On the Discovery of Success Trajectories of Authors

Saswata Pandit

On the Discovery of Success Trajectories of Authors

A thesis in partial fulfilment for the degree of

Bachelor of Technology

in

Computer Science and Engineering

Submitted By

Saswata Pandit

13/CS/34

under the supervision of

Dr. Subrata Nandi

Associate Professor
NIT Durgapur

Department of Computer Science and Engineering

National Institute of Technology, Durgapur
May 2017

This work is dedicated to my guide and my parents

Declaration

I certify that

1. the work contained in this thesis is original and has been done by me under the guidance of my supervisor.

2. the work has not been submitted to any other Institute for any degree or diploma.

3. I have followed the guidelines provided by the Institute in preparing the thesis.

4. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

5. whenever I have used materials (data, theoretical analysis, figures and text) from other sources, I have given due credit to them by citing them it the text of the thesis and giving their details in the references.

Saswata Pandit

Certificate Of Recommendation

This is to certify that the thesis entitled **"On the Discovery of Success Trajectories of Authors"**, submitted by **Saswata Pandit** for the partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science and Engineering**, is a bonafide research work under the guidance of **Dr. Subrata Nandi**. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma. In our opinion, this thesis is of the standard required for the partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology**.

Internal Guide
Dr. Subrata Nandi
Assoc. Professor, Deptt. of CSE
National Institute of Technology
Durgapur-713209, INDIA

(Counter Signed by)
Dr. Goutam Sanyal
Professor & Head, Deptt. of CSE
National Institute of Technology
Durgapur-713209, INDIA

Certificate of Approval

The foregoing thesis is hereby approved as a creditable study of technological subject carried out and presented in a manner satisfactory to warrant its acceptance as a prerequisite with degree for which it has been submitted. It is to be understood that by this approval, the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the thesis only for the purpose for which it has been submitted.

BOARD OF THESIS EXAMINERS

1. ______________________________ 2. ______________________________

3. ______________________________ 4. ______________________________

5. ______________________________ 6. ______________________________

Acknowledgements

First and foremost, I have to thank my research supervisor **Dr. Subrata Nandi**. Without his assistance and dedicated involvement in every step throughout the process, this thesis would never have been accomplished. His guidance helped me in all the time of research and writing of this thesis. I would like to thank him very much for the support, patience, motivation and understanding over the duration of the project.

I wish to acknowledge **Dr. Goutam Sanyal**, HOD of CSE, for providing supporting environment for my work. I would also like to show gratitude and respect to other faculty members of our department for their encouragement.

I would like to express deep gratitude to Phd Scholar Dinesh Pradhan for his encouragement and gracious support throughout the course of my work.

Last but not the least, I would like to thank my parents for their unconditional support and love.

Saswata Pandit

On the Discovery of Success Trajectories of Authors

Abstract

Understanding the qualitative patterns of research endeavor of scientific authors in terms of publication count and their impact (citation) is important to quantify success trajectories. In this thesis, we discover at least six different success categories (trajectories) in terms of normalised citation count. We examined the career profile of authors from the domains of computer science and physics. The observations from individual trajectories of authors helped us to categorize the authors. Further, the trajectory information can be used to predict the future success of an author at an early stage of her career.

Contents

List of Figures

Chapter 1

Introduction

> *The future is like a corridor into which we can see only by the light coming from behind.* — Edward Weyer Jr.

The career trajectory of an individual author is governed by many decisions and events and that can have significant impact on her career. Thus the trajectory depends on author's past accomplishments and can be defined with respect to different objectives such as teaching ability, funding, publications, expertise on particular domain etc. First, we should come up with a criterion that is universally acceptable for characterising the author's trajectory. Citation count of an author's scientific publications, is the most important criterion accepted universally. Second, devising an efficient and effective way of computation to report the $'success\ trajectory'$ of an author and to assign him to one of the six categories. In this work of ours, we believe that this trajectory categorization would help us to understand and model the career profile of authors and to predict future success of an author at an early stage of her career.

