Get To the Top and Stay There: A Study of Citation Rank Dynamics in Academia

Yan Wu
Department of Information Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong
yanwu@ie.cuhk.edu.hk

Srini Venkat
Virginia Bioinformatics Institute
Virginia Tech.
Blacksburg, VA 24061, USA
vsriniv@vbi.vt.edu

Dah Ming Chiu
Department of Information Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong
dmchiu@ie.cuhk.edu.hk

ABSTRACT

Citations serve as an important metric for identifying experts and opinion leaders in academic communities. In this paper, we analyze the evolution of yearly rankings of top-cited (C-list) authors in the domain of Computer Science. In searching for factors that help authors become top-cited, we also gather authors in the top-collaboration list (A-list) and top-publication list (P-list) for each year, and analyze cross-correlation of the C-list with the corresponding A-list and P-list for each year. Results show that the A-list and P-list serve as (unreliable) indicators for appearance on the C-list, but their effect is quick and short-lived. Through further case studies we find other key factors, such as the seminal importance of an author's publication and the association on an author's work with hot topic trends, may significantly affect rank dynamics. Based on the study of citation rank dynamics in academia, we then discuss the modeling of rank dynamics, specifically a model based on item visibility and item strength, and the general applicability of such a model.

Keywords

Rank Dynamics; Top-cited Author; Author Strength; Author Visibility

1. INTRODUCTION

Ranking systems are omnipresent, serving various purposes in different communities. For instance, best-seller lists of books and top-rated lists of music albums help consumers decide on their purchases. Yearly rankings of teams and individuals in various sports reflect their relative excellence in performance. Search results direct us to the most relevant web pages, and trending topics help identify the most discussed events in social media.

In the scientific community, we rank top publishing venues for different domains, and top universities in specific areas of research. Further, we can rank highly cited papers and most influential authors. Researchers publish papers to convey their ideas and findings. They also collaborate to jointly make advances, and in general to exchange ideas. The citations accumulated from published papers, help them get noticed, and serve as an indicator of their performance. Thus, among the different ways of ranking researchers (by publications, by collaborations, by citations), citation rankings serve as a reliable metric of their contributions to the community. Ranking authors based on their yearly citations help identify the experts and opinion leaders.

In this paper, we consider the yearly rankings of authors (limiting ourselves to the top 100 authors) in the Computer Science field, with respect to yearly publications, collaborators and citations, and study their dynamics. We refer to these rank lists as P-list, A-list and C-list respectively. We analyze the potential influence of having the highest number of publications and number of collaborations on becoming top-cited authors. Such an analysis will help shed light on questions such as "how does an author become part of the top-cited list?", and "how does a top-cited author compete and retain his position in the list?". The key observations are as follows:

i. For authors who appear in the C-list and the A-list/P-list, 80% of them tend to appear earlier in the A-list or P-list.

ii. When it comes to the longevity of stay in the C-list for the common authors in C-list and A-list/P-list, those who enter the C-list first tend to stay much longer (40%) on average.

iii. Even though A-list and P-list lead the C-list for the common appearances, they do not guarantee it in any way. In fact, only $20-30\%$ of the authors make it to the C-list.

iv. For authors who enter C-list through the help of appearing in A-list/P-list first, the delay is short (approximately $2-3$ years).

In short, although appearing in top-collaborations or top-publications list may affect one's appearance in the top-cited list, the impact is quick and short-lived. Besides, through case studies, we note the relevance of topical trends in the community, and its contribution to an author's citation performance.

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2890519.

Based on the analysis of citation rank dynamics in academia, we then discuss how to model the dynamics of rank lists, based on key factors that influence their evolution. We point our the limitations of some related works, and describe our idea for this research direction.

