Journal of Informetrics 9 (2015) 69–78

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Understanding and modeling diverse scientific careers of researchers

Tanmoy Chakraborty[a,*], Vihar Tammana[b], Niloy Ganguly[a], Animesh Mukherjee[a]

[a] Department of Computer Science & Engineering, Indian Institute of Technology, Kharagpur 721302, India
[b] Microsoft Corporation, Bellevue, WA, United States

ARTICLE INFO

Article history:
Received 22 September 2014
Received in revised form
16 November 2014
Accepted 18 November 2014

Keywords:
Scientific career
Diversity
Stochastic model

ABSTRACT

This paper analyzes the diverse scientific careers of researchers in order to understand the key factors that could lead to a successful career. Essentially, we intend to answer some specific questions pertaining to a researcher's scientific career – What are the local and the global dynamics regulating a researcher's decision to select a new field of research at different points of her entire career? What are the suitable quantitative indicators to measure the diversity of a researcher's scientific career? We propose two entropy-based metrics to measure a researcher's choice of research topics. Experiments with large computer science bibliographic dataset reveal that there is a strong correlation between the diversity of the career of a researcher and her success in scientific research in terms of the number of citations. We observe that while most of the researchers are biased toward either adopting diverse research fields or concentrating on very few fields, a majority of the prominent researchers tend to follow a typical "scatter-gather" policy – although their entire careers are immensely diverse with different types of fields selected at different time periods, they remain focused primarily in at most one or two fields at any particular time point of their career. Finally, we propose a stochastic model which, quite accurately, mimics the notion of field selection process observed in the real publication dataset.

1. Introduction

> *"It is really important to do the right research as well as to do the research right. You need to do 'wow' research, research that is compelling, not just interesting."*
>
> – Richard M. Reis, Stanford University

Of all the decisions we make as an emerging scientist, none is more important than identifying the right research area, and in particular, the right research topic. The success of scientific career gets determined by these two choices. Change in scientific research career can be defined as any major change in work-role requirements or work context (Brett, 1982, 1984; Nicholson, 1984) and as a process that may result in a change of job, profession, or a change in one's orientation of work while continuing in the same job (Albert, Ashforth, & Dutton, 2000; Evered & Louis, 1981). People believe that many factors act as an active role to regulate these changes. For instance, researchers might try to align themselves with the cutting-edge

* Corresponding author. Tel.: +91 9475444030.
E-mail address: its_tanmoy@cse.iitkgp.ernet.in (T. Chakraborty).

http://dx.doi.org/10.1016/j.joi.2014.11.008
1751-1577/

research at the current time and as a result of this a change in scientific research career becomes unavoidable (Passi & Mishra, 2004). On the other hand, this career shift might be described as an effect of "saturation" in the field of a researcher leading to a switch to the other fields (Pettigrew, 1990).

In this paper, we use a massive dataset of scientific publications in computer science domain and attempt to analyze the local and the global dynamics regulating a researcher's decision to select new field of research over the entire career. In particular, we investigate how the "relatively more successful" researchers make choices to select their fields of research at different points in their career. A remarkable observation is the "relatively more successful" researchers unlike the rest of the lot tend to behave in a "scatter-gather" fashion, i.e., they seem to work in diverse research areas over their entire career span; however in each time-slice of their career they pump in their most concentrated efforts in only one particular area of research. We provide extensive evidence for this through appropriate quantification of local and global diversity measures for the choice of research topics followed by rigorous empirical analysis of a large volume of citation data to corroborate the above observation.

This observation further motivates us to build a stochastic model that can reproduce the real-world phenomenon of field selection process. Evaluations of our model through the real-world data lead us to conclude that our model, quite accurately, mimics the field selection process for all the researchers present in the dataset. Note that, we use the terms "author" and "researcher" interchangeably in the rest of the paper. We make our experimental codes available in the spirit of reproducible research: https://github.com/centrality-multiplex/Modeling-ResearcherCareer.git.

2. Related work

Recently, research on citation, co-citation and co-authorship networks has gained interest in information sciences (Börner, Chen, & Boyack, 2003; Börner, Dall'Asta, Ke, & Vespignani, 2005; Chen, 2003) and in statistical physics (Barabási et al., 2002; Newman, 2001; Redner, 2005). The accumulation of published articles enables also the drawing of evolutionary tree-like structures of referencing over time. One famous example is the idea of a "historiograph" proposed by Garfield (Garfield, 1977, 2004; Garfield, Pudovkin, & Istomin, 2003).

Field mobility, or field migration (Vlachy, 1981), is defined as scientists moving into new research topics. Field mobility can be measured by identifying different research topics (fields or subfields), estimating the activity of scientists in these fields, and following the activity of scientists over time to mark the transitions. Field mobility has been investigated already since the 1980s (Le Pair, 1980; van Houten, van Vuren, Le Pairs, & Dijkhuis, 1983). Field mobility has been discussed as the driving force for the exploration of new territories in the "landscape" of science (Scharnhorst, 2001; Urban, 1982). More specifically, field mobility has been modeled as an exchange mechanism between research fields leading to a co-evolution or coupled growth of scientific specialties (Chen, Börner, & Fang, 2013; Ebeling & Feistel, 1986). Hellsten, Lambiotte, Scharnhorst, and Ausloos (2007) introduce a new approach to detecting scientists' field mobility by focusing on an author's self-citation network, and the co-authorships and keywords in self-citing articles.

