Will This Paper Increase Your *h*-index?

Scientific Impact Prediction

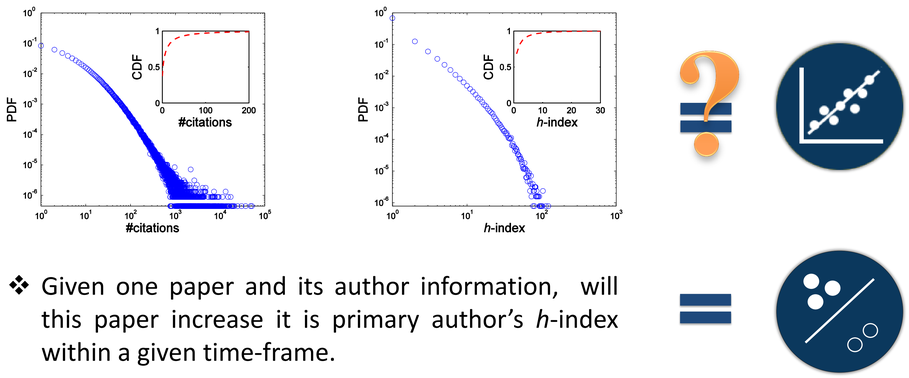
Yuxiao Dong
University of Notre Dame
ydong1@nd.edu

Reid A. Johnson
University of Notre Dame
rjohns15@nd.edu

Nitesh V. Chawla
University of Notre Dame
nchawla@nd.edu

Problem

- Predicting the number of citations of each paper.
- Predicting the *h*-index of each author.



- Given one paper and its author information, will this paper increase it is primary author's *h*-index within a given time-frame.

Academic Data

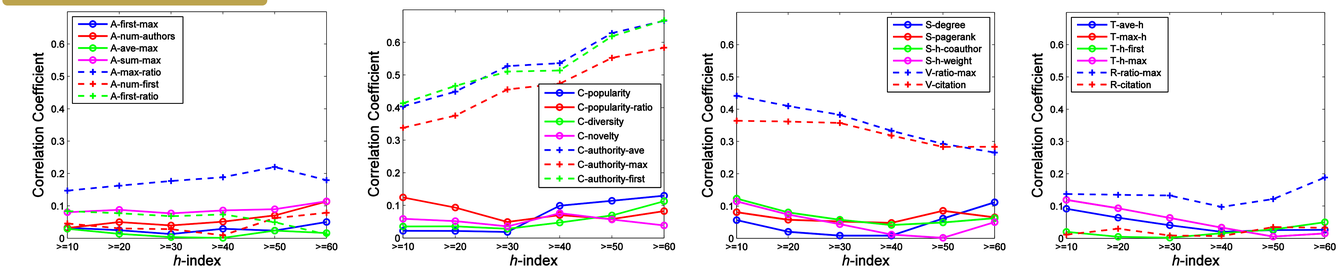
We use the real-world academic dataset from ArnetMiner, which is a free online service for academic social network analysis and mining.

- 1,712,433 authors, 2,092,356 papers, between 1950 and 2012
- 4,258,615 collaboration relationships and 8,024,869 citation relationships
- http://arnetminer.org/AMinerNetwork

