The Best Nurturers in Computer Science Research

Bharath Kumar M. Y. N. Srikant

IISc-CSA-TR-2004-10
http://archive.csa.iisc.ernet.in/TR/2004/10/

Computer Science and Automation
Indian Institute of Science, India

October 2004

The Best Nurturers in Computer Science Research

Bharath Kumar M.* Y. N. Srikant†

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 and citations as success metrics. 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. The heuristic is extended to accurately yield ranked nurturers inside a particular time period. 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 under-graduate 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?

*Author for TR correspondence. mbk@csa.iisc.ernet.in

†srikant@csa.iisc.ernet.in

2. Look up the web sites that present the most successful researchers, based on the number of publications [2], the citations they have received [3] [4], or by their Erdos Number [5].

   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.
  - 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.

The table 1 shows the output of the Nurturer-Finder heuristic; the top 50 authors based on publication count (every publication gets the author a value of $\frac{1}{\text{number of authors}}$), and the top 50 nurturers computed on the Computer Science bibliographic database DBLP [2].

2 The Nurturer-Finder's Design Principles

While it may be argued that nurturing may even happen inside the confines of a classroom, or even through well-written books, mining among associations in bibliographic databases remains the best context to 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

Rank	Top Authors		Top Nurturers: Publication Count	
	Name	Value	Name	Value
1	Bill Hancock	161.00	Jeffrey D. Ullman	144.39
2	Joseph Y. Halpern	143.23	Zohar Manna	126.91
3	Diane Crawford	137.00	Albert R. Meyer	113.88
4	Grzegorz Rozenberg	135.27	Michael Stonebraker	106.20
5	Moshe Y. Vardi	135.00	John E. Hopcroft	97.23
6	Kang G. Shin	131.57	Robert Endre Tarjan	95.72
7	Micha Sharir	131.20	Ugo Montanari	90.14
8	Christos H. Papadimitriou	129.39	C. V. Ramamoorthy	88.30
9	Hermann A. Maurer	125.08	Zvi Galil	83.51
10	Philip S. Yu	117.71	Christos H. Papadimitriou	81.95
11	Ronald R. Yager	116.95	Ronald L. Rivest	80.45
12	Hector Garcia-Molina	114.12	Kurt Mehlhorn	78.20
13	Jeffrey D. Ullman	111.37	John Mylopoulos	77.01
14	Kurt Mehlhorn	110.60	Amir Pnueli	76.27
15	Michael Stonebraker	110.48	Grzegorz Rozenberg	75.86
16	David Eppstein	110.01	Richard J. Lipton	75.00
17	Sudhakar M. Reddy	105.16	John H. Reif	74.42
18	Arto Salomaa	103.77	Adi Shamir	74.26
19	Saharon Shelah	102.67	Jacob A. Abraham	73.51
20	Manfred Broy	101.22	Leonidas J. Guibas	71.76
21	John H. Reif	99.92	Oscar H. Ibarra	69.56
22	Elisa Bertino	98.94	Jan van Leeuwen	69.37
23	Richard T. Snodgrass	98.54	Micha Sharir	69.26
24	Oded Goldreich	98.15	Shimon Even	68.78
25	David B. Lomet	97.34	Gio Wiederhold	68.09
26	Robert Endre Tarjan	96.71	Kang G. Shin	67.42
27	Gerard Salton	96.69	Ashok K. Agrawala	66.63
28	Oscar H. Ibarra	94.98	Edmund M. Clarke	66.28
29	Peter G. Neumann	94.66	Avi Wigderson	66.06
30	Gheorghe Paun	94.32	Franco P. Preparata	66.05
31	Edwin R. Hancock	93.12	Richard C. T. Lee	65.55
32	Christoph Meinel	92.49	Danny Dolev	65.12
33	Bruno Courcelle	92.00	Alberto L. Sangiovanni-Vincentelli	62.39
34	Derick Wood	91.23	Abraham Silberschatz	61.91
35	Hartmut Ehrig	89.03	Catriel Beeri	60.95
36	Ben Shneiderman	88.92	David J. DeWitt	60.78
37	Bernard Chazelle	88.22	David P. Dobkin	60.55
38	Marek Karpinski	87.79	Mike Paterson	60.29
39	Won Kim	87.53	Clement T. Yu	58.54
40	Ingo Wegener	87.07	Derick Wood	57.52
41	Jeffrey Scott Vitter	86.47	Oded Goldreich	56.94
42	Amir Pnueli	86.13	Hermann A. Maurer	56.60
43	Ugo Montanari	86.08	Azriel Rosenfeld	56.59
44	Robert L. Glass	86.07	Sartaj Sahni	55.81
45	Nancy A. Lynch	86.03	Nancy A. Lynch	54.98
46	Azriel Rosenfeld	85.87	Silvio Micali	54.59
47	Sushil Jajodia	84.40	Theo Hrder	54.53
48	Zvi Galil	83.94	Seymour Ginsburg	54.34
49	David Harel	83.91	Stefano Ceri	54.31
50	David Peleg	83.26	John L. Hennessy	52.96

Table 1: **Top 50 authors and nurturers based on publication count**

(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.

