Query Independent Scholarly Article Ranking

Shuai Ma, **Chen Gong**, Renjun Hu, Dongsheng Luo, Chunming Hu, Jinpeng Huai

SKLSDE Lab, Beihang University, China

Beijing Advanced Innovation Center for Big Data and Brain Computing

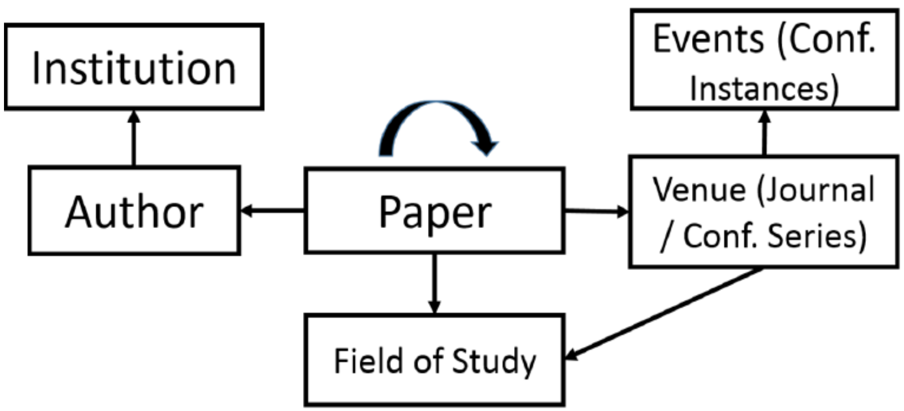
Query Independent Scholarly Article Ranking

- Goal: giving static ranking based on scholarly data only
- Applications
  - Playing a key role in literature recommendation systems, especially in the cold start scenario
  - For search engines, determining the ranking of results

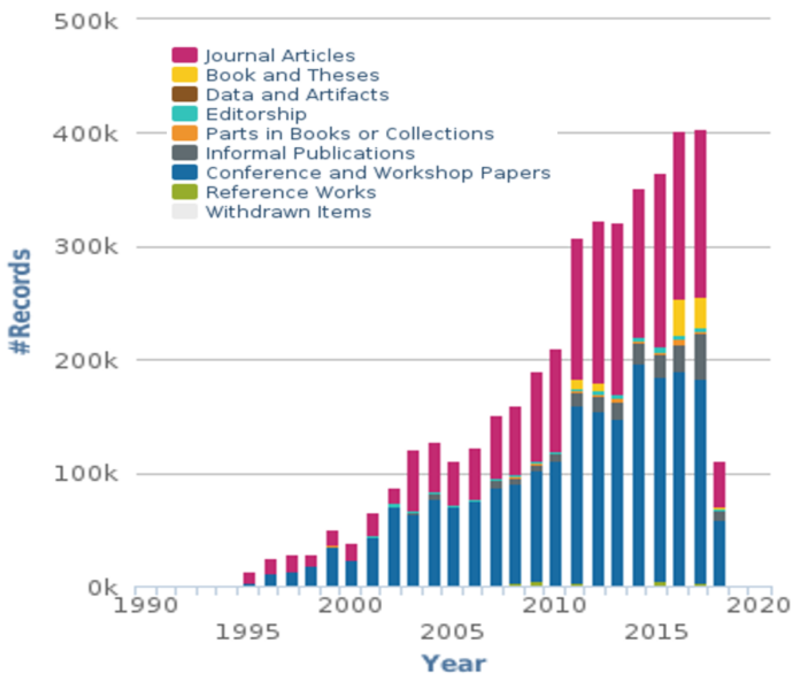
WSDM Cup 2016 http://www.wsdm-conference.org/2016/wsdm-cup.html

Challenges

- Heterogeneous, evolving & dynamic
  - Multiple types of entities involve with different contributions
  - Entities and their importance evolve with time
  - Academic data is dynamic and continuously growing



The Microsoft Academic Graph [Sinha et al. 2015]



New Records per year of dblp Database

Arnab Sinha, et al. An Overview of Microsoft Academic Service (MAS) and Applications. In WWW, 2015.
https://dblp.uni-trier.de/statistics/newrecordsperyear.html

Outline

- Ranking Model
  - Our Time Weighted PageRank
  - Ranking with Importance Assembling
- Ranking Computation
- Dynamic Ranking Computation
- Experimental Study
- Summary

Why Weighted PageRank?

- Traditional PageRank
  - Assumption of equally propagating
    - Articles are equally influenced by references
  - Bias: favor older articles while underestimate new ones

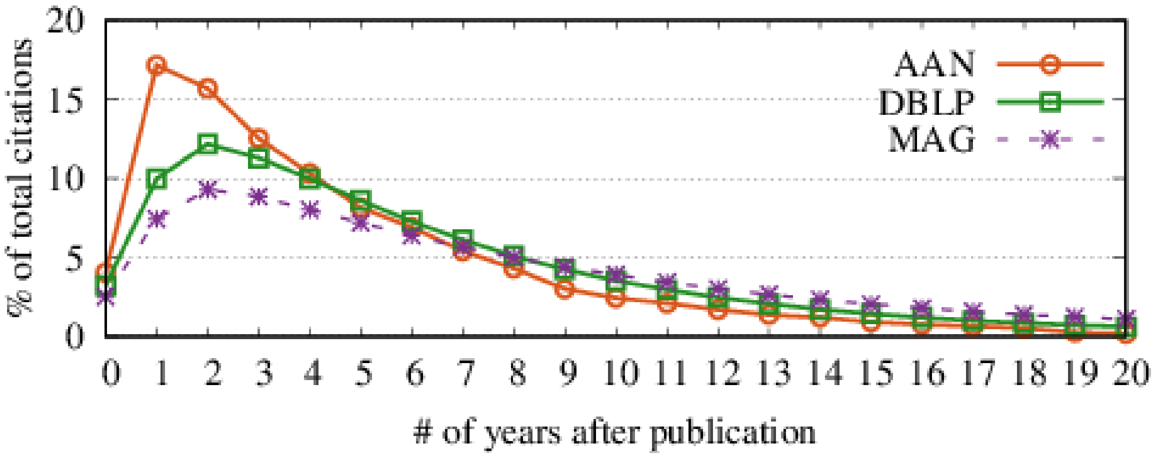
- Not all citations are equal [Valenzuela et al. 2015]
  - Different articles typically have different impacts

- Weighted PageRank
  - Key: how to determine the weights (differentiate impacts)