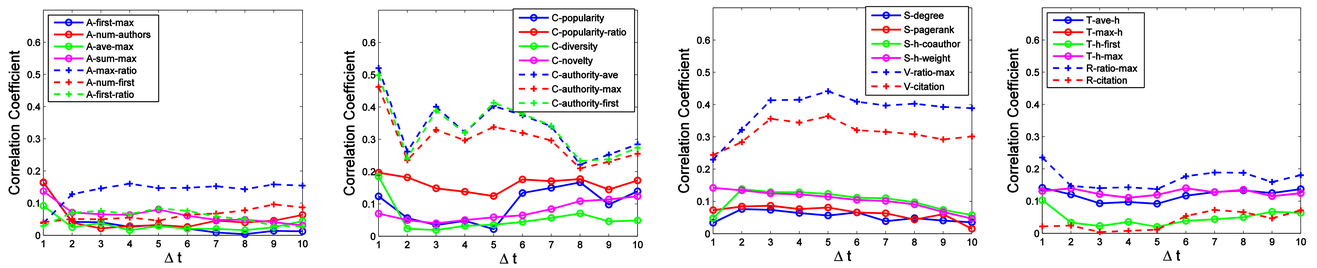
Scientific Factor

Factor Type	Factor Name	Factor Description
Author	*A-first-max*	First author's *h*-index divided by max-*h*-index
	A-ave-max	Avg. *h*-index of all authors divided by max-*h*-index
	A-sum-max	Sum of *h*-indices divided by max-*h*-index
	A-first-ratio	Ratio between max-*h*-index and number of papers by first author
	A-max-ratio	Ratio between max-*h*-index and number of papers by primary author
	A-num-authors	Number of authors of the given paper
	A-num-first	Number of papers by the first author
Content	*C-popularity*	Ave. number of citations over different topics
	C-popularity-ratio	Ave. number of citations over different topics divided by max-*h*-index
	C-novelty	Topic novelty of the paper
	C-diversity	Topic diversity of the paper
	C-authority-first	Consistence between the first author's authority and the paper
	C-authority-max	Consistence between the primary author's authority and the paper
	C-authority-ave	Avg. consistence between each author's authority and the paper
Venue	*V-ratio-max*	Ratio between #papers $\geq$ max-*h*-index citations divided by max-*h*-index
	V-citation	Avg. number of citations of all reference divided by max-*h*-index
Social	*S-degree*	Number of co-authors of the paper's authors
	S-pagerank	PageRank values of the paper's authors in the weighted collaboration network
	S-h-coauthor	Avg. *h*-index of co-authors of paper's authors divided by max-*h*-index
	S-h-weight	Weighted avg. *h*-index of co-authors of paper's authors divided by max-*h*-index
Reference	*R-ratio-max*	Ratio between number of references $\geq$ max-*h*-index and number of references
	R-citation	Avg. number of citations divided by the max-*h*-index
Temporal	*T-ave-h*	Avg. Δh-indices of the authors between now and three years ago
	T-max-h	Max Δh-indices of the authors between now and three years ago
	T-h-first	Δh-index of the first author between now and three years ago
	T-h-max	Δh-index of the max-*h*-index author between now and three years ago

Factor Correlation



The following figures present the changes of factor importance when predicting for scholars with different *h*-indices. Note that t=2007 and Δt=5 years. We observe that the author's authority on a subject and the published venue are the most highly correlated factors.



The following figures present the changes of factor correlation as the time period Δt is varied. Note that t=2007 and a minimum threshold for the primary author's *h*-index is set to 10.

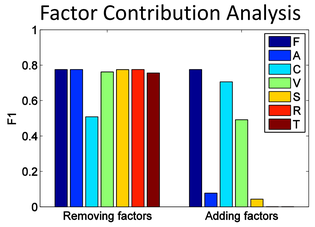
- A scientific researcher's authority on a topic is the most decisive factor in facilitating an increase in her/his *h*-index. This suggests that it is best to focus on what one is good at.
- The quality of the venue in which a given paper is published is a crucial factor in determining the probability that that paper will subsequently contribute to the authors' *h*-indices.
- People in social society often follow vogue trends. However, working on an academically "hot topic" in which one has little expertise is unlikely to further one's scientific impact, in so far as it is measured by an increase in one's *h*-index.

Predictability

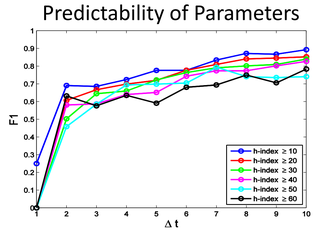
Predictive results of whether the papers published in time t will contribute to the *h*-indices of the authors within a given time period Δt.
t=2007, Δt=5 years, and *h*-index threshold is set to 10.
LRC—Logistic Regression; Random guess with half positive and half negative.

	Precision	Recall	F1	AUC	Accuracy	Pre@3	MAP
Random	0.305	0.500	0.375	0.500	0.500	0.672	0.522
LRC	0.854	0.711	0.776	0.938	0.875	0.925	0.965

Factor Contribution Analysis



Predictability of Parameters



Conclusion

- Our problem definition offers a strong potential for quantifying scientific impact.
- We find that an authors' authority on the publication topic and the published venue of a paper play the most decisive roles in determining whether a paper will contribute to its primary author's *h*-index.
- Surprisingly, we observe that the popularity of the publication topic and the co-authors' influence are not correlated to the prediction target.
- Our study also demonstrates a greater than 87.5% potential predictability for whether a paper will contribute to its primary author's *h*-index within five years.
- Overall, our findings unveil mechanisms for quantifying scientific impact and provide concrete suggestions to researchers for better expanding their scientific influence and, ultimately, for more effectively "*standing on the shoulders of giants*."

Acknowledgement

Reference

1. J. Tang, et al. ArnetMiner: Extraction and mining of academic social networks. In KDD'08.
2. J. Cheng, L. Adamic, P.A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In WWW'14.
3. J. E. Hirsch. An index to quantify an individual's scientific researcher output. PNAS 2005.