2. RELATED WORK

Study of ranking dynamics in complex systems has attracted much attention in the complex systems community. For example, [1] tried to find a general law that regulates ranking dynamics in complex systems. [3] studied the rankings in sports and tried to model the competitiveness among sport teams. [7] provided a general discussion on how to analyze and model ranking data. These studies tend to be either domain specific, or only focus on short-term ranking dynamics (based on noise) so far.

Another group of papers related to our work are those studying author's collaborations, publications or citations separately. For example, [6] studied the complete trials of coauthor network evolution. [4] demonstrated the effect of aging on researchers' publication patterns. [12] discussed the competition of preferential attachment and aging effect during an author's collection of citations. [11] quantified the different factors that might affect the citation of an paper. However, those papers only focused on the analysis of the general patterns in collaborations, publications, or citations separately and have not extensively studied the correlations between them. In our paper, we analyze the top authors in each category and focus on the correlations between citations and publications or collaborations.

Some papers study correlations between collaborations and citations. For example, [8] analyzed the temporal changes in citation and collaboration over time and the influence of coauthor network on citation network. While they did a statistical analysis on the percentage of self-citation, coauthor-citation and distant citation of papers, we focus on the analysis of an author's social visibility and influence in helping him accumulate citations. Furthermore, while [8] studies the entire community, we focus on the evolution of the top-ranked authors in each category.

3. DATASET AND PREPROCESSING

Before proceeding with the empirical analysis, in this section, we describe our dataset. Our dataset is collected from ArnetMiner [10]. We use the V5 dataset listed on the website[1]. It was constructed based on paper information in DBLP and citation information from ACM and other sources. In the dataset, each paper is assigned an index ID, with information including the paper title, author list, publication year, publication venue, and the list of references. In order to get a complete set, we removed papers published before 1960, since the origin of Computer Science, as a discipline, is traced back to early 1960s [2]. We also removed papers published after 2009 to get rid of the incomplete records of recent years, since the original dataset was updated in 2011.

3.1 Top-author Rank Lists

We focus on three top-author rank lists in this paper, i.e., top authors with the most number of coauthors, publications and citations in a given year. The size of one's collaboration network is a measure of one's social presence. The number of one's publications reflects his productivity and the citations one gets demonstrate collective influence of his publications till date.

For the counting of coauthors of an author in one year, we use a weight based counting method [9]. It is based on the assumption that authors who have written many papers together, know each other better. With this assumption, the counting proceeds as follows: for an author i with at least one publication at year t, we find his coauthor set $C_{i,t}$ at year t, and then define the collaboration weight between author i and one of his coauthor j ($j \in C_{i,t}$) as

$$w_{ij} = \sum_{p} \frac{1}{n_p - 1}, \tag{1}$$

where p is the set of papers with both author i and author j in the author list in year t, n_p is the number of authors of paper p. w_{ij} is also capped by 1. Then the number of coauthors of author i at year t is

$$S_{i,t} = \sum_{j \in C_{i,t}} w_{ij}, \tag{2}$$

Since w_{ij} is capped by 1, $S_{i,t} \leq |C_{i,t}|$ and $S_{i,t} \leq |p|$. Fig. 1 gives an example of the weight based counting method. In Fig. 1, author a has three papers in one year, labeled 1, 2, and 3. Paper 1 is written by author a, b, c and d together. Paper 2 is written by author a and b together. Paper 3 is written by author a, b and e together. For author a, his set of collaborators for that year include b, c, d and e. The weighed counting of collaborator b is $1 + 1/3 + 1/2 = 11/6$ and then capped by 1, of c is 1/3, of d is 1/3, and of e is 1/2. Then the total weighted number of coauthors for author a is $1 + 1/3 + 1/3 + 1/2 = 13/6$.

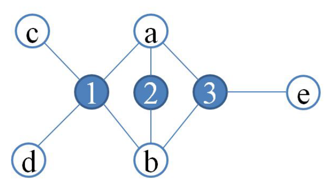


Figure 1: Example of the weight based counting method.

