Will This Paper Increase Your *h*-index? Scientific Impact Prediction

Yuxiao Dong, Reid A. Johnson, Nitesh V. Chawla

iCeNSA Interdisciplinary Center for Network Science & Applications

Interdisciplinary Center for Network Science and Applications

UNIVERSITY OF NOTRE DAME

Scientific Impact

Integral to the success of scientific research is the publication and dissemination of impactful work and findings.

Scientific Impact

*"An emerging area of interest in research on the '**science of science**' is the prediction of future impact."*

How? *What?*

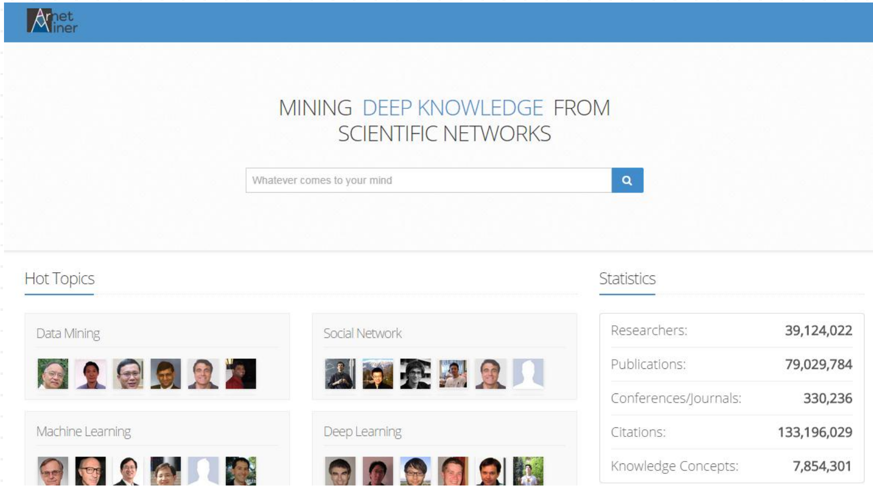
J. A. Evans.
Science 342, 2013

D. E. Acuna, S. Allesina, K. P. Kording. Future Impact: Predicting Scientific Success. Nature 489, 2012
D. Wang, C. Song, A.-L. Barabasi. Quantifying long-term scientific impact. Science 342, 2013.
B. Uzzi, S. Mukherjee, M. Stringre, B. Jones. Atypical Combinations and Scientific Impact. Science 342, 2013.
H.-W. Shen and A.-L. Barabási. **Collective credit allocation in science.** PNAS 111, 2014.

Academic Data

- A real-world academic dataset from Arnetminer
  - 1,712,433 authors
  - 2,092,356 papers
  - 4,258,615 collaborations
  - 8,024,869 citations

  - *http://arnetminer.org/AMinerNetwork*



J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su. ArnetMiner: Extraction and mining of academic social networks. KDD'08.

Scientific Impact: #citations

The number of citations of each publication

Title 1–20	Cited by	Year
SMOTE: synthetic minority over-sampling technique NV Chawla, KW Bowyer, LO Hall, WP Kegelmeyer Journal of Artificial Intelligence Research (JAIR) 16, 321-357	2471 *	2002
Editorial: special issue on learning from imbalanced data sets NV Chawla, N Japkowicz, A Kotcz ACM Sigkdd Explorations Newsletter 6 (1), 1-6	882	2004
SMOTEBoost: Improving prediction of the minority class in boosting NV Chawla, A Lazarevic, LO Hall, KW Bowyer Knowledge Discovery in Databases: PKDD 2003, 107-119	450	2003
New perspectives and methods in link prediction RN Lichtenwalter, JT Lussier, NV Chawla Proceedings of the 16th ACM SIGKDD international conference on Knowledge ...	229	2010
SVMs modeling for highly imbalanced classification Y Tang, YQ Zhang, NV Chawla, S Krasser Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 39 ...	218	2009

http://scholar.google.com/ . Accessed on Dec. 18th, 2014

#citations prediction

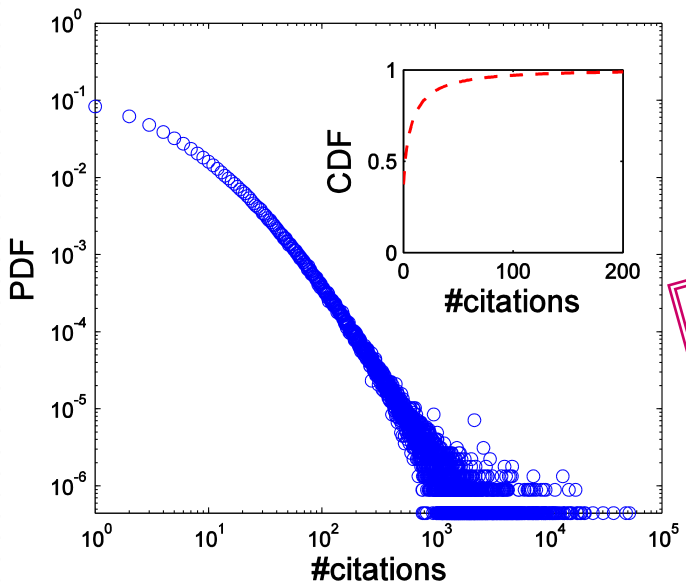
Predicting the number of citations of publications

Title 1–20	Cited by	Year
SMOTE: synthetic minority over-sampling technique NV Chawla, KW Bowyer, LO Hall, WP Kegelmeyer Journal of Artificial Intelligence Research (JAIR) 16, 321-357	2471 *	2002
Editorial: special issue on learning from imbalanced data sets NV Chawla, N Japkowicz, A Kotcz ACM Sigkdd Explorations Newsletter 6 (1), 1-	882	2004
SMOTEBoost: Improving prediction s in boosting NV Chawla, A Lazarevic, LO Hall, KW Bow Knowledge Discovery in Databases: PKDD	450	2003
New perspectives and methods in link RN Lichtenwalter, JT Lussier, NV Chawla Proceedings of the 16th ACM SIGKDD international conference on Knowledge ...	229	2010
SVMs modeling for highly imbalanced classification Y Tang, YQ Zhang, NV Chawla, S Krasser Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 39 ...	218	2009

R. Yan, C. Huang, J. Tang, Y. Zhang, and X. Li. To better stand on the shoulder of giants. JCDL'12, pp. 51-60. 2012.
D. Wang, C. Song, A.-L. Barabasi. Quantifying long-term scientific impact. Science, 342 (6154), 2013.