1.1 Issues & Challenges

Most important source of information that bears maximum importance in this respect are the citations (references) that a paper receives from other publications. One naive approach for finding research contribution of an au-

thor might be through systematically combining the performance scores of the research publications they have authored. Currently popular solutions like H-index [1] and its variants [2] mainly use this approach. However there are a number of issues in such approach that requires addressing:
1. Majority of the scientific research articles are written by multiple authors nowadays [3], which makes it difficult to find individual contribution under such a shared author regime. Significant modification to the above naive strategy for combining credits from individual publications needs to be devised. Most author-centric indices such as h-index, g-index captures either growth or saturation of research profile but fails to capture decline of success.
2. Receiving citations from peers is a dynamic and time-consuming process. The papers that is published recently would often receive less citations than those published long back, even though the former may have more research contribute and may eventually have the prospect of gaining higher popularity (citations) than the later in future.
3. Democratic view to citations may not be a decent way for computing the publication and author contribution. Citation from a Nobel Laureate or Fields Medalist may carry more information than a citation from a research beginner.
4. Citations may not be the only source of information regarding the computation of author impact in the domain. Social bonding of authors through multi-facet relationships like co-authorship, co-publishing (publishing to the same venues), co-chairing (in conferences), co-editorship (of journals), co-affiliation (in academic/research institutes), co-citation (of papers), and so on are also bears significant information about an author.

1.2 Our Contribution

In our study, we first filter out the set of authors whose publication information are available for **at least 10 years**. Then for all these authors the information related to their papers and citations are combined to get the distribution of **normalised citation count** with respect to time(success trajectory). Using 2 heuristics as explained later we categorize the author into one of the six categories according to the nature of the success trajectory. We explored two massive bibliographic datasets, CS and Physics indexed by Microsoft Academic Search.

1.3 Thesis Outline

The thesis is organized as follows.
Chapter 1 gives a brief introduction of the problem domain addressed in this thesis.
Chapter 2 discusses the related work and motivation.
Chapter 3 describes the proposed framework and experimental setup.
Chapter 4 describes the dataset used and construction of the framework.
Chapter 5 discussion of results obtained.
Chapter 6 includes conclusion and future scope.

Chapter 2

Literature Survey

2.1 Categorization of Scientific citation profiles in Computer Science

A general agreement in literature is that citation profiles of publications follows a universal pattern. Increase in number of citations in first two years with a steady peak for one to two years followed by decay. The framework suggested by Chakraborty et al[4] where they analyzed a dataset consisting of **1.5 million** computer science papers maintained by **Microsoft Academic Search** and classified the papers into six categories depending on citation count of articles over the years. The six categories are as follows:

PeakInt : Papers with **single** citation peaks in **first five** years following publication but not first year, followed by exponential decay.
PeakLate : Papers with very few citations in initial **five** years, then a **single** peak after **5** years of publication, followed by decay.
PeakMul : Papers with **multiple** citation peaks throughout its lifetime.
MonIcr : Papers with **increasing** monotonic growth from the beginning year of publication.
MonDcr : Papers with **decreasing** monotonic growth from beginning year of publication.
Oth : Apart from first five types, a large number of papers resides in this

group who receive **less than one citation** a year on average, thus no statistical inference can be made.

