Credit Allocation in Science

Chunheng Jiang

jiangc4@rpi.edu

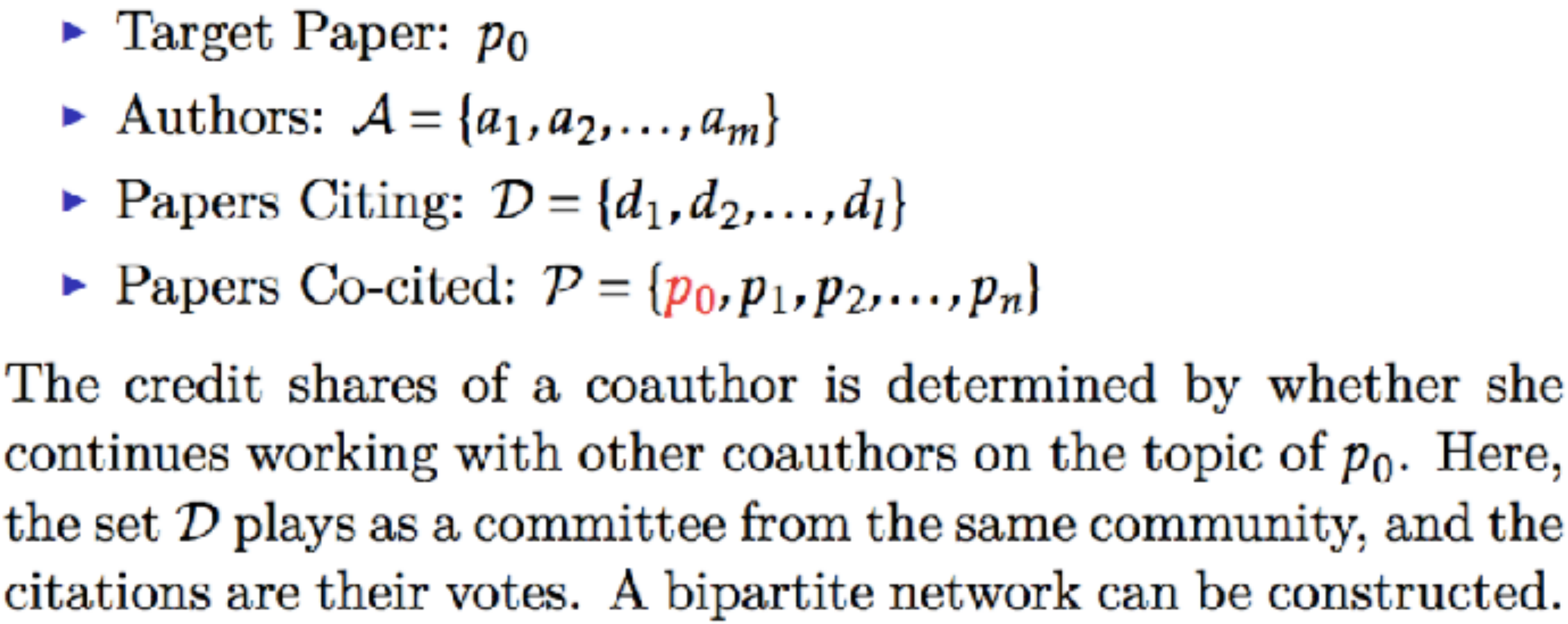
Mon, Mar 26, 2018

Background

Credit allocation (a.k.a credit assignment) is to solve the problem of allocating credits to the coauthors of a joint work according to their intrinsic contribution.

- Collaboration among researchers is a vital component in academia.
- No general guideline to place authors: *contribution, seniority or alphabetical ordering*
- Number of coauthors is increasing