Changing patterns of scientific activity have been also discussed in the context of interdisciplinarity. Attempts to measure interdisciplinarity rely on citation and publication patterns (see e.g., Rinia, van Leeuwen, Bruins, van Vuren, & van Raan, 2002). However, some studies also follow certain authors through their publication records (Pierce, 1999; Urata, 1990). Some studies use interviews and surveys to trace academic careers but this approach is restricted to rather small case studies (van Houten et al., 1983; Wagner-Döbler & Berg, 1993). Career moves of scientists are also a topic of science history or sociology research (see for e.g., an earlier research Gilbert, 1977). Currently, there are no automated techniques for quantitatively measuring scientists' field adaptation/mobility.

On the other hand, Zhou, Ji, Zha, and Giles (2006) illustrate how topic evolution and social interaction lead to build an author's research career. Biryukov and Dong (2010) make an attempt to analyze an author's scientific career through exploration of scientific communities based on some features and use them to compare the sets of top ranked conferences with the low ranked ones. Despite well-documented literature on bibliographic dataset, the last few decades have witnessed a scarcity of empirical research on career change. The key questions related to the dynamical process of career change along with the associated outcomes remain mostly uninvestigated.

In this paper, we attempt to investigate the following research questions – (i) How a researcher decides to select her field of research at different points in her career? (ii) What are the suitable quantitative indicators to measure the diversity of a researcher's scientific career? (iii) Can we mimic the real-world field selection process of researchers using a computational model? To the best of our knowledge, this is the first attempt to analyze the scientific career of researchers extensively using new quantitative indicators. Moreover a new computational model is proposed to imitate the field selection process of researchers that unfolds the real-world dynamics controlling the shift of research career.

3. Dataset

In this experiment, we used the DBLP dataset of the computer science domain developed by Chakraborty et al. (Chakraborty, Sikdar, Ganguly, & Mukherjee, 2014; Chakraborty, Sikdar, Tammana, Ganguly, & Mukherjee, 2013). The dataset contains 702,973 valid papers and 495,311 authors. The attributes of each paper are as follows: the name of the research paper, a unique index of the paper, the list of author(s), the year of publication, the publication venue, the list of research papers the given paper cites and (in some cases) the abstract and the keywords of the paper. After a series of preprocessing

Table 1
General information of the dataset.

Number of papers	702,973
Number of authors	495,311
Average number of papers by an author	3.52
Average number of authors per paper	2.609
Number of unique venue name	1,705

Table 2
Percentage of papers in various fields of computer science domain.

Fields	% of papers	Fields	% of papers
AI	12.64	Algorithm	9.89
Networking	9.41	Databases	5.18
Distributed Systems	4.66	Comp. Architecture	6.31
Software Engg.	6.26	Machine Learning	5.00
Scientific Computing	5.73	Bioinformatics	2.02
HCI	2.88	Multimedia	3.27
Graphics	2.20	Computer Vision	2.59
Data Mining	2.47	Programming Language	2.64
Security	2.25	Information Retrieval	1.96
NLP	5.91	World Wide Web	1.34
Education	1.45	Operating Systems	0.90
Embedded Systems	1.98	Simulation	1.04

steps, we retained those entries in the dataset which contain the information about the paper index, the title, the publication venue (conference/journal) of the paper, the year of publication and the references. Since the filtered dataset did not have the necessary field[1] information of the papers, we tagged them using the Microsoft Academic Search (MAS) Engine[2] (detailed description of the field tagging process can be found in Chakraborty et al., 2013). MAS categorizes papers of computer science domain into the 24 fields as noted in Table 2. We crawled the site to find the field(s) of papers present in the filtered dataset using the title of the paper. Approximately, 88.12% of the papers could be tagged with their respective fields when searched with the paper title. Fields of rest 11.88% of the papers have been inserted using the conference/journal name of the paper. About 11.23% of the papers have more than one field. Table 2 notes the percentages of papers in various fields in the tagged dataset. Some of the general information pertaining to the filtered dataset are presented in Table 1. For the current study, we have considered papers belonging to only one particular field for a better interpretation of the results. Furthermore, we have considered only those authors who have published at least five papers for the sake of measuring one of our proposed metrics, called *window-entropy* as discussed in Section 5.3. We have made the dataset publicly available at http://cnerg.org (see "Resources" tab).

