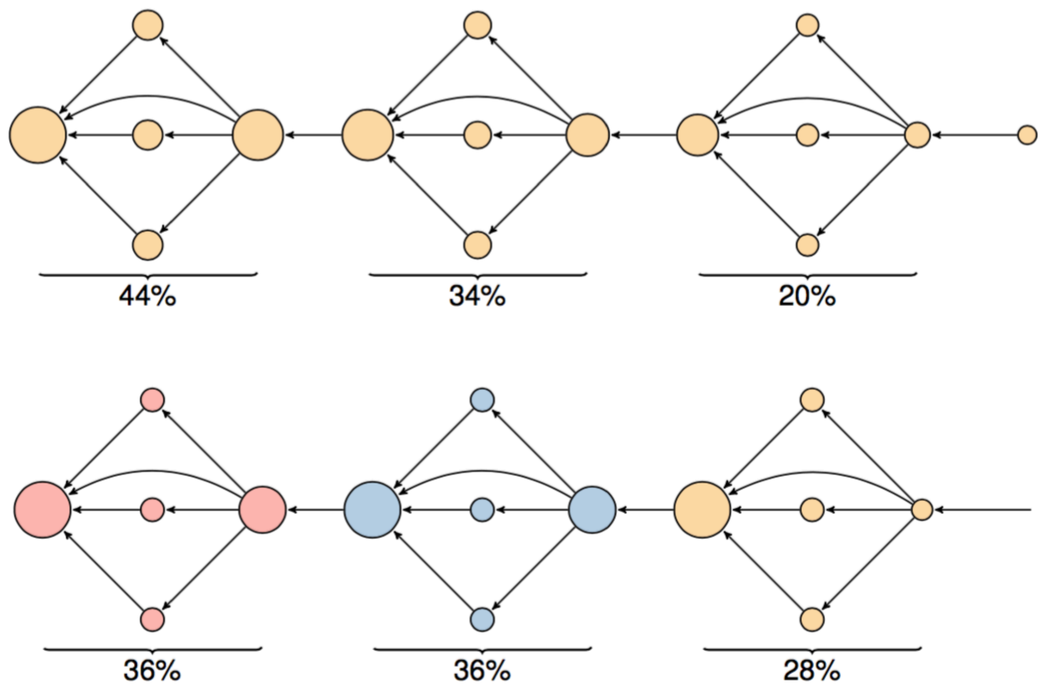
Article-Level EigenFactor (ALEF)



Jevin West, Information School, University of Washington
Ian Wesley-Smith, Information School, University of Washington
Carl T. Bergstrom, Department of Biology, University of Washington

Microsoft Research

Home Rules Leaderboard WSDM 2016 Log in

The Graph

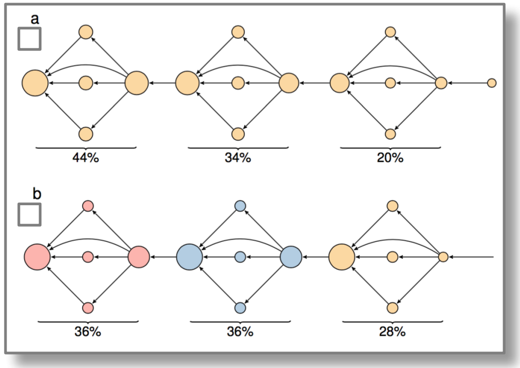
The Microsoft Academic Graph is a heterogeneous graph containing scientific publication records, citation relationships between publications, as well as authors, institutions, journal and conference "venues," and fields of study.

The Data

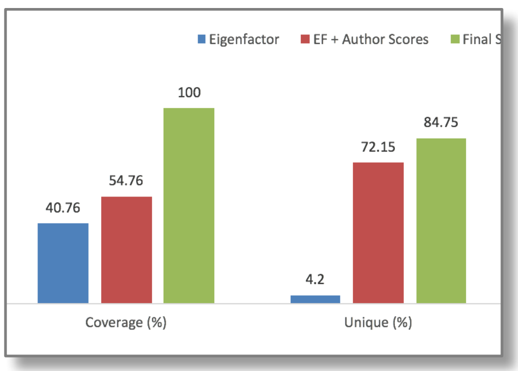
This data is available as a set of zipped text files stored in Microsoft Azure blob storage and available via HTTP. The file size (zipped) is ~30GB and may be downloaded here.

The Challenge

The goal of the Ranker Challenge is to assess the query-independent importance of scholarly articles, using data from the Microsoft Academic Graph.



Article-level Eigenfactor



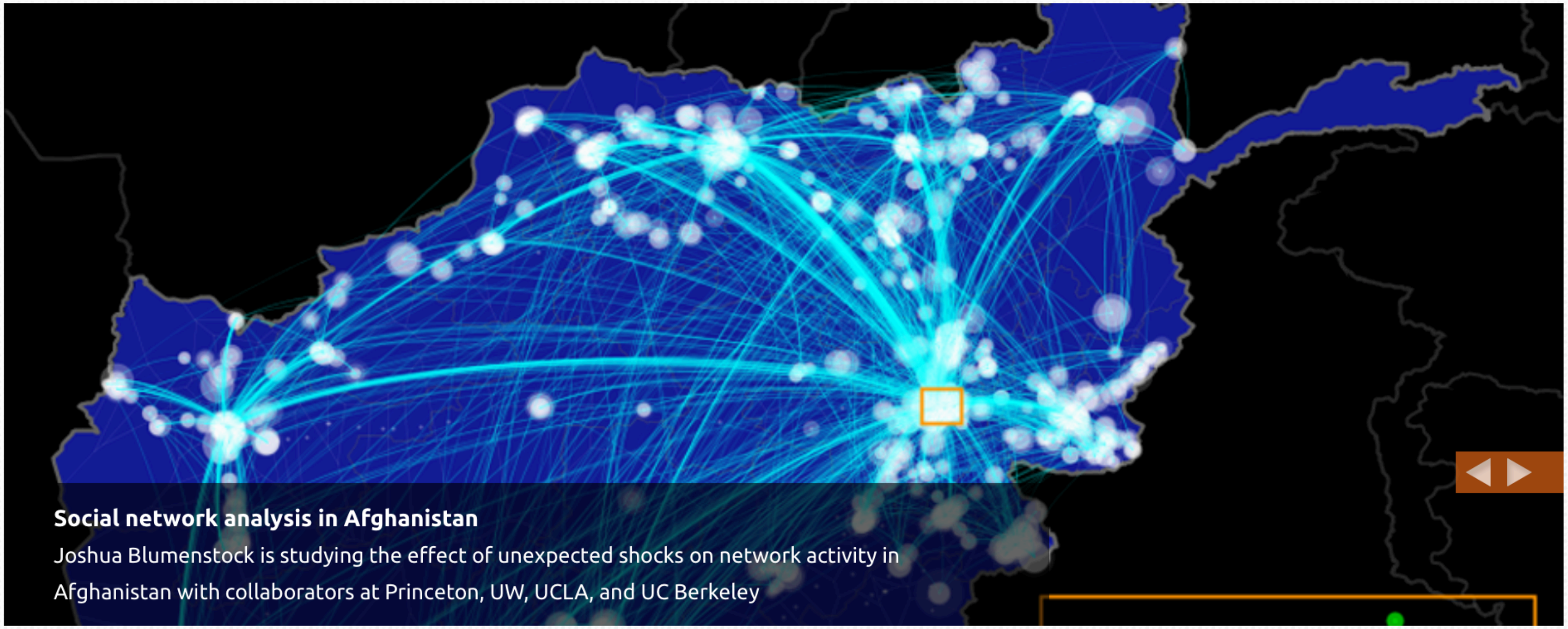
WSDM Cup Challenge

W DataLab
Data Science and Analytics Lab

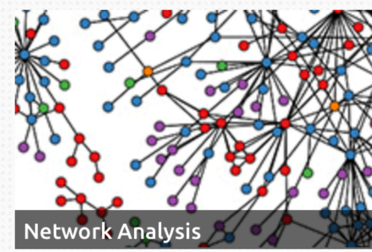
University of Washington Information School

Projects | Publications | People | About

Search



Research Focus Areas



News and Updates

28 **Blumenstock at Population Association of America**

What we do

The DataLab is the nexus for research on Data Science and Analytics at the UW iSchool. We study **large-scale, heterogeneous human data** in an

Journal Ranking

$$P = \alpha H + (1 - \alpha)\, a.e^T$$

Matrix representing the random walk over citations

Probability of not teleporting

Cross-citation Matrix dictating the structure of the citation network

Probability of teleporting to completely new journal weighted by the number of articles in that journal

$$EF = 100 \frac{H\pi}{\sum_i [H\pi]_i}$$

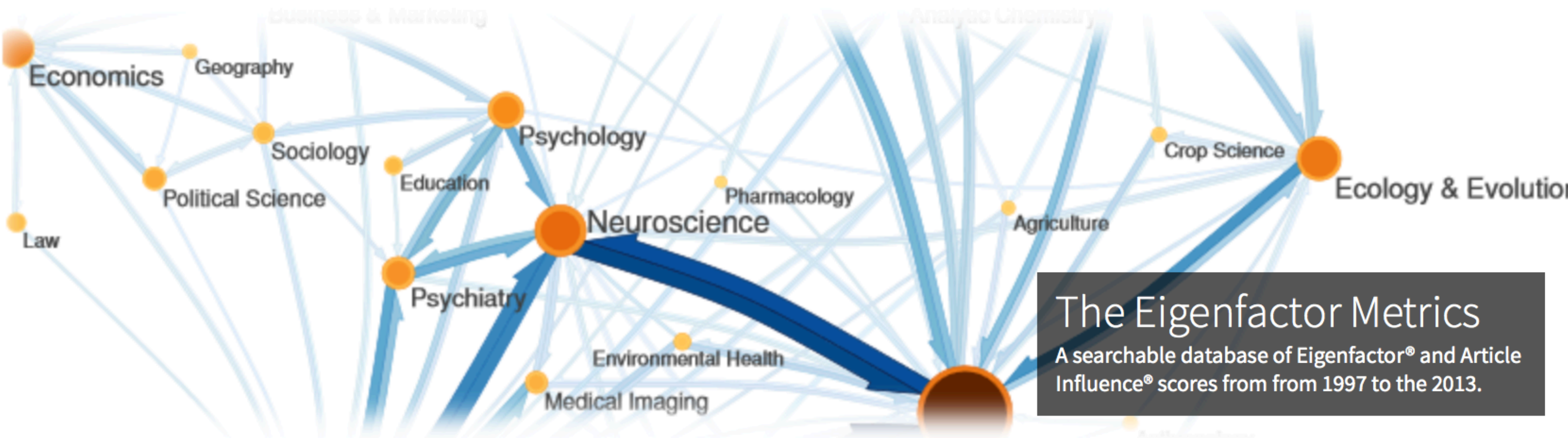
Leading eigenvector of the random walk matrix P.

Normalization

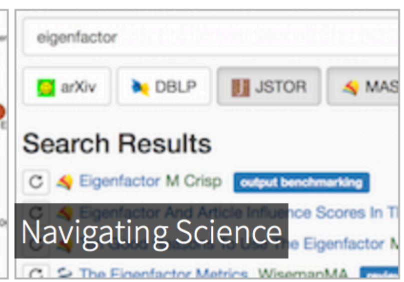
West, JD et al. (2010) *College of Research Libraries*

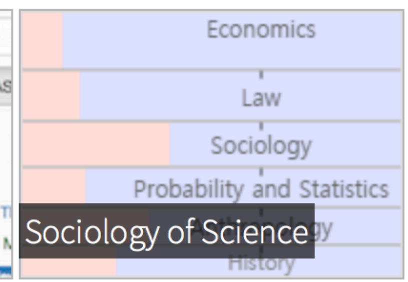
EIGENFACTOR.org

HOME | PROJECTS | PAPERS | ABOUT



RESEARCH AREAS





NEWS

23 Nov.

JEVIN WEST ON MEGAJOURNALS IN THE *CHRONICLE OF HIGHER EDUCATION*

Jevin West discusses the rise of the megajournal and our open access cost effectiveness tool in the *Chronicle of Higher Education*.

---

23 Nov.

EIGENFACTOR TEAM PLACES SECOND IN MICROSOFT RESEARCH'S WSDM CUP

The WSDM Cup Challenge asked teams to use 30GB of data from the Microsoft Academic Graph to rank the

Ranking and mapping article-level citation networks

Martin Rosvall*
*IceLab, Umeå University**

Jevin West
Information School, University of Washington, Seattle, WA 98195-1800

Daril Vilhena and Carl T. Bergstrom
Department of Biology, University of Washington, Seattle, WA 98195-1800
(Dated: August 15, 2014)

Time-directed networks pose a challenge for flow-based methods of network analysis. Such networks are acyclic or nearly acyclic and thus very far from the nearly ergodic structures that flow-based methods are designed to handle. Without suitable modification, flow-directed ranking algorithms such as the Eigenfactor score put too much weight on older documents. Flow-based methods of cluster detection, such as the map equation approach, can fail to resolve important structures. Here we show how flow-directed methods can be modified to avoid these problems and thereby perform well on time-directed networks. To demonstrate the power of the new *article level Eigenfactor* metrics, we rank the 1.8 millions articles in JSTOR. To illustrate the power of our clustering approach, we create a hierarchical citation map of the JSTOR corpus.

Science is a massively parallel human endeavor to explain and predict the nature of the physical world. Thousands of individual scholars build cumulatively upon the prior work of a yet greater number of authors, and report upon their progress through their scholarly publications. Following the conventions of scholarly citation, each author references those predecessors most important in the development of her ideas. The most innovative research opens up avenues between new ideas—novel citation trails. Other researchers follow these tracks, guided by the citations that the pioneers laid down. What emerges is a latticework of citations from which we can in principle map the geography of scientific thought and retrace the pathways along which intellectual activity has proceeded. As Derek de Solla Price famously noted in 1965, this lattice is in fact a vast network of citations, growing dynamically and organically, doubling in size every ten to twenty years [1]. Our aim is to map out the way that ideas flow through scientific communities, so that we can comprehend large-scale patterns, identify important contributions, and better navigate the literature.

