

The Best Nurturers in Computer Science Research

Bharath Kumar M.

mbk@csa.iisc.ernet.in

Dept. of Computer Science
and Automation,
Indian Institute of Science

Y. N. Srikant

srikant@csa.iisc.ernet.in

Dept. of Computer Science
and Automation,
Indian Institute of Science

Abstract

The paper presents a heuristic for mining nurturers in temporally organized collaboration networks: people who facilitate the growth and success of the young ones. Specifically, this heuristic is applied to the computer science bibliographic data to find the best nurturers in computer science research. The measure of success is parameterized, and the paper demonstrates experiments and results with publication count as a success metric. Rather than just the nurturer's success, the heuristic captures the influence he has had in the independent success of the relatively young in the network. These results can hence be a useful resource to graduate students and post-doctoral candidates. Interestingly, there is a recognizable deviation between the rankings of the most successful researchers and the best nurturers, which although is obvious from a social perspective has not been statistically demonstrated.

Keywords: Social Network Analysis, Bibliometrics, Temporal Data Mining.

1 Introduction

Consider a student Arjun, who has finished his undergraduate degree in Computer Science, and is seeking a PhD degree followed by a successful career in Computer Science research. How does he choose his research advisor? He has the following options with him:

1. Look up the rankings of various universities [1], and apply to any "reasonably good" professor in any of the top universities.

Does working with any reasonably good professor at a top university ensure that Arjun gets the training to pursue a successful research career?

2. Look up the web sites that present the most successful researchers, based on the number of publications [2] or the citations they have received [3] [4].

Arjun can then do his own analysis and find out how many of these researchers are active at the current date. He wants to ensure he does not work with a professor who's past his prime; or neglect a young and upcoming professor.

But still, does working with a top professor, who's

known for his research, imply Arjun will learn how to do good research and in due course have a successful research career?

3. Get word-of-mouth information on the social aspects of working with a particular advisor.

Arjun can talk to an advisor's past and current students, get their feedback, attribute a certain trust to what each one says, and then decide. How many people will Arjun ask? How much will he trust each individual feedback?

For Arjun, it is more important to seek a professor who will **nurture** him to become a good researcher: one who will teach him how best to do research that ends up in good publications, one who will bootstrap him into a good research network, where he hops onto a successful research career path on his own. Although being with a good researcher or in a top school does help, there is no guarantee of being nurtured. A good researcher may not be a good nurturer, and getting into a top school does not always ensure a good research career.

Arjun would benefit if:

- there is a way to summarize the nurturing ability of a researcher by mining the performance of people he nurtured, and thereby compare one nurturer with another.
- there is a way to find out the best nurturers in a given period of time.
- there is a way to find out researchers who have nurtured people to publish many papers, to obtain many citations for their papers or in a given area of research.

This paper presents a *Nurturer-Finder* heuristic that Arjun can use. When Arjun chooses to work with any of these people, he is assured that he is not just choosing them for their research prowess, but for the positive experiences people like himself had in the past. It may turn out that the nurturers also happen to be successful researchers themselves, as the results show.

2 The Nurturer-Finder's Design Principles

While it may be argued that nurturing may even happen inside the confines of a classroom, through well-written

books, or that a true nurturer may just stay behind the scenes and not even be a co-author; mining among associations in bibliographic databases remains the best context to scalably look for nurturers in research:

- Publishing is the defacto standard for evaluating good research.
- The art of scientific reporting is best taught “hands on”. Senior collaborators typically give direction on the most important aspects of the innovation, provide appropriate feedback on its capabilities and limitations, and contrast the innovation with other progress in the area.
- People who have contributed towards a research project often end up as co-authors in the subsequent publication.
- Bibliographic databases are well documented, and are already used for extensive analysis of the impact of research.

However, all publications may not have a nurturer-nurtured pair; often, publications have “almost equals” as co-authors. Hence, the heuristic must not stray in its analysis, and report any co-author pair as a nurturer-nurtured pair. In contrast, no co-author pair can be neglected, since every collaboration can potentially be a context of nurturing.

The nurturer-finder heuristic is inspired by the concept of *gurudakshina* known from ancient Indian traditions. After finishing his education, a student (*shishya*) pays tribute to his teacher (*guru*) for the knowledge he was bestowed. On the same light, whenever a person achieves some success (through a publication), he attributes a part of that success to his “gurus” proportionate to their nurturing influence on him. The gurus with the highest *gurudakshina* are the best nurturers.