4. Diversity measures

Diversity of an author's research career can be understood as the degree of variation/changes in research fields over the entire career. Since "diversity" of a sequence can be efficiently measured by Shannon's entropy (Shannon, 1948), we propose two different versions of entropy measurement to quantify diversity of an author's research career. If F is the set of unique fields of papers written by an author a, the *plain entropy* of author a (denoted by $H_p^a(F)$) is calculated over the number of times author a writes papers in a particular field in her entire research career as defined by the following equation:

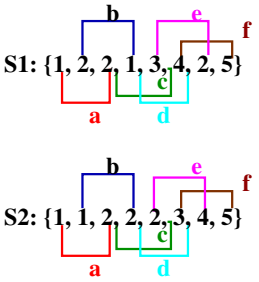
$$H_p^a(F) = -\sum_{i \in F} p_i \log(p_i) \tag{1}$$

where

$$p_i = \frac{number\ of\ papers\ written\ by\ a\ in\ field\ i}{total\ number\ of\ papers\ written\ by\ a}.$$

Zero plain entropy implies that an author worked in a single field throughout her career whereas a high value indicates that she has worked in various fields in different time spans of her career. However, the plain entropy does not capture the order information of a particular value within the sequence; it just considers probability distributions, i.e., if we interchange the position of different entries in a sequence keeping the frequency contribution of each individual field same, the plain entropy of the sequence remains constant. In our case, since our primary interest is to understand the change in research fields adopted by an author in different time periods, the ordering information of fields in the sequence turns out to be important. Therefore in order to capture the local diversity, we propose another measure of field diversity for an author a

[1] Note that, the different sub-branches like Algorithms, AI, Operating Systems, etc. constitute the different "fields" of computer science domain.
[2] http://academic.research.microsoft.com/.



	S1	S2
H_p	0.65	0.65
H_w	0.41	0.29

Entropy at different positions of sliding window

	S1	S2
a	0.27	0.27
b	0.27	0.27
c	0.48	0
d	0.48	0.27
e	0.48	0.48
f	0.48	0.48

Fig. 1. Illustrative example of the calculation of the plain (H_p) and window (H_w) entropies. In two sequences ($S1$ and $S2$), the frequency of each entry is same but their positions are different. Therefore, though the values of H_p are same for two sequences, the value of H_w is different. The positions of the sliding window (of size 3) are represented by different colors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

called the *window entropy* (denoted by $H_w^a(F)$) defined as follows – a window of size k slides over the sequence of fields in F, the plain entropy for the sequence contained within that window at each position is calculated, and the mean of all these positions is computed to measure the window entropy of the entire sequence as described in the following equation:

$$H_w^a(F) = -\frac{1}{n-k+1}\sum_{i=1}^{n-k+1} H_p^a(w_i) \tag{2}$$

where w_i is the set of fields in the ith sliding window of size k, ranging from i to $(i+k-1)$. The window entropy indicates the diversity in the selection of fields in short spans of time by considering only previous k fields in the sequence. Fig. 1 illustrates the calculation of these two measures. The motivation behind these measures is to understand whether the author is working simultaneously in diverse fields throughout the career or she is following the "scatter-gather" policy, i.e., while working in diverse fields at the macro scale, at the micro scale, concentrating on only one particular field within a given time slice. Low $H_w^a(F)$ indicates that the author indeed follows a "scatter-gather" policy; whereas high values indicate that the author has a tendency to work in many different fields simultaneously within short periods of time.

5. Experimental results

5.1. Statistical analysis of authors' careers

We first plot in Fig. 2 the distribution of fields selected by the authors over their entire career. It follows a truncated power-low behavior and shows that around 64% of the total authors worked only in one field, 18% of the total authors worked in two fields and so on. In Fig. 3(a), we show the average number of fields in which an author contributed in a particular year from the start of her career (i.e., after her first publication). It can be observed that as the career of an author progresses over time, the number of distinct fields she has contributed to increases till around fifteen years and then mostly stabilizes. This

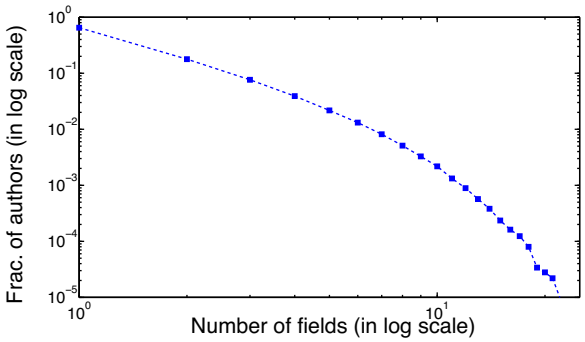


Fig. 2. Distribution of fields adopted by the authors (plotted in log–log scale). The y-value corresponding to the x-value indicates the fraction of authors contributing to x number of fields in their careers.