#citations prediction

- publications with few citations are extremely common
- publications with many citations are relatively rare

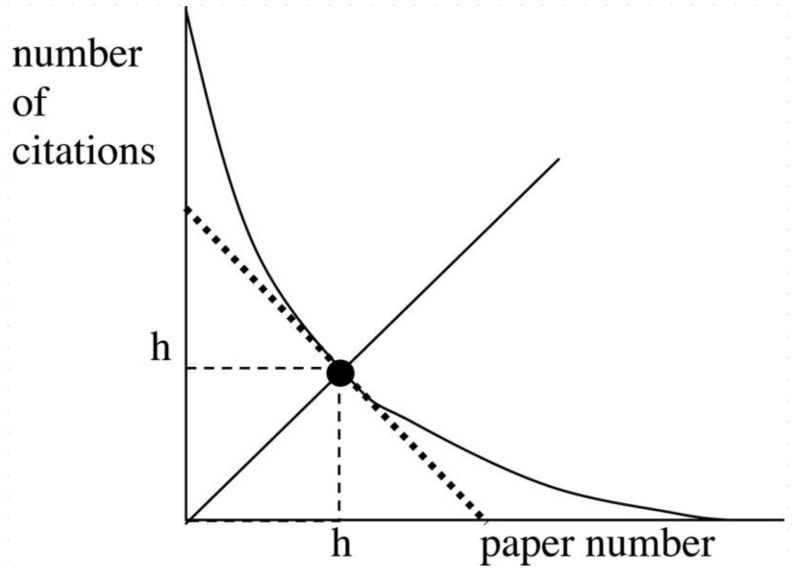


Challenge

6.91% (155k out of 2 million) of the papers obtain more than 50 citations from 1950 to 2012.

Scientific Impact: *h*-index

h-index



J. E. Hirsch. An index to quantify an individuals' scientific research output. PNAS 102(45). 2005.

Scientific Impact: *h*-index

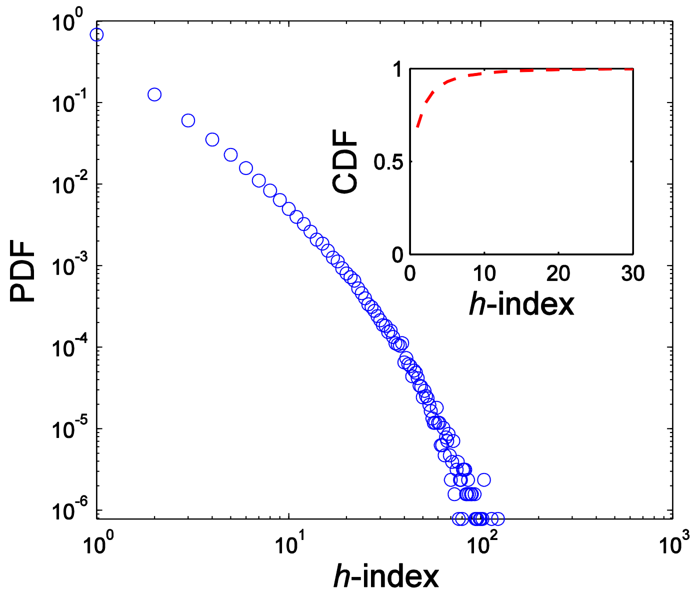
The *h*-index of each author

Experts	H-index	Rank
 Thomas Huang H-index: ***128***, #Papers: ***766***, #Citations: ***65956*** Professor, Beckman Institute at the University of Illinois	128	1
 Anil K. Jain H-index: ***128***, #Papers: ***440***, #Citations: ***90064*** Distinguished Professor, Michigan State University	128	2
 Philip S. Yu H-index: ***127***, #Papers: ***812***, #Citations: ***68713*** Professor and Wexler Chair in Information Technology, Department of Computer Science, University of Illinouis Chicago	127	3
 Jiawei Han H-index: ***116***, #Papers: ***652***, #Citations: ***90056*** Professor, Department of Computer Science, University of Illinois at Urbana-Champaign	116	4
 H. Garcia H-index: ***115***, #Papers: ***429***, #Citations: ***55531*** Professor, Departments of Computer Science and Electrical Engineering, Stanford University	115	5

http://arnetminer.org/ranks/author/hindex/ Accessed on Dec. 18th, 2014

h-index prediction

Predicting the *h*-index of each author?



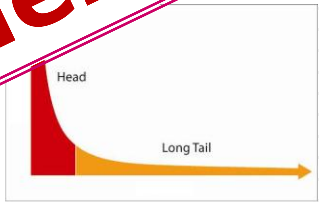
Challenge

0.0125% (159 out of 1.7 million) of the researchers have an h-index over 60

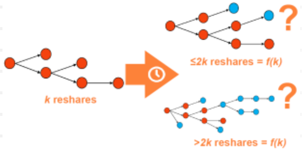
Scientific Impact

- Predicting the #citations of each paper
- Predicting the *h*-index of an author

Challenge

- Predicting whether a cascade will double in size[1]

[1]

[1] J. Cheng, L. Adamic, A. Dow, J. Kleinberg, J. Leskovec. Can cascades be predicted? In WWW'14.