  $significance_c$, 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.

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

  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,

$$collaborations^t = \{c \in collaborations | time_c < t\}$$

  Note: The collaborations that happened at the time instant t are not included in this set.

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

$$people^t = \bigcup \left\{ associates_c : c \in collaborations^t \right\}$$

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

$$cumulative\text{-}significance_p^t = \sum_{c \in collaborations^t} significance_c^p$$

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

$$^q association_p^c = significance_c^p * \frac{significance_c^q}{significance_c}$$

  The $\frac{significance_c^q}{significance_c}$ 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.

$$^q early\text{-}association_p^t = \sum_{c \in collaborations^t} \left(\frac{^q association_p^c}{cumulative\text{-}significance_p^{time_c}} \right)$$

- 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'.

$$self\text{-}establishment_p^t = {}^p early\text{-}association_p^t$$

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

$${}^q nurturing\text{-}influence_p^t =$$

$$\sum_{c \in collaborations^t} \left(\frac{{}^a association_p^c * \left(nurtureship_q^{time_c} \right)^\alpha}{cumulative\text{-}significance_p^{time_c}} \right)$$

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

  Since nurtureship is collected based on tributes from other people, it has a tendency to grow faster than the cumulative significance of a person. The cumulative significance typically grows linearly since a person can only put a relatively constant amount of effort every year. The selection of α determines the domination nurtureship has over cumulative significance. For higher values of α, a person with higher nurtureship imparts a bigger nurturing-influence, even if the person is well into his career. 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.

  ${}^p nurturing\text{-}influence_p^t$ 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

$$tribute_c^p = significance_c^p \quad * \quad \left(1 - \frac{self\text{-}establishment_p^{time_c}}{cumulative\text{-}significance_p^{time_c}} \right)$$

- 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

$$^{q}tribute_{c}^{p} =$$

$$\begin{cases} \dfrac{tribute_{c}^{p} * {}^{q}nurturing\text{-}influence_{p}^{time_c}}{\sum\limits_{r \in (people^{time_c} - p)} {}^{r}nurturing\text{-}influence_{p}^{time_c}} & \\ & \text{if } q \notin associates_c \\ 0 & \text{if } q \in associates_c \end{cases}$$

  The tribute is thus appropriated proportionate to the nurturing influence.

- The $nurtureship_{p}^{t}$ of a person is the cumulative sum of the tributes collected by p from other associates until time t. The term $nurtureship_{p}^{t'}$ 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.

$$nurtureship_{p}^{t'} = nurtureship_{p}^{t} + \sum_{\substack{c \in collaborations; \\ time_c = t}} \sum_{q \in associates_c} {}^{p}tribute_{c}^{q}$$

  and

$$nurtureship_{p}^{0} = 1$$

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

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

$$^{q}tribute_{p}^{t} = \sum_{c \in collaborations^{t}} {}^{q}tribute_{c}^{p}$$

  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. $^{p}tribute_{p}^{t} = 1.0$ This accounts for the default value of $nurtureship_{p}^{0}$.

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 with the significance measure for each publication being a constant, and the number of citations it received, respectively. Since the DBLP does not have a comprehensive list of all citations, the results based on citations are not as accurate as the other one, based on publication count.

The algorithm to implement the Nurturer-Finder heuristic used a few optimizations. Nurturing-influence is only tracked between a pair of people who have had an association already. This conserved the space and time needed during calculations. The intermediate values of tributes and nurturing-influences, for every year given by a person to another are stored, to facilitate calculating the nurturers over different time slices. Re-runs for different time slices can then give lists of nurturers without having to mine the whole database again. The algorithm was implemented using Java, and used the DBL Browser libraries [6] 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 later in the paper.

4.1 Nurturing for Publication Count

The top 50 authors and the top 50 nurturers are reported based on publication count as the significance measure. Every entry in the DBLP has a significance of 1, and an author's significance for participation is $\frac{1}{|associates|}$. Thus, people with the highest sum, based on the fraction of their participation in each publication, are reported as the top researchers. 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 still acts as a good measure to compare the results of the best authors with the best nurturers. Table 1 displays the top authors according to their cumulative fractional publication count, and the top nurturers according to the cumulative tributes they have got. Although the semantics of top authors and top nurturers