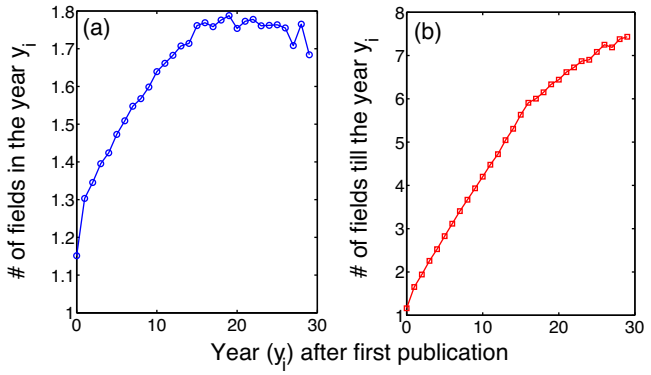


Fig. 3. (a) Average number of fields to which an author contributes in a year after the first publication and (b) the total number of fields in which an author contributes till a particular year after the first publication. Both of these are calculated as an average over all authors present in the data set.

may be understood as an author's career profile, in the initial years a scientist is usually more actively developing skillset in a field and is keen setting up collaborations in the field as well as in the closely associated ones. Eventually, after fifteen years of her scientific career, she would generally tend to focus on a fixed set of fields where she has considerable expertise; thus a decline forward at the end of the curve is observed in Fig. 3(a). In Fig. 3(b), we plot the total number of fields contributed by an author (on an average) up to a certain year after her first publication. It is to be noted that this plot cannot be obtained directly by cumulating the result in Fig. 3(a) since we measure the total number of distinct fields in an author's career in Fig. 3(b); whereas cumulating over Fig. 3(a) might count a field more than once. It can be observed in Fig. 3(b) that the total number of fields to which an author contributes increases till around fifteen years and at a relatively lower rate afterward. This plot also indicates the average number of years that an author usually takes to start contributing to the *i*th field. For example, an author starts contributing to the third field in about 6–7 years from the start of her career.

5.2. Diversity of author's scientific career

Next, we analyze the diversity of an author's scientific career at different time points in terms of two proposed entropy-based measures, namely the plain entropy ($H_p^a(F)$) and the window entropy ($H_w^a(F)$). Fig. 4(a) shows the plain entropy probability histogram of all authors, i.e., fraction of authors with plain entropy ranges divided into several buckets. It can be observed that high fraction of authors tend to have small plain entropy since most of them have worked in very few research fields. Fig. 4(b) shows the average plain entropy of sequence of fields of an author till a certain number of publications. We take an empirical cutoff of 80 publications while computing the entropy measures because most of the authors fall in this region. The publications of authors after this cutoff are ignored for the entropy calculation. One can observe the increase, relative to the number of fields selected by an author which is faster at the early stages following a gradual stabilization toward the end.

Similarly, we plot both these figures for the window entropy in Fig. 5. Interestingly, the major concentration of authors lies in 0.5–1.2 window-entropy region plus a significant amount of mass in the first bucket. From both the histograms of plain and window entropies, one can conclude that though the average behavior indicates that an author tends to select only a few research fields in her entire research career, she seems to prefer working simultaneously in multiple fields at a particular time point. We observe that such fields are quite related to each other. If we would have taken more coarse-grained field classification scheme, these similar fields might have come within one particular field. For instance, Graphics and Multimedia are so closely related fields in the computer science domain that one can assume these as two subfields under the same

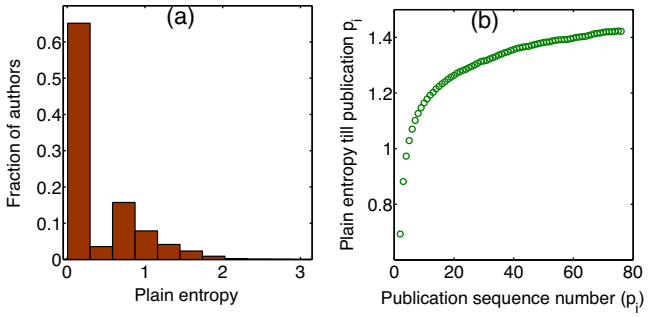


Fig. 4. (a) Plain entropy distribution histogram of authors and (b) plain entropy till a certain number of publications (averaged over all the authors).

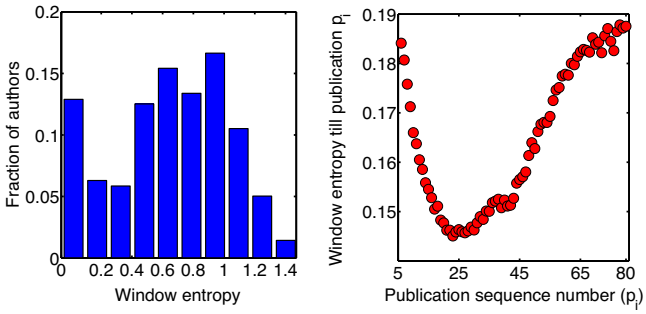


Fig. 5. (a) Window entropy distribution histogram of authors and (b) window entropy till a certain number of publications (averaged over all the authors).

top level field called "Image Processing". Since we have adopted the field categorization proposed by Microsoft Academic Search, we restrict our entire analysis to this finer categorization scheme.

The scatter plot in Fig. 5(b) indicates that researchers, at the beginning of their research, tend to select a number of fields simultaneously thus making the average entropy higher. The reason could be that they are initially not very confident which particular fields of research they should select in order to survive in the scientific community. Gradually with experience, they tend to get stabilized after publishing 20–25 papers, thus reaching a lower entropy at the middle of the curve. Following this, again a steady growth of the curve indicates that possibly they start collaborating with other researchers from the other fields and hence, by doing this, they tend to advocate interdisciplinary research toward the end of their careers.