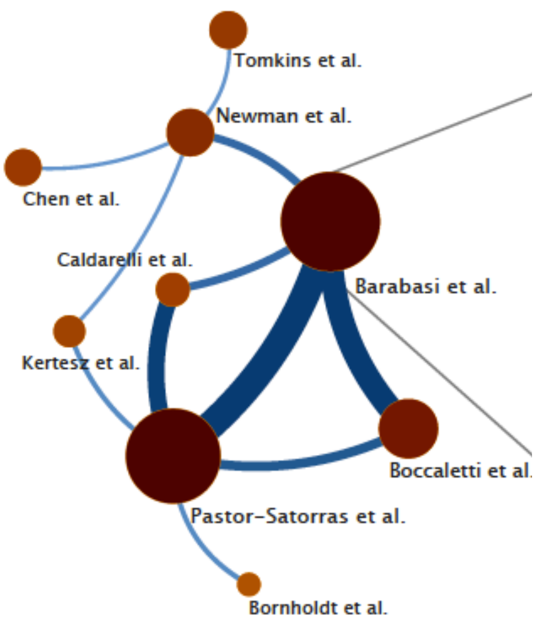
Network theory offers a rich set of tools for ranking and trails of citations move inexorably back in time.

In this paper, we show how flow-based methods of network analysis can be adapted for use with time-directed and acyclic networks. This will allow us to apply flow-based methods to map science at the scale of individual articles. To use network analysis to better navigate the literature, we also present a method to label the structures that we uncover at every scale, using the textual content of the articles themselves. We will need to do all of this in a way that is scalable to the full universe of scholarly publication, and that is updatable, so that researchers may always be navigating with maps that are current not to years, but to days. These methods will also be suitable for studying patent networks, court case networks, and other similarly time-directed structures.

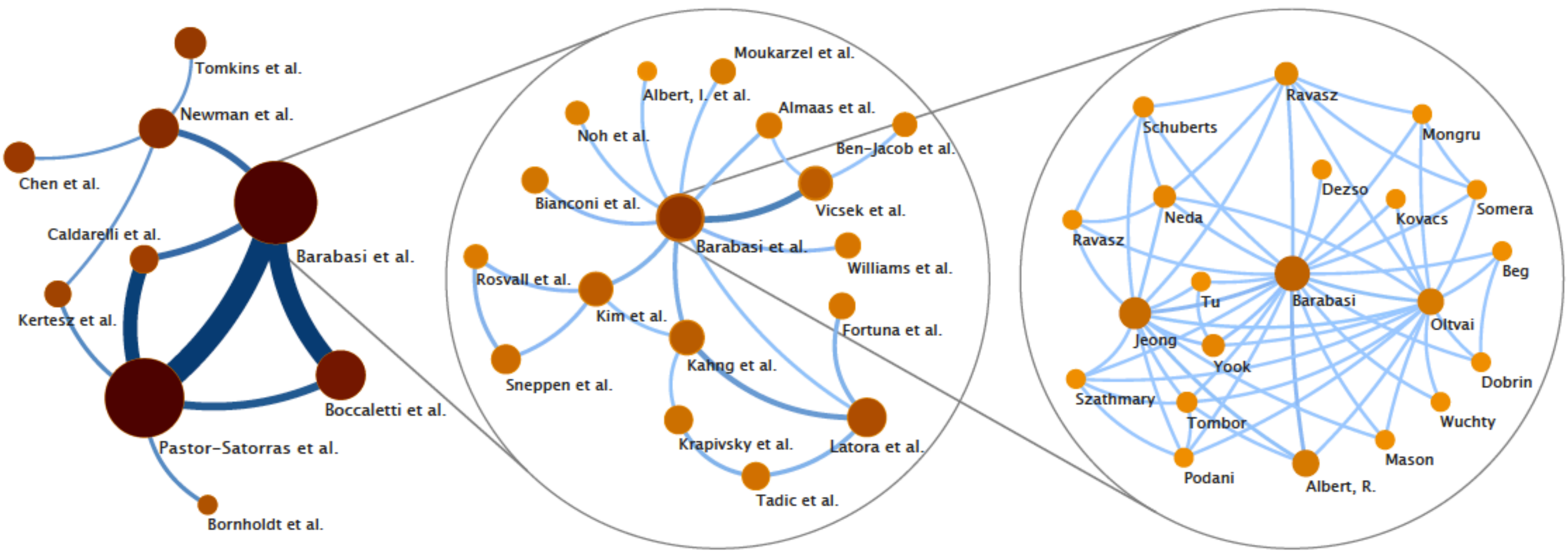
I. RANKING

The standard PageRank algorithm can be viewed as tracing the path of a random walker on a directed, and possibly weighted, network. Most of the time, the random walker