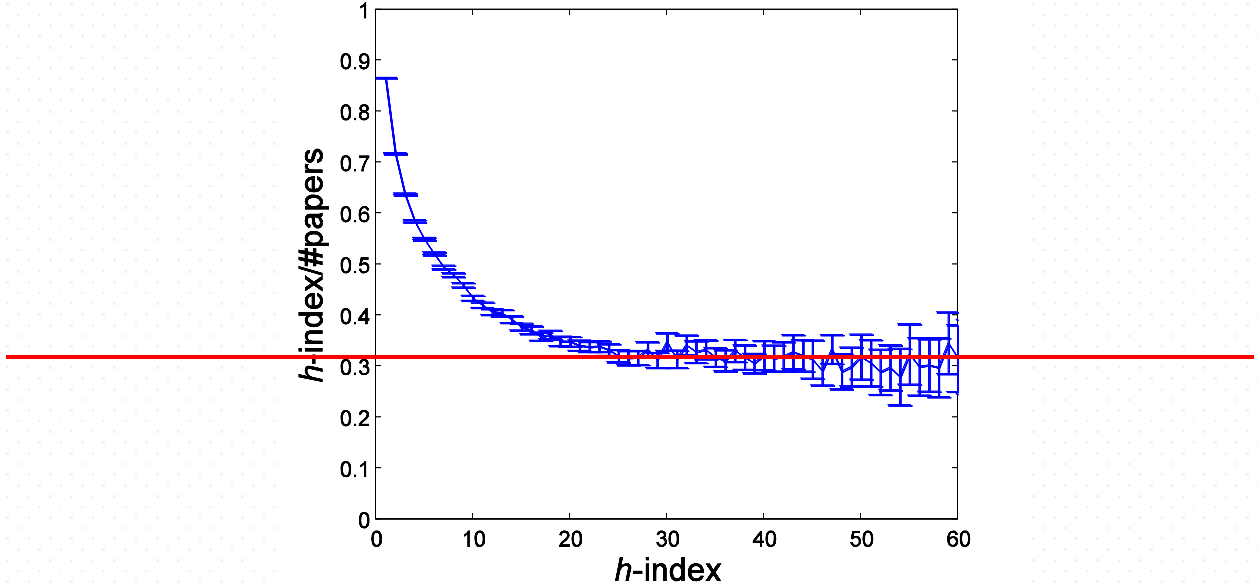
Scientific Impact Prediction Problem

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

the author of the given paper with the highest *h*-index.

Scientific Impact: *h*-index

h-index vs. *h*-index/#papers



The ratio between one's *h*-index (≥20) and her/his number of papers stabilizes at **0.3**.

Scientific Impact Prediction Problem

Can Cascades be Predicted?

Justin Cheng
Stanford University
jcccf@cs.stanford.edu

Lada A. Adamic
Facebook
ladamic@fb.com

P. Alex Dow
Facebook
adow@fb.com

Jon Kleinberg
Cornell University
kleinber@cs.cornell.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

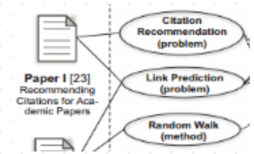
primary author*
h-index: 81

Given this paper at $t=2014$ and its primary author, the task is to predict whether it will get at least 81 citations within $\Delta t=5$ years.

* The determination of the primary author is based on information accessed on Dec. 18th, 2014.

Factors driving scientific impact

Content



Reference

Author

Paper

Temporal

Collaboration social network

Venue

Factors --- author

Author
7 factors

Can Cascades be Predicted?

first author

Justin Cheng
Stanford University
icccf@cs.stanford.edu

Lada A. Adamic
Facebook
ladamic@fb.com

P. Alex Dow
Facebook
adow@fb.com

primary author
h-index: 81

Jon Kleinberg
Cornell University
kleinber@cs.cornell.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

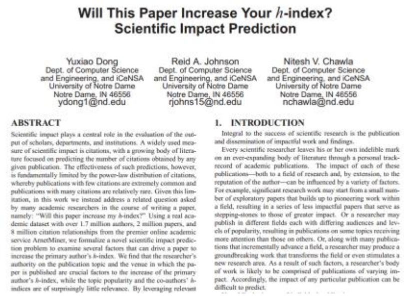
all authors
average author

* The determination of the primary author is based on information accessed on Dec. 18th, 2014.

Factors --- content

Content

7 factors



Will This Paper Increase Your h-index? Scientific Impact Prediction

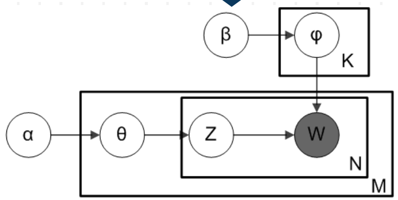
Yuxiao Dong, Dept. of Computer Science and Engineering, and iCeNSA, University of Notre Dame, Notre Dame, IN 46556, ydong1@nd.edu

Reid A. Johnson, Dept. of Computer Science and Engineering, and iCeNSA, University of Notre Dame, Notre Dame, IN 46556, rjohns15@nd.edu

Nitesh V. Chawla, Dept. of Computer Science and Engineering, and iCeNSA, University of Notre Dame, Notre Dame, IN 46556, nchawla@nd.edu

ABSTRACT

1. INTRODUCTION



scientific impact: 0.5
science of science: 0.4
social network: 0.1

topic popularity

deep learning is hot!

topic novelty

divergence of topics between this paper and its reference

topic diversity

divergence of topics of this paper

topic authority

authors’ authority on the topics of this paper

Factors --- venue

npg Science AAAS

Venue
2 factors

Top publications - Data Mining & Analysis   Learn more

	Publication	h5-index
1.	ACM SIGKDD International Conference on Knowledge discovery and data mining	69
2.	IEEE Transactions on Knowledge and Data Engineering	57
3.	ACM International Conference on Web Search and Data Mining	54
4.	ACM Conference on Recommender Systems	36
5.	IEEE International Conference on Data Mining (ICDM)	36
6.	SIAM International Conference on Data Mining	35