5.3. Correlation between plain and window entropies

In Table 3, we present a confusion matrix indicating the correlation between the two diversity measures. We classify the entire population of authors into four parts corresponding to the four cells of the matrix. Note that, the low (high) entropy values correspond to values below (above) the median value of the respective type of entropy. Each region on the matrix corresponds to different types of career profile. For instance, the region corresponding to low window entropy and high plain entropy indicates that the authors here do not work simultaneously in multiple fields; rather they choose to work in fields one after the other. On the other hand, the region indicating high plain and window entropies corresponds to those authors who have worked in diverse areas and also contributed simultaneously to multiple fields at any particular time point. Table 3 shows the population density (in percentage) of authors in each region. In parallel, what would be more interesting to investigate is the importance of each such region, i.e., what would be the preferred strategy a new author should adopt in order to acquire higher importance in scientific community. In any bibliographic dataset, a raw measure to quantify the importance of an author is usually the number of citations she has received by publishing papers. Therefore, we measure the importance of each region by calculating the average citations an author of the corresponding region has received (the value within parenthesis in each region of Table 3 indicates this importance). The key observation is that the region with high plain entropy and low window entropy has high average citation value compared to all other regions. This indicates that the authors having relatively higher citations follow a "scatter-gather" policy, i.e., they work in diverse fields over their entire career but remain confined to a few fields in each time slice. A deeper inspection shows that this region has the lowest density of authors that essentially implies that such authors are rare. The authors in the region with high plain entropy and high window entropy have least average citation count which indicates that the authors who have worked in a large number of fields in the entire career as well as in each shorter time slice get low citations. It is worth noting that in this experiment, we consider only those authors who have published at least five articles because the size of the sliding

Table 3
Confusion matrix indicating the population density of authors (in %) and average number of citations obtained by an author (in parenthesis) in different regions.

		Window entropy	
		Low	High
Plain entropy	Low	43.37 (11.61)	6.6 (11.21)
	High	6.61 (13.36)	43.41 (10.38)

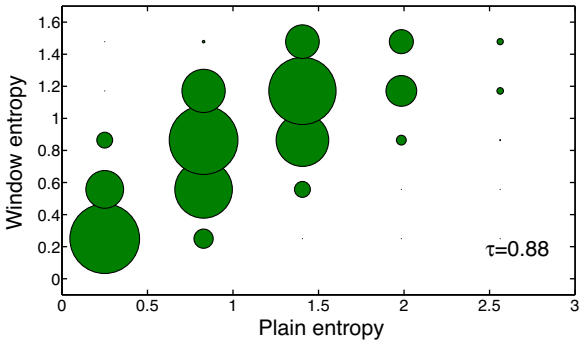


Fig. 6. Scatter plot showing the correlation between plain entropy and window entropy with size of the circle proportional to number of points in the region surrounding the center of the circle.

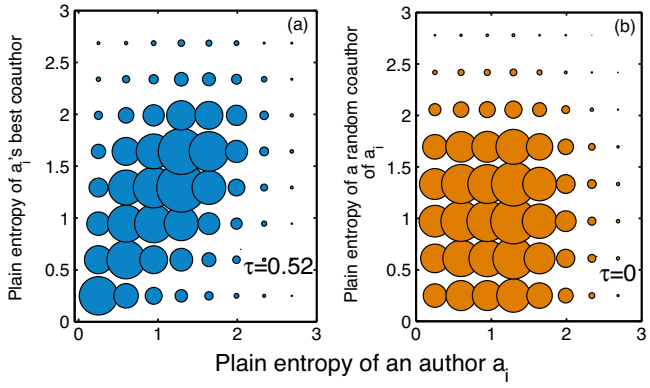


Fig. 7. Scatter plot of the plain entropy of an author (a) with her strongest collaborator and (b) with any arbitrary coauthor.

window for calculating the window entropy is assumed to be five. We further vary the window size, but similar results are observed in all the regions for different window values.

Next we measure the correlation between plain and window entropies using Pearson correlation coefficient (τ)[3] as shown in Fig. 6. The size of the circle is proportional to number of points in the region surrounding the center of the circle. The correlation is very high ($\tau = 0.88$) between these two entropies which indicates a strong dependency between them. These results once again agree with the proportion of authors shown as two most popular regions on the principal diagram of the confusion matrix (i.e., low plain entropy and low window entropy, high plain entropy and high window entropy). This correlation increases with the increase in the size of the window since k-window entropy with large k value gradually approaches the plain entropy. Furthermore, it would be interesting to understand to what extent an author is influenced by her other colleagues (coauthors) in selecting a new field. More particularly, we intend to measure the correlation between the entropy of an author with her strongest collaborator (with whom she has published the largest number of papers) and to what extent this correlation changes when compared to any arbitrary coauthor. Fig. 7(a) shows the scatter plot of the plain entropy of an author with the plain entropy of her strongest collaborator. The Pearson correlation between them is quite high (0.52) compared to any arbitrary coauthor ($\tau = 0$ as shown in Fig. 7(b)). This result can have two implications: (a) an author either tries to align herself in the direction of her strongest collaborator in choosing research fields or (b) an author chooses such a collaborator with whom the research interests have maximum alignment. Similar correlation exists for the case of window entropy (since plain and window entropies are highly correlated as discussed in Fig. 6).