The design principles are elucidated as follows:

1. The effect of nurturing manifests in the **post-associative period**.

Any amount of success a person may have with his nurturer, it is still not indicative whether he has been successfully nurtured. The nurturing is true and complete, when he tastes success “on his own” in the absence of his nurturer. This period is hence termed as post-associative, and is used as the context to decide the extent of the nurturing.

2. The more **self-made** a person is, the less he attributes his success to his past associates.

People who have seen success on their own, without associating with too many people, especially early in their career, can be termed as *self-made*. They are the self-motivated people, who probably were not nurtured at all by someone else. It is fair that these people attribute less of their success to their past associates.

3. The success achieved by a person at any time is considered to be influenced by all his past associates. However it is **tributed** to only those who do not have a direct pay-off in the current collaboration.

While contributing towards a publication, an author may be acting upon the influence he’s had from many of his past and current associates. However, all the current associates (the co-authors) in the publication still have their own pay-offs from it. So, the tribute for one’s success is only given away to past associates who have helped influence him to be successful in a current venture without a motive of their own.

4. The tribute is appropriated among the past associates in proportion to their estimated **nurturing influence** on the person.

Nurturing happens most when a person is still young in his career - and the people who associated with him earlier are more important (in terms of a nurturing influence) than the ones he associates with later in his career. This can be termed as the *strength of early association*. As an aside - while the strength of early association of a person with his nurturer will be high, the reverse need not hold, since the nurturer is expected to be already relatively mature in his career.

A person need not have been nurtured equally by all people he had good early associations with. The ones who nurtured him more are most likely those who were termed to have a good *nurturing ability* by other people as well.

Thus, an associate’s nurturing influence on a person is proportional to the strength of early association with this person and the associate’s own nurturing ability. The tribute can then be appropriated to each past associate in accordance to the proportion of their nurturing influence.

The above principles guide the design of the Nurturer-Finder heuristic, which works based on the following outline. Publications are processed in temporal sequence, at some granularity, either grouped by years or by months.

1. As every person publishes, his strength of early association with his associates, and their nurturing influence on him are tracked.

2. Every time he achieves a certain success from a publication, it is tributed to his past associates for influencing him in his “formative” years, in accordance to their nurturing influence. The more self-made a person is, the less is his tribute.

3. Every person collects the tribute he gets from others.

4. The person with the highest tribute is the best nurturer. People can also be sorted on the tributes they have, to arrange them in non-increasing order of their nurturing abilities.

3 The Formulation of the Nurturer-Finder Heuristic

The heuristic is abstractly formulated, allowing for reuse in domains outside of bibliographic databases.

A publication is an instance of a collaboration, and happens at a certain discrete instant in time. The bibliographic database is termed as the set of *collaborations*.

A collaboration c has the following properties,

$associates_c$, the set of people involved in the collaboration c .

$time_c$, the time at which the collaboration happened.

sig_c , the quantifier representing the significance of the collaboration, which could be equal to 1, the impact factor of the conference or journal where it was published, or the number of citations the publication has received.

Each associate p in a collaboration gets a certain significance measure to himself: his share of success. In the model used here, the success is equally shared among the associates.

$$sig_c^p = \begin{cases} \frac{sig_c}{|associates_c|} & \text{if } p \in associates_c \\ 0 & \text{if } p \notin associates_c \end{cases}$$

Other models, for instance, can give importance to the position of the author's name in the list, while deciding the significance of each associate.

The set of all collaborations that have happened till time t , is given by,

$$collabs^t = \{c \in collabs \mid time_c < t\}$$

The set of all people involved in all collaborations till time t is represented by,

$$people^t = \bigcup \{associates_c : c \in collabs^t\}$$

The cumulative significance of each person until time t is represented by,

$$cum-sig_p^t = \sum_{c \in collabs^t} sig_c^p$$

A measure of the degree of association a person q had in the significance a person p achieved during a collaboration c is given by,

$${}^q assoc_p^c = sig_c^p * \frac{sig_c^q}{sig_c}$$

The $\frac{sig_c^q}{sig_c}$ factor is indicative of q 's involvement in c . Higher q 's involvement, higher is his association with p 's significance.

