based indices also suggest that the practice of interdisciplinarity in citations occurs mainly between related scientific communities and this phenomenon has undergone a much more modest increase over the years. Few fields such as Data Mining, WWW, Natural Language Processing, Computational Biology, Computer Vision, Computer Education provide clear indications of interdisciplinarity in terms of all the metrics proposed here. Moreover, for already very interdisciplinary fields, such as Data Mining, the indicators may have a certain "saturation" effect forcing it towards the core region of the computer science domain.

The analysis of the computer science domain presented here is not only interesting on its own right to track the evolution of interdisciplinarity over time, but it also provides essential benchmarks for future investigations. Now-a-days, top funding agencies like NSF[5] and EU[6] have already started promoting interdisciplinary research in their own separate ways. Therefore, a quantitative and effective method of identifying interdisciplinarity could be crucial in making important funding decisions. Furthermore, the proposed classification scheme can also be effective to identify interdisciplinary research papers among all the published papers. This experiment is the first attempt to measure the interdisciplinarity of a field and also a fundamental step for building up a specialized recommendation system aiming to predict future combination of fields generating new interdisciplinary areas of research. As a final remark, we would like to stress on the fact that the study presented here is generic and the entire methodology can be used exactly in its current form to identify interdisciplinarity in any other domain of research.

References

[1] P. van den Besselaar and G. Heimeriks, "Disciplinary, multidisciplinary, interdisciplinary - concepts and indicators," The 8th conference on Scientometrics and Informetrics ISSI2001, Sydney. Australia, July 16-20, 2001.

[2] F. Morillo, M. Bordons, and I. Gómez, "An approach to interdisciplinarity through bibliometric indicators," *Scientometrics*, vol. 51, no. 1, pp. 203–222, Apr. 2001.

[3] C. S. Wagner, J. D. Roessner, K. Bobb, J. T. Klein, K. W. Boyack, J. Keyton, I. Rafols, and K. Börner, "Approaches to understanding and measuring interdisciplinary scientific research (idr): A review of the literature," *J. Informetrics*, vol. 5, no. 1, pp. 14–26, 2011.

[4] D. J. de Solla Price, "Networks of Scientific Papers," *Science*, vol. 149, no. 3683, pp. 510–515, Jul. 1965.

[5] A. L. Porter, A. S. Cohen, J. D. Roessner, and M. Perreault, "Measuring researcher interdisciplinarity," *Scientometrics*, vol. 72, no. 1, pp. 117–147, 2007.

[6] L. Leydesdorff, "Betweenness centrality as an indicator of the interdisciplinarity of scientific journals," *J. Am. Soc. Inf. Sci. Technol.*, vol. 58, no. 9, pp. 1303–1319, Jul. 2007.

[7] J. Klein, *Interdisciplinarity: History, Theory, and Practice*. Wayne State University Press, 1990.

[8] N. Metzger and R. N. Zare, "SCIENCE POLICY:Interdisciplinary Research: From Belief to Reality," *Science*, vol. 283, no. 5402, pp. 642–643, Jan. 1999.

[9] R. K. Pan, S. Sinha, K. Kaski, and J. Saramäki, "The evolution of interdisciplinarity in physics research," *Nature Scientific Reports 2*, 2012.

[10] 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.

[11] T. Chakraborty, S. Sikdar, V. Tammana, N. Ganguly, and A. Mukherjee, "Computer science fields as ground-truth communities: Their impact, rise and fall," in *ASONAM*, 2013 (accepted).

[12] C. Research, N. Sciences, N. Engineering, and I. Medicine, *Facilitating Interdisciplinary Research*. National Academies Press, 2005.

[13] P. Chen and S. Redner, "Community structure of the physical review citation network," *J. Informetrics*, vol. 4, no. 3, pp. 278–290, 2010.

[14] J. Xie, B. K. Szymanski, and X. Liu, "SLPA: Uncovering Overlapping Communities in Social Networks via a Speaker-Listener Interaction Dynamic Process," in *ICDM Workshops*, 2011, pp. 344–349.

[15] L. Waltman, N. J. van Eck, and E. C. M. Noyons, "A unified approach to mapping and clustering of bibliometric networks," *J. Informetrics*, vol. 4, no. 4, pp. 629–635, 2010.

[16] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir, "A model of Internet topology using k-shell decomposition," *PNAS*, vol. 104, no. 27, pp. 11 150–11 154, Jul. 2007.

[5] http://www.nsf.gov/od/iia/additional_resources/interdisciplinary_research/

[6] http://cordis.europa.eu/fp7/ict/content-knowledge/projects_en.html#project-list_