6. Stochastic model

As a following step, we propose a stochastic model to reproduce the field selection process of the authors. Then we evaluate the predictions of our model by comparing the outcomes with real data. Essentially, given a set of papers arranged in the chronological order of the year of publication, the model automatically tags the papers with the appropriate fields such that the field-diversities of the authors in the model closely match with the results obtained from the real-world dataset.

[3] http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient.

The model works as follows – the input to the model are: (i) a set of papers each having information of its list of authors, year of publication and the number of citations it received so far; (ii) a set of 24 research fields (listed in Table 2); (iii) the number of bins (n) and (iv) the number of bins (k) whose constituent papers need to be tagged to bootstrap the model. All the papers are chronologically ordered based on the year of publication and are categorized into n number of equal-size bins. To avoid cold-start problem (Schein, Popescul, Ungar, & Pennock, 2002), we tag the papers of first k bins with field information obtained from the real-world data. For each paper p in the rest of the $n-k$ bins, we calculate the likelihood of p getting tagged by a field using the previous field and citation information of the papers written by the author(s) of p. Then the most likely field is preferentially selected and used to tag p. If all the authors of a paper are appearing for the first time in the system, i.e., the paper under consideration is the first paper for all the authors, then we tag the paper with a random field selected among the set of 24 fields. We then update the data structures for the successive iterations. Once all these tasks are completed, the next paper is added into the system and the above steps are repeated. If all the bins are completed, then the simulation process is over and the predictions are ready for evaluation. Note that, we are not interested to tag the papers with the exact field of computer science present in the dataset, but to annotate the papers in such a way that it indeed captures the field selection process of a researcher that in turn resembles the diversity of a researcher's scientific career. The pseudo-code of this stochastic model is shown in Algorithm 1.

Algorithm 1. Stochastic model for capturing the diversity of an author's scientific career

Input: P = a set of all publications having information of the authors, the year of publication and number of citations; F = a set of 24 research fields; n = number of bins; k = number of bins whose constituent papers are to be tagged earlier from the real dataset.

Output: Papers in P are tagged with fields.

Order all the papers chronologically based on the year of publication.
Categorize the papers into n equal-size bins, namely $b_1, b_2, \ldots, b_n$ such that b_1 and b_n contain earliest and latestpapers respectively.
Tag all the papers in first k bins ($b_1, b_2, \ldots, b_k$) with fields from the real-world dataset.
for all $b_t \in \{b_{k+1}, \ldots, b_n\}$ **do**
  for all paper $p_i \in b_t$ **do**
    for all $a_j \in Auth(p_i)$ **do** ▷ $Auth(p_i)$ = set of authors of paper p_i **do**
      $W(a_j) = \frac{\text{Citations received by } a_j \text{ till } t-1 \text{ bins}}{\sum_{a_k \in Auth(p_i)} \text{Citations received by } a_k \text{ till } t-1 \text{ bins}}$
    end for
    for all field $f_l \in F$ **do**
      $C(f_l) = \sum_{a_j \in Auth(p_i)} (W(a_j) \times N_{a_j}(f_l))$; where $N_{a_j}(f_l)$ = number of papers in field f_l
      written by a_j **do** ▷ indicating likelihood of paper p_i being tagged by field f_l
    end for
    if $\sum_{f_l \in F} C(f_l) == 0$ **then**
      Randomly select a field f_l
    else
      Preferentially select a field with probability defined by $p(f_l) = \frac{C(f_l)}{\sum_{f_l \in F} C(f_l)}$
    end if
    Tag paper p_i with field f_l
  end for
end for

6.1. Evaluation of the dynamical model

Finally, we evaluate our stochastic model with the real-world dataset by comparing the two diversity measures. The evaluation is conducted with all papers categorized into hundred bins (n = 100 in Algorithm 1) with the papers in initial ten bins (k = 10) tagged with real data. We ignore all such authors present in the dataset who have published one paper since the diversity measures for them would match even for any random selection of field. We plot the probability distributions of authors having different plain (Fig. 8(a)) and window (Fig. 8(b)) entropies obtained from the real-world dataset and from the proposed model. The correlation coefficient between the model and the real points are 0.97 and 0.95 for plain and window entropies respectively which are quite significant. Moreover, the field probability density correlation, i.e., the correlation between the frequency distribution of the fields in the entire real-world dataset (shown in Fig. 2) with the results obtained from the model is 0.88.