average citations of papers in this venue

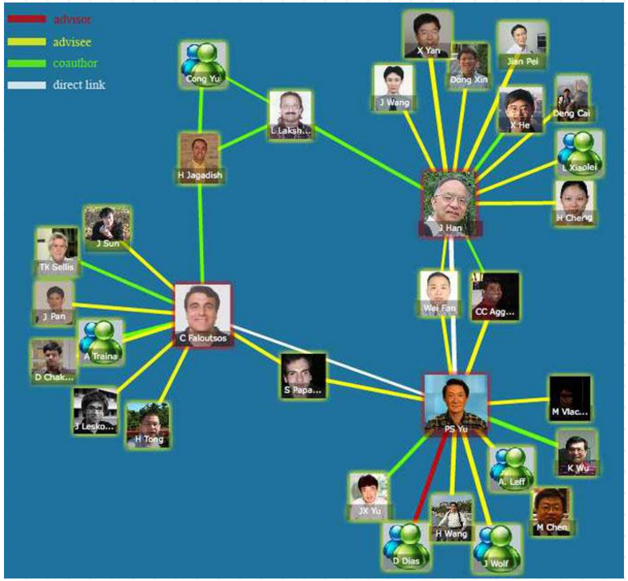
h-index contribution ratio of papers in this venue

http://scholar.google.com/ . Accessed on Dec. 18th, 2014

Factors --- social

Collaboration social network

4 factors

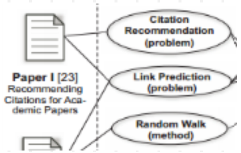


degree

Pagerank

coauthors’ h-indices

Factors --- reference



Reference

2 factors

7. REFERENCES

[1] M. Ahmed, S. Spagna, F. Huici, and S. Niccolini. A peek into the future: Predicting the evolution of popularity in user generated content. In *WSDM '13*, pages 607–616. ACM, 2013.

[2] S. Bethard and D. Jurafsky. Who should I cite: Learning literature search models from citation behavior. In *CIKM '10*, pages 609–618. ACM, 2010.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[4] C. Castillo, D. Donato, and A. Gionis. Estimating the number of citations using author reputation. In *SPIRE '07*, pages 107–117. Springer, 2007.

[5] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *WWW '14*, pages 925–936, 2014.

[6] E. Garfield. Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122(3159):108–111, 1955.

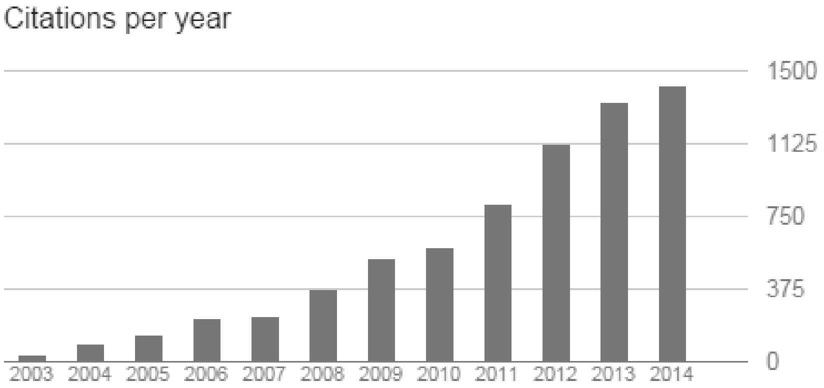
citations of references

h-index of references

standing on the shoulder of giants

Factors --- temporal

Temporal
4 factors



authors' *h*-index increasing rate

Factor Definition

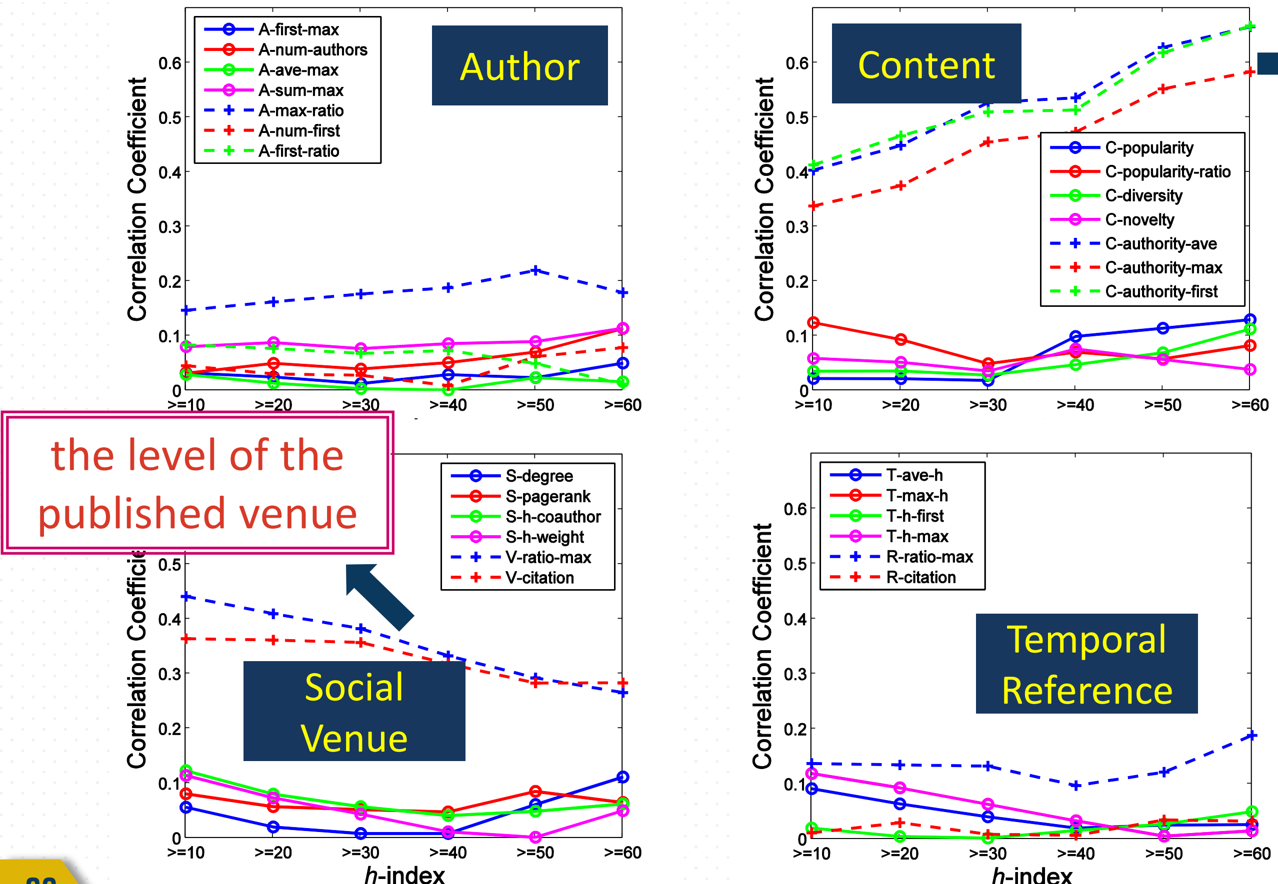
Table 1: Factor Definition. We employ six categories of factors, comprised of author, topic, reference, social, venue, and temporal attributes. max-h-index denotes the h-index of the primary author (i.e., the author with the maximum h-index) of a given paper.