The **early association** q had with p , until time t is representative of the successful collaboration p had with q early in his career.

$${}^q early-assoc_p^t = \sum_{c \in collabs^t} \left(\frac{{}^q assoc_p^c}{cum-sig_p^{time_c}} \right)$$

A measure of how self-made a person is, is also useful - to determine his independence on his associates for his success. This measure also considers the *earliness* of his *self-establishment*. The intuition being that, a person who gets independent success later in his career, but after collaborating with people early on, is not as self-made as a person who

was independent right from the start. It is likely that a self-made person was not nurtured by too many people at all, and hence he must attribute less of his success to his 'mentors'.

$$self-estab_p^t = {}^p early-assoc_p^t$$

The nurturing influence a person q has had on p , (where $p \neq q$) until time t is given by

$${}^q ni_p^t = \sum_{c \in collabs^t} \left(\frac{{}^q assoc_p^c * (nship_q^{time_c})^\alpha}{cum-sig_p^{time_c}} \right)$$

The term $nship_p^t$, which is detailed later, is indicative of the nurturing ability of a person p until time t .

Most people have a relatively constant research output every year. Their cumulative significance would thus grow linearly. However, the nurtureship of a person, which is made up by collecting tributes from collaborators, can grow faster than the cumulative significance. In that event, when a person p collaborates with another person q who has a large nurtureship a little late in p 's career, he still concedes a large nurturing influence to q . This may sideline the earlier nurturers of p . To manage this effect the parameter α is introduced, which is used to control the dominance of nurtureship over cumulative significance. For smaller values of α , the earliness factor dominates the nurturing influence. Increasing the value of α makes the people with higher nurtureship "richer" at the cost of the others. In the experiments reported in the paper, α was hand-engineered to 0.5, for satisfactory results.

${}^p ni_p^t$ is not defined. A person does not nurture himself.

The tribute given away to past associates everytime a person p achieves a certain significance through a collaboration c , is given by

$$trib_c^p = sig_c^p * \left(1 - \frac{self-estab_p^{time_c}}{cum-sig_p^{time_c}} \right)$$

The tribute a person p gives to an associate q , (where $p \neq q$), because of achieving a certain significance through a collaboration c is given by

$${}^q trib_p^c = \begin{cases} \frac{trib_c^p * {}^q ni_p^{time_c}}{\sum_{r \in (people^{time_c} - p)} {}^r ni_p^{time_c}} & \text{if } q \notin associates_c \\ 0 & \text{if } q \in associates_c \end{cases}$$

The tribute is thus appropriated proportionate to the nurturing influence.

The $nship_p^t$ of a person is the cumulative sum of the tributes collected by p from other associates until time t . The term $nship_p^t$ is used to represent the nurturing ability of a person right after time t , inclusive of the collaborations

that happened in that time instant. This is incrementally calculated.

$$nship_p^{t'} = nship_p^t + \sum_{\substack{c \in collabs; \\ time_c = t}} \sum_{q \in associates_c} ptrib_c^q$$

and $nship_p^0 = 1$

Thus, the best nurturer is one who has the highest $nship_p^{t'}$ where t is the current time.

The total tribute a person p gives to an associate q until time t is represented by,

$${}^q trib_p^t = \sum_{c \in collabs^t} {}^q trib_c^p$$

This is used to present a drill down of the nurtureship of each person, showing the extent of tribute each of their nurtured give them. ${}^p trib_p^t = 1.0$ This accounts for the default value of $nship_p^0$.

4 Some Experiments on the DBLP Database

The Digital Bibliography and Library Project (DBLP) [2] provides digital information on major computer science journals and publications, and indexes more than 520000 articles. Citations are also available for a subset of the articles indexed. The DBL-browser offers an interface to access the compressed database containing the article information. The Nurturer-Finder heuristic was applied on the DBLP in two sets of experiments to find nurturers for publications count, and for citations. For lack of space, this paper only reports a few top results based on publication count. A more detailed account of these experiments, some special handling needed for some outliers and a method to yield nurturers in time slices is discussed in a technical report [6].

The algorithm was implemented using Java, and used the DBL Browser libraries [5] for accessing the publication records. The algorithm is incremental in nature, and parses each publication in the database exactly once. Every time a publication is processed, all past associates of every co-author are processed, to be assigned tributes.

α is chosen as 0.5 in the following experiments. A discussion on the choice of α is considered in the technical report [6].