Hierarchical Mapping
without ALEF

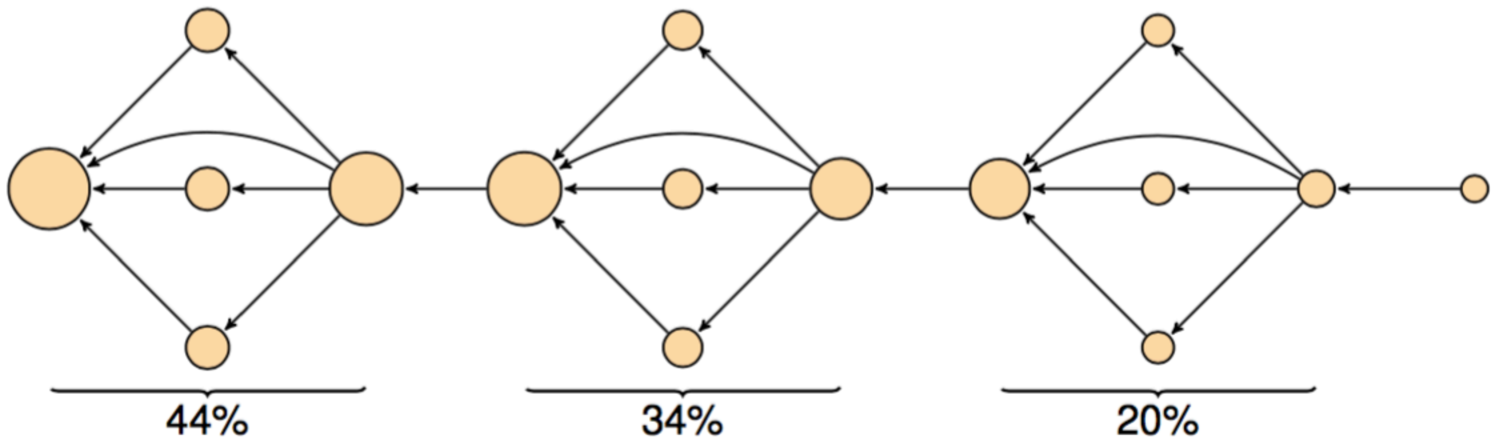


Hierarchical Mapping *with ALEF*

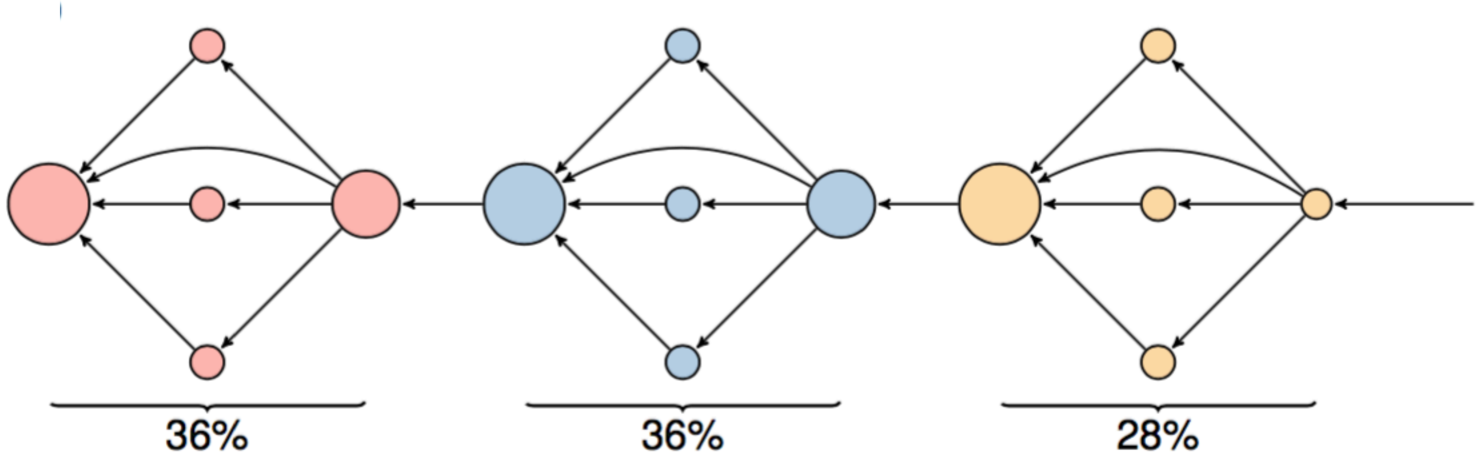


Flow Distribution

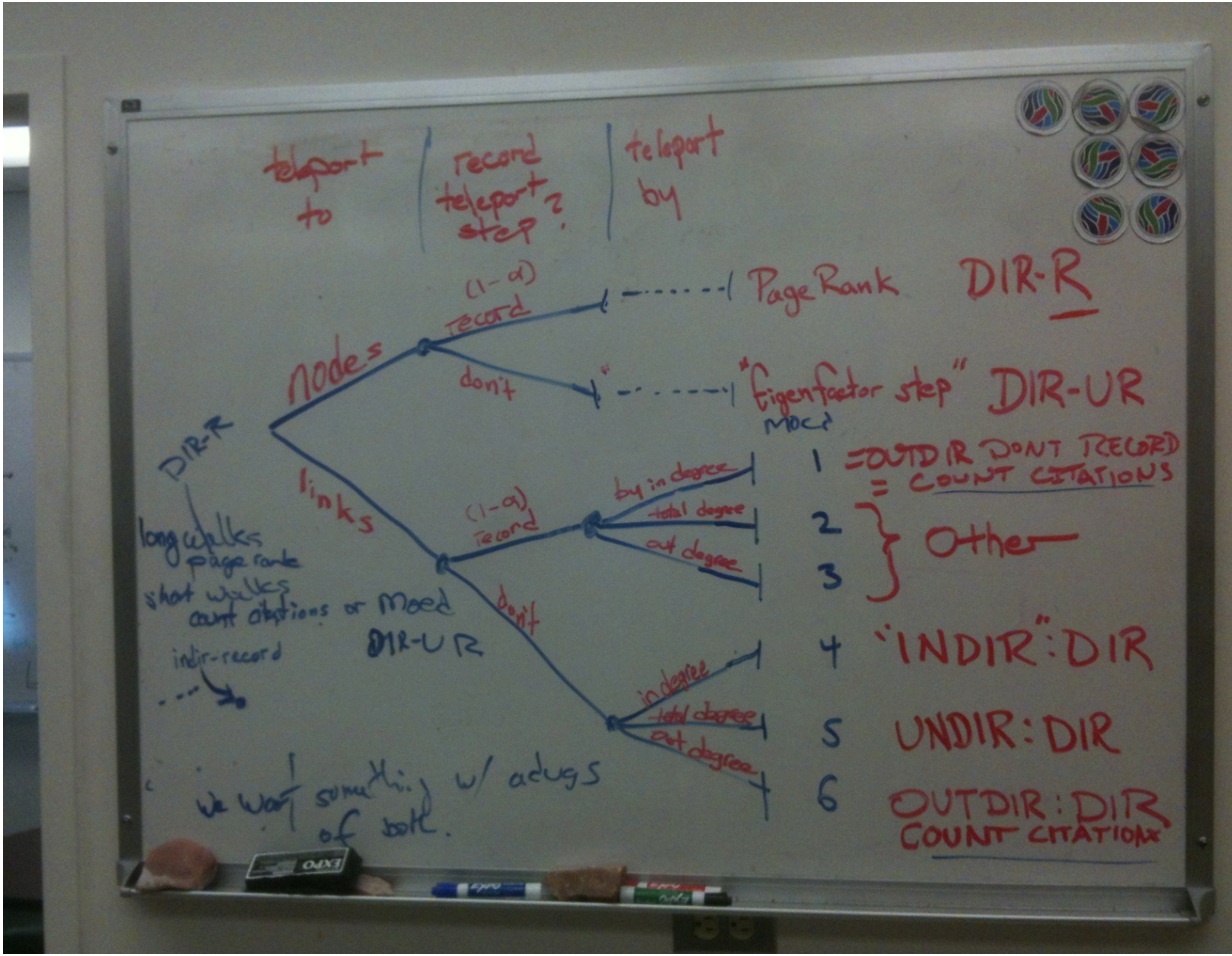
PageRank



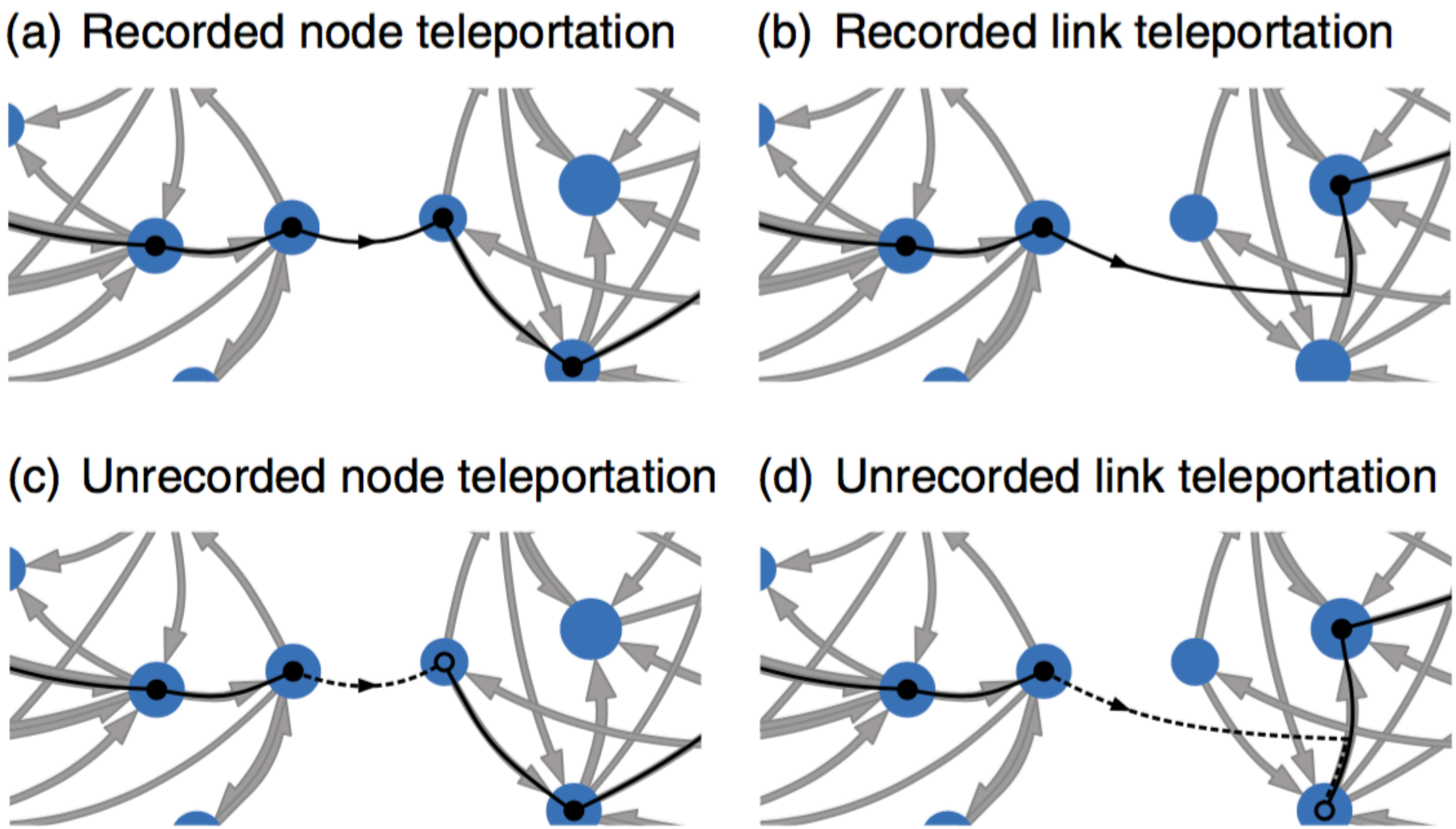
ALEF



Time


teleport to
record teleport step?
teleport by
(1-α) record
don't
PageRank
DIR-R
"Eigenfactor step"
DIR-UR
nodes
links
DIR-R
MOed
1 =OUTDIR DONT RECORD = COUNT CITATIONS
(1-α) record
by in degree
total degree
out degree
2
3
Other
don't
4 "INDIR":DIR
in degree
total degree
out degree
5 UNDIR:DIR
6 OUTDIR:DIR COUNT CITATIONS
long walks page rank
short walks count citations or MOed
DIR-UR
indir-record
we want something w/ advugs of both.
EXPO

Smart Teleportation



Lambiotte & Rosvall (2012) *PhysRevE*

Mechanics

adjacency matric

1. calculate step weight

$$w_i = \sum_j^n (Z_{ij} + Z_{ij}^T)$$

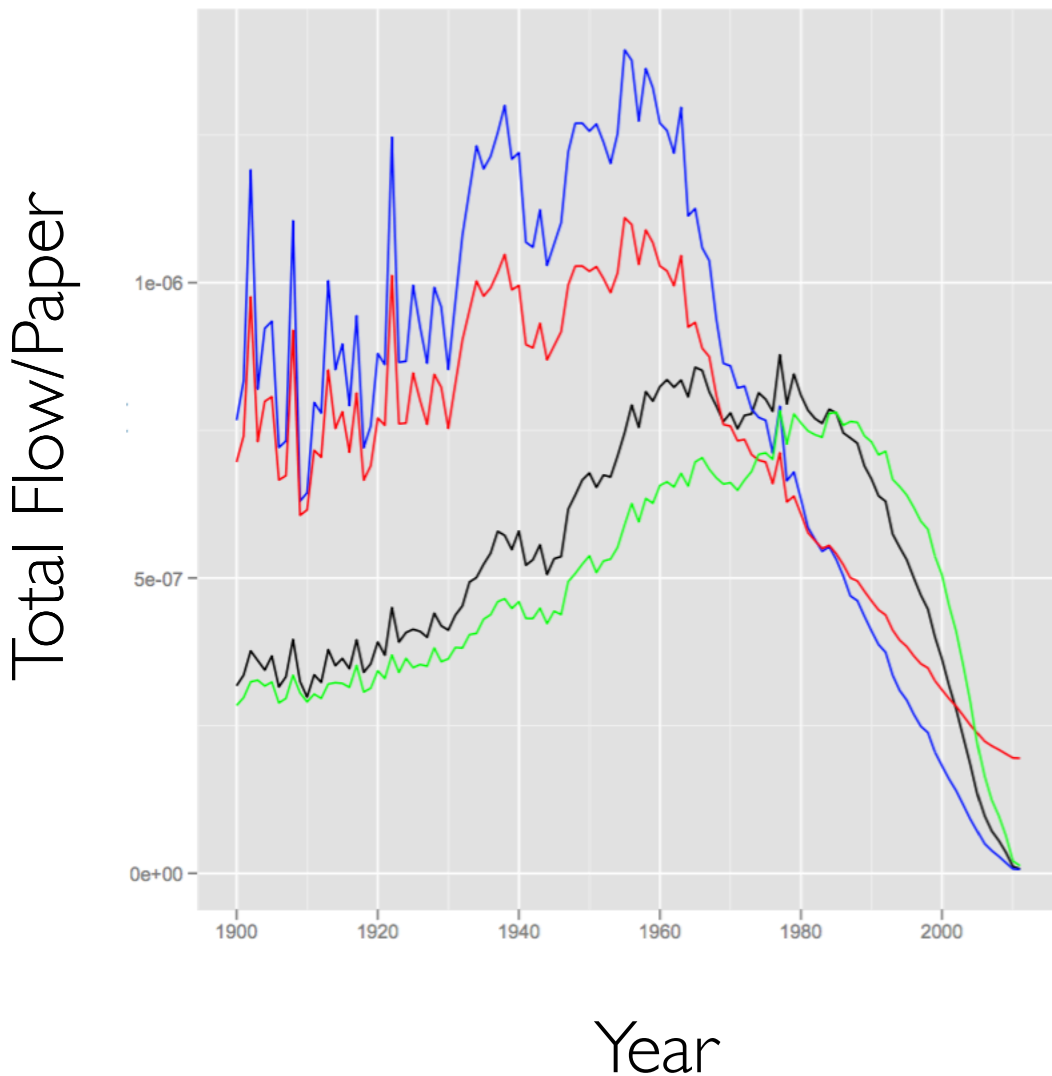
2. make row stochastic

$$\mathbf{H}_{ij} = \frac{\mathbf{Z}_{ij}}{Z_i}$$

3. one-step on network

$$\mathrm{ALEF} = n \frac{\mathbf{H}_{ij}^T . w_i}{\sum_j [\mathbf{H}_{ij}^T . w_i]_j}$$

Flow Distribution (JSTOR)



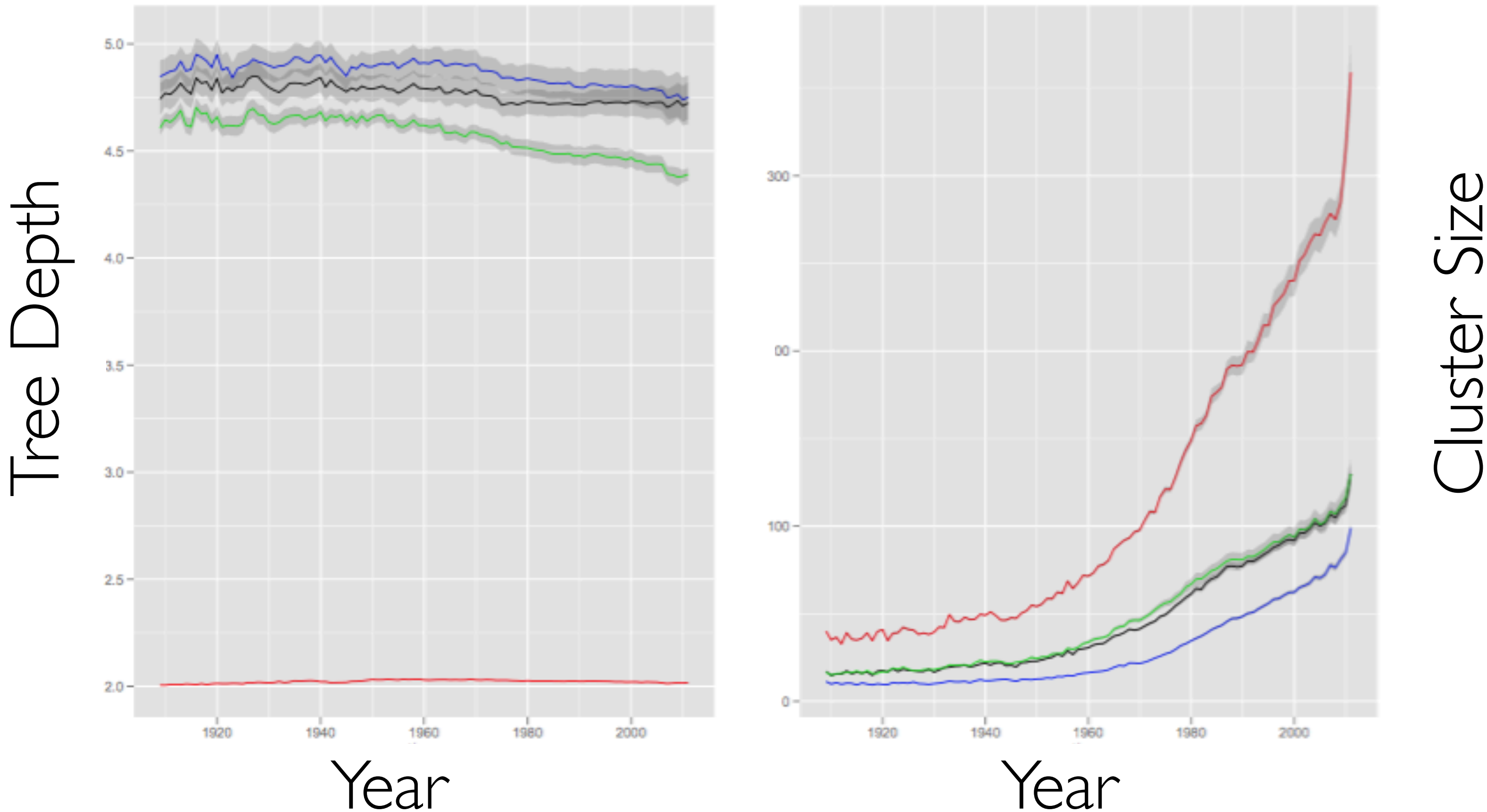
Black = ALEF

Green = citations

Red = PageRank

Blue = unrecorded teleport

Tree Depth and Cluster Size



Black = ALEF

Green = OUTDIR

Red = DIR-R

Blue = DIR-UR

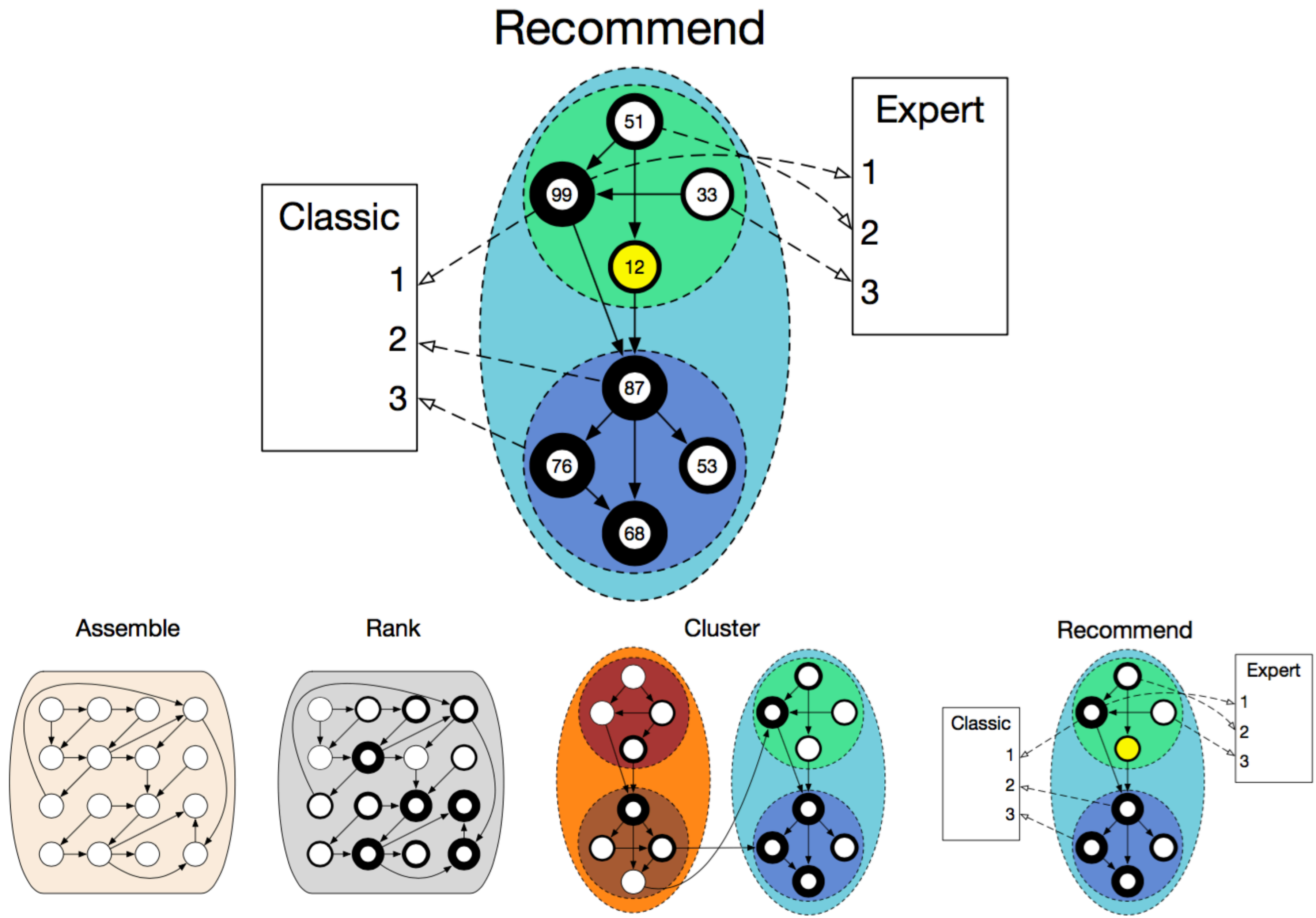
ALEF Strengths

Performs well

Simple mechanics

Fast calculation

High resolution partitions



West, Wesley-Smith, Bergstrom (2016) A recommendation system based on hierarchical clustering of an article-level citation network. *IEEE, Transactions on Big Data*