M. Valenzuela, V. Ha and O. Etzioni. Identifying Meaningful Citations. In AAAI Workshop, 2015.

Intuitions of Impacts of Articles

- Time decaying



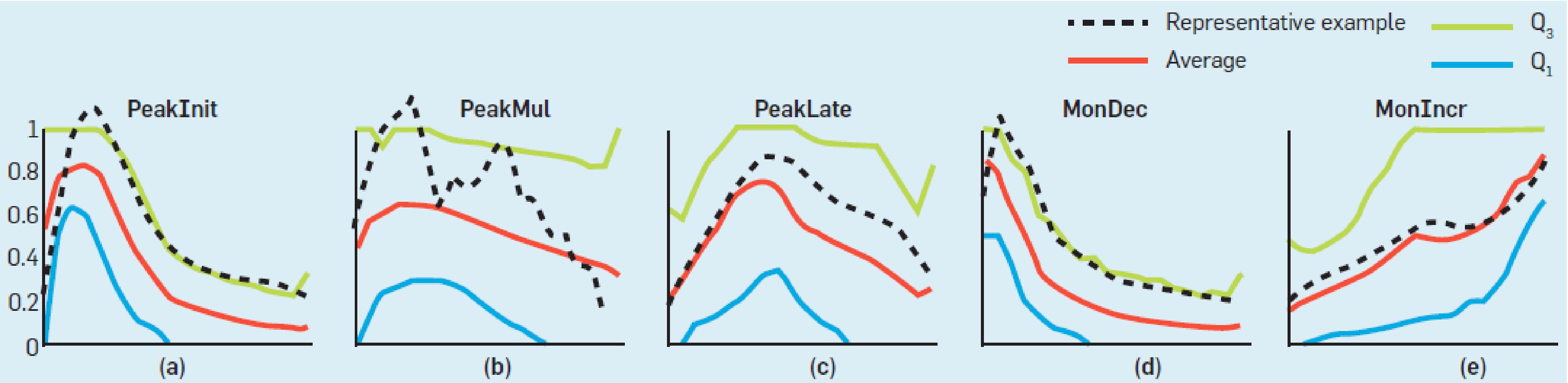
- Most previous work simply decays exponentially [1-4]

When to decay?

[1] X. Li, B. Liu and P. Yu. Time sensitive ranking with application to publication search. In ICDM, 2008.
[2] Y. Wang et al. Ranking scientific articles by exploiting citations, authors, journals and time information. In AAAI, 2013.
[3] H. Sayyadi and L. Getoor. Future rank: Ranking scientific articles by predicting their future pagerank. In SDM, 2009.
[4] D. Walker et al. Ranking scientific publications using a model of network traffic. Journal of Statistical Mechanics: Theory and Experiment, 2007.

When to Decay

- Different patterns for different articles [Chakraborty et al. 2015]
  - Categorized by when articles reach their citation peaks
  - PeakInit, PeakMul, PeakLate, MonDec, MonIncr, Other



Different Citation Patterns[Chakraborty et al. 2015]

Decaying only after the peak time of each individual article

Tanmoy Chakraborty, Suhansanu Kumar, Pawan Goyal, Niloy Ganguly, et al. On the categorization of scientific citation profiles in computer sciences. *Commun. ACM* 2015.

Our Time-Weighted PageRank

- Importance propagation based on time-weighted impacts
- Time-weighted impact

$$w(u, v) = \begin{cases} 1, & T_u < Peak_v \\ e^{\sigma(T_u - Peak_v)}, & T_u \geq Peak_v \end{cases}$$

T_u: time of paper u, $Peak_v$: peak time of paper v, σ: decaying factor

  - Decaying with time only after the peak time
  - Each individual article has its own peak time

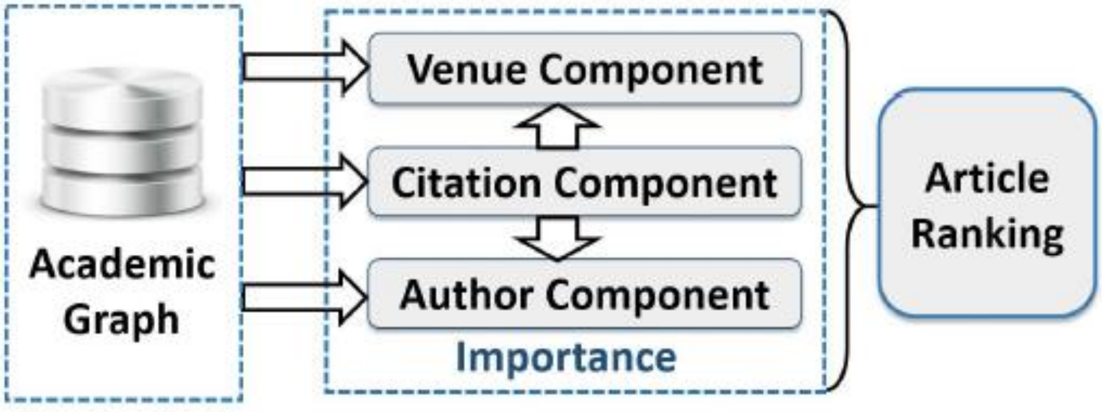
- Remarks
  - Considering the temporal information and dynamic impacts
  - Alleviating the bias through decayed time-weighted impacts

Outline

- Ranking Model
  - Our Time Weighted PageRank
  - Ranking with Importance Assembling
- Ranking Computation
- Dynamic Ranking Computation
- Experimental Study
- Summary

Why Importance Assembling?

- Cold start case: ranking new articles
  - No citations yet: only using citation information fails
  - Venue and author information should be incorporated
- Observation
  - Multiple types of entities involve with different contributions
- Assembling the different contributions of citation, venue and author components



Ranking with Importance Assembling

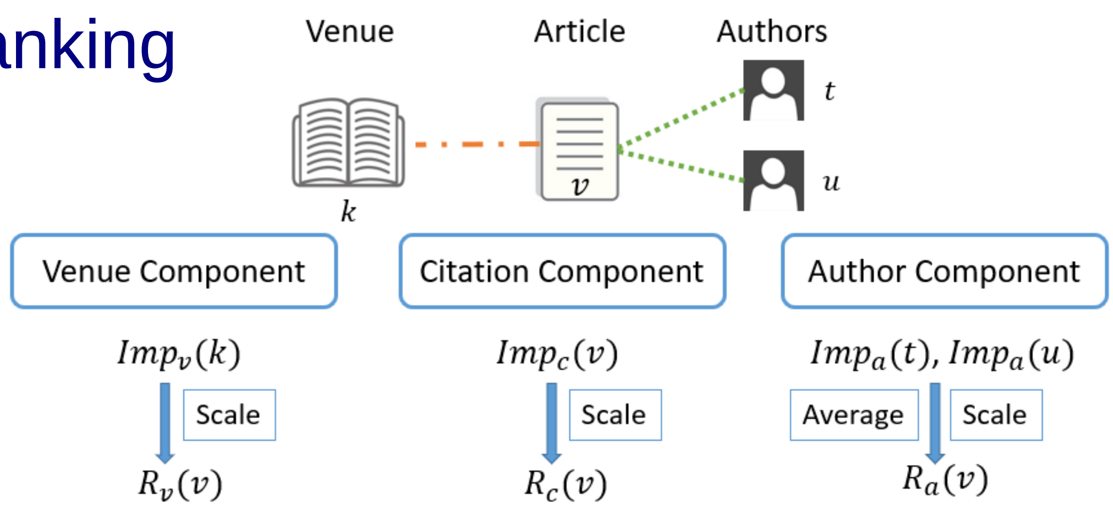
- Importance is defined as a combination of the prestige and popularity