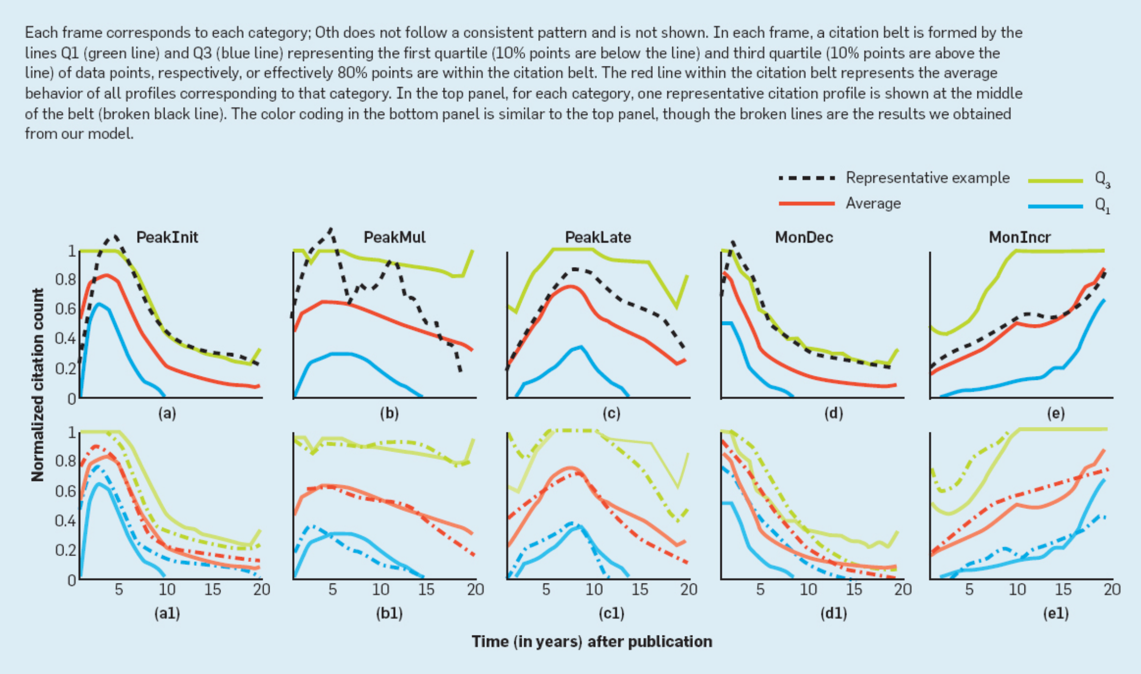

Each frame corresponds to each category; Oth does not follow a consistent pattern and is not shown. In each frame, a citation belt is formed by the lines Q1 (green line) and Q3 (blue line) representing the first quartile (10% points are below the line) and third quartile (10% points are above the line) of data points, respectively, or effectively 80% points are within the citation belt. The red line within the citation belt represents the average behavior of all profiles corresponding to that category. In the top panel, for each category, one representative citation profile is shown at the middle of the belt (broken black line). The color coding in the bottom panel is similar to the top panel, though the broken lines are the results we obtained from our model.

Figure 2.1: (a) PeakInt; (b) PeakMul; (c) PeakLate; (d) MonDcr; (e) MonIcr;

2.2 Motivation

As seen from above that the citation profiles of articles can be categorised depending on the citation counts received over a considered period of time. Further as Chakraborty et al[4] discussed the different models and methodologies to characterize and study each citation profile and to form dynamic model of growth based on some mathematical formulation. The same can also be attempted on authors. Corresponding to authors we have all paper and citation related information. Thus the success trajectory of author with respect to time be defined as the distribution of success (normalised citation count). The hueristics used by Chakraborty et al in [4] can also be used or modified to detect peaks in the success trajectory of authors to see if such distinct categories as in the case of papers, is also possible or not for authors.

This will help us to better judge the quality of research of scientific authors, provide insight into the career timeline of authors. The major reasons that significantly affects an author's career can be explored deeply. Further research into each of the discovered author profiles.

Chapter 3

Proposed Framework

3.1 Experimental Setup Outline

To begin with, we first extract the selected set of authors (whose publication information are available for 10 years) along with their publications and citations. Unlike a paper, an author requires a few-years time frame to set up her career. Thus an initial three year buffer window is provided to each author and therefore we consider the fourth year of the career timeline of the author as beginning of the logical year of her career profile.

We quantify the success of an author **a** at time **t** as S_a^t. It is the ratio between the number of citations received till time time **t** by author **a** (termed as C_a^t and the number of papers published till time **t** by author **a** (termed as P_a^t).

Now that we have obtained the longitudinal information of success with respect to time, multiple data processing steps are carried out.

1. The longitudinal data points are smoothen out by using five-years moving average filtering.
2. The data points are scaled by normalising them with the maximum value present in the time series.

The reason behind the above two data processing steps is to ensure that ***peaks*** (highest point/data value) can easily be detected in the success trajectory plot of an author using simple heuristics.

3.2 Heuristics for trajectory discovery

The following two heuristics are used to detect peaks in the trajectory :-

1. The height of a peak should be at least **seventy-five** percent of the maximum peak-height.
2. Two consecutive peaks should be separated by **more than 2** years, otherwise treated as single peak.

Chapter 4

Dataset and Peak Detection in Success Trajectory

4.1 Dataset Collection and Filtering

In our present study, we are using two massive bibliographic datasets **CS** and **Physics** developed and maintained by Microsoft Academic Search[1] and is publicly available[2]. The original dataset has 126,909,021 papers and 528,682,289 citation relationships. Along with citation relationships the papers are also associated with related information such as title, authors, year of publication, publication venue, journal/conference etc. The overall size of the dataset is around 100 **GB** and is separated across multiple text files. The paper related information are stored in these text files where records of each paper are row separated.