6 groups

	Factor	Description
Author	*A-first-max*	The first author's h-index divided by the max-h-index.
	A-ave-max	The average h-index of all authors divided by the max-h-index.
	A-sum-max	The sum of h-indices divided by the max-h-index.
	A-first-ratio	The ratio between max-h-index and the number of papers attributed to the first author.
	A-max-ratio	The ratio between max-h-index and the number of papers attributed to the primary author.
	A-num-authors	The number of authors of the given paper.
	A-num-first	The number of papers by the first author.
Content	*C-popularity*	The average number of citations over different topics (see Eq. 1).
	C-popularity-ratio	The average number of citations over different topics divided by the max-h-index.
	C-novelty	The topic novelty of this paper (see Eq. 2).
	C-diversity	The topic diversity of this paper (see Eq. 3).
	C-authority-first	The consistence between the first author's authority and this paper (see Eq. 4).
	C-authority-max	The consistence between the primary author's authority and this paper.
	C-authority-ave	The average consistence between each author's authority and this paper.
Venue	*V-ratio-max*	The ratio between the number of papers $\geq$max-h-index citations divided by the max-h-index.
	V-citation	The average number of citations of all references divided by the max-h-index.
Social	*S-degree*	The number of co-authors of the paper's authors.
	S-pagerank	The PageRank values of the paper's authors in the weighted collaboration network.
	S-h-co-author	The average h-index of co-authors of the paper's authors divided by the max-h-index.
	S-h-weight	The weighted average h-index of co-authors of the paper's authors divided by the max-h-index.
Reference	*R-ratio-max*	The ratio between the number of references $\geq$max-h-index and the total number of references.
	R-citation	The average number of citations divided by the maximum h-index.
Temporal	*T-ave-h*	The average Δh-indices of the authors between now and three years ago.
	T-max-h	The maximum Δh-index between now and three years ago.
	T-h-first	The Δh-index of the first author between now and three years ago.
	T-h-max	The Δh-index of the max-h-index author between now and three years ago.

26 factors

Factors Correlation



authors' authority on the topics of this paper

t = 2007

Δt = 5

X-axis:
primary author's *h*-index

Y-axis:
correlation coefficient

Factors Correlation

A scientific **researcher's authority** on a topic is the most decisive factor in facilitating an increase in his or her *h*-index.

Factors Correlation

The level of **the venue** in which a given paper is published is another crucial factor in determining the probability that it will contribute to its authors' *h*-indices.

Factors Correlation

Publishing on an academically “hot” but unfamiliar topic is difficult to further one's scientific impact, at least as measured by an increase in one's *h*-index.

Prediction: predictability

Is Scientific Impact Predictable?

Prediction: predictability

t = 2007 Δt = 5
21,519 papers

On average, 30.5% of papers successfully contributed to their primary author's *h*-indices in 2012.

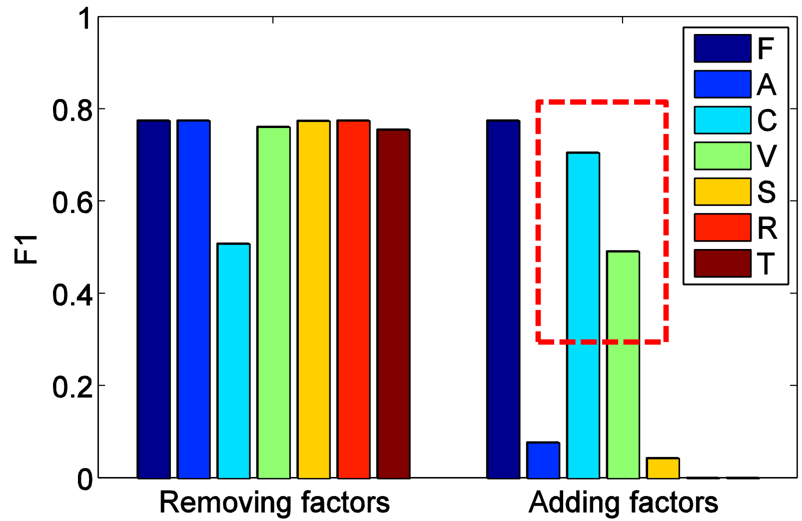
Task: predict whether the number of citations for each paper published in 2007 is larger than or equal to the primary author's *h*-index in 2012

R: Random guess
LRC: Logistic regression

Features: 26 factors
Half training, half test

Method	Pre.	Rec.	F_1	AUC	Accu.	Pre@3	MAP
R	0.305	0.500	0.375	0.500	0.500	0.674	0.522
LRC	0.854	0.711	0.776	0.938	0.875	0.925	0.965

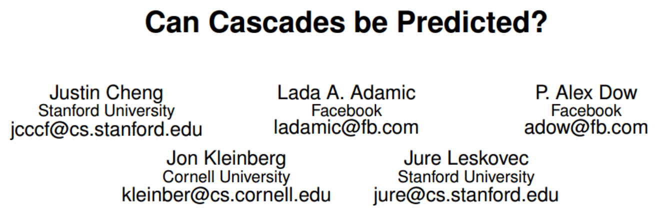
Prediction: factor contribution



F: Full factors
A: Author
C: Content
V: Venue
S: Social
R: Reference
T: Temporal

t = 2007
Δt = 5
Logistic regression

Prediction: predictability


Can Cascades be Predicted?