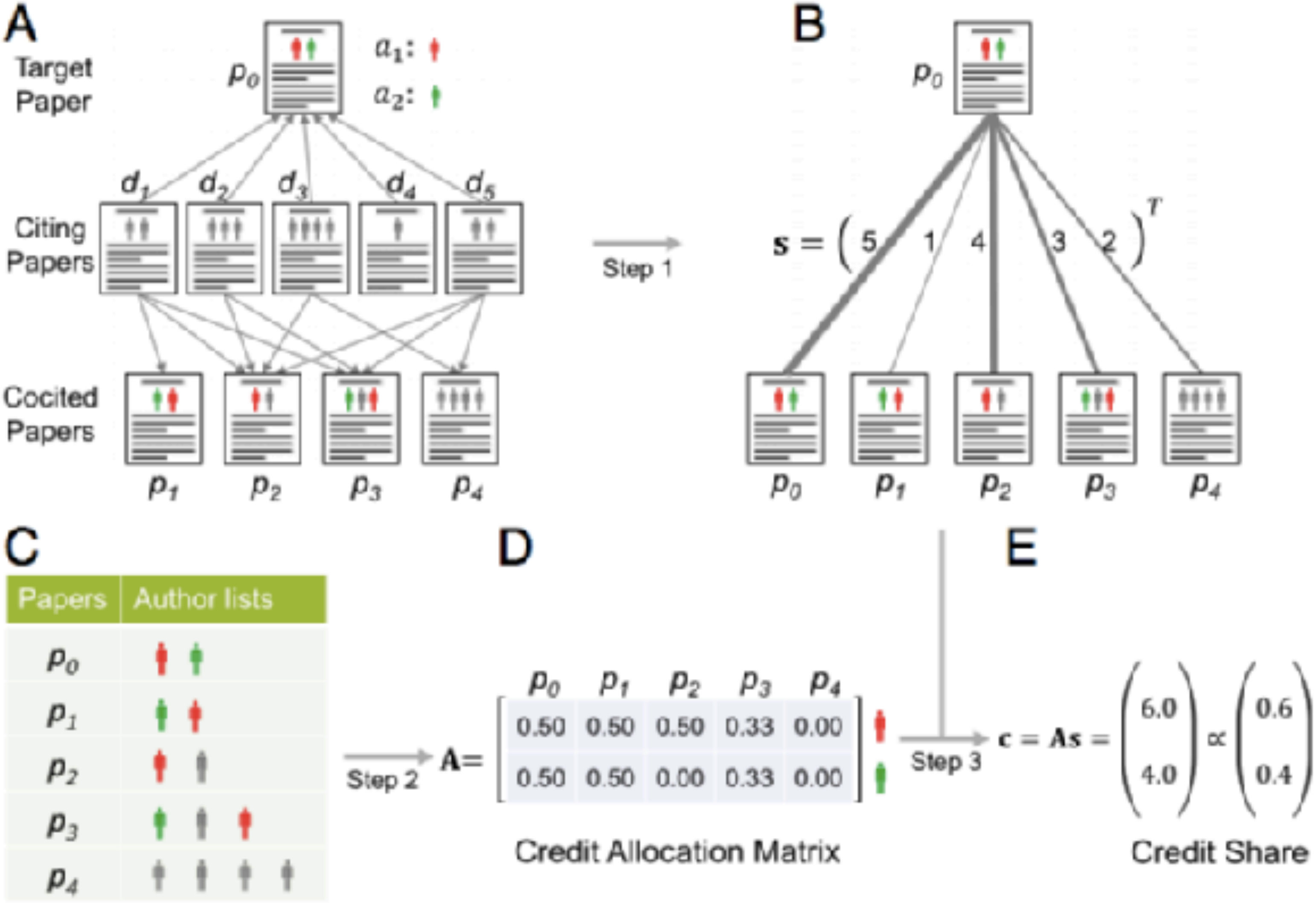
Shen's Method



- Target Paper: p_0
- Authors: $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$
- Papers Citing: $\mathcal{D} = \{d_1, d_2, \ldots, d_l\}$
- Papers Co-cited: $\mathcal{P} = \{p_0, p_1, p_2, \ldots, p_n\}$

The credit shares of a coauthor is determined by whether she continues working with other coauthors on the topic of p_0. Here, the set $\mathcal{D}$ plays as a committee from the same community, and the citations are their votes. A bipartite network can be constructed.