Rank	Nurturer Nurtured	Value
1	Jeffrey D. Ullman	144
	Henry F. Korth	8
	Yehoshua Sagiv	8
	Fereidoon Sadri	7
	Alberto O. Mendelzon	6
	Sam Toueg	6
	Ravi Sethi	5
	David Maier	5
	Joan Feigenbaum	5
2	Zohar Manna	126
	Martn Abadi	23
	Amir Pnueli	21
	Adi Shamir	15
	Nachum Dershowitz	11
	Shmuel Katz	6
	Thomas A. Henzinger	6
	Jean Vuillemin	5
	Luca de Alfaro	5
	Ashok K. Chandra	5
3	Albert R. Meyer	113
	Joseph Y. Halpern	38
	John C. Mitchell	11
	Nancy A. Lynch	7
	David Harel	7
4	Michael Stonebraker	106
	Marti A. Hearst	8
	Michael J. Carey	7
	Akhil Kumar	7
	Timos K. Sellis	6
	Sunita Sarawagi	5
	Joseph M. Hellerstein	5
	Margo I. Seltzer	5
5	John E. Hopcroft	97
	Jeffrey D. Ullman	24
	Robert Endre Tarjan	14
	Richard Cole	12
	Steven Fortune	5
	Joachim von zur Gathen	5
	Gordon T. Wilfong	5
6	Robert Endre Tarjan	95
	Thomas Lengauer	11
	Haim Kaplan	6
	Jeffery Westbrook	6
	Andrew V. Goldberg	6
	David R. Cheriton	5
7	Ugo Montanari	90
	Roberto Gorrieri	7
	Andrea Corradini	7
	Francesca Rossi	6
	Vladimiro Sassone	6
	Alberto Martelli	6
	Pierpaolo Degano	5
	Giorgio Levi	5

Table 2: **Publication Count: Top nurturers and their nurtured**

are very different, it is noticable that a few people in the top authors do not exist in the top nurturers table.

The tables 2 and 3 show the drill down for a few top nurturers. The drill down lists people who were 'nurtured' by them, and the value of the tributes they gave away to the nurturer. These nurtured people are those who co-authored with the nuturers early in their careers, and then went onto 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.

The figure 1 shows a few typical ways the nurtureship of different people has grown year by year. Assuming that the number of publications a person can yield over a year by and large remains a constant, a closer inspection of the curves reveals the following phases in growth:

1. *Quadratic growth*: A quadratic growth phase implies that during this period, the nurturer is collecting tributes from people he nurtured in the past, and also that he continues to nurture newer people.

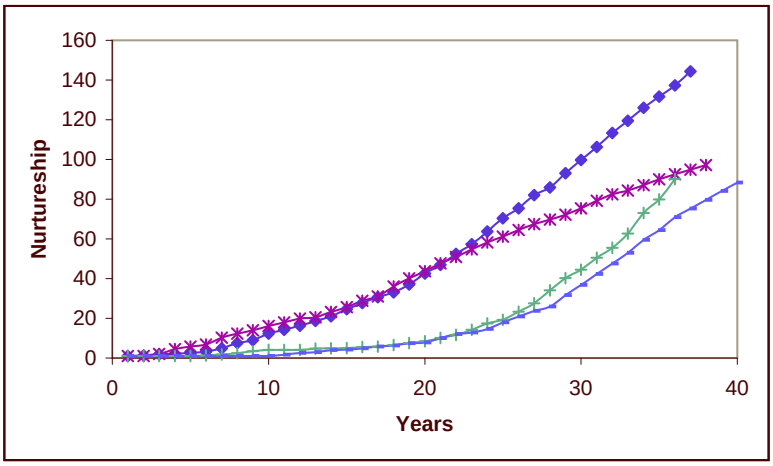

Figure 1: Some typical Nurtureship Growth curves

2. *Linear growth*: In a linear growth phase, the tributes he is receiving have evened out, and he receives a constant tribute each year. This happens if the number of people he has nurtured that are still actively publishing remains a constant, most likely since he has stopped nurturing newer people.

3. *Negligible growth*: The nurturer has stopped nurturing newer people, and the people he had nurtured in the past too have stopped publishing actively.

The figure 1 shows phases of both quadratic and linear growth in a few curves, with the linear growth normally occuring towards the latter stages. For these curves to level out, we may have to wait for a few more years when the second generation of computer science researchers stop being active. Signs of leveling out can be seen in at least one of the curves in the figure. A more formal study of the phases in each curve, a means to automatically identify the phase of nurtureship a person is in, will provide very useful information on how actively a person is still nurturing.

4.2 Nurturing for Citations

DBLP makes citations available for a subset of articles in the database. The database is pre-processed once and the number of times a particular article is cited, into the entire future available, is computed. Results of nurturers obtained based on citations is biased towards the earlier articles, since many of

the recently published articles would not have reached their fullest "citation potential".

Using citations as the significance measure exposed a potential flaw in the heuristic. The number of citations an article receives can be huge, and thus, all its glory may be tributed to just any earlier associate. Some of these associates would get tributes from too few people, but large ones from those few people, which cast some doubts if they were "false positives". They typically cropped up quite often in cases where people's early research works received tremendous citations. This phenomenon was particularly not troublesome in the earlier significance measure, since each publication just had a significance of 1, and to climb up the nurturer charts, a nurturer had to repeat his "nurturing" many times over with different people. However, with the case of citations, to have significantly high positions in the nurturer charts, it was enough to be associated early with just one person who ended up having lots of citations later. This phenomenon can be termed as *tribute dominance (TD)*, where the total tribute obtained is heavily dominated by the tributes given by too few people. A measure of the tribute dominance is computed and used to weed out the false positives.