Justin Cheng
Stanford University
jcccf@cs.stanford.edu

Lada A. Adamic
Facebook
ladamic@fb.com

P. Alex Dow
Facebook
adow@fb.com

Jon Kleinberg
Cornell University
kleinber@cs.cornell.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

Published at 2014

Δt = 5 years

Δt = 10 years

Is a paper more predictable given a long or short timeframe Δt?

Prediction: predictability

Published at 2014

Inferring User Demographics and Social Strategies in Mobile Social Networks

Yuxiao Dong†, Yang Yang‡, Jie Tang‡, Yang Yang†, Nitesh V. Chawla†
†Department of Computer Science and Engineering, University of Notre Dame
†Interdisciplinary Center for Network Science and Applications, University of Notre Dame
‡Department of Computer Science and Technology, Tsinghua University
ydong1@nd.edu, yyang.thu@gmail.com, jietang@tsinghua.edu.cn, yyang1@nd.edu, nchawla@nd.edu

Can Cascades be Predicted?

Justin Cheng, Stanford University, jcccf@cs.stanford.edu
Lada A. Adamic, Facebook, ladamic@fb.com
P. Alex Dow, Facebook, adow@fb.com
Jon Kleinberg, Cornell University, kleinber@cs.cornell.edu
Jure Leskovec, Stanford University, jure@cs.stanford.edu

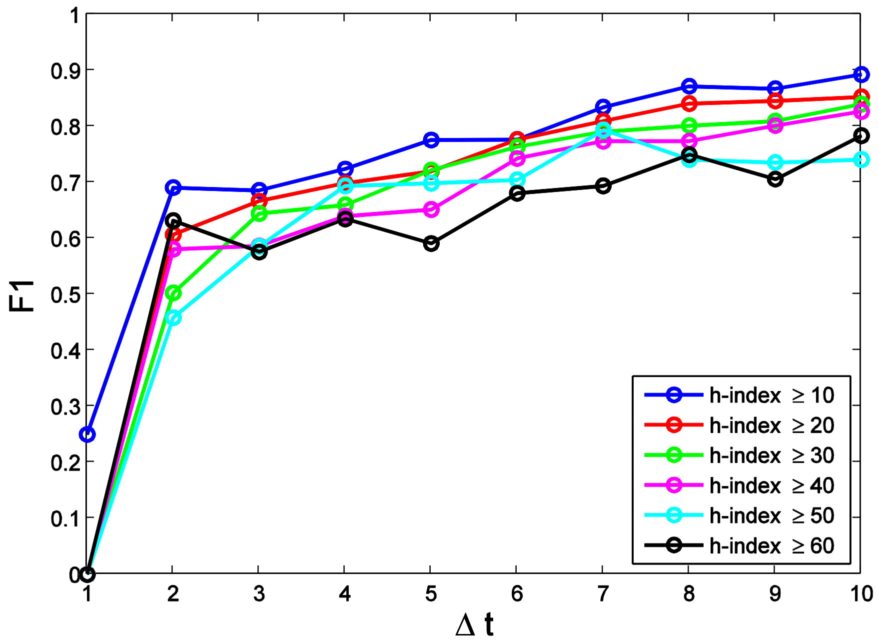
Primary author's *h*-index: 33

Primary author's *h*-index: 81

Is a primary author with a high or a low *h*-index more predictable?

The determination of the primary author is based on information accessed on Dec. 18th, 2014.

Prediction: predictability



$t + \Delta t = 2012$
Logistic regression

1. more difficult for papers with a high *h*-index primary author
2. more difficult when given a shorter timeframe Δt.

Future work

1. Only work on computer science domain

TODO: physics, mathematics, biology …

2. Authors' h-indices evolve within Δt

TODO: co-evolution of authors' h-indices and #citations

When a measure becomes a target, it ceases to be a good measure

---Charles Goodhart

Acknowledgements

Army Research Laboratory
(ARL)

U.S. Air Force Office of Scientific Research
(AFOSR)

Defense Advanced Research Projects Agency
(DARPA)

National Science Foundation
(NSF)