favoring those with recent citations

$$Imp(v) = Prs(v)^{\lambda} Pop(v)^{1-\lambda}$$, λ: importance weighing factor

favoring those with citations soon after publication

- Final ranking

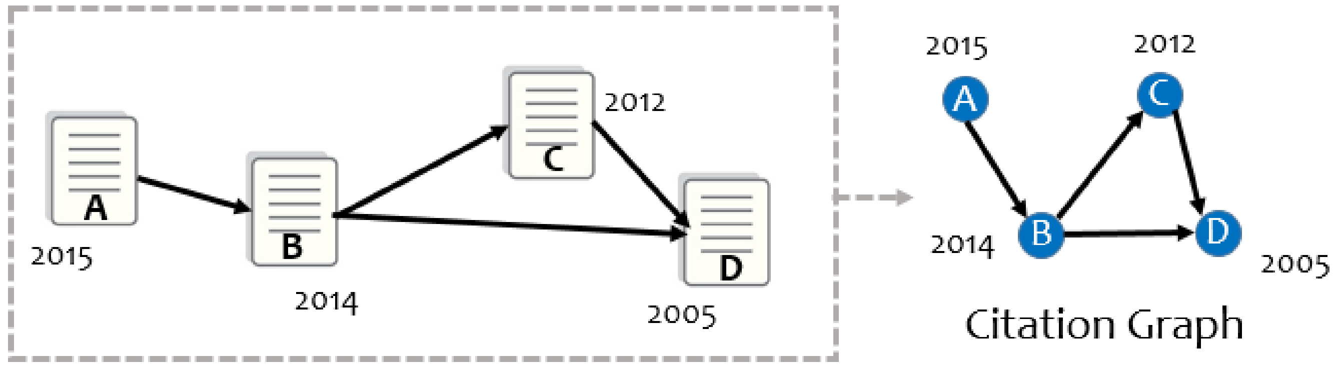


$$R(v) = \alpha R_c(v) + \beta R_v(v) + (1 - \alpha - \beta) R_a(v)$$

α and β: aggregating parameters

Importance Computation

- Citation component



  - Prs_c of article v is its TWPageRank score on the citation graph
  - Pop_c of article v is the sum of its citation freshness

$$Pop_c(v) = \sum_{(u,v)\in E} e^{\sigma(T_0 - T_u)}$$

T_0: current year, T_u: time of u, σ: decaying factor

- Venue component
  - Constructing a venue graph and computing in similar way
- Author component
  - Using average prestige and popularity of his/her published articles

Outline

- Ranking Model
  - Our Time Weighted PageRank
  - Ranking with Importance Assembling
- Ranking Computation
- Dynamic Ranking Computation
- Experimental Study
- Summary

Batch Algorithm batSARank

- Importance

$$Imp(v) = Prs(v)^{\lambda} Pop(v)^{1-\lambda}$$

- Popularity computation

$$Pop_c(v) = \sum_{(u,v)\in E} e^{\sigma(T_0 - T_u)}$$

  - Can be done by scanning all citations once

- Prestige computation
  - Traditionally computed by TWPageRank in an iterative manner and is the most expensive computation
  - Adopting block-wise computation method batTWPR [Berkhin 2005]
    - Treating each strong connected component (SCC) as a block
    - Processing blocks one by one following topological orders
    - The edges between blocks are only scanned once

P. Berkhin. Survey: A survey on pagerank computing. Internet Mathematics, vol. 2, no. 1, pp. 73–120, 2005.

Why Adopting Block-wise Method?

- Observation:
  - citations obey a natural temporal order
  - SCC edge ratios are small for citation and venue graphs

Graphs	Nodes	Edges	Largest $\|SCC\|$	SCC edge ratio
citation-AAN	18,041	82,944	20	0.9%
citation-DBLP	3,140,081	14,260,658	23	1.6%
citation-MAG	126,909,021	526,498,920	351	0.1%
venue-AAN	565	22,527	18	2.8%
venue-DBLP	56,370	7,094,231	1,467	2.1%
venue-MAG	584,298	162,431,575	10,473	1.8%
web-BS	685,230	7,600,595	334,857	59.51%

Based on statistics of scholarly data,
block-wise method is a good choice for TWPageRank

- Taking t=100 for example, algorithm batTWPR only needs to scan 4|E| edges on citation and venue graphs, but over 59|E| edges on Web graphs.

Outline

- Ranking Model
  - Our Time Weighted PageRank
  - Ranking with Importance Assembling
- Ranking Computation
- Dynamic Ranking Computation
- Experimental Study
- Summary

Incremental Algorithm incSARank

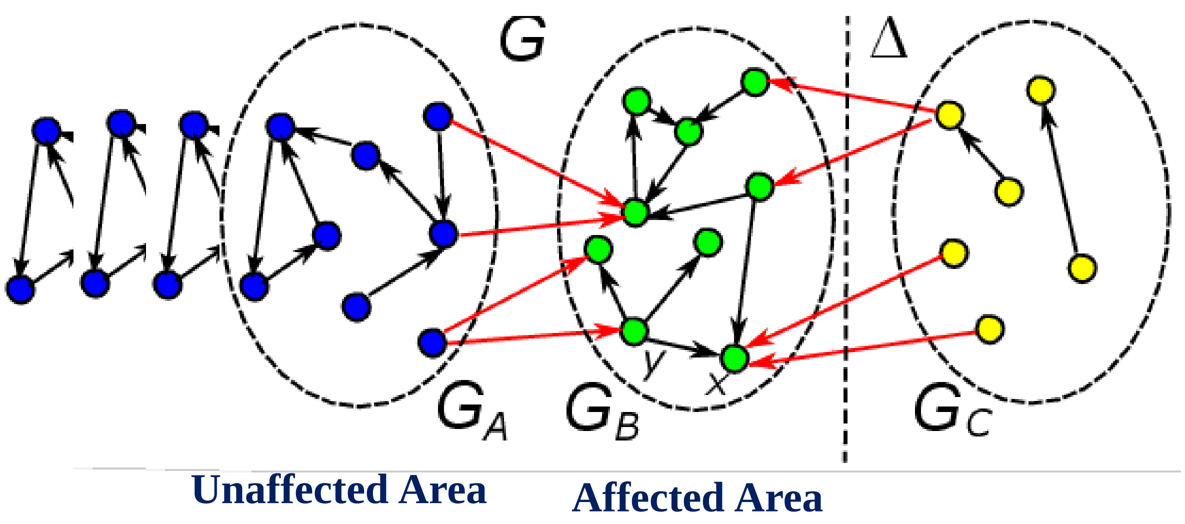
- Observation on scholarly data
  - Data only increases without decreasing
  - Citation relationships obey a natural temporal order

The original block-wise graph and topological order do NOT change
The existing popularity simply needs to be scaled

- Data structure maintenance
  - Only new SCCs and new topological order need to be computed
- Popularity computation
  - Computing freshness of new citations
- Prestige computation
  - Incremental TWPageRank algorithm incTWPR
  - Partitioning graph G into affected and unaffected areas
  - Employing different updating strategies for different areas

