Noname manuscript No.
(will be inserted by the editor)

Measuring Academic Influence Using Heterogeneous Author-Citation Networks

Fen Zhao · Yi Zhang · Jianguo Lu · Ofer Shai

Received: date / Accepted: date

Abstract Academic influence has been traditionally measured by citation counts and metrics derived from it, such as H-index and G-index. PageRank based algorithms have been used to give higher weight to citations from more influential papers. A better metric is to add authors into the citation network so that the importance of authors and papers are evaluated recursively within the same framework. Based on such heterogeneous author-citation academic network, this paper gives a new algorithm for ranking authors. It is tested on two large networks, one in Heath domain that contains about 500 million citation links, the other in Computer Science that contains 8 million links. We find that our method outperforms other 10 methods in terms of the number of award winners identified in their top-k rankings. Surprisingly, our method can identify 8 Turing award winners among top 20 authors. It also demonstrates some interesting phenomenons. For instance, among the top authors, our ranking negatively correlates with citation ranking and paper count ranking.

Keywords Heterogeneous Network · Author Ranking · PageRank · Scholarly Data

1 Introduction

Academic influence is inherently difficult to measure. Citation count has been used widely since 1927 (Gross and Gross, 1927). H-index was introduced to simplify the citation count by disregard papers that are less cited (Hirsch, 2005). However, H-index treats papers equally once they pass a threshold

Fen Zhao · Yi Zhang · Jianguo Lu
School of Computer Science, University of Windsor, ON, Canada
E-mail: {zhao15m, zhang18f, jlu}@uwindsor.ca

Ofer Shai
Chan Zuckerberg Initiative Inc.
E-mail: ofer.shai@gmail.com

value (the H-index), measures unfavorably for authors who publish one or two very highly cited papers. G-index ameliorates this problem by giving credits for citation counts of each paper that pass a threshold value (the G-index) (Egghe, 2006).

Citations are not independent and acting alone. Instead, they form a complex network in which papers and authors interact with each other. In such network, not every citation is equal. A citation from an influential paper should have a higher weight than others. Thus, Bonacich (1972) proposed that the principal Eigenvalues of a citation matrix should be used for the importance of the papers. This idea also inspired the well-known PageRank algorithm when applied on the Web network (Brin and Page, 1998). Intuitively, the influence of a node is proportional to the probability of being visited in a random walk on the graph.

Despite the successful application of the PageRank algorithm in the Web domain, we have not seen a wide application of the algorithm in bibliometrics where the very idea originated. This is due to two significant differences between the academic network and the Web. Firstly, citation networks are mostly acyclic: papers only cite papers in the past, not the ones to be published in the future. Although occasionally there are loops due to the merge of different versions of a paper, most citations form a chain chasing down to earlier papers. Secondly, academic networks are inherently heterogeneous. In the Web network where PageRank is used, there is only one type of node (web pages) and one type of links (hyperlinks). In the academic network, there are at least two kinds of nodes, i.e., papers and authors.

To solve the first problem, Chen et al. (2007) proposed to employ a lower damper factor (α) in the PageRank algorithm. It can be interpreted as a higher random jump probability ($1 - \alpha$) in the random walk interpretation. They propose to use $\alpha = 0.5$ in contrast to normal practice which is $\alpha = 0.85$. A high random jump probability implies that every node/paper will receive credits from random sources. Hence, author ranking will be highly correlated with paper counts as we will demonstrate in the Experiment section.

To solve the second problem, there are at least two approaches. One approach is to work on an author network that is derived from the heterogeneous network. Then, the PageRank algorithm is applied to the author network. The difference is how the author network is induced. West et al. (2013) derived an author-citation network that is induced from the paper citation network. In the induced author network, author A has a weighted link to author B if A cites a paper written by B. The weight reflects the division of credits to multiple authors and multiple references.

The second approach develops algorithms directly on the heterogeneous academic network. Zhou et al. (2007) proposed Co-Ranking method to run random walks on three different networks-a social network between authors, a citation network connecting papers, and a bipartite network between authors and papers. Sun et al. (2009) use a heterogeneous network to represent the academic network, where authors, papers, and conferences are nodes in the graph. First, they apply their RankClus framework to generate clusters based on conferences, then use the Authority Ranking rule on each conference cluster.

Regardless of the approach, there is no objective evaluation to compare the resulting ranking. Evaluation of the existing methods is mostly anecdotal, citing a few well-known authors being ranked high by their methods. The data is also fragmented, consisting of small networks in a narrow area.

This paper proposes a new ranking method, called APR (Author-PageRank), which applies to heterogeneous academic networks. Papers can only cite older papers, therefore random walks can only go from older papers to newer ones. APR handles the acyclic network problem by adding links between papers and authors. When a new paper and an old paper are written by one author, the random walks can start from the old paper to its author, then go to the new paper; thereby random walks can visit newer papers. Different from the large jump probability used in (Chen et al., 2007) that transfers much of the weight to random papers, it transfers only 15% weight of random jump. It tackles the second problem (the heterogeneous network problem) by combining the author and paper networks together. Instead of random walking on different networks and aggregating the results (Zhou et al., 2007), one random walk is performed on the entire network. Additionally, rather than working on an induced author network, the entire network is maintained so that information is not lost or skewed during the network transformation as in (West et al., 2013).

We test our method on two large data sets. One is a large academic network in health domain that is collected by us. It contains 15 million papers, 12 million authors, about 500 million citations. The other is the well-known AMiner (ArnetMiner Academic Social Network) network in computer science developed by Tang et al. (2008). We evaluate our methods based on the number of Nobel Prize winners for the Health data, and the number of Turing Award winners for the CS data. Our method outperforms all other methods consistently for both datasets. Among the top 50 CS authors ranked by APR, there are 16 Turing Award Winners. Our ranking result is also very different from that of existing methods in terms of Spearman rank correlation. One interesting result is that APR is negatively correlated with paper count, H-index, and G-index among top authors.

2 Related Work

Measuring academic influence for papers, journals, authors et al. has been studied for decades. The most straightforward method is citation count (Gross and Gross, 1927), which is still been widely used recently. An entity with more citations would be ranked higher. H-index (Hirsch, 2005) and G-index (Egghe, 2006) are introduced to treat papers differently according to citation count.

PageRank algorithm is applied to address the academic ranking problem because not every citation is equal. It is firstly applied on citation network to identify the most influential papers in (Chen et al., 2007) and they also find that a paper's citation number and its PageRank value are closely correlated. Amjad et al. (2015b) proposes a new informative metric called Topic-based Heterogeneous Rank which measures the impact of a scholarly data with respect to a given topic in a heterogeneous scholarly network containing authors, papers and journals. One of the main limitations of the proposed method is the computational complexity and high memory usage. Su et al. (2011) studies how missing data in the PageRank algorithm influences the result of papers ranking and proposes PrestigeRank algorithm on that basis, but there is insufficient evidence to make a definite conclusion that PrestigeRank is better than PageRank or citation counts. Zhou et al. (2016) introduces a preferential mechanism to the PageRank algorithm when aggregating resource from different nodes to enhance the effect of similar nodes. Though the method in this paper can more accurately predict papers future degree than PageRank, the prediction for small or zero degree nodes is still not satisfactory. Yan (2014) proposes topic-based PageRank, when applied to a data set on library and information science publications. Another two methods CiteRank (Walker et al., 2007) and FutureRank (Sayyadi and Getoor, 2009) are introduced to rank papers and predict the future citation number. PageRank is also used to rank journals (Bollen et al., 2006) (Su et al., 2011) (Dellavalle et al., 2007) (González-Pereira et al., 2010), and even scientific contribution of countries (Ma et al., 2008).