Papers

- J.D. West, M. Rosvall, C.T. Bergstrom (2016) Ranking and mapping article-level citation networks, *in prep*
- *J.D. West, I. Wesley-Smith, C.T. Bergstrom (2016)* A recommendation system based on hierarchical clustering of an article-level citation network. *IEEE, Transactions on Big Data*
- *I. Wesley-Smith, C.T. Bergstrom, J.D. West (2016) Static Ranking of Scholarly Papers using Article-Level Eigenfactor (ALEF)*, WSDM Conference: Entity Ranking Challenge Workshop
- I. Wesley-Smith, J.D. West (2016) Babel: A platform for research in scholarly article recommendation. *WWW Conference, Workshop on Big Scholarly Data*

Babel Home About API

babel.eigenfactor.org

oren etzioni

DBLP JSTOR MAS PLOS PubMed

Ian Wesley-Smith

...s

...l Methods For Analyzing Speedup Learning Experiments O Etzioni **satisfaction programs**

...Face And Computer-Mediated Communities Amitai Etzioni, Oren Etzioni 1998 **resources sustained**

...ument Clustering O Zamir **document clustering**

... Communities: Virtual Vs. Real A Etzioni 1996 **implications internet**

...l Methods For Analyzing Speedup Learning Experiments. O Etzioni 1993 **scheduling problems**

...l Methods For Analyzing Speedup Learning Experiments O Etzioni 1993 **generating abstractions**

Get Related Web Document Clustering: A Feasibility Demonstration O Zamir 1997 **document clustering**

Get Related Web Document Clustering: A Feasibility Demonstration. O Zamir 1997 **browsing large**

Get Related Sound And Efficient Closed- World Reasoning O Etzioni **proving problem**

Get Related Appears In Comm. OfACM O Etzioni **scalable comparison-shopping**

« Previous 1 2 3 4 5 6 7 8 9 10 Next »

Papers related to Statistical Methods For Analyzing Speedup Learning Experiments O Etzioni **satisfaction programs**

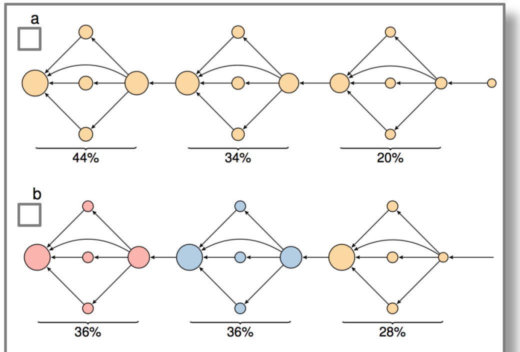
Get Related Automatically Configuring Constraint Satisfaction Programs: A Case Study S Minton 1995 **satisfaction programs**

Get Related Abstraction Via Approximate Symmetry T Ellman 1992 **satisfaction programs**

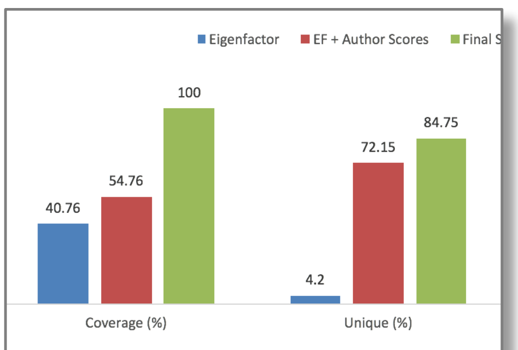
Get Related Integrating Heuristics For Constraint Satisfaction Problems: A Case Study S Minton 1992 **satisfaction programs**

Get Related An Analytic Learning System For Specializing Heuristics S Minton 1992 **satisfaction programs**

Get Related Automated Synthesis Of Constrained Generators W Braudaway 1988 **satisfaction programs**