Affected and Unaffected Area Analysis

- Affected area
  - Nodes that are reachable from newly added nodes
  - Nodes with outgoing edges having weight changes
  - Nodes that are reachable from other affected nodes
- The rest of the original graph is unaffected area



Time Complexity Analysis

- Data structure maintenance
  - Saving $O(|V| + |E|)$ time (about 90%)
- Popularity computation
  - Saving $O(|E|)$ time (about 90%)
- Prestige computation
  - Saving $O(|E_A \cup E_{AB}|)$ time (about 30%)

Cost: $O(|V|)$ space for affected/unaffected areas

	Statis.	**Citation graphs on** AAN	DBLP	MAG
V	$\|V_A\|$	47.4%	52.3%	69.2%
	$\|V_B\|$	46.8%	40.0%	26.3%
	$\|V_C\|$	5.8%	7.8%	4.5%
E	$\|E_A\|$	3.0%	2.4%	0.9%
	$\|E_{AB}\|$	26.5%	30.2%	26.6%
	$\|E_B\|$	59.8%	59.3%	65.5%
	$\|E_{CB}\|$	10.4%	7.2%	7.0%
	$\|E_C\|$	0.3%	0.9%	0.1%

Outline

- Ranking Model
  - Our Time Weighted PageRank
  - Ranking with Importance Assembling
- Ranking Computation
- Dynamic Ranking Computation
- Experimental Study
- Summary

Experimental Settings

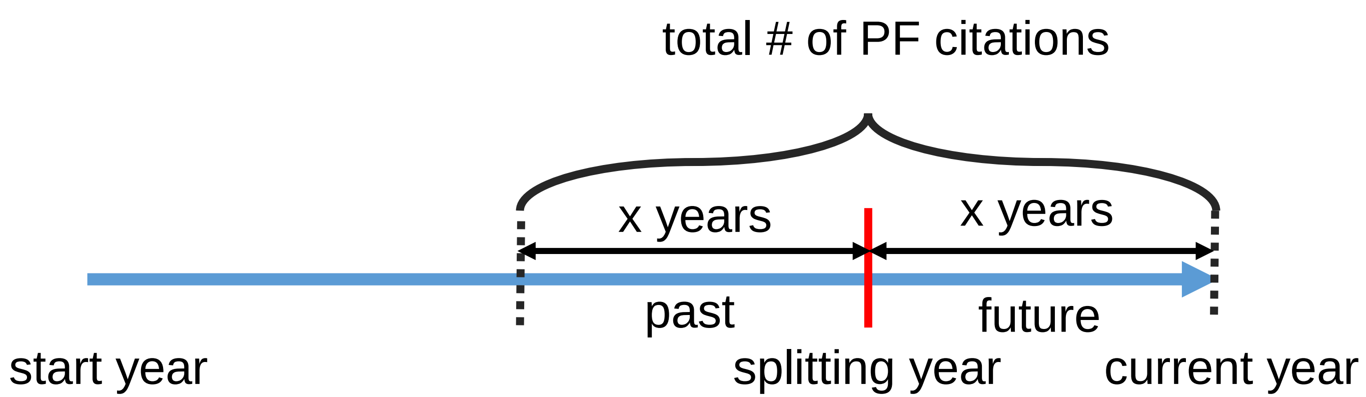
- Datasets:
  - AAN [Liang et al. 16], DBLP [Tang et al. 08], MAG [Sinha et al. 15]
- Metric: pairwise accuracy
  - $\text{PairAcc} = \frac{\#\text{ of agreed pairs}}{\#\text{ of all pairs}}$
- Algorithms
  - PRank [Brin et al. 98]: PageRank on the article citation graph;
  - FRank [Sayyadi et al. 09]: using citation, temporal and other heterogeneous information;
  - HRank [Liang et al. 16]: using both citation and heterogeneous information based on hyper networks;
  - SARank: our method;

R. Liang and X. Jiang, Scientific ranking over heterogeneous academic hypernetwork, in AAAI, 2016.
J. Tang, J. Zhang, L. Yao, et al., Arnetminer: Extraction and mining of academic social networks, in KDD, 2008.
A. Sinha, Z. Shen, Y. Song, et al., An overview of microsoft academic service (MAS) and applications, in WWW, 2015.
S. Brin and L. Page, The anatomy of a large-scale hypertextual web search engine, Computer Networks, 1998.
H. Sayyadi and L. Getoor, Future rank: Ranking scientific articles by predicting their future pagerank, in SDM, 2009.

Experimental Settings

➢ Ground-truth:

- RECOM [Liang et al. 16], which assumes articles with more recommendations are more important
- PFCTN for article ranking in a concerned year (splitting year)
  - Simply using citation numbers for fair evaluation
  - Past and future citations contribute equally
  - Articles in the same pairs must be in similar research fields and published in the same years
  - Articles with more PF citations are more important



Effectiveness with RECOM

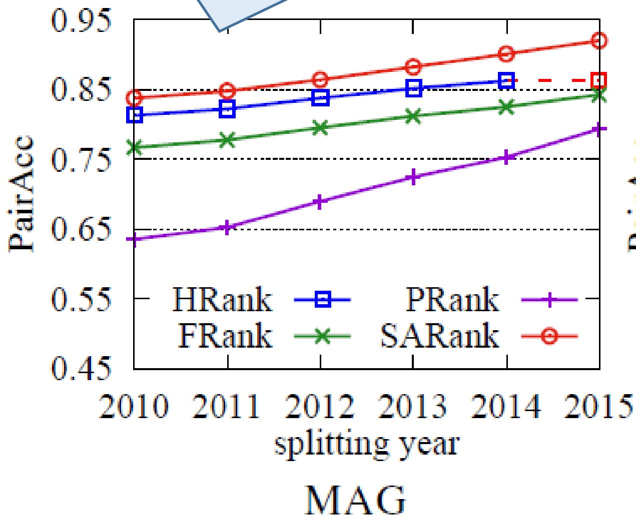
Datasets	PRank	FRank	HRank	SARank
AAN	0.671	0.738	0.758	**0.805**
DBLP	0.651	0.729	0.730	**0.778**
MAG	0.615	0.655	0.658	**0.680**

SARank consistently ranks better with RECOM

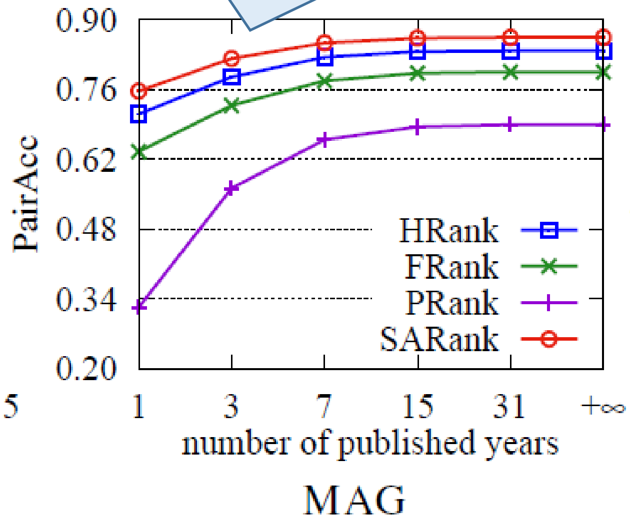
Note: RECOM is originally given on AAN, and we extend it to DBLP and MAG through exact title matching.