Such a weight based counting method not only captures the relative size of one's coauthors, but also takes the collaboration strength into consideration. Moreover, it can help us get rid of the bias of papers with many authors (which will make all the authors of that paper rank high in the list when using the simple directed counting method $|C_{i,t}|$).

The counting of publications and citations of an author in one year is relatively straightforward, i.e, if one is listed as an author of a paper, that paper is counted as his publication; if one paper is cited by another, each author of that paper gets one increment for citation.

After the counting, we set the cutoff at 100 to obtain the top-100 author rank list. For authors at the 100 borderline, we retain the authors who have appeared in the top rank list of previous years, and then proceed to choose randomly among the others. By doing so, we ensure the ranking dynamics is smoother. However, this does not affect most of our results, since it only pertains to the borderline authors. Table 1 describes the general information of the three rank

[1]http://aminer.org/billboard/citation

lists. We call the top list based on yearly number of coAuthors as A-list, the top list based on yearly number of Publications as P-list and the top list based on yearly number of Citations as C-list respectively. In Table 1 we can see that there are fewer distinct authors in C-list than in A-list and P-list, which indicates that C-list is relatively more stable than the other two.

Table 1: Rank List Description

Rank list name	Time coverage (year)	#distinct authors
A-list	1980-2004	1050
P-list	1980-2004	1022
C-list	1980-2009	655

3.2 Author Name Disambiguation

Since we use author name as the identifier of different authors, an inevitable consequence is that authors with similar names are merged into a single author. Therefore, we do a rough author name disambiguation manually after we get the three original rank lists[2]. We choose some special cases of ambiguity, which help us in cleaning the dataset.

Note that for the general author name ambiguity problem, there exist two cases [5]: i) polyseme, where different authors with the same name is merged into one author; and ii) duplicate, where one author is split into multiple authors. Here in this paper, since our dataset has been used as the backbone of Arnetminer, we believe the website developers have already cleared the raw data, thus the second case seldom exists.

In general, an ambiguous entry appearing in the ranklist could be a result of (a) merging of one dominating author (truly linked to most of the publications) with several other normal authors, or (b) merging of several normal authors. In the former case, the ranking results will not be affected much, since the dominating author will appear in the ranklist. For the latter case, such a phenomenon could cause a spurious name to occur in the ranklist for a considerable consecutive period of time. In this paper, we set the consecutive year length to be 5 and we manually disambiguate the author names which appear in a list for five or more consecutive years. This heuristic also ensures that, errors if any, will not affect the long-term trends in the ranking dynamics.

Then we do author name disambiguation manually for those author name candidates by checking their DBLP pages. We think the information under one author name is correct and thus the counting of the annual number of collaborators, publications and citations are right under the following cases:

i. An author's DBLP page is listed with a link for reference when he shows his publications in his homepage.

ii. DBLP shows a link to one's homepage and there is no other authors with the same name listed on DBLP.

iii. DBLP lists the institution history of one author and there is no other authors with the same name listed on DBLP.

iv. DBLP shows a link to one's Wikipedia page, or ACM author profile page, or verified Google Scholar page.

v. The merging of papers from another author does not affect the final ranking result. For example, the merged papers are published after 2009.

For authors not in the cases listed above, we check all the papers from that author and split that author name into multiple authors. We then calculate the annual number of collaborators, publications and citations for each new author after splitting to decide whether an author drops out of the rank list. If an author with updated information drops out of the top 100 rank list, we will add new authors to the list based on the mechanism mentioned before, to keep the length of the rank list always 100.

After that we also do author name disambiguation manually for the top 5 authors each year in the three rank lists. For A-list and P-list in 2005-2009, they are affected heavily by the name ambiguity problem. The lists are mostly occupied by merging Chinese names. Therefore, we only consider the A-list and P-list in 1980-2004. Since all the three kinds of top lists cover 1980-2004, which spans for 25 years, we believe it is long enough to demonstrate the dynamic changes and correlations among the three lists.