4.2.1 Nurtureship Buffering based on Tribute Dominance

A person's nurtureship is built based on contributions from several tributes, from different people. Given the nature of the heuristic, its not just the truly nurtured who pay tributes, but almost every associate gives a tribute, albeit in very small amounts. Using the mean across the tributes would be misleading, since a person may have got almost zero tributes from a lot of people, and yet be a good nurturer. Further, the variance will also dominated by the large number of small values in the distribution. A good measure to find the tribute dominance, should be invariant of the number of small values that make up the sum, and yet be able to find out if a particular nurtureship is dominated by too few tributes. A measure based on partial sums is used:

Consider, ${}^{p}tribute_q^t$ the total tribute q has given to p until time t,where ${}^{p}tribute_p^t$ is made equal to 1.0, since the default value for nurtureship is 1.0.

The total tribute, the

$$nurtureship_p^t = \sum_{q \in people^t} {}^{p}tribute_q^t$$

Next, let ${}^{p}T^t$ be an array of size $|people^t|$, sorted in non-increasing order, with elements $\frac{{}^{p}tribute_q^t}{nurtureship_p^t}$ for all $q \in people^t$. T essentially contains the tributes of all people normalized by the total nurtureship, in sorted order.

Next, let ${}^pPS^t$ be an array made of partial sums of ${}^pT^t$ such that,

$$ {}^pPS_i^t = \sum_{j=i}^{|people^t|} {}^pT_i^t \qquad \text{for} \quad i = 1 \ldots |people^t| $$

The tribute dominance,

$${}^pTD^t = \frac{1}{\sum_{i=1}^{|people^t|} {}^pPS_i^t}$$

For people, with a high tribute dominance, i.e. just a few people contributing to most of the nurtureship, ${}^pTD^t$ converges on 1.0. For those with a good distribution among their tributes, i.e. comparable tributes coming from many sources, ${}^pTD^t$ tends to be smaller. The upper bound on ${}^pTD^t$ is 1, and the lower bound is $\frac{2}{|people^t|+1}$ (for the case where every person contributes an equal amount) which is relatively impractical. In these experiments, the lowest tribute dominance observed was 0.09.

Tributes to a person with high tribute dominance cannot be discarded right away. It may be the case that this person is just starting his tenure as a nurturer, and got successful in making a successful person for the first time. Such cases cannot be penalized although they may start off with a high tribute dominance. For this, tributes made to people with high tribute dominance, are stored in a "buffer", and when in due course more people give tributes to them, and the tribute dominance goes under an "acceptable" threshold, the buffered tributes are "matured". While a person's tribute is buffered, it is not made known to other people, and hence does not influence the proportioning of the tribute among different nurturers. i.e. The buffered tribute is not considered while calculating the nurturing influences on a person.

Empirically, 0.5 was chosen as a reasonable tribute dominance threshold in these experiments. Although tribute dominance is not a concern in the experiments dealing with publication count, they were used there too, to be safe.

Using the tribute buffering technique along with the calculation of tribute dominance, removed most false positives from the results, and admitted into the nurturer ranks, only those people who had nurtured at least a few people to reasonable amounts.

4.2.2 The results based on citations

Table 4 shows the top cited people, and the top nurturers for citations for a subset of publications in the DBLP. Again, as earlier, a few top cited authors

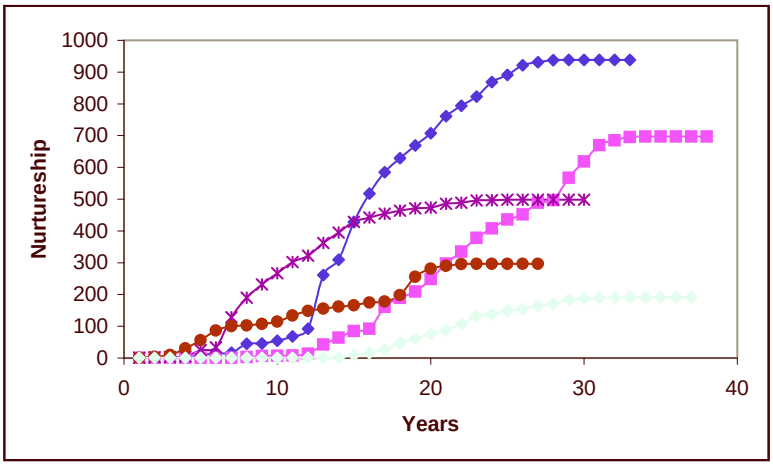

Figure 2: Some typical Nurtureship Growth curves for citations

do not have any presence in the list of top nurturers. The tables 5 and 6 show the drill down of the tributes of a few top nurturers showing the people who gave the highest tributes to each nurturer.