Shen’s Method



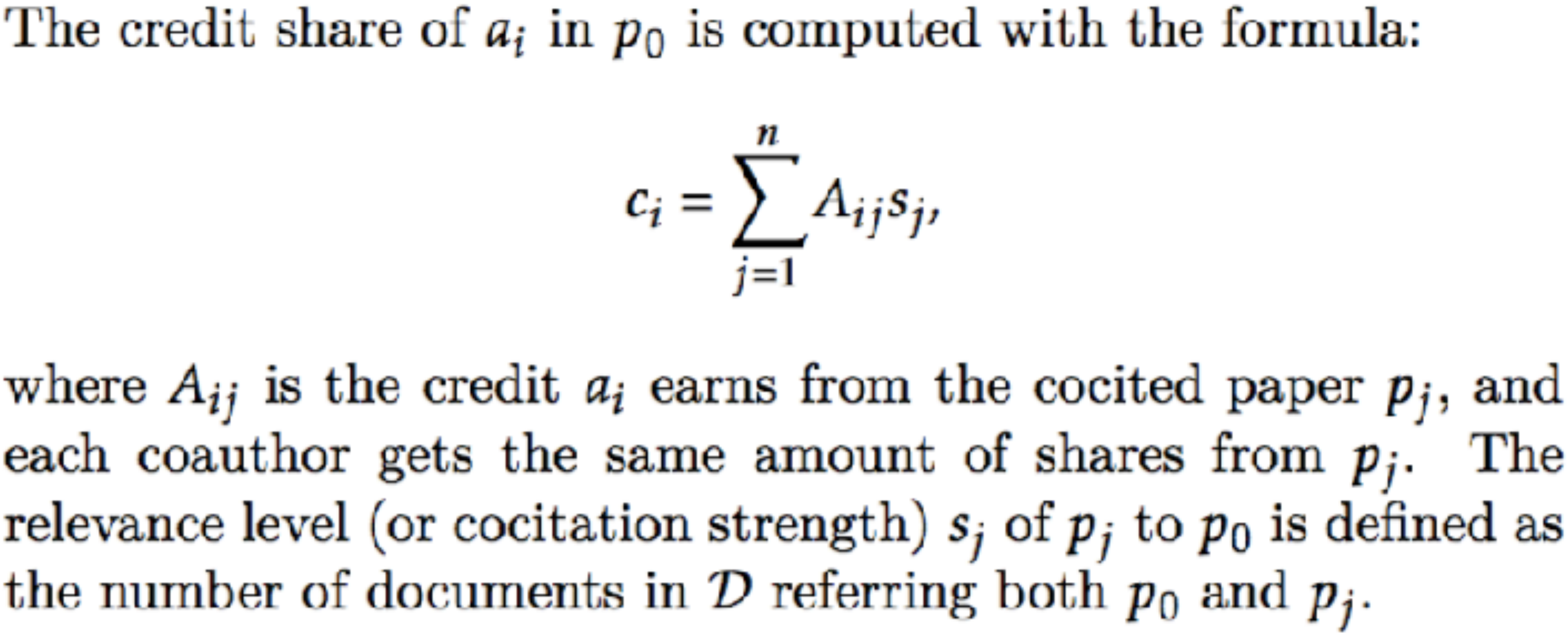
$$\mathbf{s} = \begin{pmatrix} 5 & 1 & 4 & 3 & 2 \end{pmatrix}^T$$

Papers	Author lists
p_0	🟥 🟩
p_1	🟩 🟥
p_2	🟥 ◼
p_3	🟩 ◼ 🟥
p_4	◼ ◼ ◼ ◼

$$\mathbf{A}= \begin{array}{ccccc|c} p_0 & p_1 & p_2 & p_3 & p_4 & \\ 0.50 & 0.50 & 0.50 & 0.33 & 0.00 & 🟥 \\ 0.50 & 0.50 & 0.00 & 0.33 & 0.00 & 🟩 \end{array}$$

$$\mathbf{c} = \mathbf{A}\mathbf{s} = \begin{pmatrix} 6.0 \\ 4.0 \end{pmatrix} \propto \begin{pmatrix} 0.6 \\ 0.4 \end{pmatrix}$$