Effectiveness with PFCTN

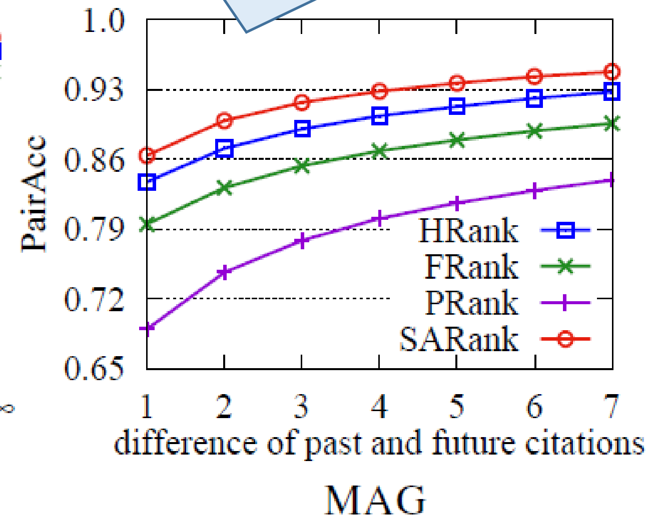
+16.7%, +7.2%, +2.9%



+23.6%, +8.3%, +3.2%



+13.4%, +6.0%, +2.4%



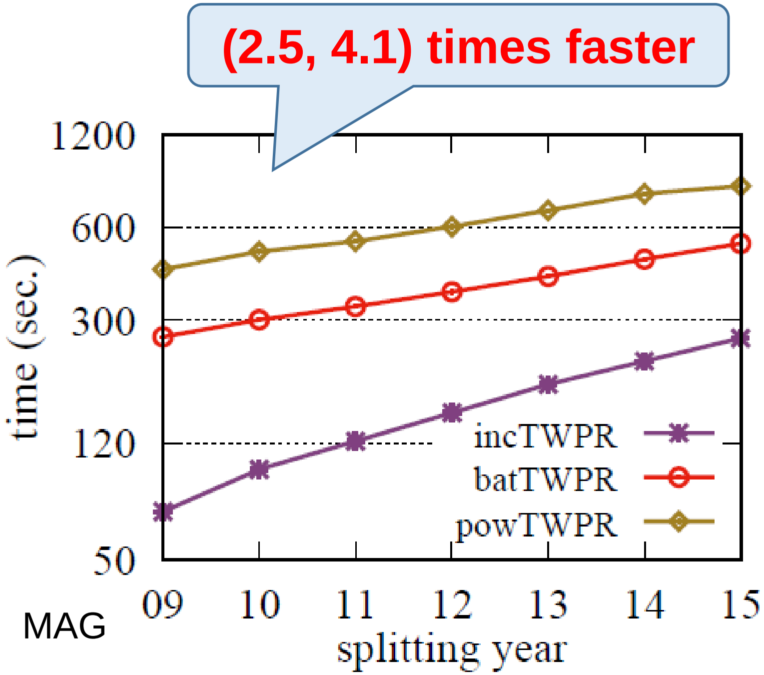
# of published years

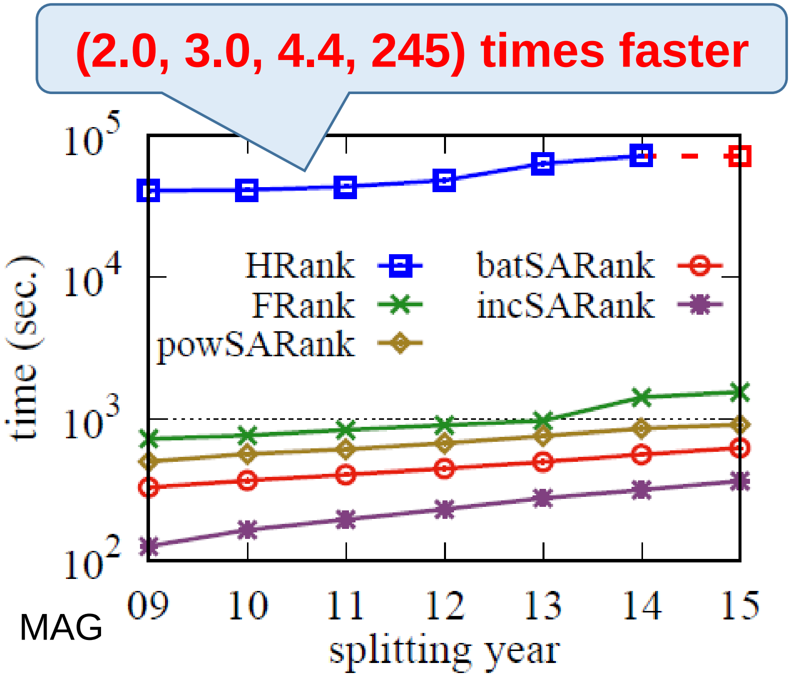
SARank consistently ranks better with PFCTN

start year article published splitting year current year

ranking data

Efficiency





Batch and incremental algorithms are more efficient

Outline

- Ranking Model
  - Time Weighted PageRank
  - Ranking with Importance Assembling
- Ranking Computation
- Dynamic Ranking Computation
- Experimental Study
- Summary

Summary

- Proposing a scholarly article ranking model SARank
  - Time-Weighted PageRank algorithm
  - Assembling the importance of articles, venues and authors

- Developing efficient ranking computation algorithms
  - Block-wise computation for TWPageRank
  - Incremental algorithm by affected/unaffected area division