Thanks

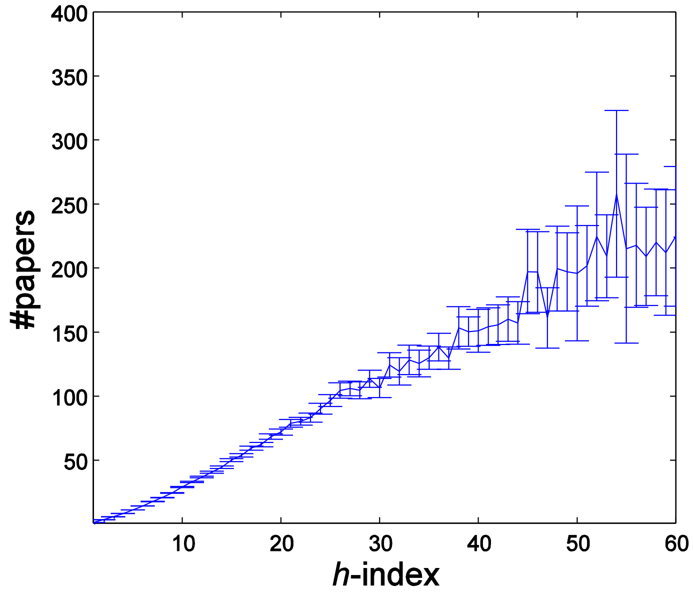
Standing on the shoulders of giants

--- Isaac Newton

Q & A

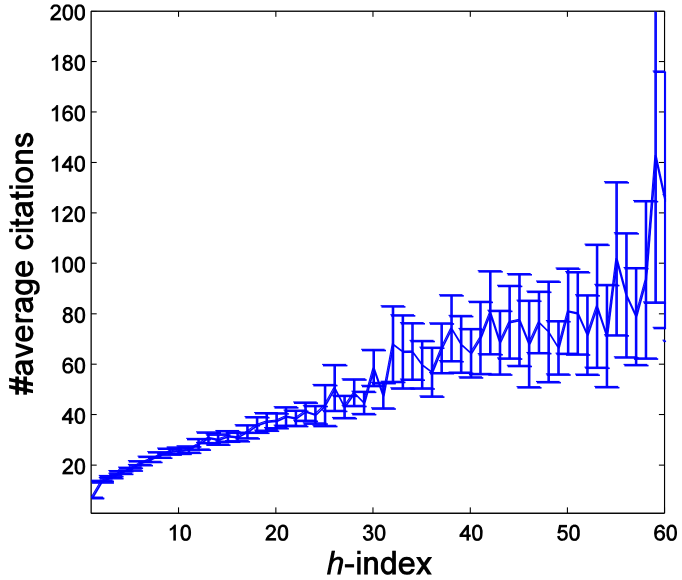
Scientific Impact: *h*-index

h-index vs. #papers



Scientific Impact: *h*-index

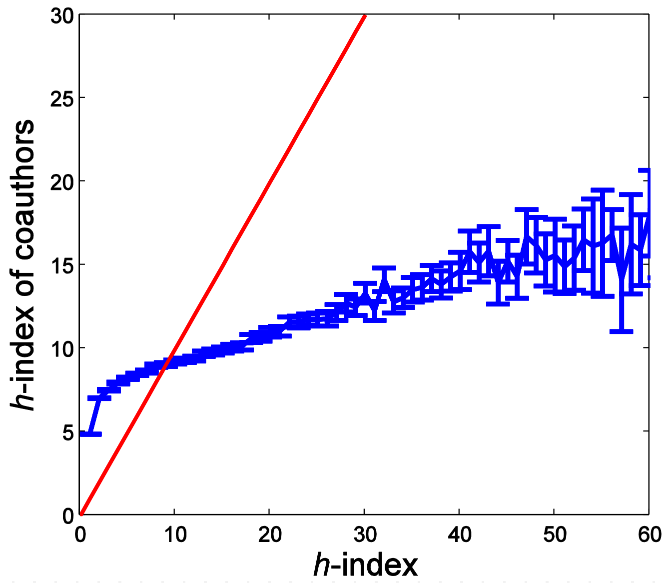
h-index vs. #average-citations



The average number of citations for each author is larger than her/his *h*-index.

Scientific Impact: *h*-index

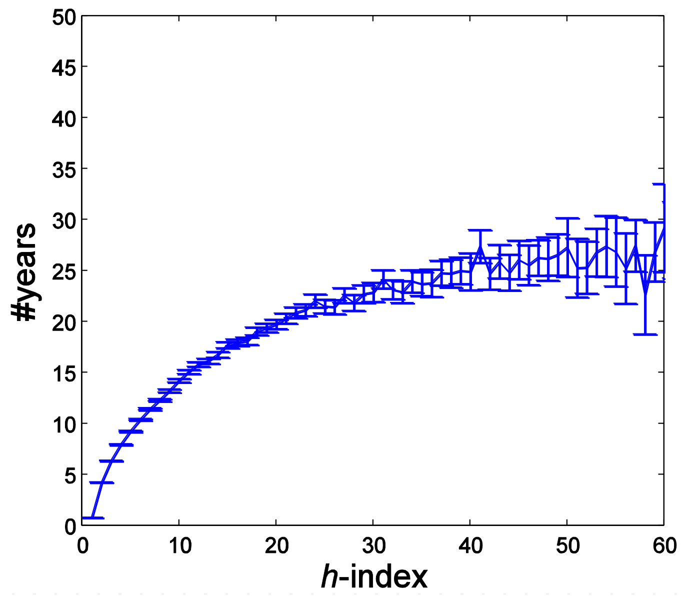
h-index vs. average *h*-index of coauthors



Typically, the author's *h*-index becomes larger than the co-authors' *h*-indices at the expected point of the author's Ph.D. graduation.

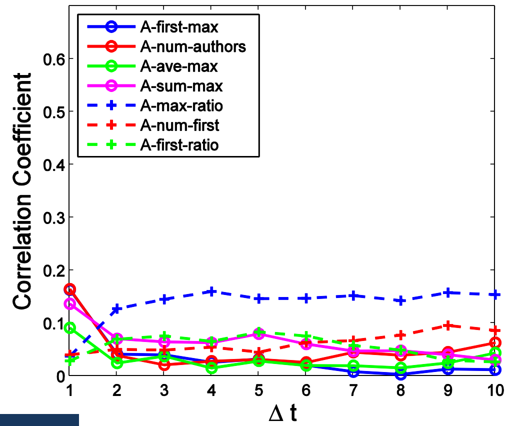
Scientific Impact: *h*-index

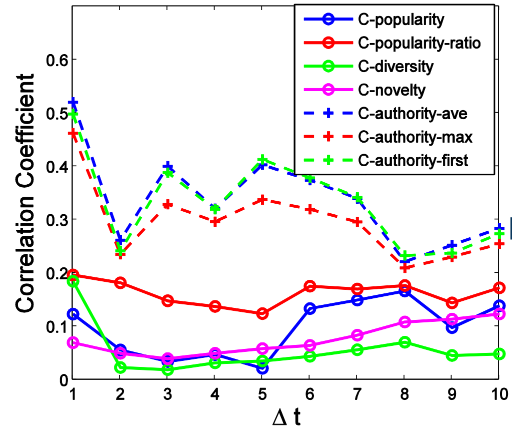
h-index vs. #career years



The rate at which the h-index increases itself increases as the length of time spent in academia becomes longer (i.e., the rich get richer).