4. DYNAMICS OF TOP-CITED AUTHOR LIST

In this section, we use the three top author rank lists extracted from our dataset to show how the rank dynamics of top-cited authors is affected by different factors. We use the dynamics of one's appearance in C-list to represent his citation dynamics.

4.1 Getting into the Top-cited List

We now show how an author gets into the top-cited list through his number of collaborations or publications, i.e, we empirically analyze an author's getting into C-list following his appearance in A-list or P-list. We focus on the common authors in C-list that also appear in A-list or P-list. Table 2 shows the intersection of distinct authors in any of the two rank lists, i.e., the number of common authors appearing in two different kinds of rank lists.

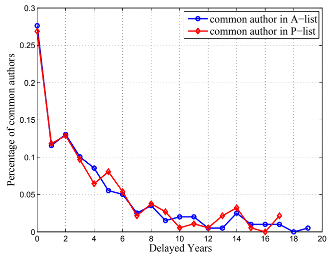
Table 2: Intersection of Distinct Authors

Intersection	A-list	P-list	C-list
A-list	1050	690	252
P-list	690	1022	234
C-list	252	234	655

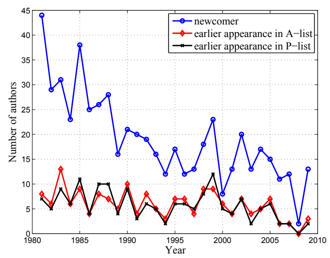
We first study the time that an author first gets into the three top lists. We extract the first appearing year in two lists respectively of the 252 common authors in C-list and A-list, and that of the 234 common authors in C-list and P-list. Results show that among the common 252 authors, 199 authors (about 78.97%) appear later in C-list than in A-list. Also among the common 234 authors, 186 authors (about 79.49%) appear later in C-list than in P-list[3]. This

[2]Table 1 shows the list information after we do author name disambiguation on the original lists extracted directly from the dataset.

[3]Note that since our data is at the yearly resolution, it is impossible to distinguish the order of citations and publications within a given year. Therefore, if an author appears in the same year in both the C-list and the A-list or P-list, we treat it as a later arrival in the C-list.



(a) Distribution of delayed years



(b) Change of different sets of newcomers over time

Figure 2: Getting into the top-cited list

shows that for common authors in both C-list and A-list (or P-list), most of them get into C-list with the help of his appearance in A-list or P-list first. So A-list and P-list may serve as indicators of authors in later C-list. We also plot the delay distribution in Fig. 2(a) for those common authors who appear later in C-list. Note that part of the reason for the high percentage of common authors with delayed year 0 is that we start the three rank lists at the same year 1980. However, even after we remove all the common authors (around 20) in both C-list and A-list (or P-list) in 1980, the percentage of common authors with short delayed years is still relatively high. Based on the delay distribution, we find that the average delay is 3.67 years for authors who enter A-list first and then enter C-list. For those who enter P-list first and then enter C-list, the average delay is 3.76 years. It indicates that being in A-list or P-list has a quick effect in helping authors get into C-list.

Besides the common authors, we also take analysis on the newcomers each year in 1981-2009 in C-list. For those newcomers, part of them are also common authors appearing earlier in A-list or P-list. We plot the change of different sets of newcomers in Fig. 2(b). We see that among those newcomers, less than half of them have ever appeared in A-list or P-list before. This result actually corresponds to the relatively small intersection percentage (35%) of C-list with A-list or P-list shown in Table 2. It indicates that there are also other factors that may affect an author's citation performance, which we will discuss later.

4.2 Staying in the Top-cited List

After an author gets into the C-list, he needs to compete with other top-cited authors in order to stay in the list. In other words, he needs to retain his relatively high competitiveness. We now analyze the behaviors of those top-cited authors after they get into the C-list and try to find the correlation of their behaviors in C-list with that in A-list or P-list.