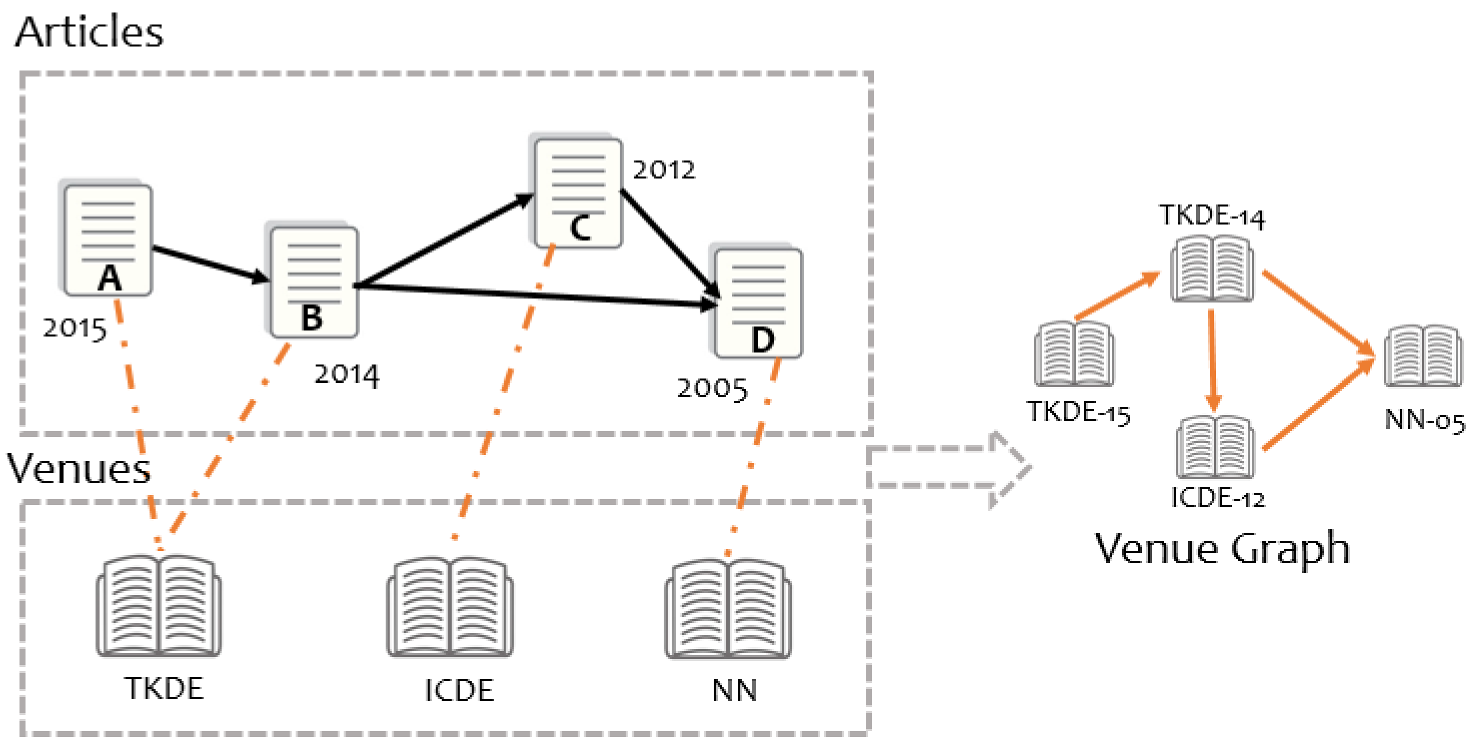
- Experimentation study
  - SARank consistently ranks better
  - Batch and incremental algorithms are more efficient
  - PFCTN, a new benchmark for article ranking

Thanks!

Q&A

Components Computation

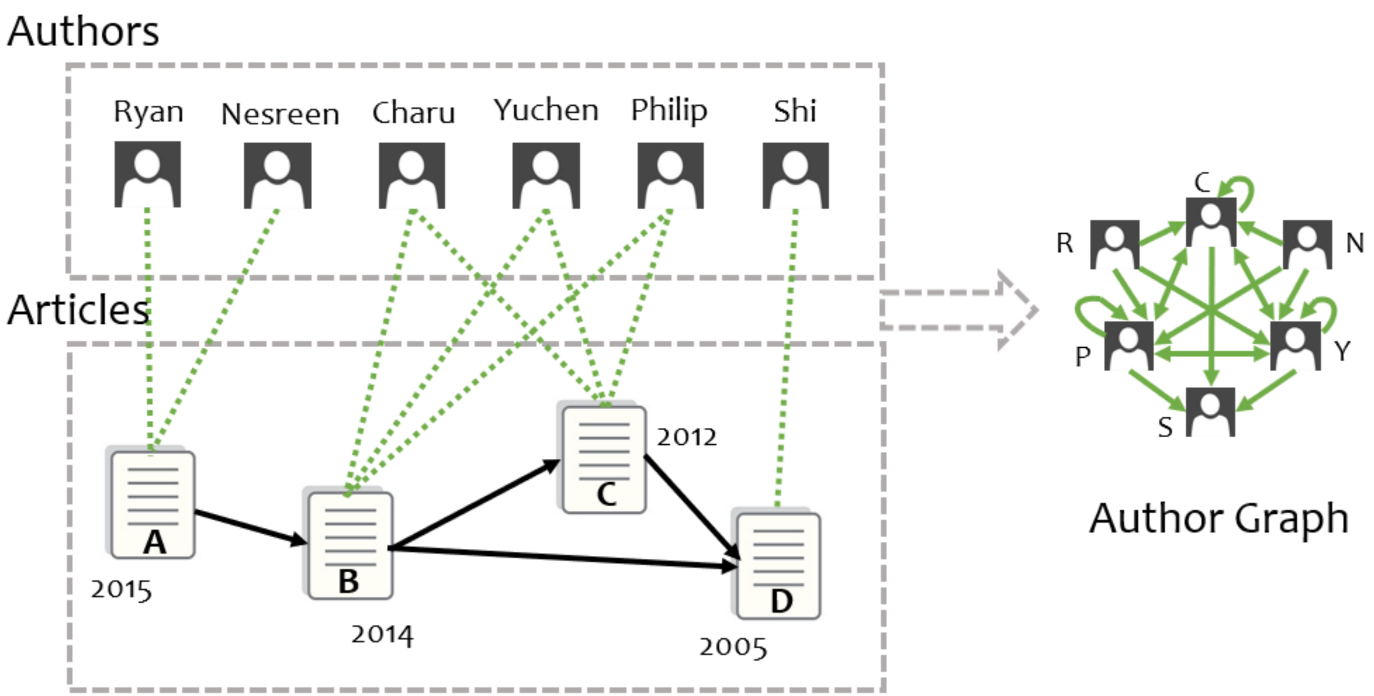
- Venue component
  - Treating the venue in each year individually and its importance is the sum of importance in all individual years



- Prs_v of venue k is its TWPageRank score on the venue graph
- Pop_v of venue k is the average popularity of its articles