There are several different kinds of academic networks for author ranking Amjad et al. (2018). PageRank is first applied to authors ranking in (Liu et al., 2005). They propose AuthorRank, a weighted PageRank algorithm, in a co-authorship network. If any two authors co-authored a paper, an undirected edge with unit weight is created between these two authors. Ding et al. (2009) use an author co-citation graph, which is same as the co-authorship network in (Liu et al., 2005). They focus on evaluating the impact of various damping factors. Their methods differ from our work in that they use a homogeneous network that consists of only papers or authors. Besides the co-authorship network, they claim that the importance of authors can be derived from papers. Sidiropoulos and Manolopoulos (2006) proposes a new version of PageRank. They first rank papers on the citation network consisting only papers and choose the same number of papers for each author. Then authors are ranked by computing the average score of all their papers. A similar research is introduced in (Fragkiadaki and Evangelidis, 2016). Another interesting work is called PR-index (Gao et al., 2016), which combines h-index and PageRank together to obtain an objective evaluation criteria. They first rank papers by PageRank, then replace the h-index's citation component with the PageRank score. Therefore, PR-index considers both productivity and popularity of an author. The importance of an author is determined only by the influence of his/her papers, without considering the relation of coauthors and the impact from authors to papers. Author citation network is also widely used. Liang and Jiang (2016) generate an author citation network based on paper citations. A paper citation results in several author citations, each of which links a citing author to a cited author in the author citation network. Yan and Ding (2011) also use an author citation network and a weighted PageRank algorithm to get the importance of authors. A similar paper is proposed in (Radicchi et al., 2009). They create a weighted author citation network from paper citation network. A weighted PageRank algorithm is then used to calculate the score of each author in the network. Another method is proposed in (West et al., 2013). They propose the Eigenfactor score on author network matrix. It is based on Eigenvector and gives more weight to the highly cited authors. Author network is easy to obtain and is effective when the network size is small, but not scalable for large data. The complexity of the PageRank algorithm depends on the number of edges in the network. The author citation network is a dense network compared with paper-author heterogeneous network, and PageRank is not expected to be executed on such a network when the number of links explodes. Paper-paper citation link is crucial for author ranking, since the influence of an author should be evaluated by his/her papers. A heterogeneous network, consisting of papers and authors, is first used by Zhou et al. (2007). There are three networks in the framework, citation network, author social network and paper-author network. They used two separate networks(i.e. citation network and author social network), and random walks are performed independently on these two networks, then the ranking are integrated afterwards. In contrast, we delete the directed links between authors and use such a heterogeneous network to rank authors for the first time. Sun et al. (2009) combine clustering and ranking together. They rank authors within each conference cluster. The reputation of conference can affect an author's influence. Basically, they use paper-author and conference-author links, but not paper-paper citation links.

Besides using the PageRank algorithm on academic networks, centrality is also applied to obtain the author importance. Bibi et al. (2018) use various centrality measures to represent the importance of authors. They also find the centrality measures are significantly correlated with the citation count and h-index. Citation count and H-index are still widely used in recent years. Steinbrüchel (2018) divides authors into two groups: PIs (principal investigators) and non-PIs. The author then introduces a new index h_{pi} based on h-index, where PIs will obtain more weight than non-PIs. Amjad and Daud (2017) first use Latent Dirichlet Allocation (LDA) to split authors into different domains, then allocate paper citations to coauthors according to their topic probability. Daud et al. (2017) try to find new influential researchers by considering the co-authors' citations, the order of appearance and the citation number of co-author venues. Similarly, Usmani and Daud (2017) obtain the ranking scores for papers and venues, then generate authors' scores accordingly. Another work (Amjad et al., 2015a) suggests authors should receive citations according to their productivity and author position.

Based on the literature review, several works derive author's importance from papers, without considering the impact between them. Most researchers use co-authorship network and author-citation network to rank authors without paper information. This kind of network is also dense and cannot be large scaled. The heterogeneous network proposed by Zhou et al. (2007) may be a progress, but they treat citation and author ranking separately. In our work, we believe that integrating authors and papers in a coherent network is a better attempt. The importance of an author is determined by not only his/her published papers, but also coauthors. Besides, papers with influential authors will attract more attention. This paper thus aims to measure the academic influence on such an academic network, and propose the APR method and make some comparison with some existing methods.

3 APR Framework

3.1 Problem Definition

Measuring academic influence is to evaluate authors quantitatively. We use a heterogeneous network to represent the academic data. The network consists of two types of nodes, i.e., authors and papers. There are two types of links – the citation link between papers, and authorship link between a paper and an author. We define the importance of an author as the probability of the author being visited by a long random walk in the heterogeneous network.

Definition 1 ***(Heterogeneous Author-citation Network)***
Given a set of authors $\boldsymbol{a} = \{a_1, a_1, \ldots, a_m\}$ *and a set of papers* $\boldsymbol{p} = \{p_1, p_2, \ldots, p_n\}$*. Let* E_{PP} *denote the citation links between papers;* E_{PA} *denote the authorship relation between a paper and an author. The heterogeneous author-citation network is a graph* $G = (\boldsymbol{a} \cup \boldsymbol{p}, E_{PP} \cup E_{PA})$*.*

For a network contains m papers and n authors, the graph can be represented by a binary $(m+n) \times (m+n)$ adjacency matrix A:

$$A = \begin{pmatrix} A_{PP} & A_{PA} \\ A_{AP} & \mathbf{0} \end{pmatrix}, \tag{1}$$

where A_{PP} is the citation matrix between papers, A_{PA} and A_{AP} represent paper-author relations. $A_{PA} = A_{AP}^T$, since the relation between papers and authors are symmetric. Note that in our graph, there are no direct relations between authors.

Given a heterogeneous author-citation network $G = (\mathbf{a} \cup \mathbf{p}, E_{PP} \cup E_{PA})$, our goal is to obtain a vector r for the network G, where r can reflect the importance/influence of authors a (and papers p).

Fig. 1 gives an example of a heterogeneous author-citation network. In this network, isolated components (p_6 and p_7) would receive very low weight if they were evaluated in citation network only. Now it is connected with the main citation network via author a_4. Besides, a random walk can also go up stream from p_4 to p_1 via author a_1.

It differs other paper/author networks such as the one proposed in (Zhou et al., 2007),(Sun et al., 2009), and (West et al., 2013). In our heterogeneous graph, there are no edges between authors. Author relations can be induced

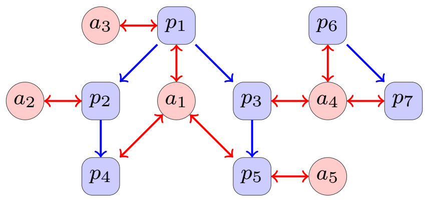


Fig. 1: An example of the heterogenous author-citation network structure.

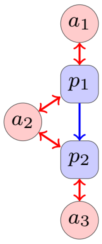
from several sources, such as co-authoring a paper (Zhou et al., 2007), citation of one author to another (West et al., 2013), or even publishing in the same conference (Sun et al., 2009). Such induced relations lost information during the graph transformation. Moreover, the induced graph normally expands in size, sometimes in orders of magnitude. For instance, if an author writes m papers, each cites n papers on average, and each paper has k coauthors, then there will be $m \times n \times k$ induced author-citation edges. Direct links between authors may also make author social network dominating the ranking system. Coauthors of a paper form a clique. Random walk traffic will be directed to such cliques, especially when the cliques size is large. The ranking should be decided mainly by papers, not author relations. Therefore, we excluded the edges between coauthors in the graph. Although direct edges are not presented, coauthor relation still plays a major role in the ranking system: the weight of an author is passed indirectly to his co-author via their papers.