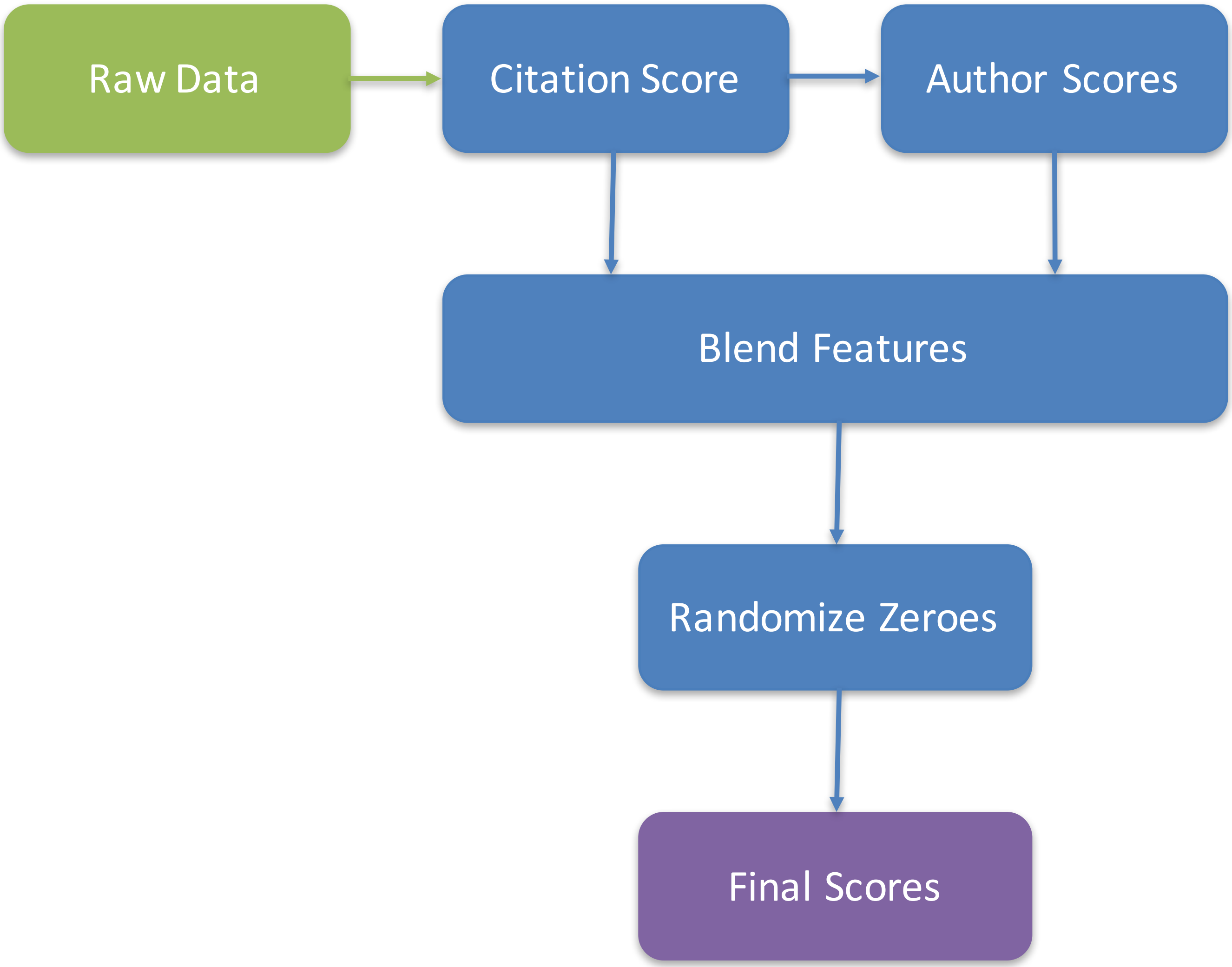


Article-level Eigenfactor

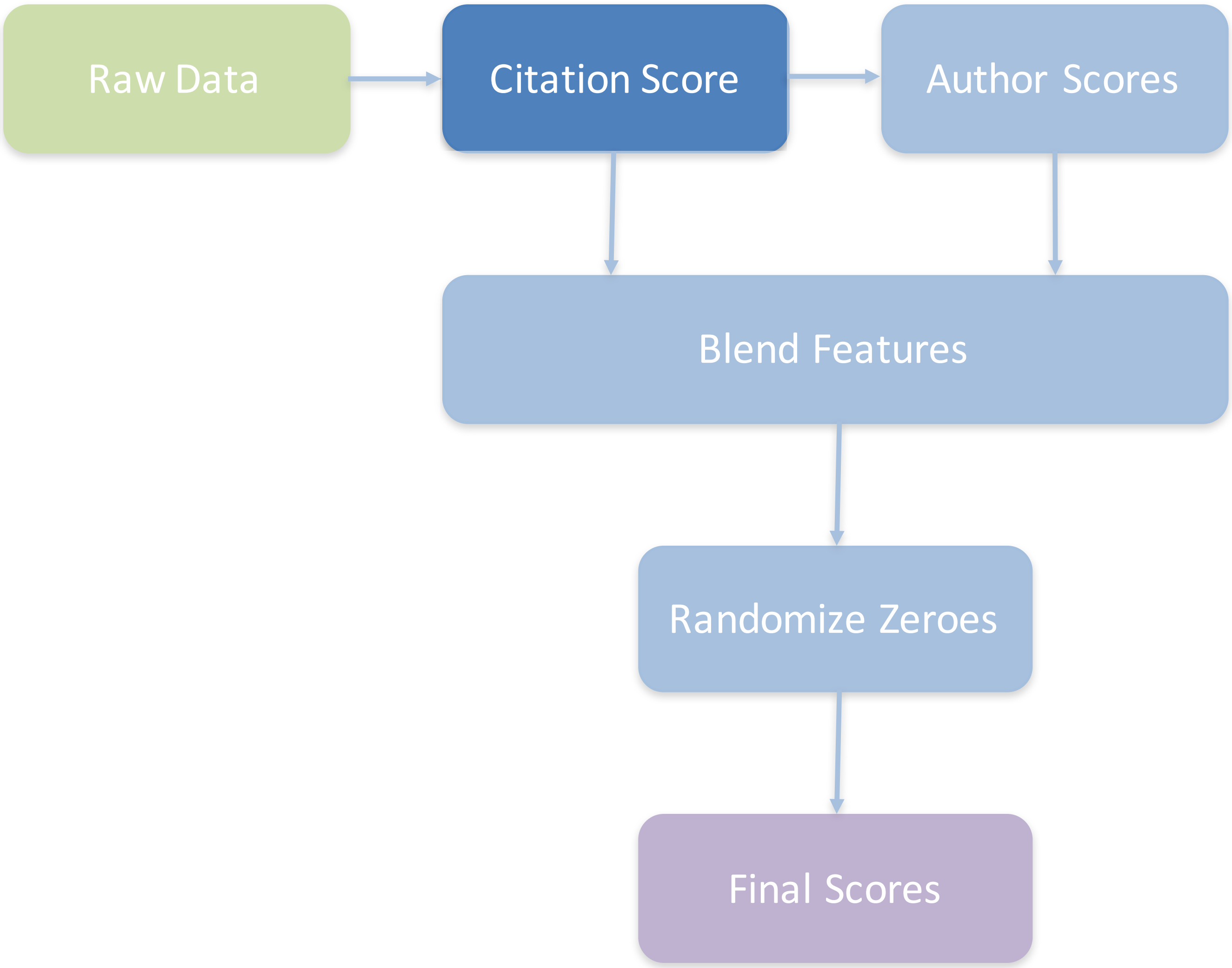


WSDM Cup Challenge

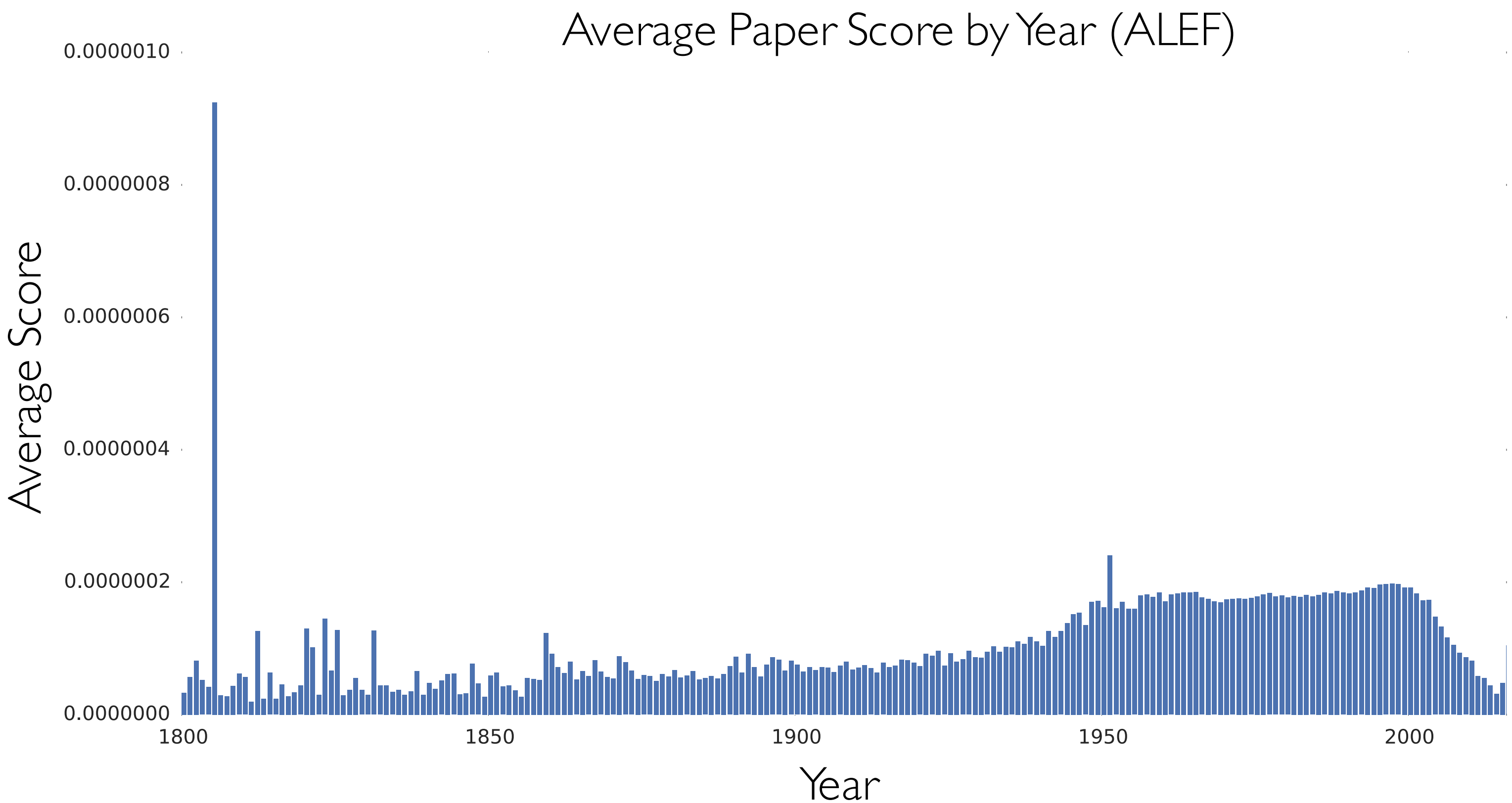
Data Pipeline



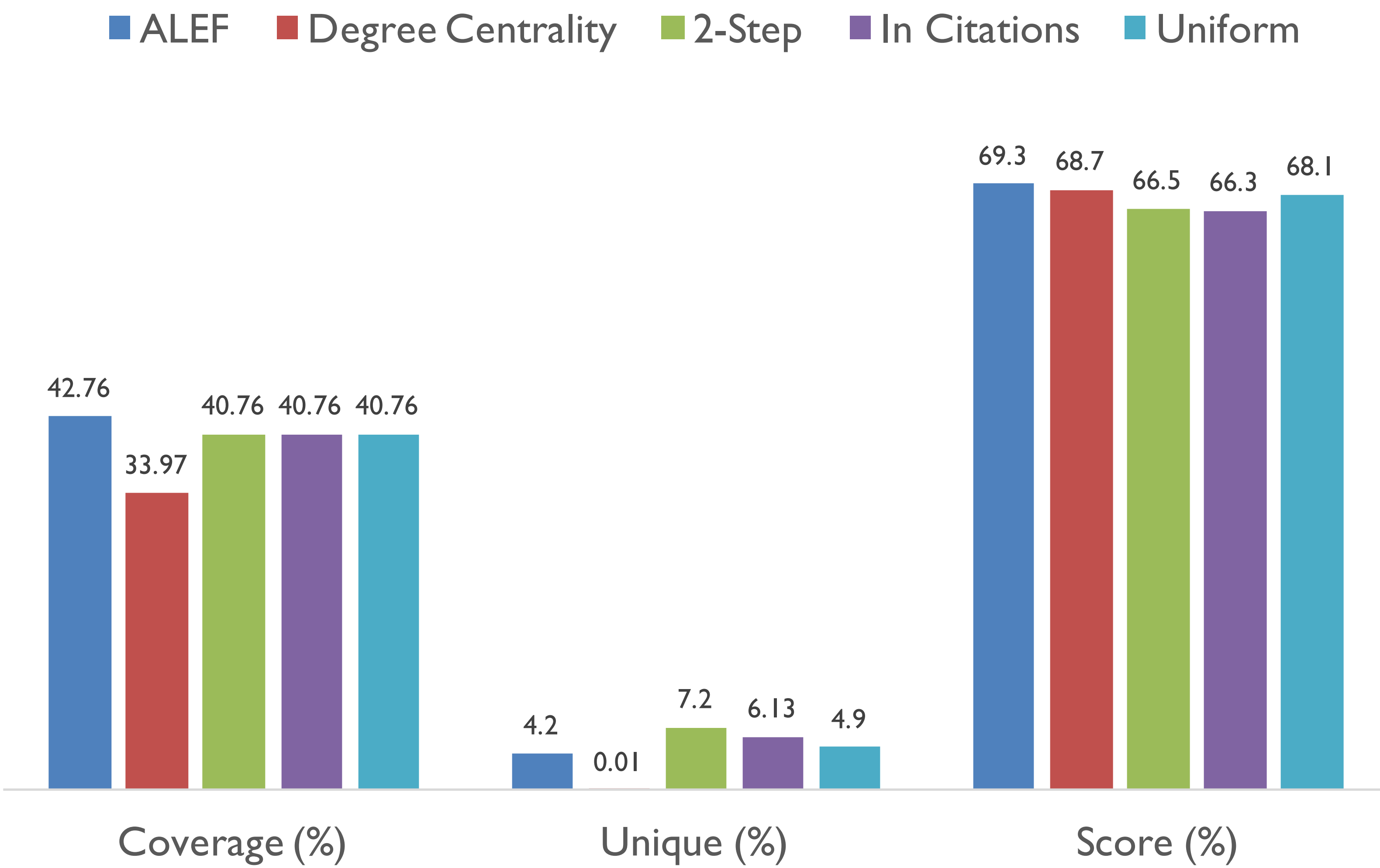
Citation Scores



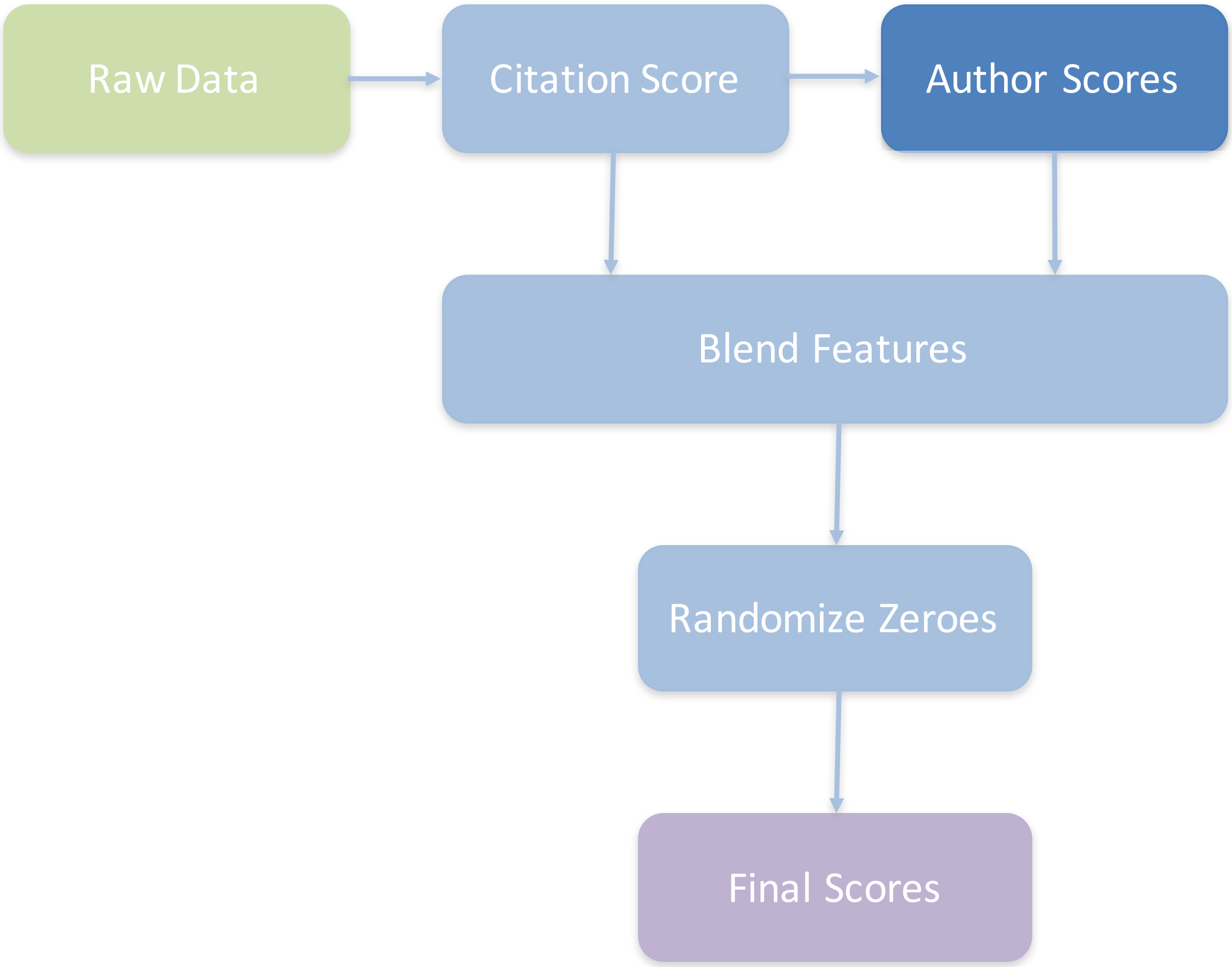
Citation Scores



Citation Variants



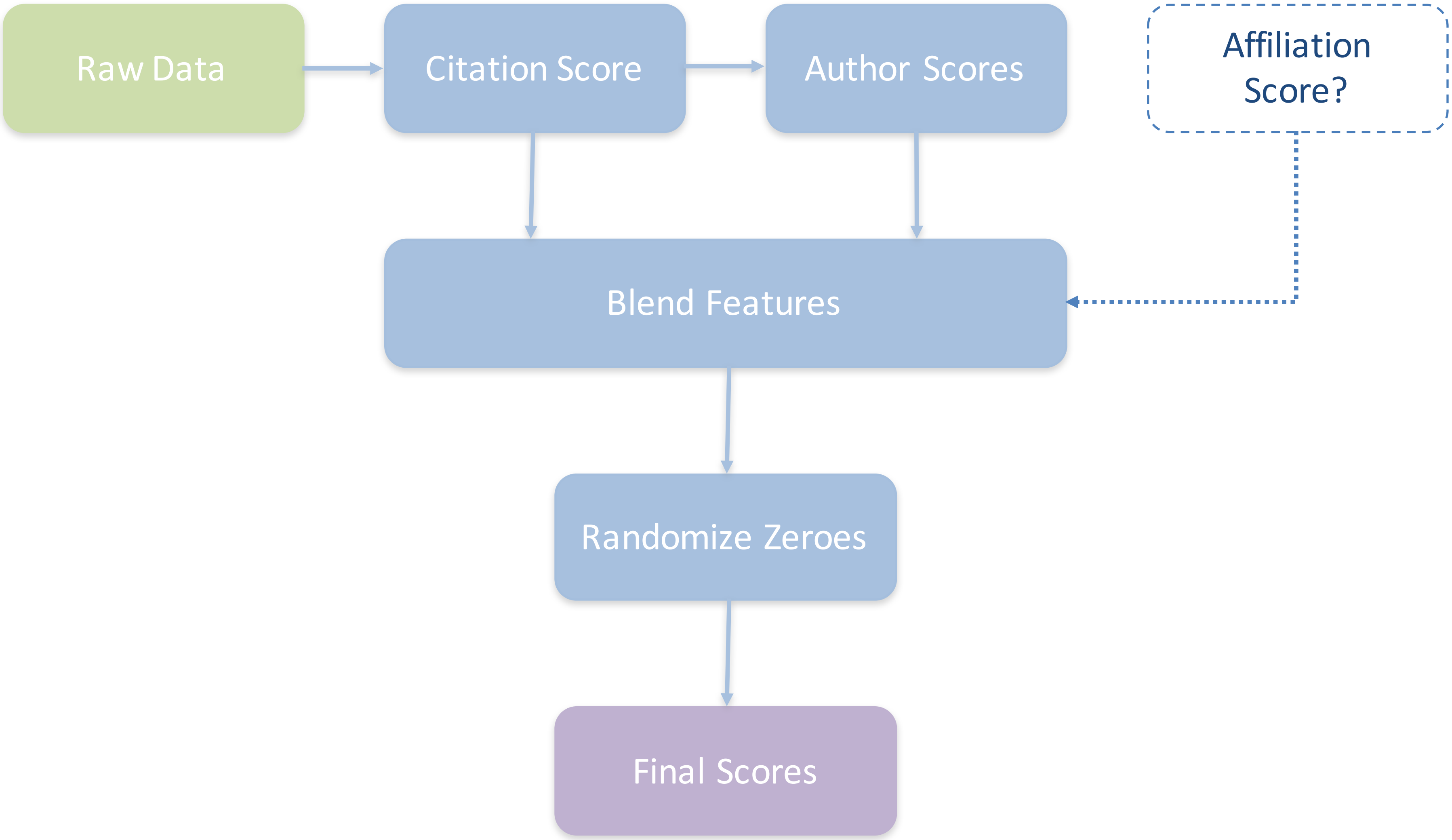
Author Scores



Author Scores

- Author Score = Average citation score of all papers
- How should paper credit be assigned?
  - Equally or Fractional?
- Why not sum?
  - Unique Scores: 72.15% vs 28.27%