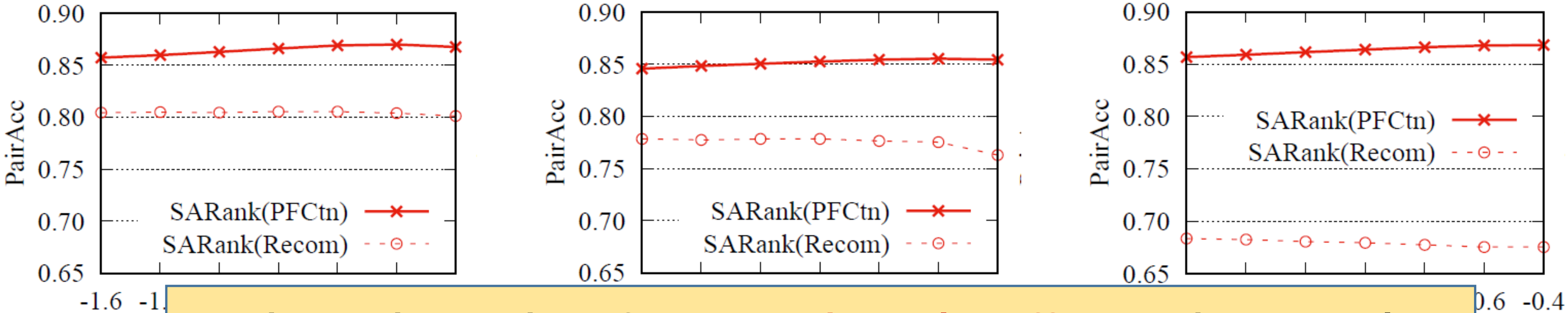
Components Computation

➢ Author component

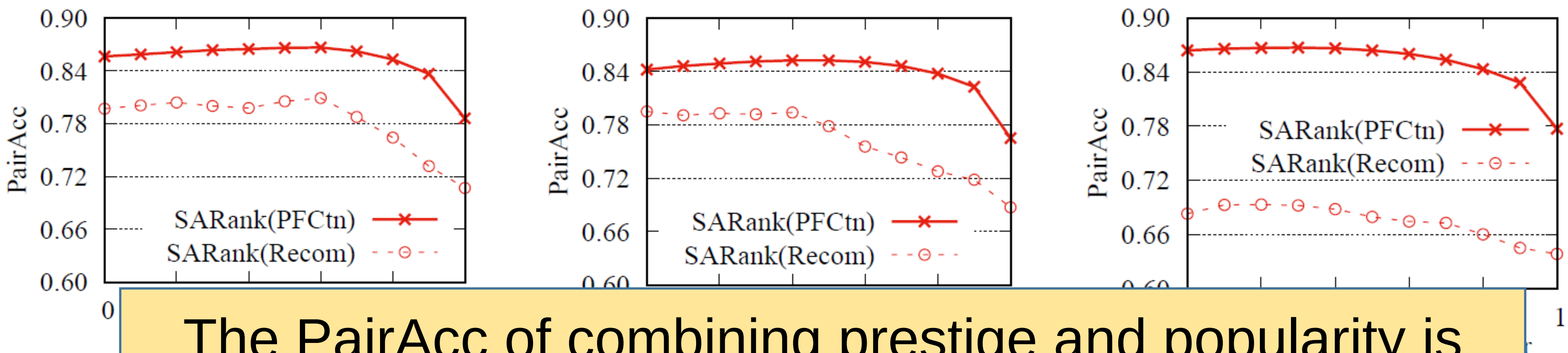


- Compute the TWPagerank on the author citation graph is computationally expensive
- Prs_a of author u is the average prestige of his/her articles
- Pop_a of author u is the average popularity of his/her articles

Impacts of Parameters

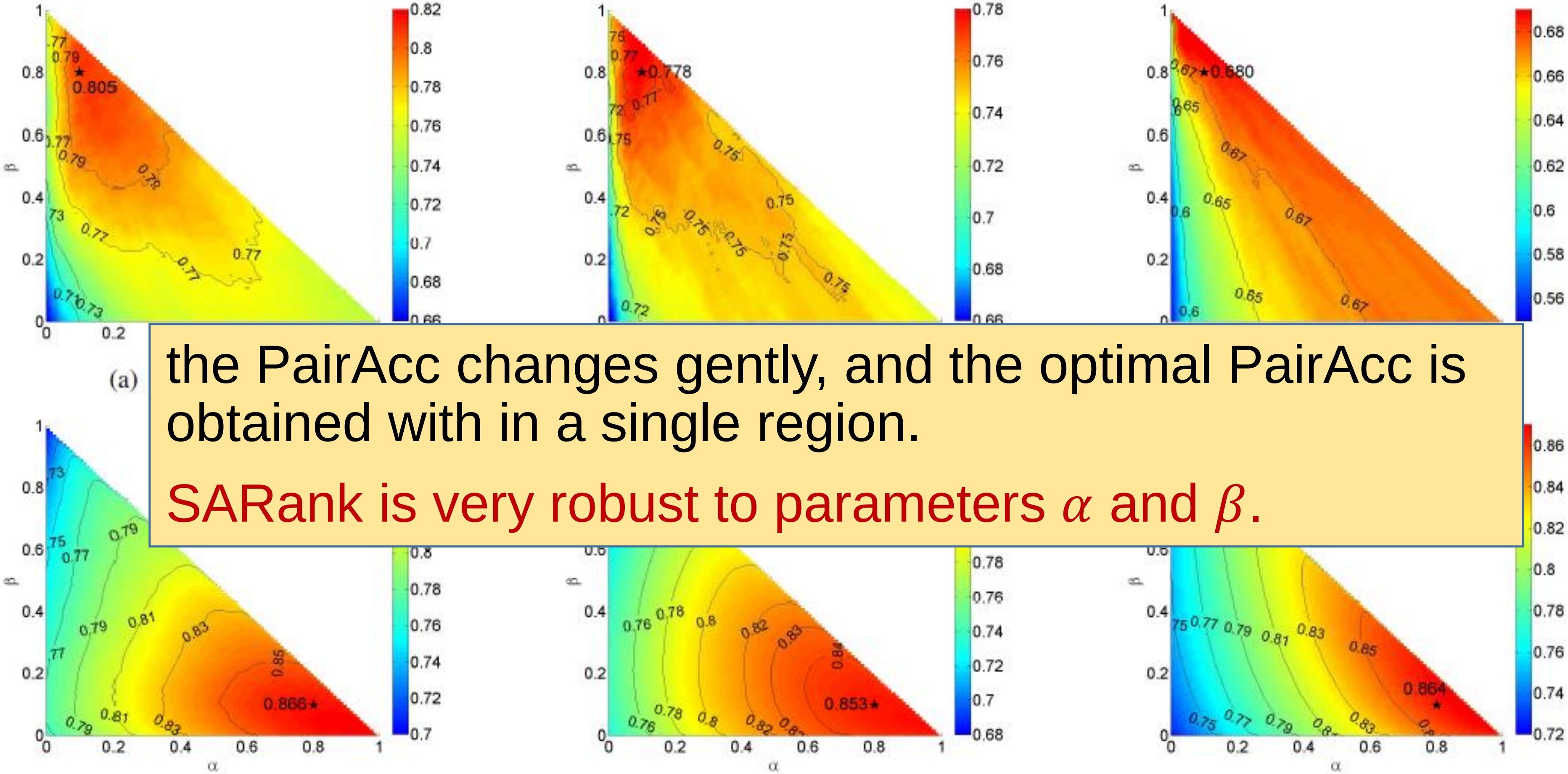


Time decaying factor σ barely affects the result



The PairAcc of combining prestige and popularity is generally better than using prestige or popularity alone

Impacts of Parameters α and β



(a)