4.1 Nurturing for Publication Count

To compute nurturers and their nurtured based on publication count, every entry in the DBLP is assigned a significance of 1, and an author's significance for participation is $\frac{1}{|associates|}$. This metric in itself is not semantically very accurate due to the disparity in quality among the journals and conferences indexed by the DBLP, but acts as a good first measure nevertheless.

The table 1 lists the top few nurturers, their nurtureship value and the people who were 'nurtured' by them, and the tribute each of them gave away to the nurturer. These nurtured people are those who co-authored with the nurturers early in their careers, and then went on to be prolific on their own as authors, even in the absence of their nurturers. Only people who gave away tributes greater than or equal to the value 5 are listed. A person may appear as "nurtured" by more than one nurturer, if he gave away reasonably big tributes to all of them.

4.1.1 Interpreting the results

- The heuristic attempts to recognize the social trait of nurturing through statistical analysis, and hence acceptance of the validity of the findings is possible only by common perception of readers conversant with the who's who of the computer science research community.

- While it is questionable whether there exists a strict nurturer-nurtured distinction in the results, if the border is blurred to mean a nurturing influence, which can be mutual too at times, the results become easier to digest.

- The list of nurturers, on its own, has successful researchers. The authors found this phenomenon most interesting because the calculation of nurtureship does not take into account any publication of the nurturer himself, and considers only post-associative success of people who co-authored with them early in their career.

- The results also suggest the ability of these people to sight talent: people who would later end up doing very well on their own. Good nurturers are also good talent sighters.

5 Discussion on related work

Barabasi et. al in [7] show the existence of preferential attachment during addition of new nodes into the collaboration network.

"For a new author, that appears for the first time on a publication, preferential attachment has a simple meaning: it is more likely that the first paper will be co-authored with somebody that already has a large number of co-authors (links) than with somebody less connected. As a result "old" authors with more links will increase their number of co-authors at a higher rate than those with fewer links."

Does this imply that the best nurturers are simply the best collaborators? When Barabasi et. al. consider the addition of new nodes, they do not track the longevity and success achieved by that new node in the collaboration network. While good collaborators may be the context for addition of newer nodes, they need not be contexts where people who perform well in the long term may be added. To answer this, a new set of experiments were conducted to identify the best collaborators. Barabasi's

Nurturer	Val	Nurturer	Val	Nurturer	Val	Nurturer	Val
Nurtured		Nurtured		Nurtured		Nurtured	
1) Jeffrey D. Ullman	144	5) John E. Hopcroft	97	9) Zvi Galil	83	15) Grzegorz Rozenberg	75
Henry F. Korth	8	Jeffrey D. Ullman	24	Moti Yung	10	Dirk Vermeir	7
Yehoshua Sagiv	8	Robert Endre Tarjan	14	David Eppstein	7	Robert Meersman	6
Fereidoon Sadr	7	Richard Cole	12	Kunsoo Park	7	16) Richard J. Lipton	75
Alberto O. Mendelzon	6	Steven Fortune	5	Nimrod Megiddo	6	Dan Boneh	8
Sam Toueg	6	Joachim von zur Gathen	5	Dany Breslauer	5	Lawrence Snyder	7
Ravi Sethi	5	Gordon T. Wilfong	5	10) Christos H. Papadimitriou	81	David P. Dobkin	5
David Maier	5	6) Robert Endre Tarjan	95	Joseph S. B. Mitchell	10	17) John H. Reif	74
Joan Feigenbaum	5	Thomas Lengauer	11	Paris C. Kanellakis	6	Paul G. Spirakis	17
2) Zohar Manna	126	Haim Kaplan	6	John N. Tsitsiklis	5	Sanguthevar Rajasekaran	8
Martn Abadi	23	Jeffery Westbrook	6	Mihalis Yannakakis	5	Philip N. Klein	7
Amir Pnueli	21	Andrew V. Goldberg	6	11) Ronald L. Rivest	80	Sandeep Sen	6
Adi Shamir	15	David R. Cheriton	5	Robert E. Schapire	10	18) Adi Shamir	74
Nachum Dershowitz	11	7) Ugo Montanari	90	Avrim Blum	9	Uriel Feige	16
Shmuel Katz	6	Roberto Gorrieri	7	Benny Chor	5	Amos Fiat	9
Thomas A. Henzinger	6	Andrea Corradini	7	Jon Doyle	5	Eli Biham	8
Jean Vuillemin	5	Francesca Rossi	6	Sally A. Goldman	5	Yossi Matias	5
Luca de Alfaro	5	Vladimiro Sassone	6	12) Kurt Mehlhorn	78	Moshe Tennenholtz	5
Ashok K. Chandra	5	Alberto Martelli	6	Michael Kaufmann	11	19) Jacob A. Abraham	73
3) Albert R. Meyer	113	Pierpaolo Degano	5	Majid Sarrafzadeh	6	Prithviraj Banerjee	18
Joseph Y. Halpern	38	Giorgio Levi	5	Norbert Blum	5	Kaushik Roy	11
John C. Mitchell	11	8) C. V. Ramamoorthy	88	13) John Mylopoulos	77	Abhijit Chatterjee	7
Nancy A. Lynch	7	Benjamin W. Wah	11	James P. Delgrande	10	W. Kent Fuchs	5
David Harel	7	Vijay K. Garg	9	Hector J. Levesque	7	20) Leonidas J. Guibas	71
4) Michael Stonebraker	106	K. Mani Chanday	9	Nick Roussopoulos	6	John Hersberger	9
Marti A. Hearst	8	Jaideep Srivastava	9	Alexander Borgida	5	Jack Snoeyink	6
Michael J. Carey	7	K. H. Kim	8	14) Amir Pnueli	76	Andrew M. Odlyzko	6
Akhil Kumar	7	Shashi Shekhar	7	Dennis Shasha	9	21) Oscar H. Ibarra	69
Timos K. Sellis	6	Wei-Tek Tsai	5	David Harel	5	Tao Jiang	15
Sunita Sarawagi	5	Atul Prakash	5	Doron Peled	5	Louis E. Rosier	6
Joseph M. Hellerstein	5			Oded Maler	5	Shlomo Moran	5
Margo I. Seltzer	5					Hui Wang	5