The figure 2 shows a few typical nurtureship growth curves. Unlike, the figure 1, no quadratic growth can be observed here, which indicates that nurturing new people to continually get highly cited is not easy, and it happens only once in a while. Most phases of growth tend to be linear, although with different slopes. The negligible growth seen towards the end are due to recent publications not yet reaching their fullest citation potential. People with extended phases of negligible growth can be considered not to be actively nurturing.

5 The Best Nurturers in a Time Period

Until now, the heuristic has been used to compute nurturers in the entire time period for which article information is available. Of interest to people, and especially students, is to find nurturers that are currently active and approachable. The heuristic in itself is biased towards the nurturers of earlier periods, since they get tributes from their nurtured for much longer periods in time. This causes a shadowing of the recent nurturers. Computing the best nurturers in smaller time periods does not boil down to using a subset of publications just pertaining to the time period - for the following reasons:

- If the computation starts at a time instant t_1, then all the past history of nurturers until then is lost.

- Since his history has been wiped out, a bigger nurturer may get smaller shares of tributes, thus leading to a misappropriation of the tributes.
- The senior researchers themselves may end up giving substantial tributes to the younger ones.

Further, the nurtureship over a given time period cannot be calculated just by adding up the tributes given to a person in that time period. A person who has been nurtured by another, continues to give tributes to him throughout his career and hence these tributes may not be representative of the nurturing influence imparted during a particular time period.

Similarly, the nurtureship cannot consider only the new associations that were formed during that time period, since older associations may get nurtured too.

Hence, a measure of the nurturing influence imparted on a person **in that time frame** is used, and each tribute in that time period is added up proportionate to the nurturing influence a person received then.

$$nurtureship_p^{t_1-t_2} = \sum_{\substack{c \in collaborations; \\ t_1 <= time_c <= t_2}} \sum_{q \in associates_c} {}^{p}tribute_c^q \quad * \left(\frac{{}^{p}nurturing\text{-}influence_q^t - {}^{p}nurturing\text{-}influence_q^{t_1}}{{}^{p}nurturing\text{-}influence_q^t} \right)$$

The tables 7 and 8 list the top nurturers based on publication count and citations in the time period (1992-2004) respectively.

5.1 On the selection of α

The figure 3 shows the frequency distributions of the top 1000 nurtureship values on a logarithmic scale for varying α, {0, 0.25, 0.5, 1, 2}. The distribution of the nurtureship values is according to the power law, for all these values of α.

In the Nurturer-Finder heuristic, the appropriation of tributes is biased, based on the nurtureship values of associates. This way, people with higher nurtureship values are said to have a greater nurturing influence than the others. α controls the extent to which this biasing happens. Larger values of α will make the "bigger" nurturers get a bigger share of tribute each time,

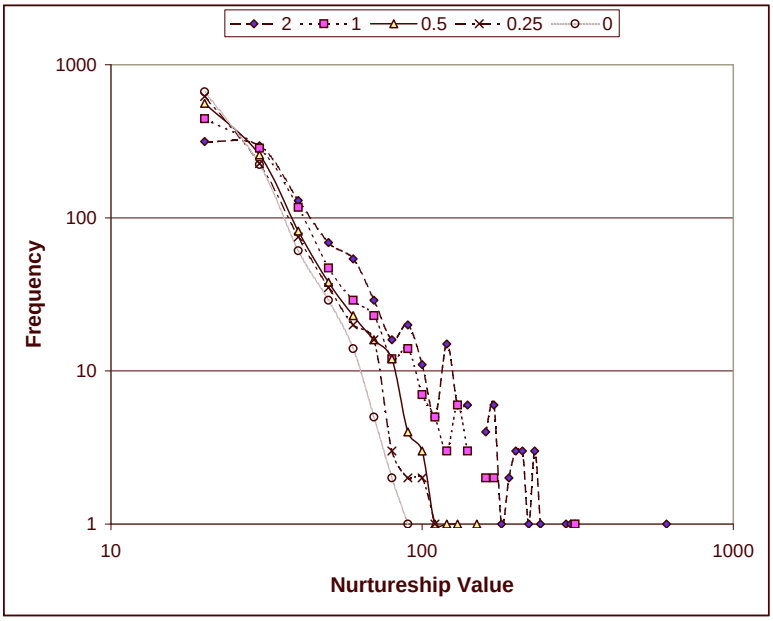

Figure 3: Nurtureship frequency distributions with varying α

at the cost of the smaller nurturers. At the same time, if α is 0, then there is no biasing and tributes are appropriated solely based on the strength of early association. More people get good nurtureship values this way. The frequency distribution for $\alpha = 2$ is skewed to the right. This stretch is fuelled by decreased values of nurtureship among the rest of the people, more so among the people who have small nurtureship values.

The α was chosen as 0.5 based on trial and error to engineer semantically acceptable results in some subsets of the database.

