On the Discovery of Success Trajectories of Authors

Dinesh Pradhan, Saswata Pandit,Subrata Nandi
Dept. of CSE, NIT Durgapur, India
{dineshkrp, saswata.pandit94, subrata.nandi}@gmail.com

Tanmoy Chakraborty
Dept. of CSE, IIT Kharagpur, India
its_tanmoy@cse.iitkgp.ernet.in

ABSTRACT

Scientific career of individuals undergo myriad vicissitudes, resulting different patterns of success trajectories. Understanding the qualitative patterns of research endeavor of individuals in terms of productions and impact is important in order to quantify such trajectories. Here, we examine the career profile of authors in computer science and physics domains and discover at least six different success trajectories in longitudinal scale. Initial observations of individual trajectories lead to characterizing the authors in each category. We further leverage this trajectory information to build a two-stage stratification model in order to predicting future success of an author at the early stage of her career. We anticipate that this discovery would be useful to model the career profiling of authors.

1. INTRODUCTION

An author's career trajectory is governed by a plenty of decisions and unforeseen events, that can significantly alter her career. As a result, the career trajectory is subjected to an author's past accomplishments and can be of different shapes. A career trajectory can be defined with respect to different objectives, such as research publications, funding, teaching ability etc. However, most important criterion accepted universally is the acceptance of an author's scientific publications to the research community in terms of *citations*.

Although there has been a constant effort devoted to understand the impact of journals and scientific articles, little has been explored at the scale of individual authors [3, 4]. Most of the author-centric indices, such as h-index, g-index consider either growth or saturation of research profiles, and therefore ignore the decline of success. In short, these indices are prone to capture different patterns of success trajectories of authors.

Here, we explore two massive bibliographic datasets constituting papers related to computer science and physics domains, and analyze the success trajectory of authors in terms of the (normalized) of citations in longitudinal scale (over the years). Interestingly, we discover at least six distinct categories of success trajectories, which to the best of our knowledge is revealed for the first time in the granularity of individual authors. The massive datasets further allow us to characterize the authors in individual trajectories. Finally, we leverage this trajectory information to build a system which predicts the future success of an author at the early stage of her career. Experimental results show a significant improvement over a baseline system. We believe that this trajectory categorization would help us in understanding and modeling the career profile of individuals.

2. EXPERIMENTAL SETUP AND RESULTS

In this section, we briefly describe the datasets and experimental framework, followed by the results in details.

Datasets. We crawled following two massive bibliographic datasets: (i) `CS`, consisting of 2,473,171 papers of computer science domain published between 1960-2010 and indexed by Microsoft Academic Search[1], (ii) `Physics`[2], containing 425,399 articles published in Physical Review journals from 1975 to 2008. In addition to the citation information, each paper is associated with a set of matedata information such as the title, the authors, the year of publication, the publication venue etc. After preprocessing, we consider 1,549,317 and 295,311 authors respectively from `CS` and `Physics` datasets whose publication informations are available for at least 10 years[3].

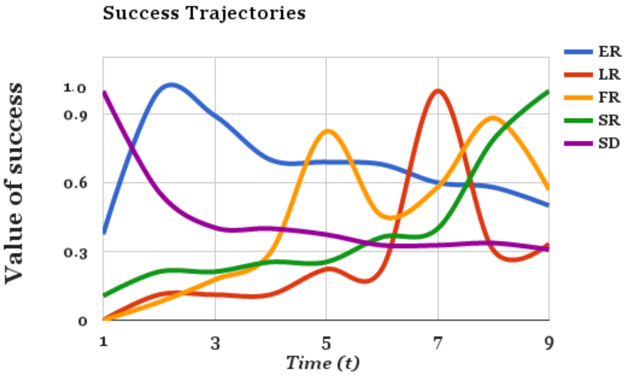
Heuristics for trajectory discovery. To begin with, we take the selected sets of authors with the information of their publications and citations over time. An initial three-year buffer window is provided to each author with the assumption that unlike for a paper, a few-years time frame is always required for an author to set up her career. Therefore, we consider the fourth year of the career timeline of an author as the beginning of the logical year of her career profile. Then we quantify the *success* of an author a at year t (termed as S_a^t) by the ratio between the number of citations received by a till t (termed as C_a^t) and the number of papers published by A till t (terms as P_a^t). For each author, thus we obtain the longitudinal information of success. This is followed by a series of data processing: firstly, to smoothen the longitudinal data points corresponding to an author, we use five-years moving average filtering; secondly, we scale the data points by normalizing them with the maximum value present in the time series; finally, we use following two heuristics to detect the peaks in the trajectory: (i) the height of a peak should be at least 75% of the maximum peak-height, and (ii) two consecutive peaks should be separated by more than 2 years; otherwise they are treated as a single peak.

Categories of success trajectories. Remarkably, we observe six different pattens of success trajectories of authors based on the count and the position of peaks present in a trajectory (see Figure 1(left)): (i) `Early-risers` (`ER`): authors whose career peaks once within initial 5 years (but not in the first year) followed by a decay; (ii) `Late-risers` (`LR`): authors whose career peaks once after least 5 years since she has published first paper, followed by a decay; (iii) `Frequent-risers` (`FR`): authors whose career peaks multiple times over the years; (iv) `Steady-risers` (`SR`): authors hav-

[1] academic.research.microsoft.com

[2] http://journals.aps.org/datasets

[3] The datasets are publicly available at cnerg.org.