3.2 APR method

The adjacency matrix represented by A can be turned into a column stochastic matrix B, where each column sums up to one. Now the network can be viewed as a Markov chain, and the influence of authors are defined as the stationary distribution of the Markov process. In other words, an author's importance is interpreted as the probability of a random surfer visiting the node. Because not every Markov chain has a stationary distribution, it is necessary to modify the network so that stationary distribution is guaranteed. We follow the normal practice, which is to add virtual links to every pair of nodes with an equal but small transition probability. I.e., a new stochastic matrix $\boldsymbol{M}$ is introduced by adding every cell with a small transition probability:

$$M = \alpha B + (1 - \alpha)\frac{1}{n}ee^T, \tag{2}$$

where $\boldsymbol{e}$ is a vector of $\mathbf{1}$'s, α is the damping factor that is normally chosen to be a value around 0.85 (Brin and Page, 1998). n is the length of the matrix. Now, the Markov process represented by $\boldsymbol{M}$ is guaranteed to be strongly connected and aperiodic, and its stationary distribution is guaranteed. The author (and paper) ranking is also the principal Eigen vector r of the matrix $\boldsymbol{M}$, which



$$\Rightarrow \quad A = \left(\begin{array}{cc|ccc} 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ \hline 1 & 0 & & & \\ 1 & 1 & & \mathbf{0} & \\ 0 & 1 & & & \end{array}\right) \quad \Rightarrow \quad B = \left(\begin{array}{cc|ccc} 0 & 0 & 1 & 1/2 & 0 \\ 1/3 & 0 & 0 & 1/2 & 1 \\ \hline 1/3 & 0 & & & \\ 1/3 & 1/2 & & \mathbf{0} & \\ 0 & 1/2 & & & \end{array}\right)$$

$$\Rightarrow M = \alpha \left(\begin{array}{cc|ccc} 0 & 0 & 1 & 1/2 & 0 \\ 1/3 & 0 & 0 & 1/2 & 1 \\ \hline 1/3 & 0 & & & \\ 1/3 & 1/2 & & \mathbf{0} & \\ 0 & 1/2 & & & \end{array}\right) + (1-\alpha) \left(\begin{array}{cc|ccc} 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ \hline 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{array}\right) \Rightarrow r = \left(\begin{array}{c} r_{p_1} \\ r_{p_2} \\ \hline r_{a_1} \\ r_{a_2} \\ r_{a_3} \end{array}\right) = \left(\begin{array}{c} 0.20 \\ 0.32 \\ \hline 0.09 \\ 0.22 \\ 0.17 \end{array}\right)$$

Fig. 2: An example of the APR framework.

can be computed by the following equation:

$$Mr = r. \tag{3}$$

Fig 2 gives an example. In this network, there are two papers and three authors. p_1 cites p_2. p_1 is written by a_1 and a_2 and p_2 is written by a_2 and a_3. We first derive the network into the adjacency matrix A, then turn it into the column stochastic matrix B. By adding virtual links to the network, a new stochastic matrix M can be deduced. In this example, we set $\alpha = 0.85$. The importance of authors and papers is the principal Eigen vector of M. As expected, a_2 is the most influential author, who writes two papers and one paper has citation. Although a_1 and a_3 both write only 1 paper and share a same coauthor, a_3's paper has citation. In this case, a_3 will gain more importance.

In our work, the largest network contains 12 million authors and even larger number of papers. Despite the large size of the matrix ($10^7 \times 10^7$), fortunately, it is sparse, and we do not actually store the virtual links during the computation. Hence, we can use the 'power iteration method' (Brin and Page, 1998) to compute the principal Eigenvector of matrix M. In our implementation, we iterate 100 times to guarantee the convergence.

4 Experiments

Experiments are conducted on two large academic networks, one is in Health domain and another is on Computer Science(hereafter CS) domain. The statistics are tabulated in Table 1. The Health data are mostly from PubMed. It focuses on academic papers in health domain only. In total there are 26,249,870 papers. After removing papers that have neither citations nor references, our citation network contains 15,366,456 papers and 479,358,572 citation links. Compared with the largest public available citation network, the Microsoft Academic Graph (MAG[1]), our data is more complete in the health domain.

[1] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

MAG data covers all the areas, hence contains much more papers (100 million). However, it has the same amount of links as our data. Compared to MAG, our data is much more complete in the health area as is demonstrated by the higher average degree of the graph (31 vs. 5).

Table 1: The academic networks for the Health and CS domains. Note that the data size is reduced due to the removal of isolated nodes.

	Node and Links	**Counts**
Health	Paper	15,366,456
	Author	12,682,306
	Paper-Paper link	479,358,572
	Author-Paper link	139,435,221
CS	Paper	1,286,254
	Author	1,004,536
	Paper-Paper link	8,024,869
	Author-Paper link	6,945,771

The second is the CS data from ArnetMiner Academic Social Network (AMiner) (Tang et al., 2008). The CS data integrates publications from DBLP and citation links from ACM Digital Library, CiteSeer, and other sources . It consists of 2,092,356 papers, 8,024,869 citation links, and 1,712,433 distinct authors. The numbers in the network are reduced after removing isolated papers.

Fig 3 shows the reference and citation distributions of papers in CS and Health dataset. As expected, both citations and references have a long tail that resembles a power-law distribution. We use the maximum likelihood estimation (Clauset et al., 2009) to estimate the power-law exponents. The probability density function (PDF) is computed and plotted on the figure as α.

Our experiments are carried out on two servers. Each one equips with 24-core CPU and 256GB memory. The complexity of PageRank-based algorithms depends on the number of edges in the graph. The largest graph in our experiment contains about 600 million edges, which can be loaded into the memory easily. The code and data can be accessed on our webpage[2].

4.1 Compared Metrics

We compare our method with *Co-Ranking*(Zhou et al., 2007), P(paper count), C(citation count)(Gross and Gross, 1927), H(H-index)(Hirsch, 2005), G(G-index)(Egghe, 2006), SPR(summation of PageRank)(Fragkiadaki and Evangelidis, 2015), and their weighted versions C_w and SPR_w(Lindsey, 1982). Weighted metrics split credits among co-authors. For an author a, P is the total number of papers that an author has published. Other indexes for a are defined as:

[2] `http://zhao15m.myweb.cs.uwindsor.ca/apr/`

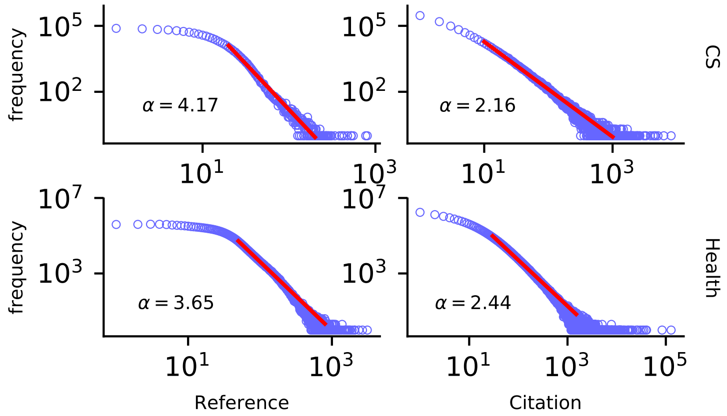


Fig. 3: Reference and citation distributions of CS and Health dataset. The x-axes represent the number of references or citations. The y-axes denote the frequency of references or citations.

$$C(a) = \sum_{p \in a} CitationCount_p \tag{4}$$

$$C_w(a) = \sum_{p \in a} \frac{CitationCount_p}{AuthorCount_p} \tag{5}$$

$$SPR(a) = \sum_{p \in a} PR_p \tag{6}$$

$$SPR_w(a) = \sum_{p \in a} \frac{PR_p}{AuthorCount_p} \tag{7}$$

Here PR_p is the PageRank value for the paper p in citation network.

For SPR, two damping factors are tested (0.85 and 0.5). 0.85 is the empirically best damping factor suggested by Brin and Page (1998) for web page ranking. $\alpha = 0.5$ was suggested by Chen et al. (2007) to offset the acyclic problem in citation network. Our APR method uses the default $\alpha = 0.85$ since there are already loops in our heterogeneous network. For the sake of simplicity, we adopt this parameter in consistency with $SPR0.85$ method. Authority Ranking in (Sun et al., 2009) is not compared because it uses co-author and co-conference links. Citation information is not included.