6 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 longetivity 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 experiments consider only the degree of a node to qualify the best collaborators. Here, the best collaborators were said to be those who collaborated with other best collaborators, and had many instances of the same. Thus, it is a weighted sum of the associations had 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.

In continuation to the experiments reported in the paper, the weighted tribute graph formed among the authors can be analysed for transitivity and neighborhoods to discover regions that have nurtured the most people. 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.

7 Conclusion

The paper presents a new heuristic based on paying tributes for post-associative success, to mine for nurturers in collaboration networks, and uses the same to find the best nurturers in computer science research through parameterizable success measures. Certain boundary conditions due to instant success in some significance measures are carefully handled using the measure of tribute dominance. A way to slice the calculation for a period of time is also presented.

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] *Erdos Number Project*, http://www.oakland.edu/enp/

[6] *DBL Browser*, http://dbis.uni-trier.de/DBL-Browser/

[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.

Rank	Nurturer Nurtured	Value
8	C. V. Ramamoorthy	88
	Benjamin W. Wah	11
	Vijay K. Garg	9
	K. Mani Chandy	9
	Jaideep Srivastava	9
	K. H. Kim	8
	Shashi Shekhar	7
	Wei-Tek Tsai	5
	Atul Prakash	5
9	Zvi Galil	83
	Moti Yung	10
	David Eppstein	7
	Kunsoo Park	7
	Nimrod Megiddo	6
	Dany Breslauer	5
10	Christos H. Papadimitriou	81
	Joseph S. B. Mitchell	10
	Paris C. Kanellakis	6
	John N. Tsitsiklis	5
	Mihalis Yannakakis	5
11	Ronald L. Rivest	80
	Robert E. Schapire	10
	Avrim Blum	9
	Benny Chor	5
	Jon Doyle	5
	Sally A. Goldman	5
12	Kurt Mehlhorn	78
	Michael Kaufmann	11
	Majid Sarrafzadeh	6
	Norbert Blum	5
13	John Mylopoulos	77
	James P. Delgrande	10
	Hector J. Levesque	7
	Nick Roussopoulos	6
	Alexander Borgida	5
14	Amir Pnueli	76
	Dennis Shasha	9
	David Harel	5
	Doron Peled	5
	Oded Maler	5
15	Grzegorz Rozenberg	75
	Dirk Vermeir	7
	Robert Meersman	6
16	Richard J. Lipton	75
	Dan Boneh	8
	Lawrence Snyder	7
	David P. Dobkin	5
17	John H. Reif	74
	Paul G. Spirakis	17
	Sanguthevar Rajasekaran	8
	Philip N. Klein	7
	Sandeep Sen	6

Table 3: **Publication Count: Top nurturers and their nurtured (contd)**

Rank	Top Authors		Top Nurturers - Citations	
	Name	Value	Name	Value
1	Jeffrey D. Ullman	2003.03	Michael Stonebraker	938.33
2	E. F. Codd	1448.00	Jeffrey D. Ullman	697.26
3	Michael Stonebraker	1292.94	Catriel Beeri	547.26
4	Jim Gray	826.00	David J. DeWitt	515.27
5	Philip A. Bernstein	781.27	Philip A. Bernstein	498.18
6	David J. DeWitt	762.07	Yehoshua Sagiv	296.82
7	Peter P. Chen	733.83	David Maier	270.27
8	Serge Abiteboul	722.28	Nathan Goodman	266.62
9	David Maier	701.49	Michael J. Carey	191.56
10	Won Kim	623.57	Gio Wiederhold	190.33
11	Yehoshua Sagiv	565.92	Rakesh Agrawal	175.84
12	Hector Garcia-Molina	549.05	Dennis Tsichritzis	166.20
13	Catriel Beeri	547.55	Raymond A. Lorie	163.13
14	Nathan Goodman	517.61	Christos H. Papadimitriou	159.81
15	Ronald Fagin	507.51	Eugene Wong	155.88
16	Umeshwar Dayal	503.76	Georges Gardarin	149.22
17	Rakesh Agrawal	503.16	Franois Bancilhon	144.55
18	Richard Hull	486.42	Bruce G. Lindsay	138.07
19	Michael J. Carey	471.13	Michael Hammer	134.55
20	Moshe Y. Vardi	461.31	Serge Abiteboul	132.92
21	Carlo Zaniolo	449.46	Donald D. Chamberlin	130.38
22	Franois Bancilhon	443.54	Stefano Ceri	116.64
23	Raghu Ramakrishnan	430.50	Alberto O. Mendelzon	114.17
24	Christos Faloutsos	400.79	Dennis McLeod	113.14
25	Jennifer Widom	390.76	Timos K. Sellis	112.83
26	Donald E. Knuth	383.50	Joachim W. Schmidt	107.43
27	Goetz Graefe	382.68	Morton M. Astrahan	105.91
28	C. J. Date	378.00	Abraham Silberschatz	104.22
29	Raymond A. Lorie	373.87	Ronald Fagin	102.86
30	Richard T. Snodgrass	366.00	Raghu Ramakrishnan	100.57
31	Shamkant B. Navathe	355.71	Hans-Jrg Schek	100.52
32	Patrick Valduriez	350.44	Mihalis Yannakakis	92.69
33	Stefano Ceri	345.61	Nick Roussopoulos	92.34
34	Yannis E. Ioannidis	342.50	Umeshwar Dayal	92.09
35	Christos H. Papadimitriou	341.48	Kapali P. Eswaran	88.79
36	S. Bing Yao	339.25	James P. Fry	87.52
37	Nick Roussopoulos	338.67	Peter Buneman	85.93
38	Per-ke Larson	331.79	Shamkant B. Navathe	84.93
39	H. V. Jagadish	328.92	Hector Garcia-Molina	84.53
40	C. Mohan	319.29	Moshe Y. Vardi	82.77
41	Antonin Guttman	315.63	Clement T. Yu	79.20
42	Eugene Wong	314.22	Alfred V. Aho	74.80
43	Jeffrey F. Naughton	313.76	Frank Wm. Tompa	73.95
44	David W. Shipman	310.67	Theo Hrder	72.58
45	Alberto O. Mendelzon	308.98	Yannis E. Ioannidis	72.09
46	Gio Wiederhold	308.89	Carlo Batini	71.31
47	Abraham Silberschatz	298.96	Jim Gray	71.09
48	Timos K. Sellis	288.60	George P. Copeland	70.67
49	Arie Shoshani	286.18	Haran Boral	70.46
50	Tomasz Imielinski	282.59	Paris C. Kanellakis	68.30