Table 1: Publication Count: Top nurturers and their nurtured

experiments consider only the degree of a node to qualify the best collaborators. Here, the good collaborators were said to be those who collaborated frequently with other good collaborators. This is similar to page rank computation [8], although the weights were computed iteratively year by year. It also differed from the nurturer-finder in that, there was no consideration for earliness, and post-associative significance.

The rankings for top collaborators showed changes when compared to the top nurturers, although the correlation with top collaborators was better than the correlation with the top authors. This suggests that the trait of nurturing is perhaps in some way related to the trait of collaborating. Looking at this the other way, it could also indicate that young people, the new entrants in the network have a preference for good collaborators. Good collaborators typically have good social networks which come in handy for the new.

Further, the weighted tribute graph can be analysed for transitivity and hence discover ‘nurturing’ neighborhoods. It is useful to mention [9], where Newman evaluates several social network measures on scientific coauthorship networks. The connectedness of a scientist is measured based on his reachability on a weighted collaboration graph.

The above mentioned references and [10] can be classified as means to infer different roles played by people in collaboration networks. The current work on nurturers can also be grouped alongside.

References

- [1] *USNews*, <http://www.usnews.com>
- [2] *DBLP*, <http://www.informatik.uni-trier.de/~ley/db/>
- [3] *ISIHighlyCited*, <http://www.isihighlycited.com>
- [4] *Most cited authors in Computer Science*, <http://citeseer.ist.psu.edu/mostcited.html>
- [5] *DBL Browser*, <http://dbis.uni-trier.de/DBL-Browser/>
- [6] Bharath Kumar M., Y. N. Srikant. *The Best Nurturers in Computer Science Research*. CSA, IISc Technical Report, 2004, <http://archive.csa.iisc.ernet.in/TR/2004/10/>
- [7] A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. *Evolution of the social network of scientific collaboration*. *Physica A*, 311(3–4):590–614, 2002.
- [8] S. Brin, L. Page, R. Motwani, and T. Winograd. *The page rank citation ranking: Bringing orer to the web*. Tech. Rep. 1999-66, Stanford Digital Libraries Working Paper, 1999, <http://dbpubs.stanford.edu:8090/pub/1999-66>.
- [9] M. E. J. Newman. *Who is the best connected scientist? a study of scientific coauthorship networks*. *Physics Review*, E64, 2001
- [10] J. Kleinberg. *Authoritative sources in a hyperlinked environment*. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.