We first analyze all the authors in C-list. Fig. 3(a) shows the minimum number of yearly citations an author should obtain in order to maintain his top position, i.e., roughly the citations gathered by the author ranked around 100. We see slow increase in earlier years and quick increase in recent years, attributable to increasing publishing rate, and the number of references in each paper. Besides the yearly citation threshold, we also plot the distribution of longevity of top-cited authors staying in C-list in Fig. 3(b). For comparison, the distribution of the longest consecutive year length is also plotted. We observe that most of the top-cited authors have a short longevity staying in C-list.

Next we do the longevity distribution analysis on the common authors in C-list and A-list (or P-list). We compare the distribution of different sets of common authors who get into C-list at the first time earlier or later than the other two lists. The result is shown in Fig. 3(c) and 3(d). We find that for common authors appearing earlier in C-list than in A-list, the average year length is 8.21 years, while that for the remaining common authors is 5.89 years. We get similar findings when doing the comparison on common authors in C-list and P-list, where the average year length for common authors appearing earlier in C-list than in P-list is 8.98 years and that for the remaining common authors is 6.02 years. It demonstrates that authors who get into A-list or P-list with the help of their top-cited work tend to stay longer in C-list. Combined with previous findings, we see that although appearing in A-list or P-list can help one enter C-list quickly, those who enter C-list first tend to stay in the list longer.

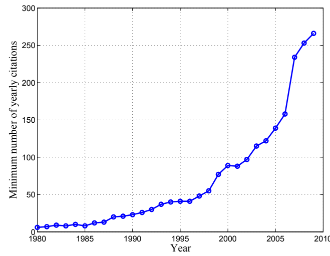
4.3 Indicators for Top-cited Authors

Now after the study of the patterns of common authors in C-list and A-list or P-list, we would like to see as a general top author in A-list or P-list, how can one become a top author in C-list.

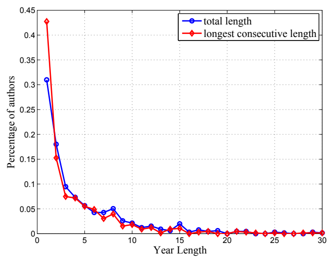
Based on the statistics in Table 2, we can see that among the 1050 authors in A-list, only 199 (about 18.95%) appear later in C-list, and among the 1022 authors in P-list, only 186 (about 18.20%) appear later in C-list. It indicates that although getting into A-list or P-list might help one enter C-list, such possibility is relatively low.

Besides the overall percentage, we also plot the change over time in Fig. 4(a), where we show the number of common authors appearing earlier in A-list or P-list each year among the 100 top authors in A-list or P-list. We find that less than half of the authors appear in C-list later in 1980-2000. The number of such authors keep fluctuating. However, from 2001 on, that number keeps decreasing, which is due to the boundary effect and the termination of A-list and P-list at 2004.

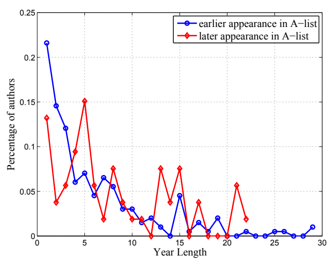
The distribution of the number of times a common author appears in A-list or P-list before he enters C-list at the first time is also plotted in Fig. 4(b). We see that most of those common authors only appear a small number of times. In general, the average number of times that a common author appears in A-list before he enters C-list is 2.37 and the average times that a common author appears in P-list before he enters C-list is 2.35. It again shows our previous finding that appearing in A-list or P-list has a quick effect in helping one enter C-list.