Table 4: **Top 50 authors and nurturers based on citations**

Rank	Nurturer Nurtured	Value
1	Michael Stonebraker	938
	Antonin Guttman	163
	Michael J. Carey	126
	Timos K. Sellis	101
	Eric N. Hanson	78
	Yannis E. Ioannidis	75
	Leonard D. Shapiro	58
	Joseph M. Hellerstein	36
	David J. DeWitt	29
	Lawrence A. Rowe	29
	Eugene Wong	28
	Sunita Sarawagi	21
	Daniel R. Ries	17
	Frank Olken	14
	Randy H. Katz	13
	Erich J. Neuhold	13
	John K. Ousterhout	12
2	Jeffrey D. Ullman	697
	Yehoshua Sagiv	89
	David Maier	73
	Alberto O. Mendelzon	66
	Catriel Beeri	64
	Ashish Gupta	45
	Allen Van Gelder	41
	Gabriel M. Kuper	34
	Yannis Papakonstantinou	29
	Fereidoon Sadri	19
	Arthur M. Keller	18
	Moshe Y. Vardi	17
	M. R. Garey	17
	Anand Rajaraman	14
	Dallan Quass	13
	Franois Bancilhon	11
3	Catriel Beeri	547
	Moshe Y. Vardi	99
	Raghu Ramakrishnan	72
	Michael Kifer	68
	Henry F. Korth	43
	Tova Milo	33
	Philip A. Bernstein	24
	Serge Abiteboul	23
	Jeffrey D. Ullman	23
	Nathan Goodman	17
	Gerhard Weikum	16
	Ron Obermarck	11
	David Maier	11
	Alberto O. Mendelzon	11
4	David J. DeWitt	515
	Rakesh Agrawal	118
	Goetz Graefe	90
	Haran Boral	33
	Hong-Tai Chou	32
	Shahram Ghandeharizadeh	19
	Jim Gray	18

Table 5: **Citations: Top nurturers and their nurtured**

Rank	Nurturer Nurtured	Value
	Michael J. Carey	18
	M. Muralikrishna	18
	Donovan A. Schneider	17
	Eugene J. Shekita	15
	Dina Bitton	12
5	Philip A. Bernstein	498
	Umeshwar Dayal	148
	David W. Shipman	82
	Nathan Goodman	54
	Catriel Beeri	46
	Marco A. Casanova	35
	Harry K. T. Wong	35
	Christos H. Papadimitriou	29
	Barbara T. Blaustein	11
	John Mylopoulos	11
6	Yehoshua Sagiv	296
	Mihalis Yannakakis	48
	Alon Y. Levy	47
	David Maier	41
	Jeffrey D. Ullman	28
	Anand Rajaraman	20
	Alberto O. Mendelzon	19
	Jeffrey F. Naughton	17
7	David Maier	270
	David Scott Warren	44
	George P. Copeland	40
	Jeffrey D. Ullman	24
	Alberto O. Mendelzon	19
	Goetz Graefe	19
	Jacob Stein	15
	William J. McKenna	14
	Albert Croker	13
	Yehoshua Sagiv	13
8	Nathan Goodman	266
	Oded Shmueli	70
	Catriel Beeri	31
	Christos H. Papadimitriou	29
	Johann Christoph Freytag	28
	Dennis Shasha	22
	Umeshwar Dayal	21
	Randy H. Katz	19
	Philip A. Bernstein	15
9	Michael J. Carey	191
	Rakesh Agrawal	48
	Hongjun Lu	31
	Miron Livny	22
	Michael J. Franklin	19
	David J. DeWitt	11
10	Gio Wiederhold	190
	Ramez Elmasri	52
	Xiaolei Qian	26
	Hector Garcia-Molina	24
	Stefano Ceri	18
	Domenico Sacc	11