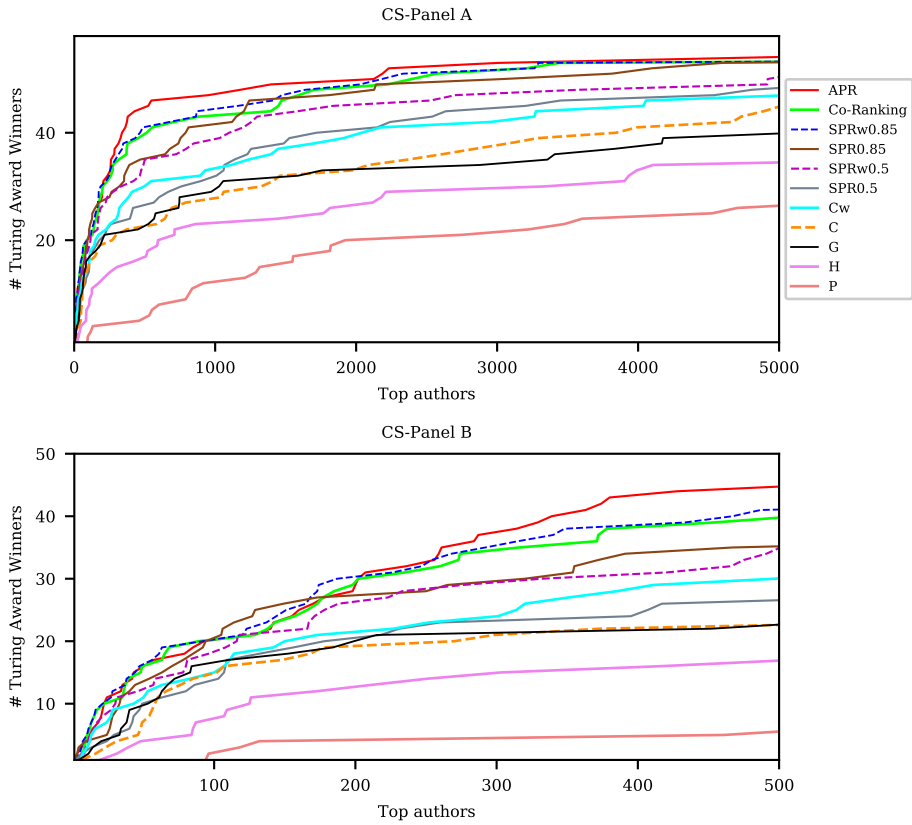


Fig. 4: Number of award winners among top-k authors on CS dataset.

5 Results

5.1 How Good Is APR?

We evaluate the ranking results using the number of award winners within top-k authors in Fig 4 and Fig 5. To quantify the difference among these methods, we treat each line as a ROC(Receiver Operating Characteristic) curve (Hanley and McNeil, 1982), then AUC(Area Under the Curve) can be calculated from each curve. The AUC values of the ROC curves in Fig 4 Panel A and Fig 5 Panel A are listed in Table 2. The awards are Nobel Prize for the health data and Turing Award for the CS data. In the figure, Panel A is the global view of top authors. Panel B is a zoom-in for the starting section that contains the top 500 for CS and top 10,000 for the Health data. We can see that APR outperforms all other methods consistently in both CS and Health data. Table 3 is the number of Turing and Nobel winners within top authors on two

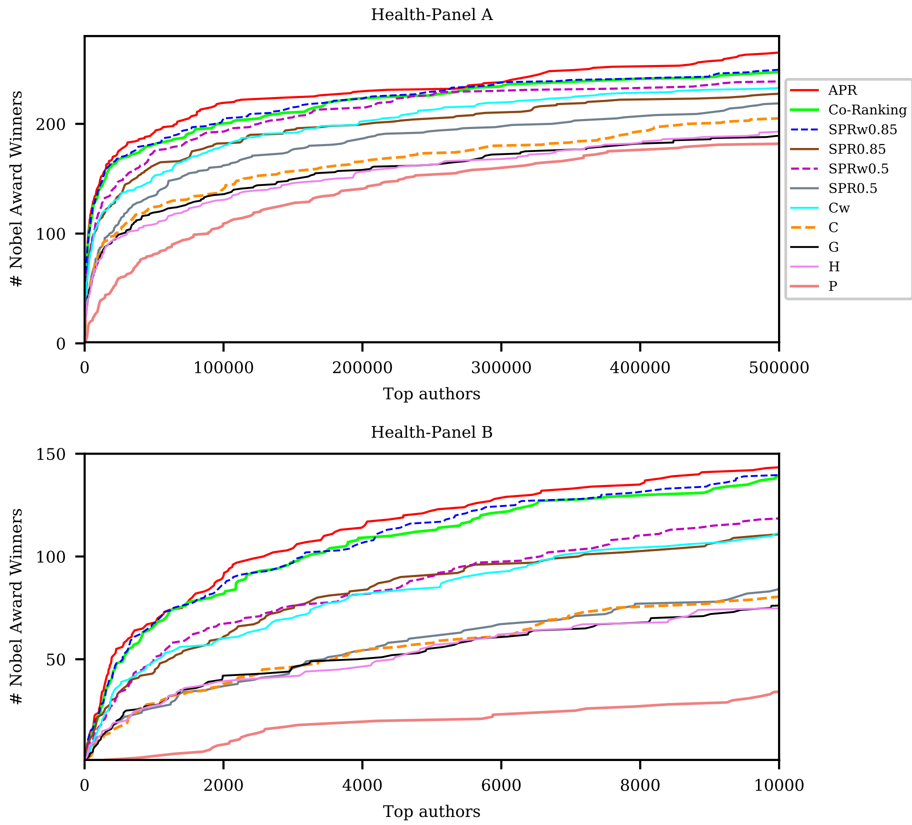


Fig. 5: Number of award winners among top-k authors on Health dataset.

datasets. APR performs best almost within every range. Table 4 and 5 list the top 25 authors ranking by APR and their indexes in other metrics on CS and Health dataset.

From the plots, especially Panel B of the Health data, we can see that the methods fall into roughly four groups. The baseline is P. Without question, it gives the lowest performance. Above that, we see a group that contains of H, G and C, which are citation-based methods. As expected, G-index is indeed an improvement of H-index. Both G and H cannot compete with C in most cases, probably because they over-simplified the citation data.

PageRank-based algorithms outperform citation-based algorithms with $\alpha = 0.85$. Sitting in between Citation-based and PageRank-based method are $SPR0.5$, which is a special case of PageRank algorithm with high random jumping probability ($\alpha = 0.5$). We shall understand that PageRank is a spectrum algorithm. When α is smaller, there is a higher probability of random jump. Thus the

Table 2: AUC values of the ROC curves in Fig 4 Panel A and Fig 5 Panel A.

Methods	CS		Health	
	AUC (1e3)	Improvement	AUC (1e6)	Improvement
APR	246.8	–	114.8	–
P	91.8	169.07%	70.2	63.53%
C	170.1	45.10%	82.9	38.49%
C_w	195.1	26.51%	100.5	14.14%
H	135.0	82.90%	77.7	47.72%
G	163.4	51.13%	78.6	45.91%
SPR0.85	231.0	6.87%	99.0	15.87%
$SPR_w0.85$	240.5	2.65%	110.4	3.93%
SPR0.5	198.6	24.30%	91.1	25.97%
$SPR_w0.5$	216.9	13.83%	106.0	8.27%
Co-Ranking	238.1	3.66%	109.4	4.88%

Table 3: Number of Turing/Nobel award winners within top authors.

Methods	CS					Health				
	50	100	200	500	1,000	100	500	1,000	2,000	10,000
APR	**16**	**20**	29	**45**	**48**	**15**	**55**	**67**	**90**	**143**
PN	0	2	4	5	12	0	1	2	8	34
C	8	14	19	22	27	3	17	28	38	80
C_w	11	15	21	30	33	8	37	48	60	111
H	4	7	12	16	23	9	19	27	39	74
G	9	16	20	22	29	4	20	27	42	76
SPR0.85	13	**20**	27	35	41	12	34	43	61	110
$SPR_w0.85$	16	**20**	**30**	41	44	**15**	48	**67**	86	139
SPR0.5	10	13	20	26	31	8	19	26	36	84
$SPR_w0.5$	12	18	26	34	38	10	34	50	67	118
Co-Ranking	**16**	**20**	29	39	43	13	48	66	81	139

algorithm favors more authors with more papers or citations. For C, $SPR0.5$, and $SPR0.85$, their weighted versions are consistently better.