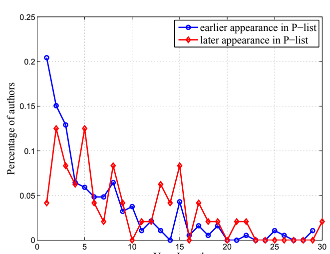
(a) Yearly citation threshold



(b) Distribution of longevity and longest consecutive length

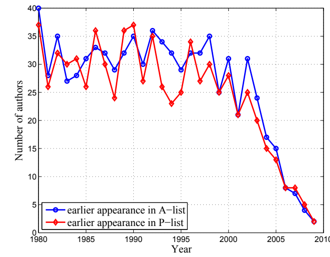


(c) Behavior of common authors in C-list and A-list

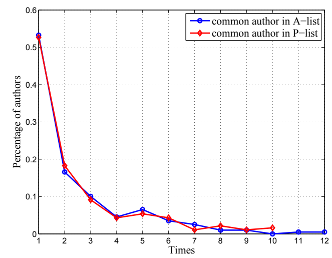


(d) Behavior of common authors in C-list and P-list

Figure 3: Staying in top-cited list



(a) Change of different sets of common authors over time



(b) Distribution of times of common authors appearing in A-list/P-list before entering C-list

Figure 4: Indicators for top-cited authors

4.4 Some Case Studies

Besides the annual number of publications or collaborators discussed above, which indicates an author's visibility to the whole community, there also exist other factors that may affect one's citation performance. We give a discussion of those factors by case study here.

First, an author's citation performance may depend on the relative importance of his papers or ideas. It is pointed out in [11] that the ultimate accumulated number of citations of one paper is only determined by that paper's intrinsic quality. Therefore, when considering an author's citation performance, which largely depends on his collective citations from publications, his important papers or ideas should play a significant role in it. In our C-list from 1980-2009, Ramakrishnan Srikant is listed in 1998-2009 as a yearly top-cited author. However, he never appears in A-list or P-list. After a detailed analysis on his yearly citations accumulated in 1998-2009, we find that most of his citations come from his papers studying association rules, with the most important one named "Fast Algorithms for Mining Association Rules in Large Databases" published in 1994. Citations to those papers are larger in magnitude when compared with his other cited papers.

Second, the hotness of one's research topics may also affect an author's citation accumulation during a period. It will help an author get many citations in a year if his research topic is relatively hot in the given year. In this case, even if an author does not have many publications, he can still attract many citations for the given year. For example, Charles E. Perkins is an expert in the research area of Ad-Hoc Networking, which was relatively hot in the early 2000's. Based on our record, he gets into the C-list in 2003 and 2006, mainly due to his work on Ad-Hoc Networking. An important paper from him studying this topic was published in 1999, named "Ad-hoc On-Demand Distance Vector Routing". He is also an author never appearing in the A-list or P-list.

Although hitting hot research topics can help one become a top author in the C-list, the length that he stays in the list may depend on the duration for which the research topic stays hot. So if an author can hit the hot research topics consecutively, he will be in the C-list for a long time. One example in our dataset is David A. Patterson, who is an expert in the area of computer architecture, having a series of articles on RISC, RAID and NoW. His important paper on RISC was published in 1985, while on NoW was published in 1995. Based on our record, he is a top-cited author from 1986 to 2003 consecutively. A more detailed analysis on his yearly citations shows that it is his consecutive work on those hot topics that makes him stay in the C-list. For example, his work on RISC, which was published in 1985, began to have very little impact since 1992.

These case studies of different top-cited authors here show that the importance of one's work or its relevance to the temporal hot research topics may also affect one's citation rank dynamics much.

5. MODELING OF RANK DYNAMICS

Inspired by the analysis of rank dynamics of top-cited authors in academia, we are now working on a model for rank dynamics, for top-cited authors and potentially applicable to other ranking systems. Due to space limitation of this paper, we will only give a discussion of some existing works and the direction of our approach.