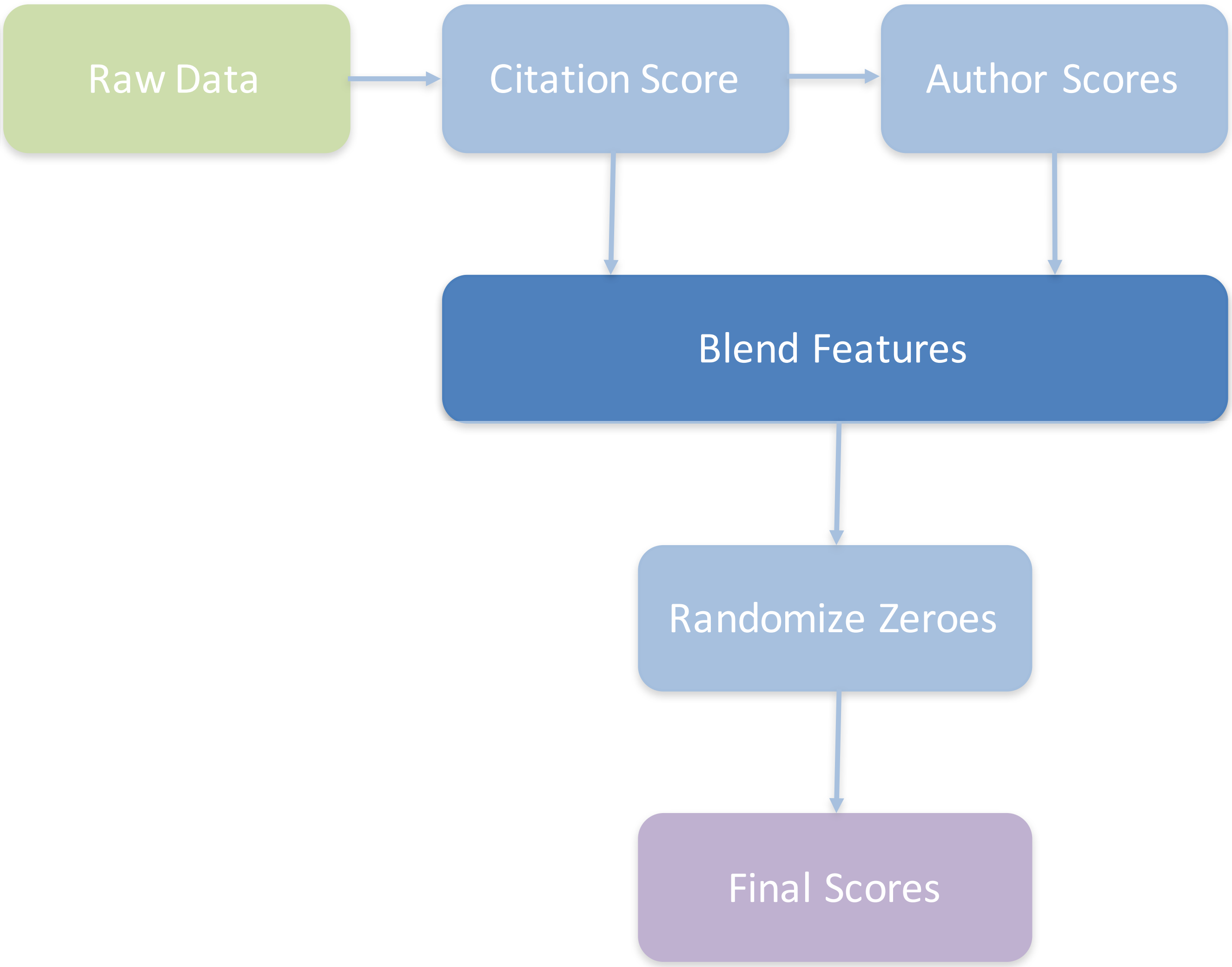
Other Features?



Other Features

- Matching datasets is hard
- Author Affiliation: University of Washington
  - george washington university
  - university of washington bioengineering
  - university of washington information school
  - university of washington school of law
  - university of washington tacoma
  - university of washington bothell
- Coverage is low: 25% of paper-author pairs have an affiliation

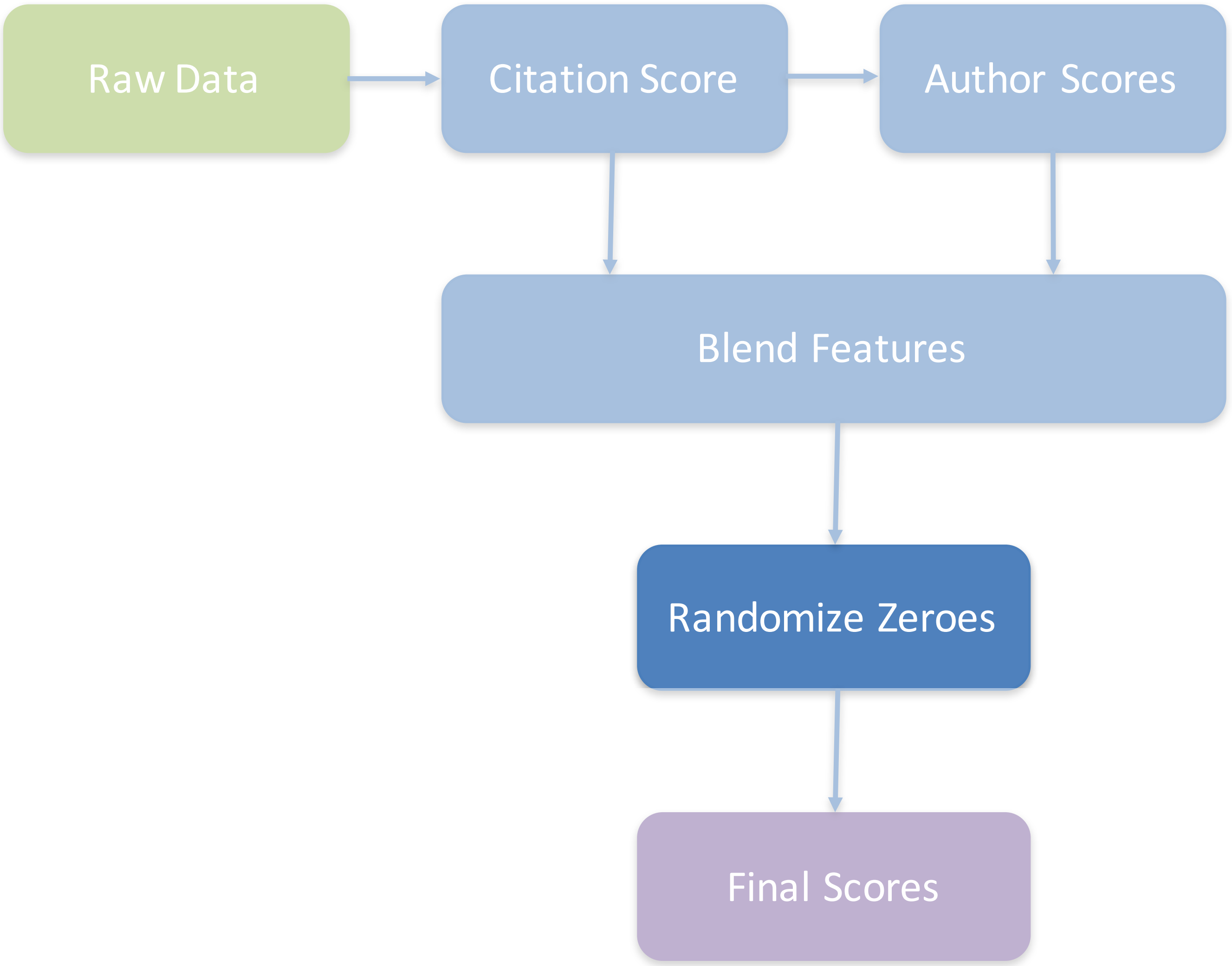
Blend Features



Blend Features

- Weighted Average
  - Weights found via manual parameter sweep
  - Citation Score: 70%
  - Author Score: 30%
- Axiom: Derived scores shouldn't outweigh the source