Fig 6 shows the top-100 APR authors along with their rankings in terms of citation count. It shows that 1) APR can identify many (20) Turing award winners; 2) Correlation between APR and C is low. For instance, Marvin Minsky is the 1461-st most cited author, but our APR rank is 35. This prompts us to explore how different APR is from other methods.

5.2 How Different Is APR?

Fig 7 shows the pair-wise Spearman's rank correlation coefficient among 11 methods for the top 100 authors. The top authors are determined by their APR values. When we extend the list to include more authors, the correlation coefficients will increase, but the pattern discussed below is similar.

We can observe that the metrics differ with each other greatly, especially with APR. APR differs from the other methods the most, probably because

Table 4: Top 25 authors by APR on CS Dataset and their indexes in other metrics. Bold names are Turing Award winners.

Rank	Name	APR (10^{-3})	P	C	C_w	H	G	SPR0.85 (10^{-3})	SPR_w0.85 (10^{-3})	SPR0.5 (10^{-3})	SPR_w0.5 (10^{-3})	CoRanking (10^{-3})
1	**Donald E. Knuth**	7.76	119	6,538	6,040	27	81	16.36	15.58	6.65	6.19	3.03
2	J. Ross Quinlan	3.49	35	8,500	8,321	19	35	8.85	8.60	4.60	4.48	1.48
3	**E. W. Dijkstra**	3.26	59	4,275	3,791	17	59	5.87	5.36	2.80	2.53	1.05
4	G. Salton	3.05	119	8,552	5,129	27	93	11.84	7.89	5.57	3.60	1.31
5	Jeffrey D. Ullman	2.57	203	14,868	6,854	49	121	15.88	7.48	8.08	3.74	1.48
6	**Judea Pearl**	2.56	130	6,819	5,764	25	83	5.50	4.60	3.79	3.13	0.93
7	V. N. Vapnik	2.50	40	10,733	7,577	24	40	8.11	5.64	5.14	3.59	1.23
8	S Haykin	2.17	28	4,770	4,698	12	28	3.75	3.67	2.89	2.80	0.78
9	David E. Goldberg	2.14	152	9,123	7,618	26	95	6.80	5.69	5.12	4.13	1.22
10	**C. A. R. Hoare**	2.10	94	6,558	5,540	22	81	7.42	6.42	3.65	3.09	1.13
11	Milton Abramowitz	1.88	1	1,450	1,450	1	1	1.99	1.99	1.36	1.36	0.42
12	**John Hopcroft**	1.87	100	6,487	2,949	22	81	9.80	4.59	4.78	2.21	0.90
13	David S Johnson	1.67	61	10,915	5,439	22	61	8.97	4.45	5.48	2.69	0.91
14	**John McCarthy**	1.64	54	1,773	1,328	13	43	6.34	3.74	2.20	1.38	0.69
15	Alfred V. Aho	1.59	63	7,862	3,358	25	63	11.09	5.05	5.11	2.30	1.00
16	John H. Holland	1.57	35	3,837	3,535	17	35	2.92	2.66	2.08	1.87	0.57
17	M. R. Garey	1.50	26	9,408	4,587	13	26	8.05	3.91	4.79	2.32	0.80
18	**Ronald L. Rivest**	1.48	141	12,177	4,284	32	111	13.60	5.10	6.94	2.60	1.02
19	Thomas M. Cover	1.48	36	4,540	2,287	11	36	3.63	1.84	2.32	1.18	0.39
20	**Leslie Lamport**	1.43	113	8,084	6,207	33	90	7.73	5.61	3.96	2.94	1.11
21	**Robin Milner**	1.39	75	7,258	5,477	29	75	4.07	3.00	2.60	1.95	0.59
22	**Allen Newell**	1.37	70	3,551	1,991	18	60	6.70	3.36	2.96	1.55	0.68
23	Nils J. Nilsson	1.37	24	1,912	1,631	14	24	2.38	2.11	1.35	1.18	0.44
24	**Niklaus Wirth**	1.36	67	2,345	1,630	21	49	4.71	3.14	2.31	1.56	0.60
25	Jakob Nielsen	1.35	85	4,077	3,420	28	64	4.56	3.70	2.79	2.27	0.80

Table 5: Top 25 authors by APR on Health Dataset and their indexes in other metrics. Bold names are Nobel Award winners.

Rank	Name	APR (10^{-4})	P	C	C_w	H	G	SPR0.85 (10^{-4})	SPR_w0.85 (10^{-4})	SPR0.5 (10^{-4})	SPR_w0.5 (10^{-4})	CoRanking (10^{-4})
1	Ulrich K Laemmli	16.88	71	136,247	131,505	41	71	25.22	24.46	12.59	12.20	5.10
2	CDC	3.03	710	8,923	8,432	38	67	1.94	1.90	1.76	1.72	0.40
3	H R EAGLE	2.34	84	6,134	4,825	29	79	4.27	3.43	1.45	1.14	0.66
4	CCP4	2.09	1	8,391	8,391	1	1	0.62	0.62	0.36	0.36	0.15
5	M M Bradford	2.01	6	83,952	83,921	4	6	7.28	7.27	4.35	4.34	1.52
6	A Robert Neurath	1.96	136	2,836	738	27	50	3.94	3.56	1.58	1.29	0.90
7	Marta Hamilton	1.91	91	19,776	18,593	21	91	2.21	2.02	1.35	1.20	0.43
8	Shelley McGuire	1.84	29	8,272	8,272	25	29	0.79	0.79	0.51	0.51	0.17
9	Jean L Marx	1.72	521	4,734	4,390	30	47	1.23	1.17	1.16	1.10	0.22
10	Robert F Service	1.65	370	5,324	5,298	33	62	1.04	1.03	0.95	0.95	0.20
11	Eliot Marshall	1.56	572	4,193	3,905	24	52	1.15	1.08	1.19	1.12	0.20
12	**John R Vane**	1.51	368	32,565	14,006	81	174	7.87	3.63	3.71	1.66	0.73
13	Richard A Kerr	1.43	698	3,452	3,370	21	38	1.11	1.09	1.28	1.26	0.19
14	**G E Palade**	1.41	176	22,365	12,188	77	149	11.56	7.51	3.53	2.07	1.37
15	**David Baltimore**	1.38	490	55,514	18,399	122	223	7.94	2.88	3.92	1.38	0.63
16	William Bernard Kannel	1.33	444	58,111	16,715	118	234	7.35	2.26	4.27	1.34	0.47
17	Edwin M Southern	1.25	81	17,377	15,297	31	81	4.75	4.33	1.85	1.62	0.93
18	Yasutomi Nishizuka	1.22	146	18,142	13,492	48	135	2.91	2.17	1.65	1.18	0.45
19	Scott Kirkpatrick	1.19	2	11,056	3,688	2	2	1.87	0.62	1.06	0.36	0.11
20	**Frederick Sanger**	1.19	72	41,887	12,930	30	72	13.48	4.20	4.67	1.47	0.90
21	John P Perdew	1.19	83	49,596	22,137	35	83	3.73	1.71	2.55	1.16	0.32
22	W J Rutter	1.14	211	25,853	6,580	66	160	6.81	1.74	2.93	0.76	0.38
23	**J L Goldstein**	1.12	375	42,660	12,554	106	200	7.46	2.26	3.66	1.09	0.49
24	Elizabeth Pennisi	1.11	392	2,971	2,887	25	39	0.71	0.69	0.76	0.74	0.14
25	Michael Scott Brown	1.10	447	49,342	14,014	117	211	7.65	2.27	3.85	1.12	0.50

it is the only method that includes authors in the heterogeneous network. For the Health data, the highest positive correlation happens between APR and Eigen-vector based methods such as *Co-Ranking* (correlation coefficient $\rho = 0.26$) and $SPR_w 0.50$ ($\rho = 0.45$). It is expected that APR correlates with these methods since all of them are based on random walk interpretation. It is surprising that the closest correlation coefficient is only 0.29(with $SPR_w 0.85$) for Health, and 0.55 for CS(with *Co-Ranking*). Both are quite low, indicating that the ranking results are very different. What is even more interesting is that in Health data, APR correlates with several indexes negatively, including