There is keen interest in the complex network community in developing models for rank list dynamics. Recent work includes notably [1]. The proposed model in [1] elegantly captures rank list dynamics caused by *noise*. In other words, the strength (fitness in their terminology) of the ranked items, in our case the authors, do not change other than statistically determined by noise. This model can only capture the short-term dynamics of rank lists; in the long run, the strength of authors surely change. There is significant literature trying to model authors' strength (productivity and influence) over time (some of them discussed in the Related Work section), but these studies are not aimed at modeling rank list dynamics. So we see it as an interesting challenge to develop such a model that can capture the forces driving the rank list dynamics.

Based on our study in the early part of this paper, we observe that there are mainly two forces that drive citation rank dynamics of authors in academia: author visibility and author strength. For authors in A-list or P-list, they either have a large number of coauthors, or a large number of publications. Large number of publications from one author will definitely result in high visibility for him during the searching phase of other authors when they try to find related works. Similarly, since coauthors are automatically assumed to be familiar about an author's work, having more coauthors also help propagate one's ideas, hence also increase one's visibility. Therefore, authors in A-list or P-list enjoy higher visibility in the community when compared with average authors. Author visibility alone, however, does not account for his citation accumulation. Based on our case studies of some top-cited authors, we find author strength and his ability to be associated with (or start) important or hot research topics, also contribute to citation accumulation, especially in the long run. While author visibility is easy to change and can be achieved in many ways by authors, the change of author strength is relatively difficult.

Based on visibility and strength, we are developing a more sophisticated epidemic-type of model for rank list dynamics. Since these two factors, visibility and strength, are quite generic, we expect that our model can have wide applicability.

6. CONCLUSION

In this paper, we study the ranking dynamics of top cited authors in the Computer Science field. Relationship of ranking dynamics of top-cited authors with their publications and collaborations is discussed in detail by a cross-correlation analysis of yearly top-author rank lists. Results show that being an author with large number of publications or collaborators has a quick but transient effect in helping one become top-cited. Besides, the importance of one's work and its relevance to the temporal hot research topics also matters. Based on the analysis of citation rank dynamics in academia, we also give a discussion on the modeling of the rank dynamics in general ranking systems, and discuss the modeling with item visibility and item strength. The detailed mathematical modeling and parameter mining are left as future work.

7. REFERENCES

[1] N. Blumm, G. Ghoshal, Z. Forró, M. Schich, G. Bianconi, J.-P. Bouchaud, and A.-L. Barabási. Dynamics of ranking processes in complex systems. *Physical review letters*, 109(12):128701, 2012.

[2] J. G. Brookshear, D. Smith, and D. Brylow. Computer science: an overview. 2012.

[3] R. Criado, E. García, F. Pedroche, and M. Romance. A new method for comparing rankings through complex networks: Model and analysis of competitiveness of major european soccer leagues. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 23(4):043114, 2013.

[4] Y. Gingras, V. Lariviere, B. Macaluso, and J.-P. Robitaille. The effects of aging on researchers' publication and citation patterns. *PloS one*, 3(12):e4048, 2008.

[5] H. Han, L. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis. Two supervised learning approaches for name disambiguation in author citations. In *Digital Libraries, 2004. Proceedings of the 2004 Joint ACM/IEEE Conference on*, pages 296–305. IEEE, 2004.

[6] D. Lee, K.-I. Goh, B. Kahng, and D. Kim. Complete trails of coauthorship network evolution. *Physical Review E*, 82(2):026112, 2010.

[7] J. I. Marden. *Analyzing and modeling rank data*. CRC Press, 1996.

[8] T. Martin, B. Ball, B. Karrer, and M. Newman. Coauthorship and citation patterns in the physical review. *Physical Review E*, 88(1):012814, 2013.

[9] M. E. Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical review E*, 64(1):016132, 2001.

[10] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 990–998. ACM, 2008.

[11] D. Wang, C. Song, and A.-L. Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

[12] Y. Wu, T. Z. Fu, and D. M. Chiu. Generalized preferential attachment considering aging. *Journal of Informetrics*, 8(3):650–658, 2014.