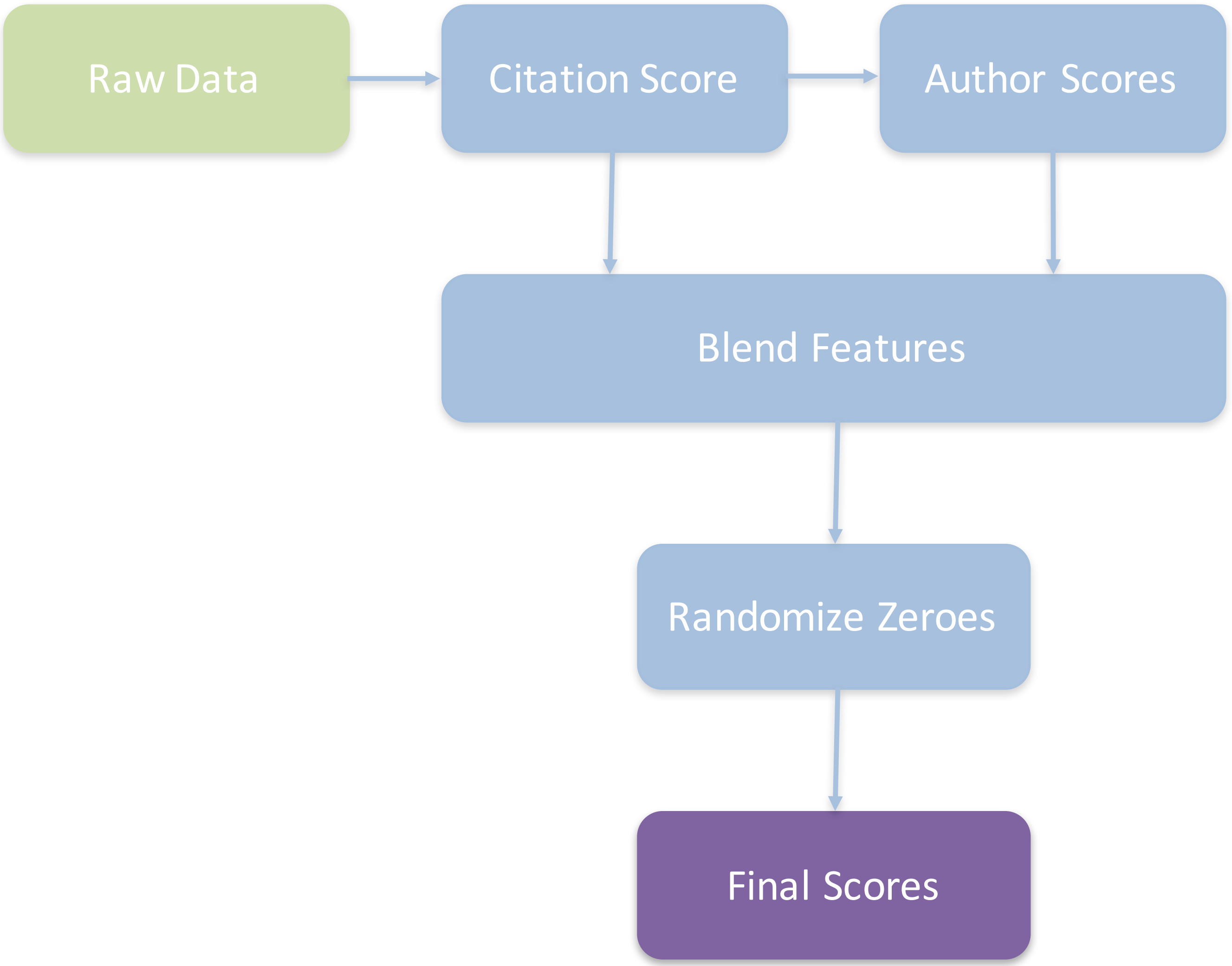
Randomize Zeroes



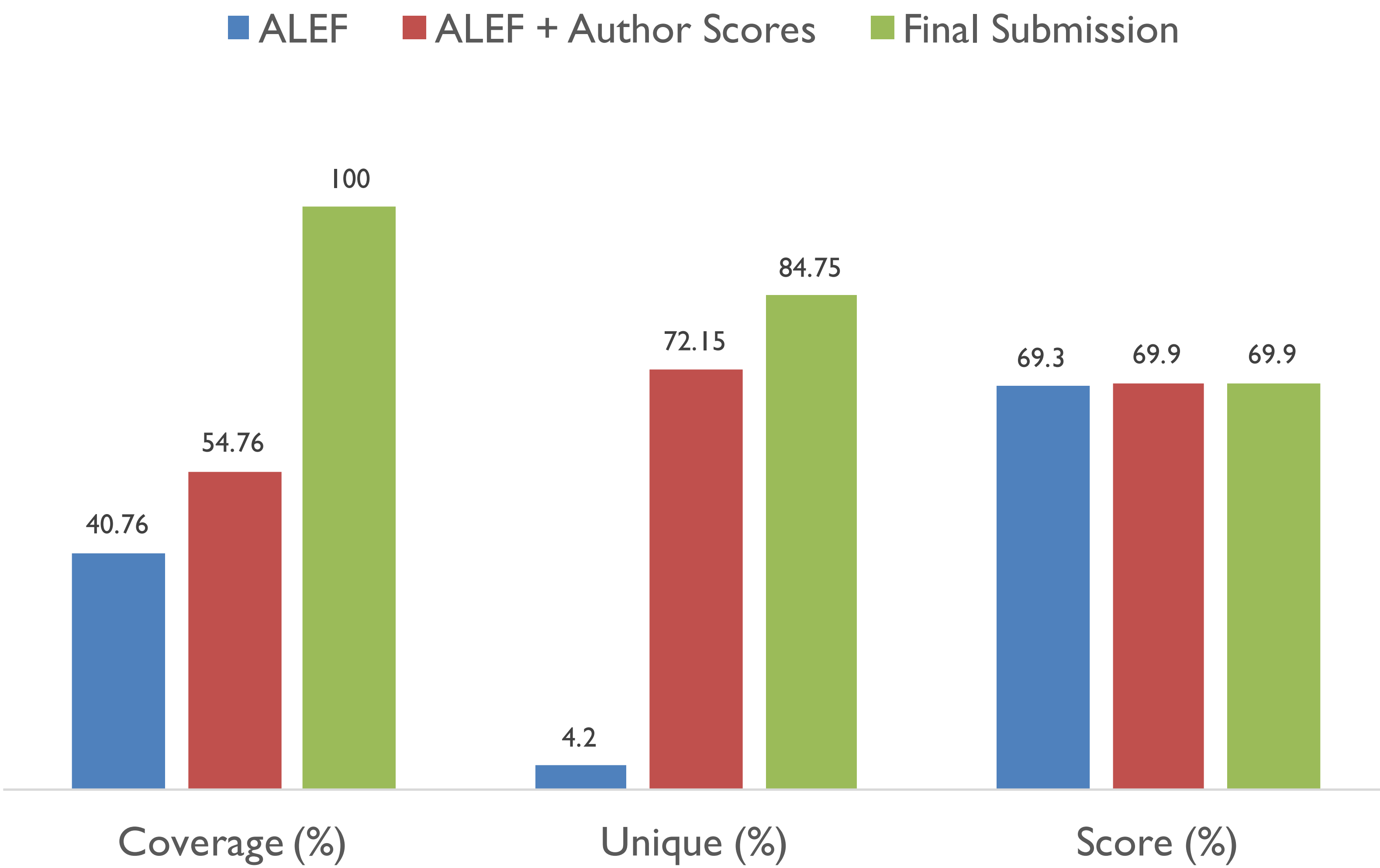
Random Chance?

- Our best isn't much better than random
  - Random: 52.6%
  - 1st: 68.3% (+30%)
- This judging is favorable to random chance
- Unscored papers assigned [0, minval * 0.999]

Phase I Results



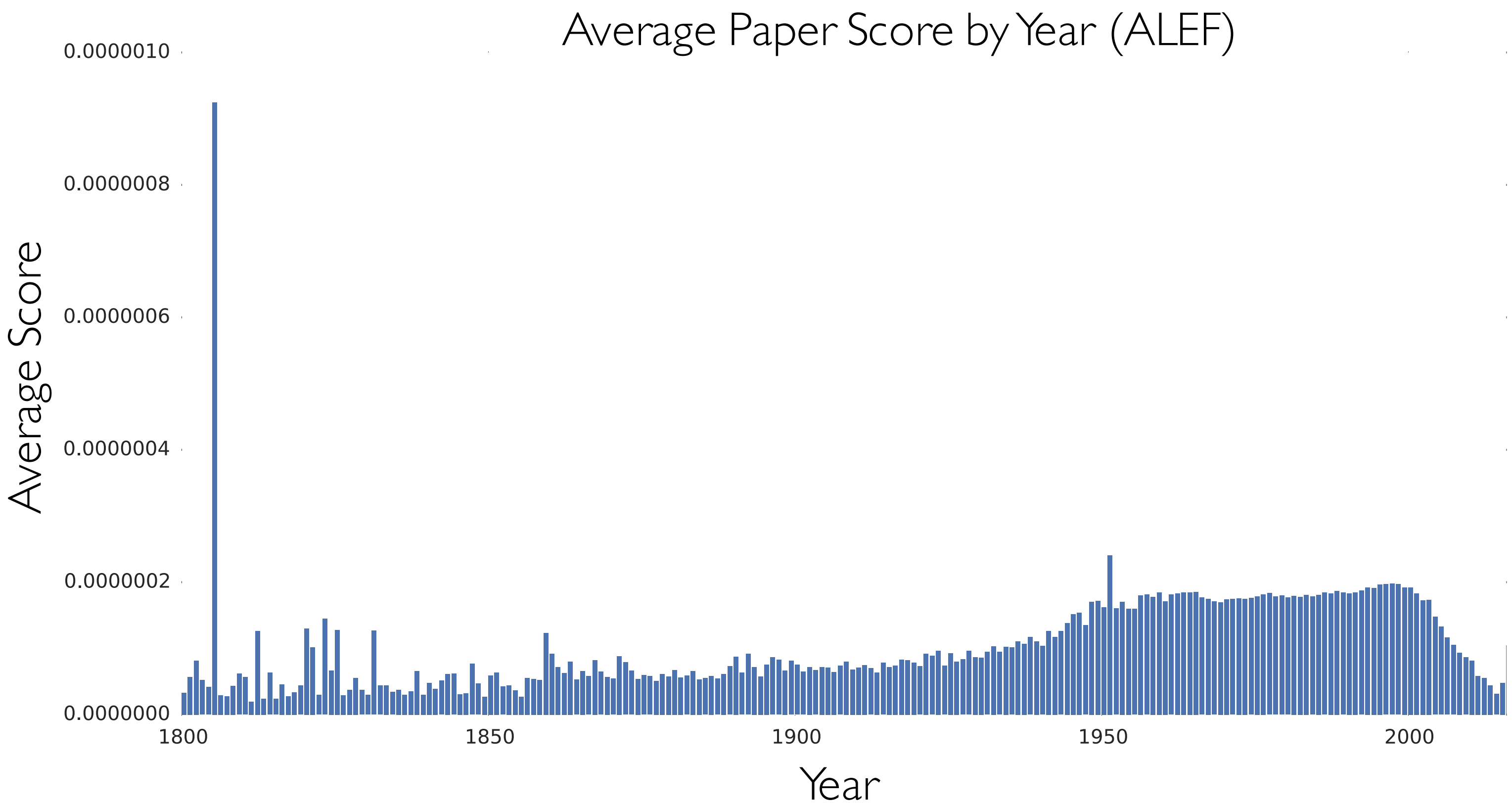
Submissions



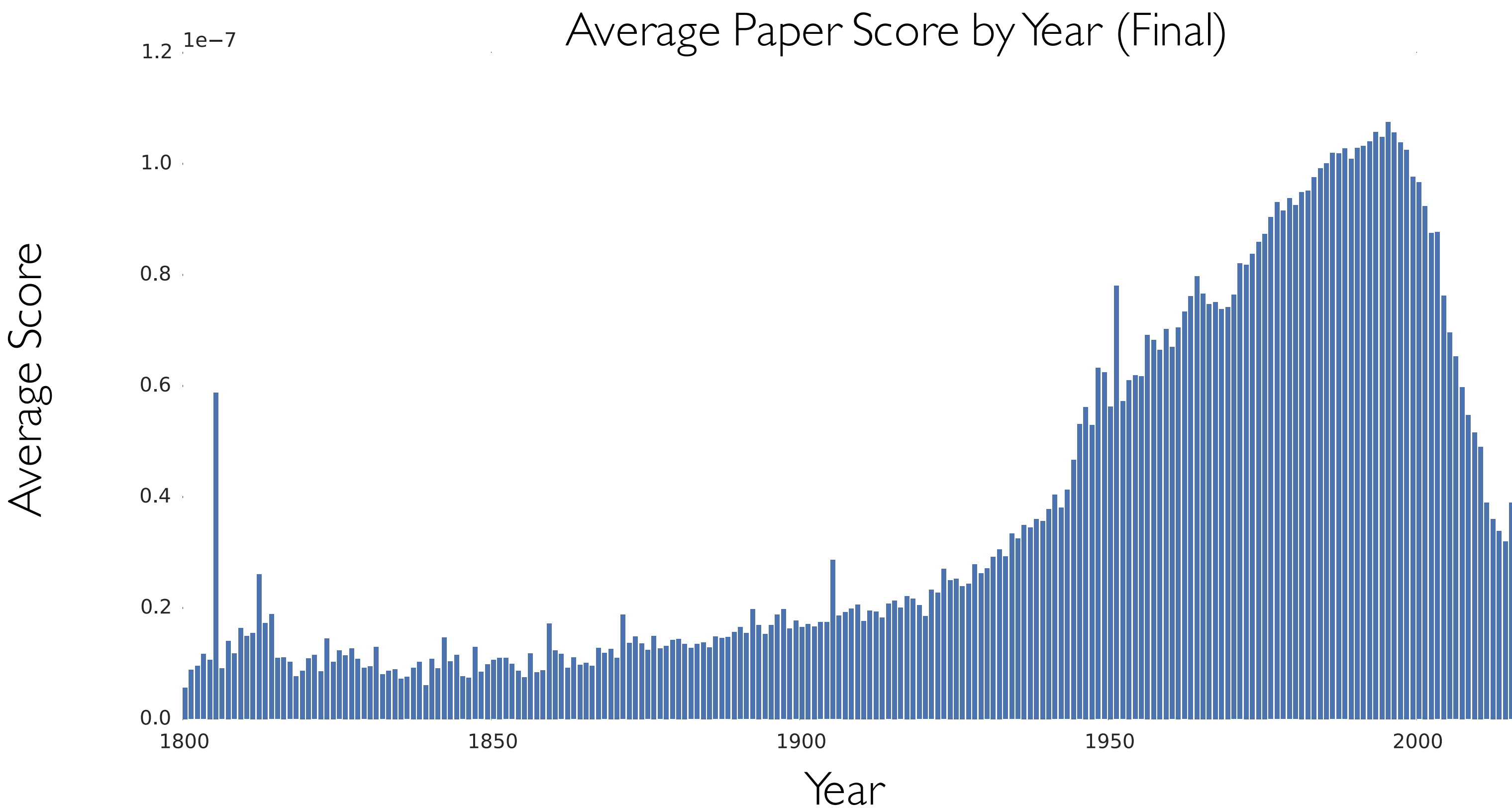
Submissions

Run	Score	Density	Coverage
1	0.693	4.20%	40.76%
18	0.307	8.29%	15.10%
19	0.699	28.27%	54.76%
20	0.693	28.37%	54.76%
22	0.641	96.15%	100.00%
23	0.44	59.25%	9.44%
25	0.699	60.72%	100.00%
26	0	100.00%	45.24%
27	0.665	7.20%	40.76%
28	0.681	4.90%	40.76%
29	0.663	6.13%	40.76%
30	0.528	5.20%	31.76%
31	0.693	71.54%	54.76%
32	0.691	72.27%	54.76%
33	0.699	72.15%	54.76%
34	0.492	14.59%	50.93%
35	0.329	21.89%	33.90%
36	0.444	100.00%	100.00%
37	0.329	21.92%	33.90%
38	0.528	62.25%	66.19%
39	0.691	73.65%	54.76%
40	0.59	73.18%	54.76%
41	0.612	72.02%	54.76%
42	0.693	3.74%	54.76%
43	0.663	72.99%	57.85%
44	0.661	19.45%	42.12%
45	0.693	5.05%	33.90%