Evaluation can also be done by comparing the population distribution of the authors in terms of the entropy measures obtained from the real-world dataset and from the model and the average citation pattern in each region as shown earlier in confusion matrices (Table 3). The population density and average citation of authors in each region obtained from the real-world data and from the model are shown in Table 4(a) and (b) respectively. It can be observed that all the properties in four regions predicted by the model have similar percentage of authors and similar citation patterns as indicated by the real data (i.e., the authors of the region with high plain entropy and low window entropy have highest average citation among all the regions; similarly, high plain entropy and low window entropy region having lowest citations as also indicated by the real-world data). Hence, our proposed model, quite efficiently, captures the proportion of authors and the average citation count obtained from the real dataset in each of the four regions based on two entropy measures.

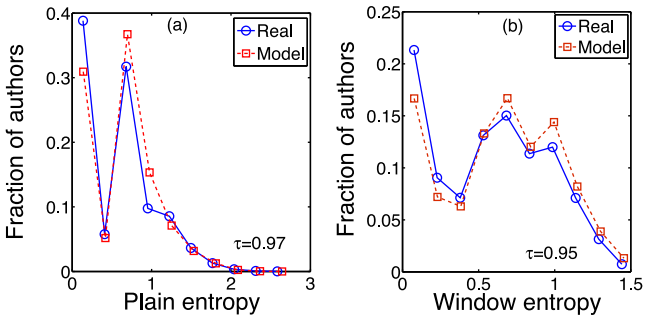


Fig. 8. (a) Plain entropy and (b) window entropy probability density plots of real-world data (blue) and the results obtained from the model (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Popularity distribution (in %) and average citation pattern (in parenthesis) of authors in each region (a) for real-world dataset and (b) for the results obtained from the model.

(a)

Plain Entropy	Window Entropy: Low	Window Entropy: High
Low	43.37 (11.61)	6.6 (11.21)
High	6.61 (13.36)	43.41 (10.38)

(b)

Plain Entropy	Window Entropy: Low	Window Entropy: High
Low	44 (11.21)	5.9 (11.15)
High	6 (12.82)	44 (10.9)

7. Conclusions

This paper has attempted to analyze the pattern of scientific field adaptation process that a researcher tends to follow in different time slices of her entire research career. We have particularly analyzed the dataset of computer science domain in order to unfold the latent characteristics of field adaptation process that indeed lead us understand how such complex adaptation process followed by the prominent researchers differs from the general behavior of authors. We conclude by summarizing our main observations and few important remarks as follows.

1. Researchers who have worked in many fields in their entire careers but remained confined in few fields in each time window get high importance in terms of citations compared to the others. Interestingly, the population of such researchers is lowest in the entire population which implies that the average population of researchers tends to avoid the typical pattern that the highly cited researchers usually follow in their research career.
2. The authors of region with high plain entropy and high window entropy have lowest average citation count which indicates that the authors who have tried various fields in the entire career as well and in each successive time period, get low citations.
3. The most popular two regions are right diagonal regions of the confusion matrix (i.e., low plain entropy and low window entropy, high plain entropy and high window entropy) which implies that generally the researchers either tend to adopt very few fields in their research career as well as in each time slice of their careers or they are very diverse in different periods of the careers thus making their entire career more diverse.
4. Finally, through the stochastic model we observe that the field selection process is indeed preferential which depends on the average expertise of the coauthors writing the paper. The expertise of an author for a particular field is usually defined by the average number of citations received by the author by publishing papers in this field.

The present study can be helpful in addressing the following applications: (i) for a partially field-tagged publication dataset, the stochastic model can be useful to tag the appropriate field information of a paper whose tags are missing, (ii) a researcher can analyze her career growth by observing different fields she has adopted so far and how this adaptation process affects her overall prestige and prominence; accordingly she can decide to modify the future career plan, (iii) a budding researcher can study this behavior prior to the beginning of her career in order to get the best pattern to make herself prominent, and (iv) most importantly, this study can lead us to develop a collaboration prediction system that can recommend a researcher the name of her appropriate collaborators. So far, our model is completely non-parametric. In future, we would like to include several other factors pertaining to the field selection process of a researcher such as the current popular research field, the effect of group collaboration into the model to make it more robust and generic. Moreover, since

the papers having multiple field-tags are excluded in the entire analysis, we intend to analyze these papers separately in order to check whether they would make further effect on the conclusions drawn here. Finally, we would like to explain how the global dynamics of scientific paradigm shift (Chakraborty et al., 2013) influences a researcher's career and vice-a-versa.