(d) AAN with PFCTN

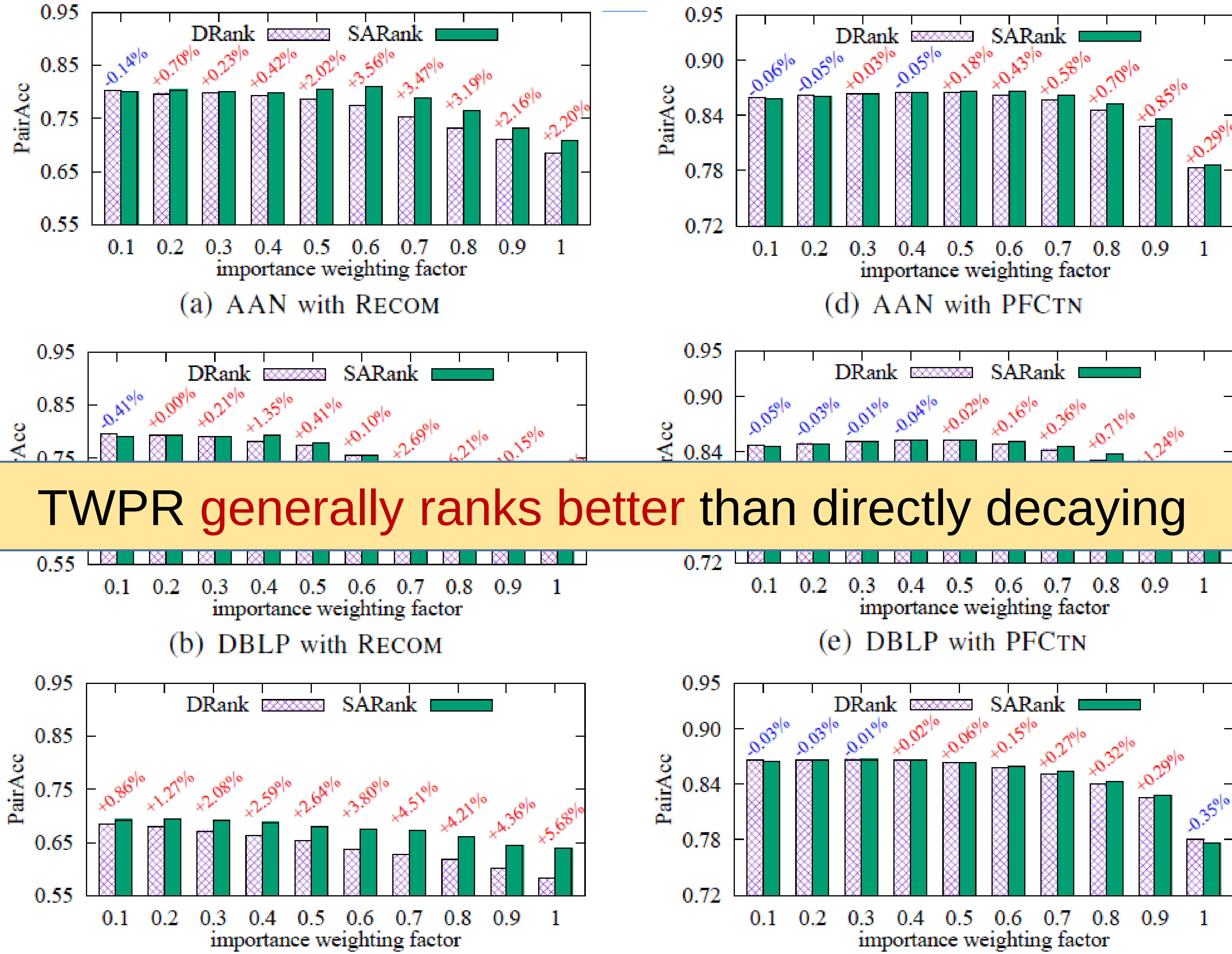
(e) DBLP with PFCTN

(f) MAG with PFCTN

the PairAcc changes gently, and the optimal PairAcc is obtained with in a single region.

SARank is very robust to parameters α and β.

SARank vs. DRank(exponentially decay directly)



(a) AAN with RECOM

(d) AAN with PFCTN

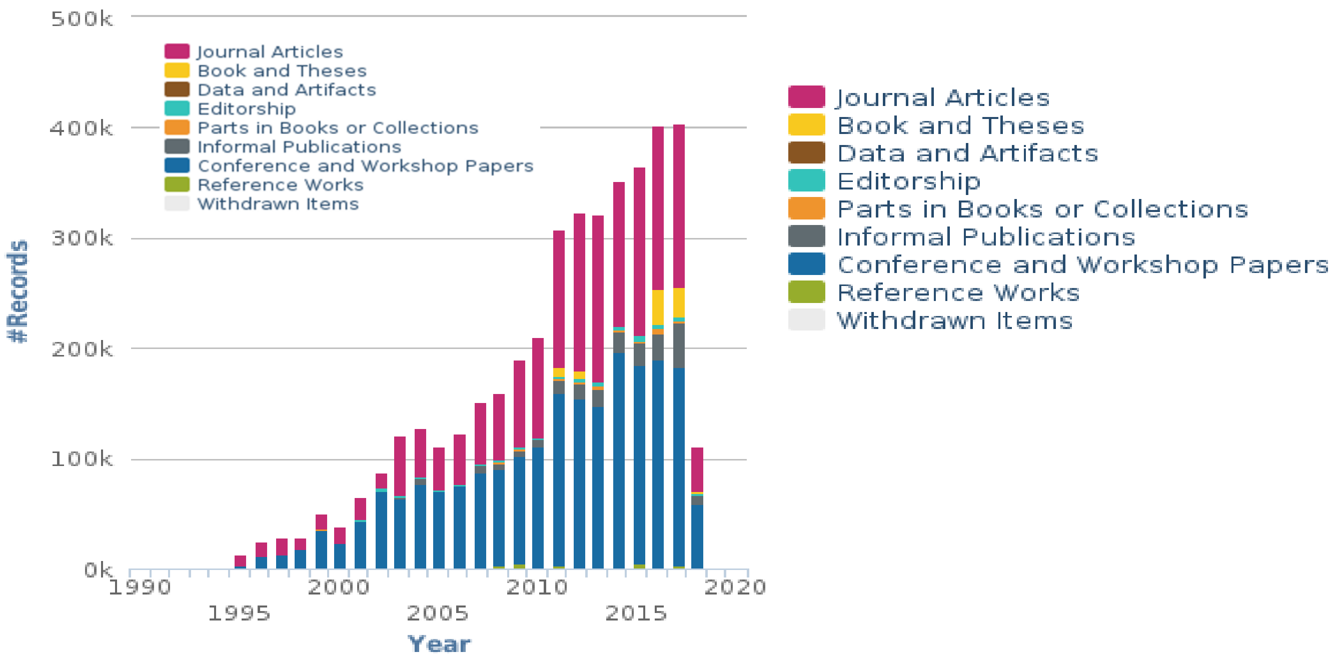
(b) DBLP with RECOM

(e) DBLP with PFCTN

(c) MAG with RECOM

(f) MAG with PFCTN

TWPR generally ranks better than directly decaying


#Records
500k
400k
300k
200k
100k
0k
1990
1995
2000
2005
2010
2015
2020
Year
Journal Articles
Book and Theses
Data and Artifacts
Editorship
Parts in Books or Collections
Informal Publications
Conference and Workshop Papers
Reference Works
Withdrawn Items