Factors Correlation 2

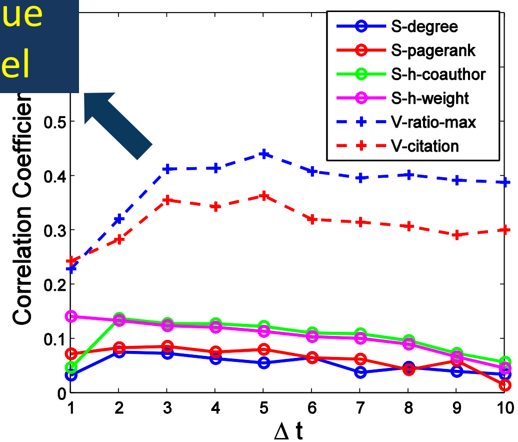


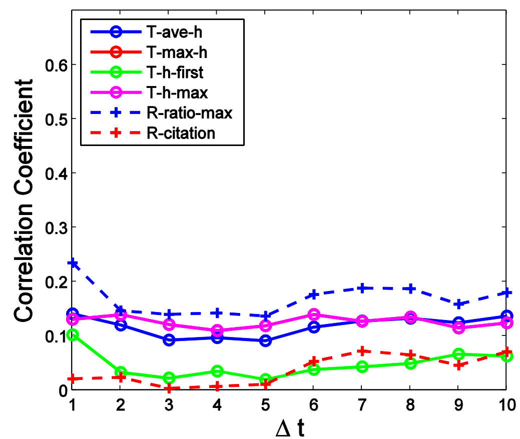


authors' authority on the topics of this paper

t = 2002

venue level

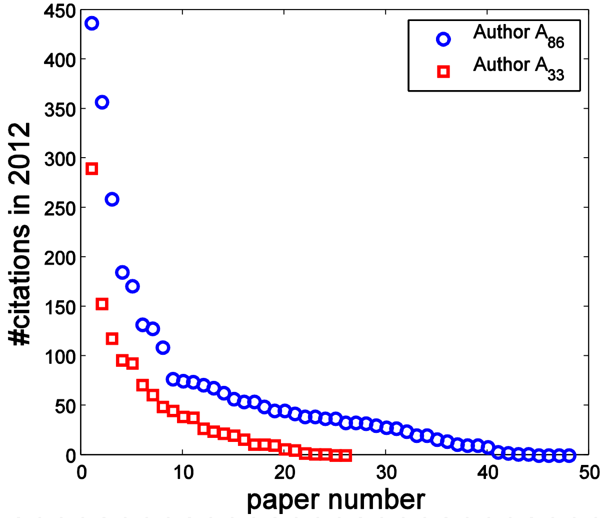




X-axis:
Δt

Y-axis:
correlation coefficient

Prediction: case study 1

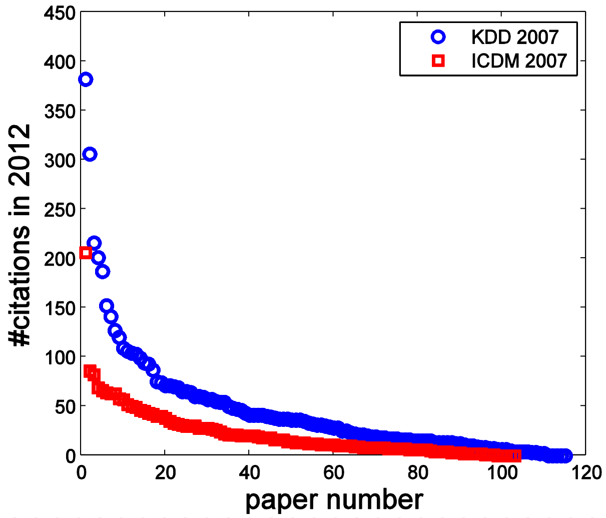


Two anonymous authors
A_{86} and A_{33}

Authors	Pre.	Rec.	F_1	AUC	Accu.	Pre@k	MAP
A_{86}	0.500	0.375	0.429	0.584	0.833	0.375	0.346
A_{33}	1.000	0.667	0.800	0.856	0.885	0.667	0.849

t = 2007
Δt = 5
Logistic regression

Prediction: case study 2



Two venues
KDD and ICDM

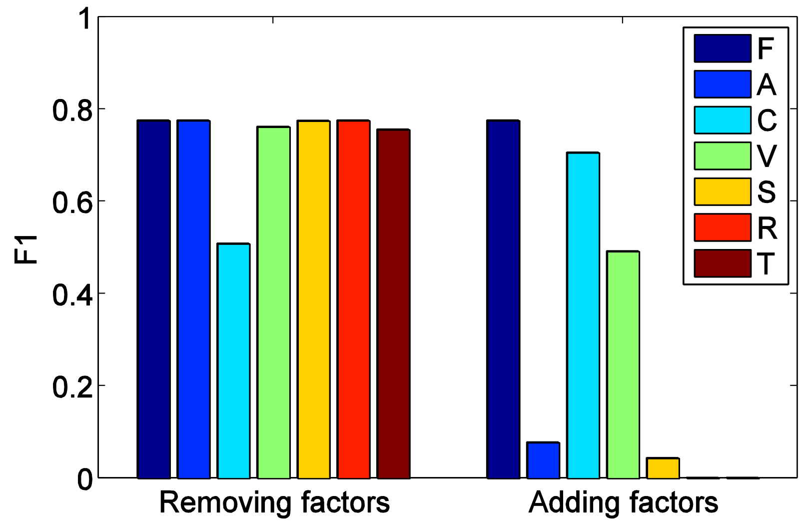
Venues	Pre.	Rec.	F_1	AUC	Accu.
KDD'07	0.800	0.889	0.842	0.884	0.818
ICDM'07	0.842	0.593	0.696	0.886	0.825

t = 2007

Δt = 5

Logistic regression

Prediction: factor contribution



F: Full
A: Author
C: Content
V: Venue
S: Social
R: Reference
T: Temporal

t = 2007

Δt = 5

Logistic regression