There are some inconsistencies in the dataset in terms of missing data fields, however all such records have been filtered out. Only those authors are selected from the datasets having at least ten years publication information from **CS** and **Physics** domain.

[1] academic.research.microsoft.com

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

4.2 Dataset Characteristics

After preprocessing and filtering, for each domain the dataset characteristics are as follow:

CS:
Number of Publications : 2,473,171
Publication Period : 1960-2010
Number of Authors having at least 10 years publication history : 1,549,317

Physics:
Number of Publications : 425,399
Publication Period : 1975-2008
Number of Authors having at least 10 years publication history : 295,311

4.3 Peak Detection

In brief the following steps are carried out for peak Detection :

1. For every author from both CS and Physics domain (having at least 10 years publication history) their success S_a^t (at time **t** for author **a**) is computed over a ten year period.

2. The data points are then smoothen, normalised and plotted with respect to each year of logical career profile.

3. Using peak detection package **peakutils** in python, the count and position of peaks in the trajectory are computed based on the heuristics[3] and finally classified.

[3]* A peak is only considered if its seventy-five percent of maximum peak-height.
* Two consecutive peaks should be separated by more than 2 years, otherwise treated as same.

Chapter 5

Result

5.1 6 variants of Success Trajectory

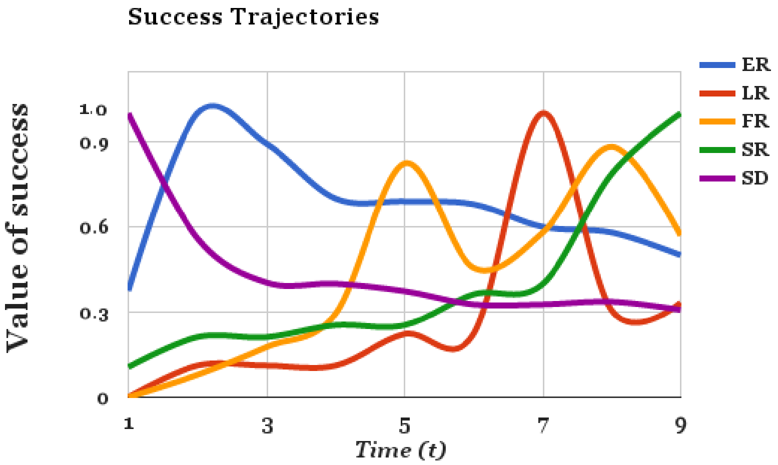


Figure 5.1: (Color online) Sample examples (taken from CS dataset) of the different success trajectories.

From the above figure, we remarkably observe six different patterns of success trajectories based on count and position of peaks in the trajectory plot detected using **peakutils**.

The categories are as follows :

1. **Early-Risers(ER)** : Authors whose career peaks **once** within **initial 5** years but not in **first** year, followed by a decay.

2. **Late-Risers(LR)** : Authors whose career peaks **once** after at **least 5** years since first publication, followed by a decay.

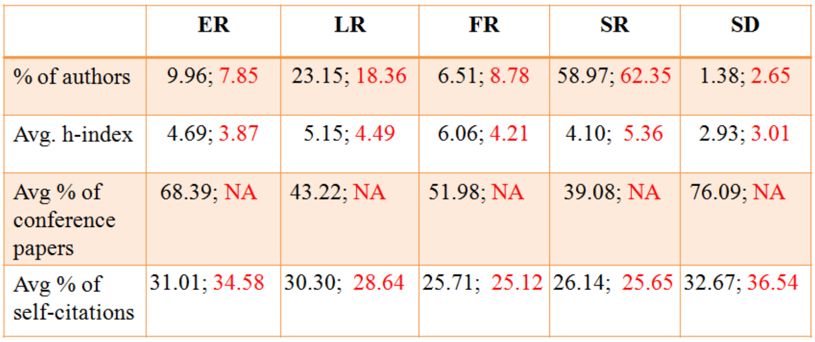
3. **Freqeunt-Risers(FR)** : Authors whose career peaks **multiple** times over the years.

4. **Steady-Risers(SR)** : Authors having a **monotonic increasing growth** over the years.