Table 6: **Citations: Top nurturers and their nurtured (contd)**

Rank	Top Nurturers for Publication Count since 1992	
	Name	Value
1	Micha Sharir	44.02
2	Alberto L. Sangiovanni-Vincentelli	38.83
3	Hector Garcia-Molina	34.41
4	Ugo Montanari	34.20
5	Kang G. Shin	33.15
6	Avi Wigderson	31.71
7	Stefano Ceri	30.76
8	Elisa Bertino	30.32
9	Fausto Giunchiglia	30.27
10	Hongjun Lu	29.79
11	Michael Stonebraker	29.02
12	Oded Goldreich	28.36
13	Jiawei Han	28.20
14	Philip S. Yu	27.21
15	Noga Alon	27.04
16	Donald F. Towsley	26.76
17	Joseph Y. Halpern	26.46
18	Thomas S. Huang	25.68
19	Leonidas J. Guibas	25.63
20	Satish K. Tripathi	25.43
21	Sushil Jajodia	25.22
22	Jeffrey D. Ullman	24.99
23	Richard R. Muntz	24.65
24	Randy H. Katz	24.45
25	Alex Pentland	24.38
26	Abraham Silberschatz	24.24
27	Edmund M. Clarke	23.36
28	Friedhelm Meyer auf der Heide	23.34
29	Richard M. Karp	23.07
30	Moti Yung	22.98
31	Roberto Gorrieri	22.88
32	John Mylopoulos	22.61
33	Zohar Manna	22.27
34	Amir Pnueli	22.21
35	V. S. Subrahmanian	22.06
36	Richard J. Lipton	22.00
37	Krithi Ramamritham	21.99
38	Ricardo A. Baeza-Yates	21.90
39	Josef Kittler	21.86
40	Christos Faloutsos	21.35
41	Rajeev Motwani	21.28
42	Timos K. Sellis	21.28
43	Kishor S. Trivedi	21.24
44	Giorgio Levi	21.23
45	David E. Goldberg	21.05
46	Paul G. Spirakis	20.91
47	Jack Dongarra	20.78
48	Mukesh Singhal	20.53
49	Robert Endre Tarjan	20.38
50	Georg Gottlob	20.34

Table 7: **Top 50 nurturers based on publication count since 1992**

Rank	Top Nurturers for Citations since 1992	
	Name	Value
1	Jeffrey D. Ullman	142.00
2	Yehoshua Sagiv	100.92
3	Michael Stonebraker	77.20
4	David J. DeWitt	54.96
5	Yannis E. Ioannidis	53.51
6	Hector Garcia-Molina	50.96
7	Peter Buneman	45.02
8	Rakesh Agrawal	34.28
9	Raghu Ramakrishnan	33.90
10	Jennifer Widom	30.96
11	Stefano Ceri	30.84
12	Timos K. Sellis	30.30
13	Michael J. Carey	29.52
14	Serge Abiteboul	26.40
15	Philip S. Yu	25.17
16	V. S. Subrahmanian	24.91
17	Ashish Gupta	24.86
18	Moshe Y. Vardi	24.78
19	Catriel Beeri	24.65
20	Abraham Silberschatz	23.77
21	Patrick Valduriez	22.15
22	Christos Faloutsos	21.91
23	Kevin Strehlo	19.00
24	Dennis McLeod	17.47
25	Arun N. Swami	17.40
26	Waqar Hasan	16.78
27	Alfons Kemper	16.22
28	Hamid Pirahesh	15.74
29	Paris C. Kanellakis	15.36
30	Miron Livny	14.59
31	Stanley B. Zdonik	13.96
32	Jeffrey F. Naughton	13.30
33	Vincent Y. Lum	13.15
34	Dallan Quass	12.91
35	Heikki Mannila	12.48
36	Hans-Jrg Schek	12.23
37	Tomasz Imielinski	12.16
38	Paolo Atzeni	11.84
39	Jan Paredaens	11.81
40	Limsoon Wong	11.65
41	Inderpal Singh Mumick	11.33
42	Bruce G. Lindsay	11.27
43	Balakrishna R. Iyer	10.99
44	Raymond T. Ng	10.89
45	Barton P. Miller	10.88
46	H. V. Jagadish	10.63
47	Ming-Syan Chen	10.59
48	A. Inkeri Verkamo	10.26
49	Michel Scholl	10.25
50	Fereidoon Sadri	10.17

Table 8: **Top 50 nurturers based on citations since 1992**