ALEF Paper Scores



Final Paper Scores



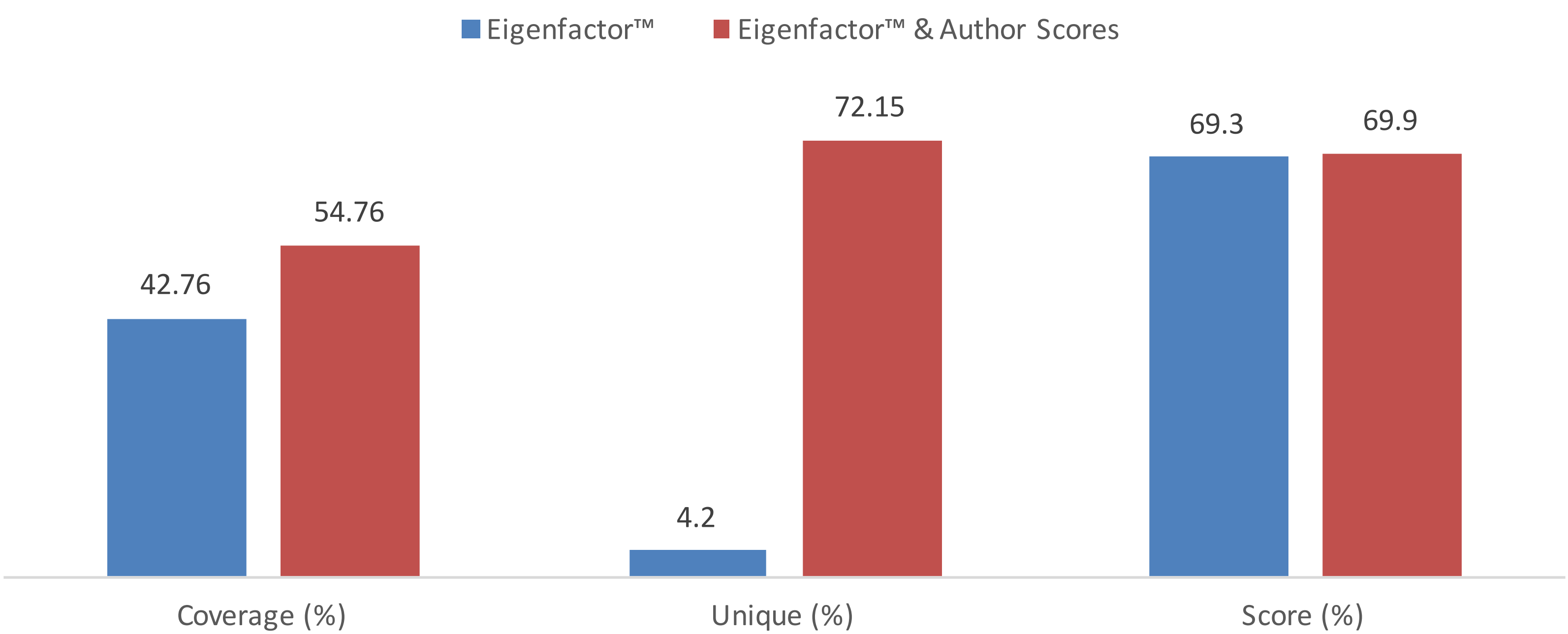
Phase I – Evaluation Results

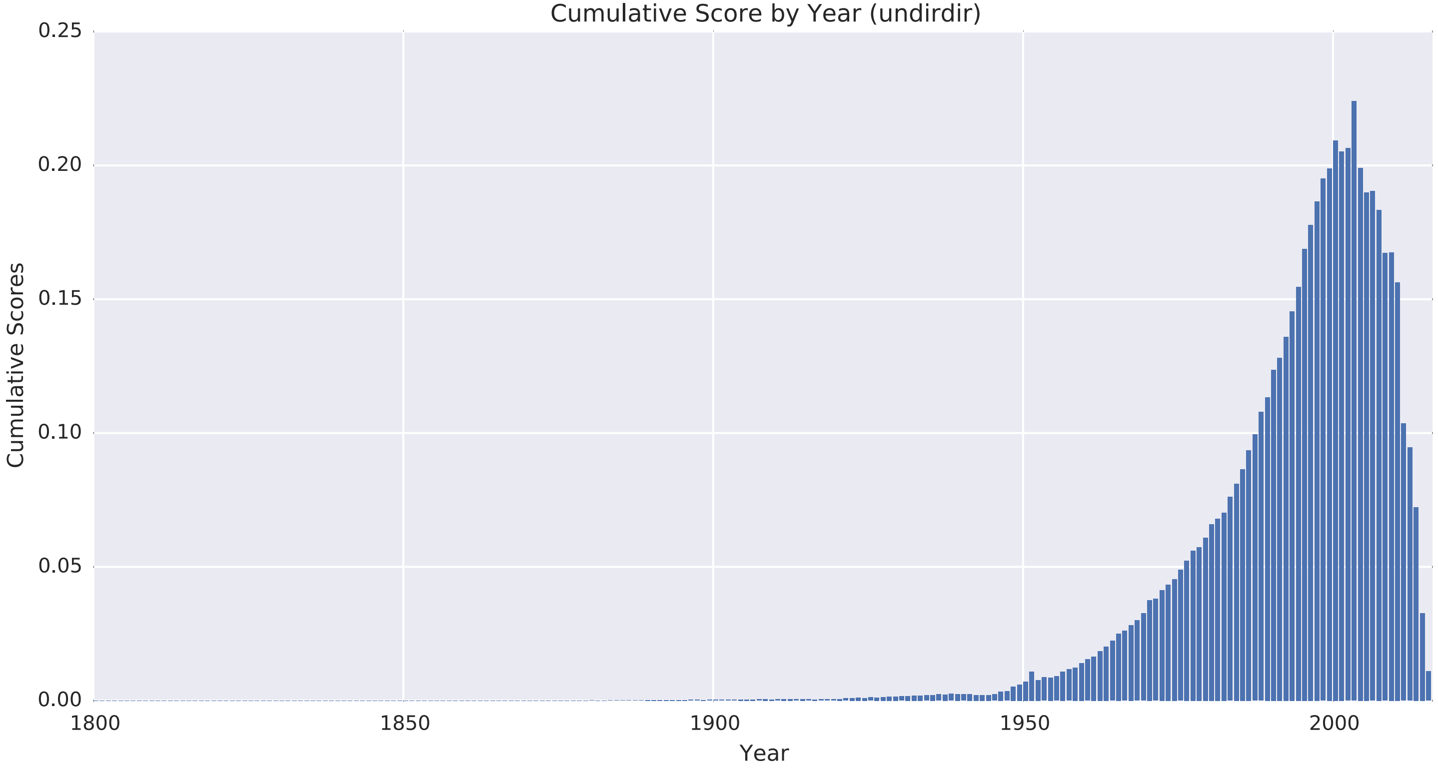
- 0.699
- 15th

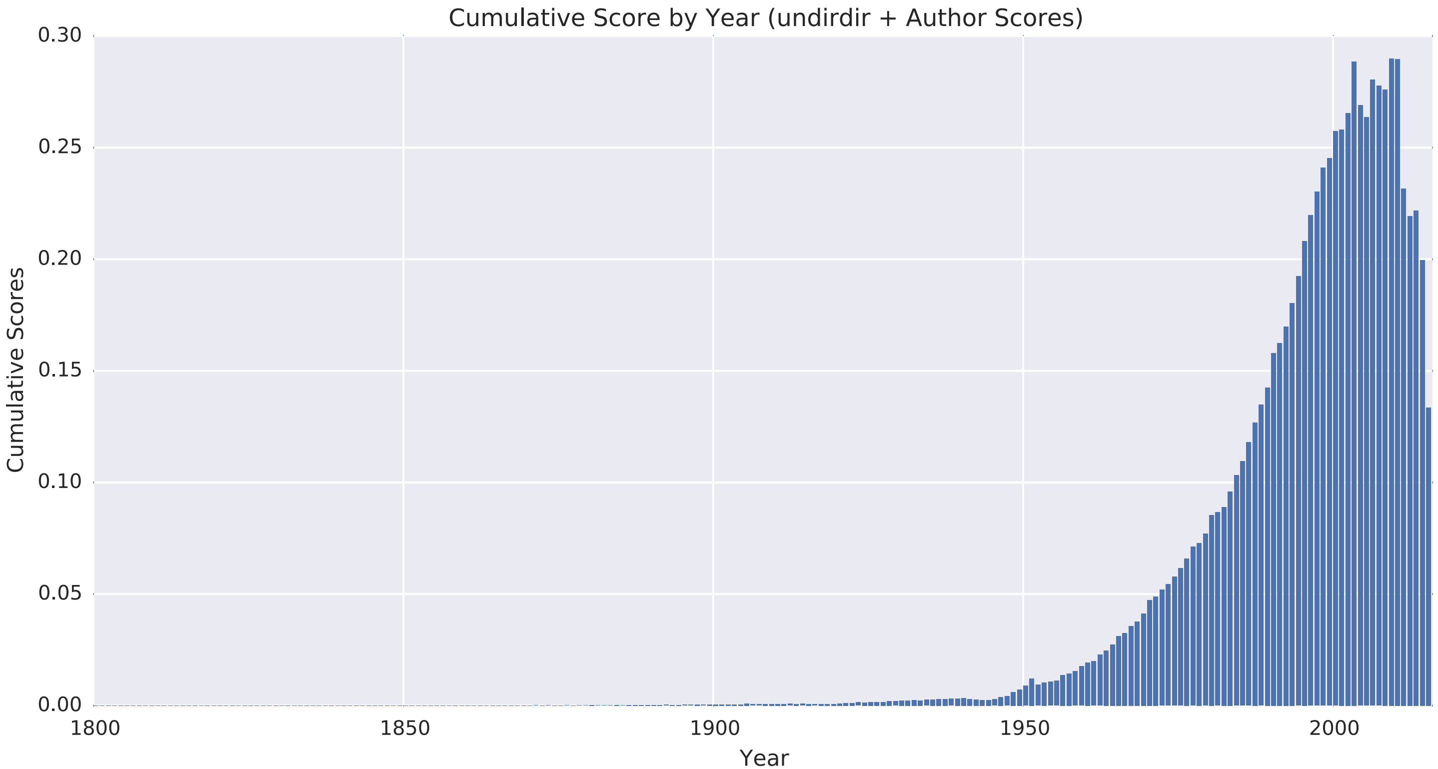
Phase I – Test Results

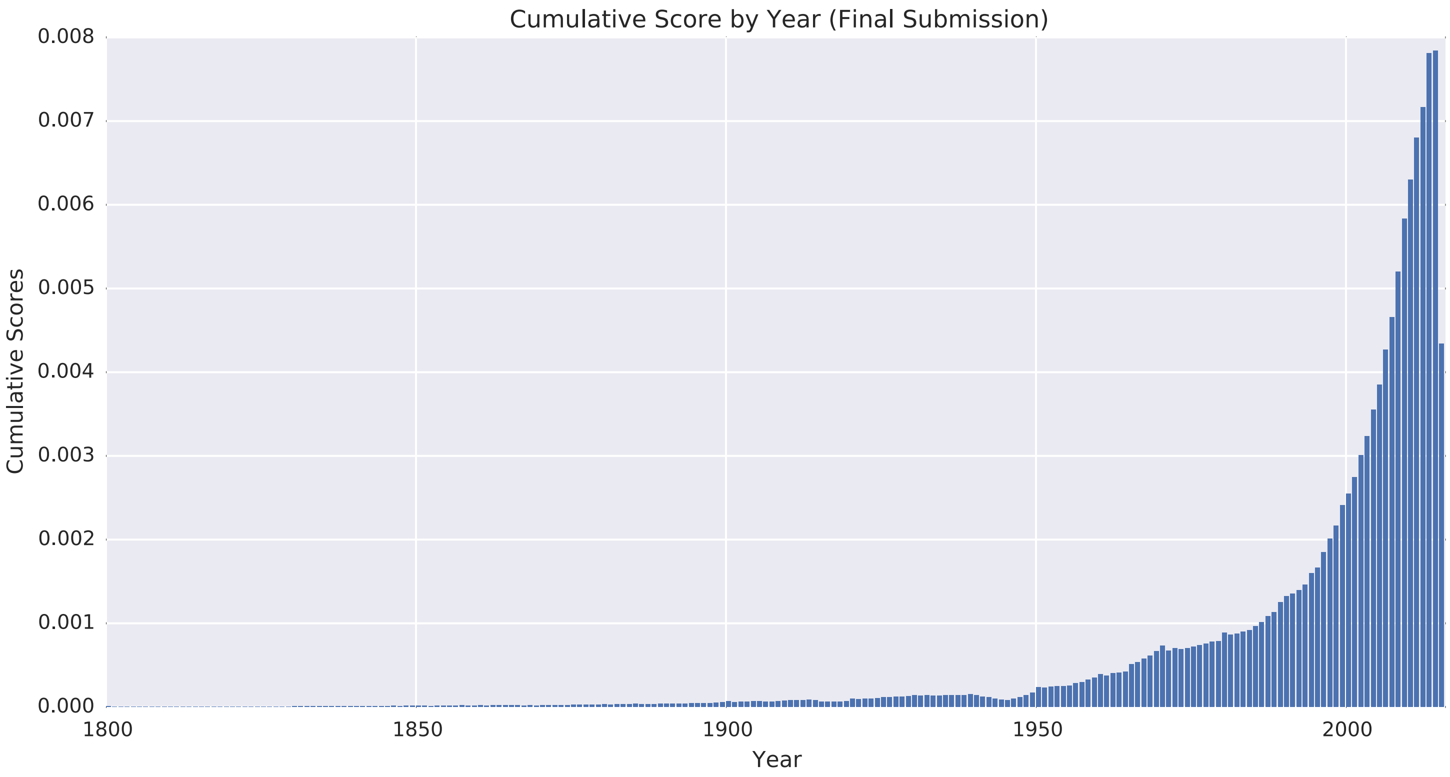
- 0.699 -> 0.676 (-3.3%)
- 15th -> 2nd

Eigenfactor™ & Author Scores




Cumulative Score by Year (undirdir)
0.25
0.20
0.15
0.10
0.05
0.00
Cumulative Scores
1800
1850
1900
1950
2000
Year


Cumulative Score by Year (undirdir + Author Scores)
Cumulative Scores
Year
0.00
0.05
0.10
0.15
0.20
0.25
0.30
1800
1850
1900
1950
2000


Cumulative Score by Year (Final Submission)
Cumulative Scores
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
1800
1850
1900
1950
2000
Year

Logistics

- Phase II
  - Verticies 49,870,036
  - Edges 949,577,946
- Calculate Citation Scores: 34 minutes
- Build Paper-Author Matrix: ~2 hours
- Calculate Author Scores: 2 minutes
- Author Score Feature: 5 minutes
- Blending: 30 seconds

ALEF Summary

- Simple, fast variant of PageRank for article-level citation networks
- Ranks and maps
- More experiments and modifications
- Data cleaning issues
- Thanks to Microsoft Academic Graph and WSDM Cup Challenge

Acknowledgements

Carl Bergstrom, Department of Biology, University of Washington

Martin Rosvall, Department of Physics, Umea University

Daril Vilhena, Department of Biology, University of Washington

Aditya Gandhi, Information School, University of Washington

SciPy.org

pandas $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

Metaknowledge Network, Templeton Foundation

NumPy

MapEquation

Resources

- Info, Data, Code - http://www.eigenfactor.org/
- Babel - http://babel.eigenfactor.org/
- J.D. West, M. Rosvall, C.T. Bergstrom (2016) Ranking and mapping article-level citation networks, *in prep*
- J.D. West, I. Wesley-Smith, C.T. Bergstrom (2016) A recommendation system based on hierarchical clustering of an article-level citation network. *IEEE, Transactions on Big Data*
- I. Wesley-Smith, C.T. Bergstrom, J.D. West (2016) Static Ranking of Scholarly Papers using Article-Level Eigenfactor (ALEF), *WSDM Conference: Entity Ranking Challenge Workshop*
- I. Wesley-Smith, J.D. West (2016) Babel: A platform for research in scholarly article recommendation. *WWW Conference, Workshop on Big Scholarly Data*
- Jevin West - http://www.jevinwest.org/
- Ian Wesley-Smith – http://iwsmith.in/