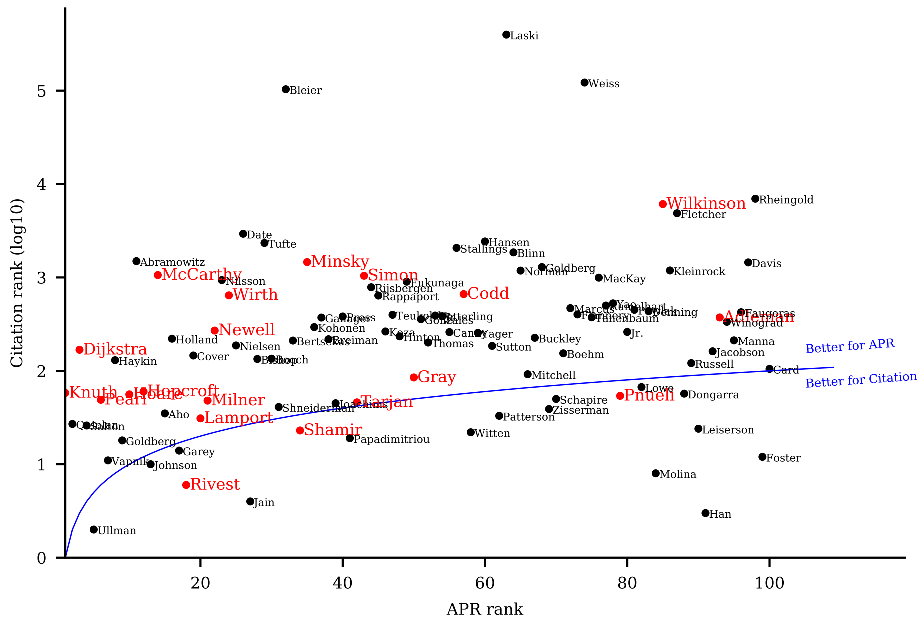


Fig. 6: Top 100 APR vs. their citation rankings.

H-index ($\rho = -0.17$), G-index ($\rho = -0.15$), and Paper count (($\rho = -0.08$). Among the top authors, the more influential you are, the fewer papers you write. This pattern also extends to the CS data.

Fig 5 illustrates the satisfied layers of the metrics. There are a few close-pairs. For instance, *Co-Ranking* and $SPR_w0.85$ correlate almost perfectly ($\rho = 0.99$). This can be explained by the fact that *Co-Ranking* runs random walks on three disparate networks. When they do the random walk on the citation network, it equals to calculate the PageRank values for the citation matrix independent of the author network. Their combination with the author network is merely summing up the PageRank values for each author from the citation network. Another group includes indexes H, G, and extends slightly to C. They are citation-based metrics.

Next, we look at the cause of the difference. In Fig 8 and 9, each subplot is the rank value against the rank for each metric. The rank values are normalized so that they sum up to one. This way we can compare them on the same scale. APR is plotted in every subplot as a reference (the red line). We can see that the weights (ranking values) of authors have a long tail distribution that resembles a power-law. That means that the top authors collect most of the weights, while a large majority of the authors have very small weights. Although the pattern is the same across all the metrics for both datasets, the slopes are different, indicating the in-equalities are different. For the CS data,

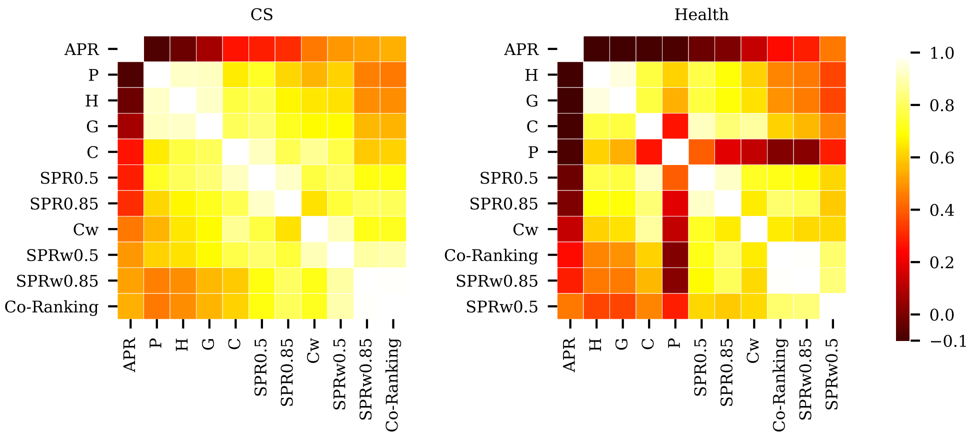


Fig. 7: Correlation between 11 ranking metrics. APR correlates with P, C, H, G negatively for the top 100 authors.

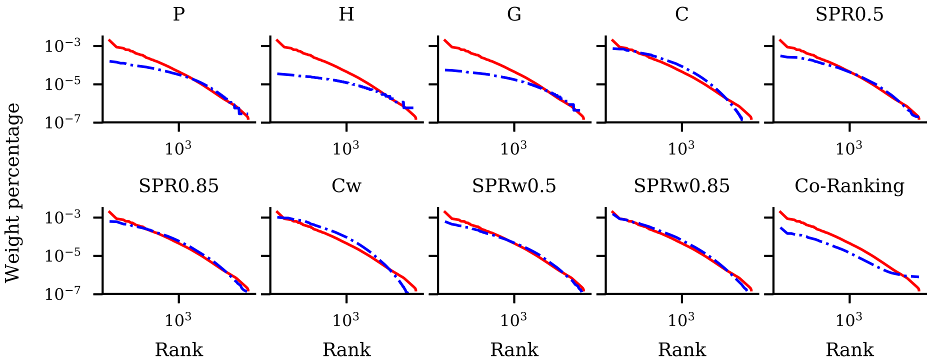


Fig. 8: Distribution of ranking values on CS dataset. APR (the solid line) vs. other methods.

the coefficient of variation (CV) of weights is 23.61 for APR, merely 1.13 for H, 2.25 for G, and 6.42 for P. When the variation of the weights are small, it won't be easy to tell the difference between the authors. That may explain why those metrics are not good. Among the top 10,000 authors, the Gini coefficient is 0.42 for APR, 0.36 for H, and 0.38 for G.

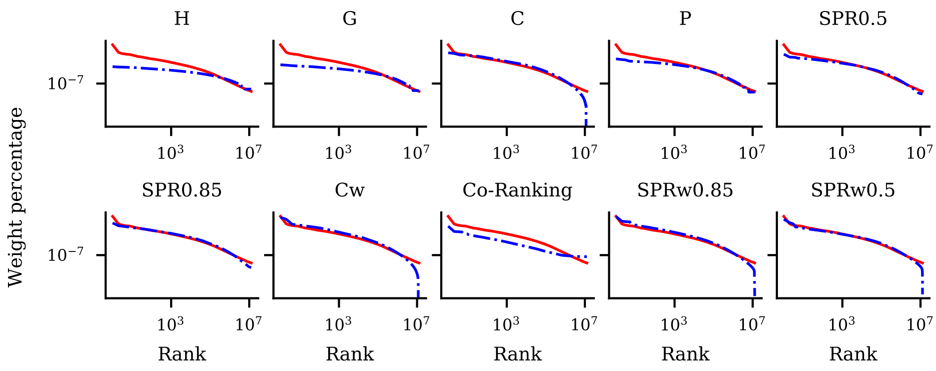


Fig. 9: Distribution of ranking values on Health dataset. APR (the solid line) vs. other methods.