	ER	LR	FR	SR	SD
% of authors	9.96; 7.85	23.15; 18.36	6.51; 8.78	58.97; 62.35	1.38; 2.65
Avg. h-index	4.69; 3.87	5.15; 4.49	6.06; 4.21	4.10; 5.36	2.93; 3.01
Avg % of conference papers	68.39; NA	43.22; NA	51.98; NA	39.08; NA	76.09; NA
Avg % of self-citations	31.01; 34.58	30.30; 28.64	25.71; 25.12	26.14; 25.65	32.67; 36.54

Figure 1: (Color online)(Left) Sample examples (taken from CS-dataset) of success trajectories; (Right) Characteristics of different trajectory categories for CS (black) and Physics (red) datasets.

ing a monotonic increasing growth of success over the years; Steady-droppers (SD): authors whose career peaks in the first year followed by a monotonic decrease over the years; and Others (OT): Apart from the above types, there exist a large volume of authors who on an average publish less than one paper per year and receive less than one citation per year. Due to the lack of proper statistical significance, we categorize them into a separate category.

Characterizing individual success trajectories. Next, we attempt to understand the authors of individual categories in more details (see Figure 1(right)). First, we calculate the percentage of authors in each category and observe that steady-risers are the major class of authors, followed by late-risers; whereas steady-droppers are rare. Second, we measure the average impact of authors in each category and notice that while in CS domain frequent-risers are the most profound authors in terms of h-index, in Physics steady-risers dominate others, the reason being that physicists prefer publishing papers in Journals (see later). However, as expected steady-droppers seem to be least prominent. Third, for CS-dataset we notice that early-risers and steady-droppers tend to publish papers mostly in conferences, while steady and frequent risers prefer publications in journals. Forth and most interestingly, we observe that early-risers and steady-droppers are mostly affected by self-citations[4]. Had the self-citation been discarded from the analysis, 53% early-risers and 63% steady-droppers have migrated to OT category.

A deeper inspection of the decay in the success trajectories of early-risers, late-risers and steady-droppers for CS (Physics) dataset revels that around 82% (79%) cases the value of S_a drops due to the enormous volume of individual publications overshadowing the effect of incoming citations. Further, we observe that during the time of decay, 46% (37%) of authors are unable to retain their most prominent collaborators (in terms of h-index), indicating that the effect of collaboration might be a reason for this decay. Interestingly, for both the datasets (CS; Physics) the rate of publications of steady-risers (2.06; 1.27) is least among others (on average 4.32; 3.29), which indicates that formers tend to emphasize on *quality*, rather than quantity. Frequent-risers seem to prefer producing quality papers between a gap of 3-5 years, which leads to sudden peaks in their success trajectories.

Leveraging trajectory information for predicting success. One crucial question one can ask in this context – how can the trajectory information be leveraged for building real applications? Here we consider the task of predicting success (defined above) of an author in future at the early stage ($\triangle t$ years after her first appearance) of her career. We consider the same set of author-centric features (along with the first two years citations and publication counts of authors) and framework discussed by Chakraborty et al. [1] where Support Vector Regression (SVR) [1] turned out to be the best learning framework. We use 10-fold cross validation technique. The baseline is designed by training SVR on the *entire* training samples and predicting the success of a query author by fitting the regression equation. On the other hand, we propose a *two-stage stratification learning* model [2]: in the first stage, a query author is mapped into one of the trajectories/strata[5] using a Support Vector Machine approach that learns from the same set of features used in the baseline; in the second stage, *only* those authors corresponding to the category of the query authors are used to train the SVR module to predict the future citation count of the query author. In this way, we remove the effect of random noise while training the regression model. Experimental results show that our model achieves 15.09% (16.3%) and 14.7% (10.5%) more accuracy in term of mean squared error and Pearson correlation coefficient respectively for CS (Physics) dataset[6].

3. CONCLUSION

We analyzed two massive publication datasets and explored six distinct categories of success trajectories of authors and characterized them. As a secondary objective, we further showed how one can leverage this category information in predicting future success of authors. As an immediate step, we shall extend this work to propose a mathematical formulation of modeling success trajectories of individuals.

[4] A citation is marked as self-citation if there is at least one author common in both citing and cited papers.

[5] Note that we know the category information of the authors present in the training set a priori, and therefore the training points are divided into six categories.

[6] The results are averaged over $\triangle\ t$, ranging from 3 to 6.

4. REFERENCES

[1] T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly, and A. Mukherjee. Towards a stratified learning approach to predict future citation counts. In *JCDL*, pages 351–360. IEEE Press, 2014.

[2] G. Haro, G. Randall, and G. Sapiro. Stratification Learning: Detecting Mixed Density and Dimensionality in High Dimensional Point Clouds. In *NIPS*, pages 553–560. MIT Press, Cambridge, MA, 2007.

[3] A. M. Petersen, H. E. Stanley, and S. Succi. Statistical regularities in the rank-citation profile of scientists. *Sci. Rep.*, 1:181, 2011.

[4] A. M. Petersen, F. Wang, and H. E. Stanley. Methods for measuring the citations and productivity of scientists across time and discipline. *PRE*, 81:036114, Mar 2010.