5. **Steady-Droppers(SD)** : Authors whose career peaks in the first year followed by a **monotonic decreasing growth** over the years.

6. **Others(OT)** : There exists a large volume of authors who on an **average publish less than one paper per year and receive less than once citation per year**. Due to lack of statistical significance, they are classified into this separate category.

5.2 Characteristics of different Success Trajectory



	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 5.2: Characteristics of different success trajectory categories for CS (black) and Physics (red) datasets.

Steady-Risers(SR) : It is the major class of authors. Authors of this class tends to publish mostly in journals. Further the rate of publication of such authors (2.06(**CS**), 1.27(**Physics**)) is least among others (on average 4.30(**CS**), 3.29(**Physics**)), indicating such authors focus on quality rather than quantity, major driving force behind their success.

Frequent-Risers(FR) : It is the class of authors who enjoys multiple successes. In **CS** dataset, authors of this class are having highest average h-index. **Frequent Risers** seem to produce quality papers in a gap of 3-5 years and prefers publications in journals.

Late-Risers(LR : Second most populated category after **Steady-Risers**. Usually enjoys success after initial **5** years. In both datasets, such authors have very high percentage of **self-citations**[1].

Early-Risers(ER : Such authors enjoy success within initial **5** years but then it decays. Mostly tends to publish in conferences, which contributes to the success decay as they are not high standard publications.

Steady-Droppers(SD) : As expected, the least prominent class. If **self-citations** are removed then, more than half of the authors from **SD** and **ER**

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

migrate to **OT** category.

5.2.1 Major Reasons for Success decay of LR, ER and SD

1. Enormous volume of individual publications overshadowing incoming citations.
2. Authors not being able to retain working relationship with their most prominent collaborators (in terms of h-index), leading to success decay.

5.3 Leveraging trajectory information for predicting future success

The crucial question which comes into play now is how can we leverage the trajectory information for predicting future success of author. Here we want to predict success at an early stage of her career, say a few years after her first publication. For this, we consider the author-centric features and framework discussed by Chakraborty et al[5] where Support Vector Regression turned out to be best learning framework, along with a two stage stratification learning model[6].

Chapter 6

Conclusion and Future Work

Conclusion

In the present work, we explored and analyzed two massive bibliographic datasets and discovered six different success trajectories and further characterized the authors in each category. As a secondary objective we showed how the trajectory information can be leveraged to predict future success of authors.

Future scope

As an future objective, we shall try to further deeply study the research history of authors in each category and try to propose a mathematical formulation of modelling success trajectories. Extending the author time line beyond 10 years to 15, 20 years. Detect whether an author migrates from one kind of success trajectory to another, the major reasons behind it.

Bibliography

[1] Hirsch JE. An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(46):16569–16572. doi:10.1073/pnas.0507655102.

[2] Alonso S, Cabrerizo FJ, Herrera-Viedma E, Herrera F. h-Index: A review focused in its variants, computation and standardization for different scientific fields. Journal of Informetrics. 2009;3(4):273 – 289. doi:http://dx.doi.org/10.1016/j.joi.2009.04.001.

[3] Barabási AL, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T. Evolution of the social network of scientific collaborations. Physica A: Statistical mechanics and its applications. 2002;311(3):590–614. doi: https://arxiv.org/abs/cond-mat/0104162

[4] T. Chakraborty, S. Kumar, P. Goyal, N. Ganguly and A. Mukherjee. On the categorization of scientific citation profiles in computer science. Communications of the ACM, Vol-58 No-9, Pages 82-90 doi : https://cacm.acm.org/magazines/2015/9/191187-on-the-categorization-of-scientific-citation-profiles-in-computer-science/fulltext

[5] 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. doi: http://cse.iitkgp.ac.in/ tanmoyc/Papers/JCDL$_2$014.*pdf*.

[6] 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. doi: https://papers.nips.cc/paper/3015-stratification-learning-detecting-mixed-density-and-dimensionality-in-high-dimensional-point-clouds.pdf

Author's Publications

[1]. Dinesh Pradhan, Tanmoy Chakraborty, Saswata Pandit and Subrata Nandi, "On the Discovery of Success Trajectories of Authors", accepted in 25th International World Wide Web Conference, April 2016, Montreal, Canada. doi: http://dl.acm.org/citation.cfm?id=2889375