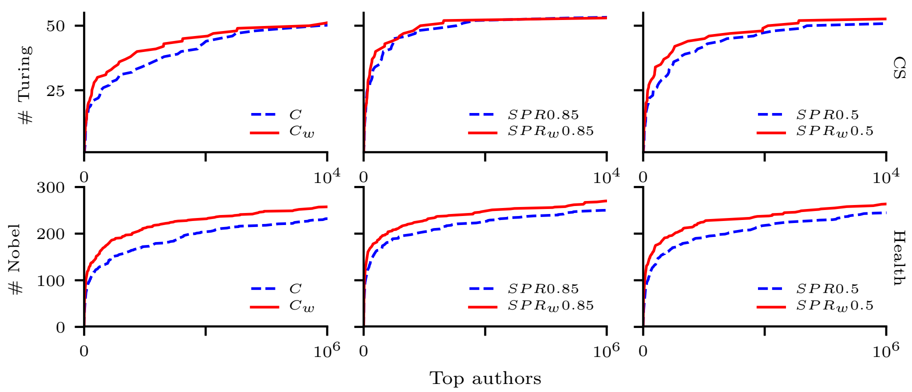


Fig. 10: Weighted ranking methods outperform the unweighted versions consistently for both CS and Health data.

5.3 Weighted vs. Unweighted Methods

Weighted methods share credits between multiple authors of a paper. They reflect authors' contribution in a more accurate way. This is verified in both data sets as illustrated Fig 10. In the figure we can see that weighted methods outperform the corresponding unweighted versions consistently along all the top authors.

6 Results of Paper Ranking

Since APR is running on the heterogeneous network, it also gives the ranking for papers. Table 6 lists the 15 highly ranked papers in CS dataset, and Table 7 lists the 15 highly ranked papers in Health dataset. We also list the citation number for each paper. Our evaluation rule is only valid for authors. It is hard to measure the quality of papers. While we still find that papers' APR values and citation numbers are highly positively correlated, with Pearson's correlation coefficient 0.78 and Spearman's correlation coefficient 0.49 in CS data; 0.82 and 0.72 in Health data.

Table 6: Top 15 papers in CS Dataset. Bold names are Turing awards Winners and their papers.

Rank	Title	Author names	Citation
1	Computers and Intractability: A Guide to the Theory of NP Completeness	M. R. Garey;David S Johnson	8,166
2	Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables	Milton Abramowitz	1,450
3	Genetic Algorithms in Search, Optimization and Machine Learning	David E. Goldberg	6,272
4	**The Design and Analysis of Computer Algorithms**	Alfred V. Aho;**John Hopcroft**	1,945
5	The nature of statistical learning theory	V. N. Vapnik	5,100
6	**A method for obtaining digital signatures and public key cryptosystems**	**Ronald L. Rivest**; **A Shamir**; **L M Adleman**	2,085
7	C4.5: programs for machine learning	J. Ross Quinlan	4,674
8	**The art of computer programming, volume 3: (2nd ed.) sorting and searching**	**Donald E. Knuth**	1,609
9	**A relational model of data for large shared data banks**	**E. F. Codd**	1,497
10	Elements of information theory	Thomas M. Cover;J. Thomas	3,341
11	Treating hierarchical data structures in the SDC Time Shared Data Management System (TDMS)	Robert E. Bleier	27
12	**Probabilistic reasoning in intelligent systems: networks of plausible inference**	**Judea Pearl**	3,230
13	Pattern Classification (2nd Edition)	R O Duda;Peter E Hart;D G Stork	4,293
14	**The art of computer programming, volume 2 (3rd ed.): seminumerical algorithms**	**Donald E. Knuth**	1,235
15	**Report on the algorithmic language ALGOL 60**	**J Backus**;**John McCarthy**; **Alan J. Perlis**...	146

7 Discussions and Conclusions

This paper proposes Author PageRank (APR) as a method for measuring academic influence of authors in a heterogeneous author-citation network. We demonstrate that it outperforms 10 other methods on two very large data sets. We also show that the ranking results differ greatly with all the other methods.

To the best of our knowledge, this is the first attempt in integrating authors and papers in a coherent academic network. In the past, various approaches have been tried to add author data into citation network. When treating citation and author ranking separately in the case of Co-Ranking, we show that their result is actually the same as the PageRank on the citation network alone. When transforming the heterogeneous network into an author-citation

Table 7: Top 15 papers in Health Dataset. Bold names are Nobel awards Winners and their papers.

Rank	Title	Author names	Citation
1	Cleavage of structural proteins during the assembly of the head of bacteriophage T4	Ulrich K Laemmli	128,117
2	A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding	M M Bradford	83,911
3	**DNA sequencing with chain-terminating inhibitors**	**Frederick Sanger**; S Nicklen...	30,432
4	Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications	T Staehelin;John Gordon;Harry Towbin	26,244
5	A short history of SHELX	George M Sheldrick	40,521
6	Detection of specific sequences among DNA fragments separated by gel electrophoresis	Edwin M Southern	13,583
7	Gapped BLAST and PSI-BLAST: a new generation of protein database search programs	Zong Hong Zhang;Thomas L Madden...	35,770
8	Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease	W J Rutter;John M Chirgwin...	11,263
9	Optimization by simulated annealing	Scott Kirkpatrick;C D Gelatt;M P Vecchi	11,051
10	The use of lead citrate at high pH as an electron-opaque stain in electron microscopy	Erica Reynolds	13,984
11	The CCP4 suite: programs for protein crystallography	Collaborative Computational Project	8,391
12	CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice	Des G Higgins;Toby James Gibson...	33,092
13	Basic local alignment search tool	Wolfgang J Miller;Eugene W Myers...	26,627
14	A new method for sequencing DNA	A M Maxam;Wendy V Gilbert	3,653
15	Interference of sodium ethylenediaminetetraacetate in the determination of proteins and its elimination	A Robert Neurath	21

network, the resulting graph can be too large to be processed. There are computational challenges when carrying out PageRank-based algorithms due to the very large size of the data. Some methods based purely on author citation relations are not scalable (e.g., (Radicchi et al., 2009), (West et al., 2013)), hence they can not deal with data sets of our size. In the author-citation network, although the node number is reduced by containing authors only, the number of links can increase in orders of magnitude, depending on the average reference number and the average number of papers per author. We solve this computational problem by replacing a dense author-author graph with a sparse author-paper-author network, hence reducing the number of edges greatly. Probably this is the reason why we never see PageRank-like algorithms run on a very large author network.

Our algorithm can be tuned in many aspects. Now we adopt the most simple approach. For instance, the damping factor is the commonly used 0.85; random jump is uniformly random to all the nodes including authors and papers; the random walk moves to the next node among all its neighbors with equal probability. Improvements can be made to have different jumping probabilities. Fig 11 shows the impact of damping factor on the overall result for α ranging from 0.05 to value very close to 1 (note that if it is one, the eigenvector may not converge). We can see that while 0.85 is better than most other values, there is still space to improve. We may also assign different weights for authors and papers, so that a random walk from a paper will have a fixed probability to cited papers, and have the remaining probability to authors.

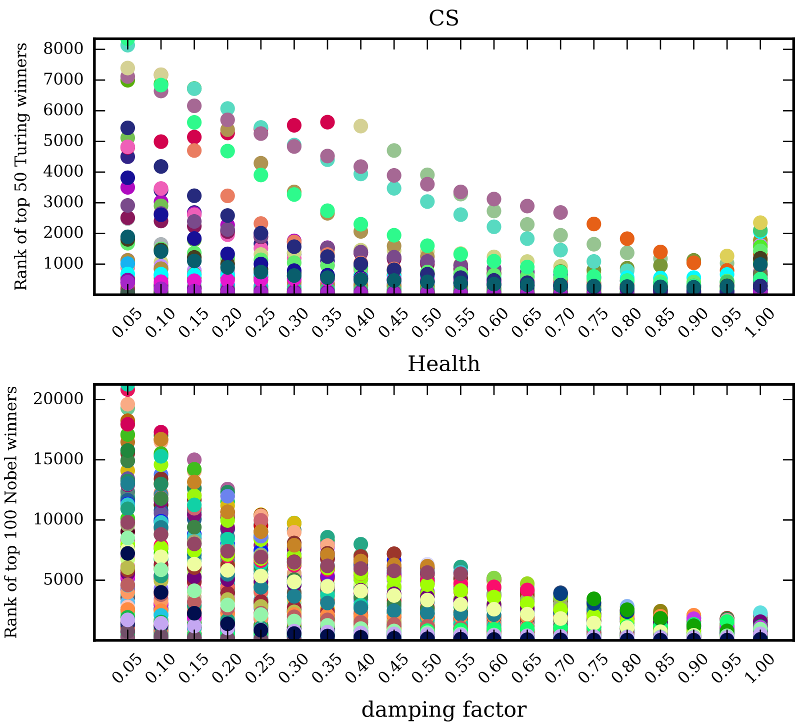


Fig. 11: The impact of damping factors on the ranking results. (top panel) 50 Turing winners in CS dataset; (bottom panel) 100 Nobel winners in Health dataset.