Acknowledgment

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

References

Albert, S., Ashforth, B., & Dutton, J. (2000). Organizational identity and identification: Charting new waters and building new bridges. *Academy of Management Review*, *25*, 13–17.
Barabási, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, *311*(3-4), 590–614.
Börner, K., Chen, C., & Boyack, K. W. (2003). . *Visualizing knowledge domains* (Vol. 37)
Börner, K., Dall'Asta, L., Ke, W., & Vespignani, A. (2005). Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. *Complex*, *10*(4), 57–67 (Research articles).
Biryukov, M., & Dong, C. (2010). Analysis of computer science communities based on DBLP. In *Proceedings of the 14th European conference on research and advanced technology for digital libraries, ECDL'10* (pp. 228–235). Berlin, Heidelberg: Springer-Verlag.
Brett, J. (1982). Job transfer and well-being. *Journal of Applied Psychology*, *67*, 450–463.
Brett, J. (1984). Job transitions and personal and role development. *Research in Personnel and Human Resources Management*, *2*(2), 155–185.
Chakraborty, T., Sikdar, S., Ganguly, N., & Mukherjee, A. (2014). Citation interactions among computer science fields: A quantitative route to the rise and fall of scientific research. *Social Network Analysis and Mining*, *4*, 187. http://dx.doi.org/10.1007/s13278-014-0187-3
Chakraborty, T., Sikdar, S., Tammana, V., Ganguly, N., & Mukherjee, A. (2013). Computer science fields as ground-truth communities: Their impact, rise and fall. In *IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)* (pp. 426–433).
Chen, C. (2003). *Mapping scientific frontiers: The quest for knowledge visualization*. Springer. URL: http://books.google.co.in/books?id=vN17wM4aPtQC
Chen, Y., Börner, K., & Fang, S. (2013). Evolving collaboration networks in scientometrics in 1978–2010: A micro–macro analysis. *Scientometrics*, *95*(3), 1051–1070.
Ebeling, W., & Feistel, R. (1986). *Physik der selbstorganisation und evolution*. Akademie-Verlag. URL: http://books.google.co.in/books?id=OQj0mgEACAAJ
Evered, R., & Louis, M. R. (1981). Alternative perspectives in the organizational sciences: "Inquiry from the inside" and "inquiry from the outside". *Academy of Management Review*, *6*, 385–395.
Garfield, E. (1977). *Essays of an information scientist*, *Vols. 1–15*.
Garfield, E. (2004). Historiographic mapping of knowledge domains literature. *Journal of Information Science*, *30*(2), 119–145.
Garfield, E., Pudovkin, A. I., & Istomin, V. S. (2003). Why do we need algorithmic historiography? *Journal of the Association for Information Science and Technology*, *54*(5), 400–412.
Gilbert, G. (1977). Competition, differentiation and careers in science. *Social Science Information*, *16*(1), 103–123.
Hellsten, I., Lambiotte, R., Scharnhorst, A., & Ausloos, M. (2007). Self-citations co-authorships and keywords: A new approach to scientists' field mobility? *Scientometrics*, *72*(3), 469–486.
Le Pair, C. (1980). Switching between academic disciplines in universities in The Netherlands. *Scientometrics*, *2*(3), 177–191. http://dx.doi.org/10.1007/BF02016696
Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(2), 404–409.
Nicholson, N. (1984). A Theory of Work Role Transitions. *Administrative Science Quarterly*, *29*(2), 172–191.
Passi, B., & Mishra, S. (2004). Selecting research areas and research design approaches in distance education: Process issues. *International Review of Research in Open and Distance Learning*, *5*(3), 329–340.
Pettigrew, A. M. (1990). Longitudinal field research on change: Theory and practice. *Organization Science Special Issue: Longitudinal Field Research Methods for Studying Processes of Organizational Change*, *1*(3), 267–292.
Pierce, S. J. (1999). Boundary crossing in research literatures as a means of interdisciplinary information transfer. *Journal of the Association for Information Science and Technology*, *50*(3), 271–279.
Redner, S. (2005). Citation Statistics from 110 Years of Physical Review. *Physics Today*, *58*(6), 49–54.
Rinia, E., van Leeuwen, T., Bruins, E., van Vuren, H., & van Raan, A. (2002). Measuring knowledge transfer between fields of science. *Scientometrics*, *54*(3), 347–362.
Scharnhorst, A. (2001). Mobility and the growth of science. *Integrative Systems Approaches to Natural and Social Dynamics*, 505–515.
Schein, A. I., Popescul, A., Ungar, L. H., & Pennock, D. M. (2002). Methods and metrics for cold-start recommendations. In *Proceedings of the 25th ACM SIGIR* ACM, New York, NY, USA, (pp. 253–260).
Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.
Urata, H. (1990). Information flows among academic disciplines in japan. *Scientometrics*, *18*(3–4), 309–319.
Urban, D. (1982). Mobility and the growth of science. *Social Studies of Science*, *12*(3), 409–433.
van Houten, J., van Vuren, H., Le Pairs, C., & Dijkhuis, G. (1983). Migration of physicists to other academic disciplines: Situation in The Netherlands. *Scientometrics*, *5*(4), 257–267. http://dx.doi.org/10.1007/BF02019741
Vlachy, J. (1981). Mobility in physics a bibliography of occupational geographic and field mobility of physicists. *Czechoslovak Journal of Physics B*, *31*(6), 669–674.
Wagner-Döbler, R., & Berg, J. (1993). *Mathematische Logik von 1847 bis zur Gegenwart: Eine bibliometrische Untersuchung*. Foundations of communication and cognition, de Gruyter. URL: http://books.google.com.au/books?id=uyOBuUcvAw8C
Zhou, D., Ji, X., Zha, H., & Giles, C. L. (2006). Topic evolution and social interactions: How authors effect research. In *2006 ACM CIKM international conference on information and knowledge management* (pp. 248–257).