Academic networks are not restricted with authors and papers. We can add other entities, such as journals, conferences, and institutions into the ranking framework (Wang et al., 2013). Ranking is not limited to author's overall influence. Better ranking could be domain dependent (Yan, 2014), given that different areas have their own style of the publication. In additional to measuring influence, there are other aspects need to be reflected, such as an author's potential and impact in the future. We also plan to extend the ranking from authors to journal (Bergstrom et al., 2008) and institutes (Liu and Cheng, 2005). Those are the topics that we will continue to work on.

Acknowledgements The research is supported by NSERC Discovery Grant.

References

Amjad T, Daud A (2017) Indexing of authors according to their domain of expertise. Malaysian Journal of Library and Information Science 22:69–82

Amjad T, Daud A, Akram A (2015a) Mutual influence based ranking of authors 34

Amjad T, Ding Y, Daud A, Xu J, Malic V (2015b) Topic-based heterogeneous rank. Scientometrics 104(1):313–334

Amjad T, Daud A, Aljohani NR (2018) Ranking authors in academic social networks: a survey. Library Hi Tech 36(1):97–128

Bergstrom CT, West JD, Wiseman MA (2008) The EigenfactorTM Metrics. Journal of Neuroscience 28(45):11433–11434

Bibi F, Khan H, Iqbal T, Farooq M, Mehmood I, Nam Y (2018) Ranking authors in an academic network using social network measures. Applied Sciences 8(10):1824

Bollen J, Rodriquez MA, Van de Sompel H (2006) Journal status. Scientometrics 69(3):669–687

Bonacich P (1972) Factoring and weighting approaches to status scores and clique identification. The Journal of Mathematical Sociology 2(1):113–120

Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30(1-7):107–117

Chen P, Xie H, Maslov S, Redner S (2007) Finding scientific gems with Google's PageRank algorithm. Journal of Informetrics 1(1):8–15

Clauset A, Shalizi CR, Newman ME (2009) Power-law distributions in empirical data. SIAM review 51(4):661–703

Daud A, Aljohani NR, Abbasi RA, Rafique Z, Amjad T, Dawood H, Alyoubi KH (2017) Finding rising stars in co-author networks via weighted mutual influence. In: Proceedings of the 26th International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee, pp 33–41

Dellavalle RP, Schilling LM, Rodriguez MA, Van de Sompel H, Bollen J (2007) Refining dermatology journal impact factors using PageRank. Journal of the American Academy of Dermatology 57(1):116–119

Ding Y, Yan E, Frazho A, Caverlee J (2009) PageRank for ranking authors in co-citation networks. Journal of the American Society for Information Science and Technology 60(11):2229–2243

Egghe L (2006) Theory and practise of the g-index. Scientometrics 69(1):131–152

Fragkiadaki E, Evangelidis G (2015) Three novel indirect indicators for the assessment of papers and authors based on generations of citations. Scientometrics pp 1–38

Fragkiadaki E, Evangelidis G (2016) Three novel indirect indicators for the assessment of papers and authors based on generations of citations. Scientometrics 106(2):657–694

Gao C, Wang Z, Li X, Zhang Z, Zeng W (2016) Pr-index: using the h-index and pagerank for determining true impact. PloS one 11(9):e0161755

González-Pereira B, Guerrero-Bote VP, Moya-Anegón F (2010) A new approach to the metric of journals' scientific prestige: The SJR indicator. Journal of Informetrics 4(3):379–391

Gross PLK, Gross EM (1927) College Libraries and Chemical Education. Science 66(1713):385–389

Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology 143(1):29–36

Hirsch JE (2005) An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences 102(46):16569–16572

Liang R, Jiang X (2016) Scientific ranking over heterogeneous academic hypernetwork. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI Press, AAAI'16, pp 20–26

Lindsey D (1982) Further evidence for adjusting for multiple authorship. Scientometrics 4(5):389–395

Liu NC, Cheng Y (2005) The Academic Ranking of World Universities. Higher Education in Europe 30(2):127–136

Liu X, Bollen J, Nelson ML, Van de Sompel H (2005) Co-authorship networks in the digital library research community. Information Processing & Management 41(6):1462–1480

Ma N, Guan J, Zhao Y (2008) Bringing PageRank to the citation analysis. Information Processing & Management 44(2):800–810

Radicchi F, Fortunato S, Markines B, Vespignani A (2009) Diffusion of scientific credits and the ranking of scientists. Physical Review E 80(5)

Sayyadi H, Getoor L (2009) FutureRank: Ranking Scientific Articles by Predicting their Future PageRank. In: Apte C, Park H, Wang K, Zaki MJ (eds) Proceedings of the 2009 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics, Philadelphia, PA, pp 533–544, DOI 10.1137/1.9781611972795.46

Sidiropoulos A, Manolopoulos Y (2006) Generalized comparison of graph-based ranking algorithms for publications and authors. Journal of Systems and Software 79(12):1679–1700

Steinbrüchel C (2018) A citation index for principal investigators. Scientometrics

Su C, Pan Y, Zhen Y, Ma Z, Yuan J, Guo H, Yu Z, Ma C, Wu Y (2011) PrestigeRank: A new evaluation method for papers and journals. Journal of Informetrics 5(1):1–13

Sun Y, Han J, Zhao P, Yin Z, Cheng H, Wu T (2009) RankClus: Integrating clustering with ranking for heterogeneous information network analysis. In: Proceedings of the 12th International Conference on Extending Database Technology Advances in Database Technology - EDBT '09, ACM Press, Saint Petersburg, Russia, p 565

Tang J, Zhang J, Yao L, Li J, Zhang L, Su Z (2008) ArnetMiner: Extraction and mining of academic social networks. In: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08, ACM Press, Las Vegas, Nevada, USA, p 990

Usmani A, Daud A (2017) Unified author ranking based on integrated publication and venue rank. International Arab Journal of Information Technology (IAJIT) 14(1)

Walker D, Xie H, Yan KK, Maslov S (2007) Ranking scientific publications using a model of network traffic. Journal of Statistical Mechanics: Theory and Experiment 2007(06):P06010–P06010

Wang Y, Tong Y, Zeng M (2013) Ranking scientific articles by exploiting citations, authors, journals, and time information. In: Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, AAAI Press, AAAI'13, pp 933–939

West JD, Jensen MC, Dandrea RJ, Gordon GJ, Bergstrom CT (2013) Author-level Eigenfactor metrics: Evaluating the influence of authors, institutions, and countries within the social science research network community. Journal of the American Society for Information Science and Technology 64(4):787–801

Yan E (2014) Topic-based Pagerank: Toward a topic-level scientific evaluation. Scientometrics 100(2):407–437

Yan E, Ding Y (2011) Discovering author impact: A PageRank perspective. Information Processing & Management 47(1):125–134

Zhou D, Orshanskiy SA, Zha H, Giles CL (2007) Co-ranking Authors and Documents in a Heterogeneous Network. In: Seventh IEEE International Conference on Data Mining (ICDM 2007), IEEE, Omaha, NE, USA, pp 739–744

Zhou J, Zeng A, Fan Y, Di Z (2016) Ranking scientific publications with similarity-preferential mechanism. Scientometrics 106(2):805–816