SB Method

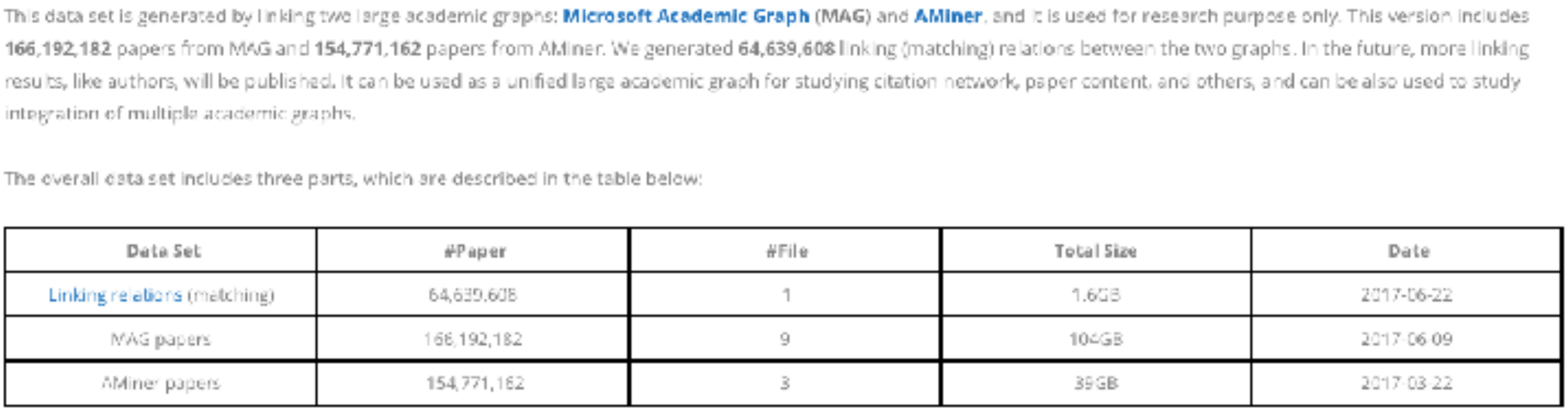


The credit share of a_i in p_0 is computed with the formula:

$$c_i = \sum_{j=1}^{n} A_{ij} s_j,$$

where A_{ij} is the credit a_i earns from the cocited paper p_j, and each coauthor gets the same amount of shares from p_j. The relevance level (or cocitation strength) s_j of p_j to p_0 is defined as the number of documents in $\mathcal{D}$ referring both p_0 and p_j.

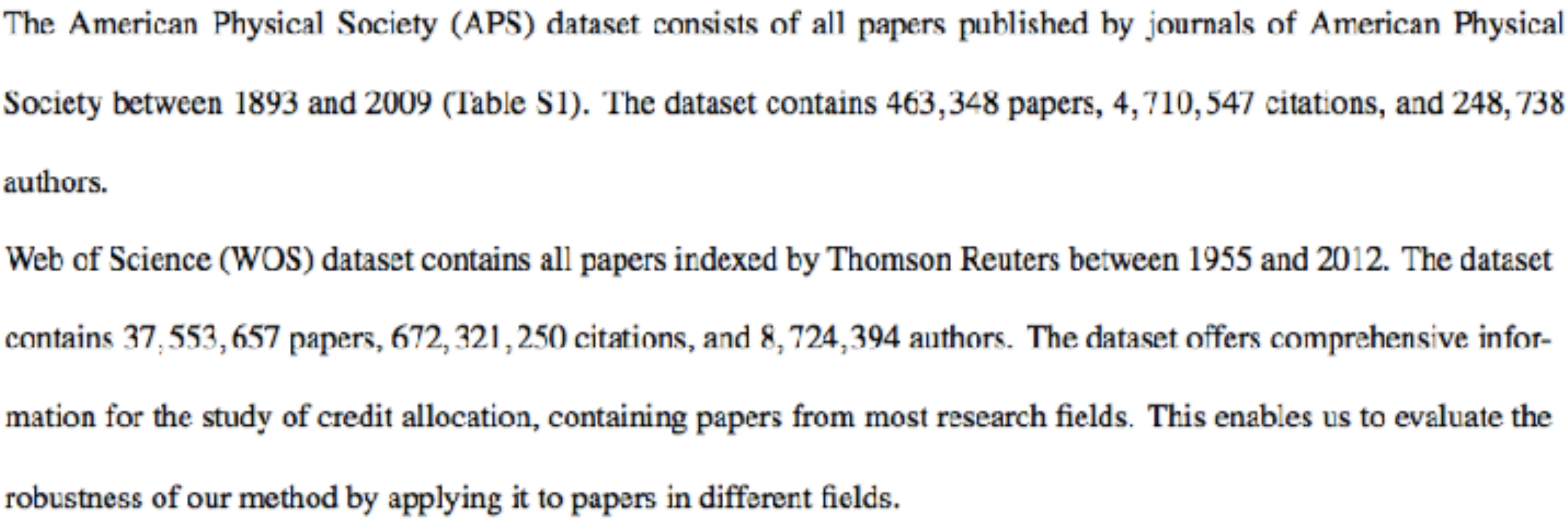
Dataset



This data set is generated by linking two large academic graphs: **Microsoft Academic Graph** (MAG) and **AMiner**, and it is used for research purpose only. This version includes **166,192,182** papers from MAG and **154,771,162** papers from AMiner. We generated **64,639,608** linking (matching) relations between the two graphs. In the future, more linking results, like authors, will be published. It can be used as a unified large academic graph for studying citation network, paper content, and others, and can be also used to study integration of multiple academic graphs.

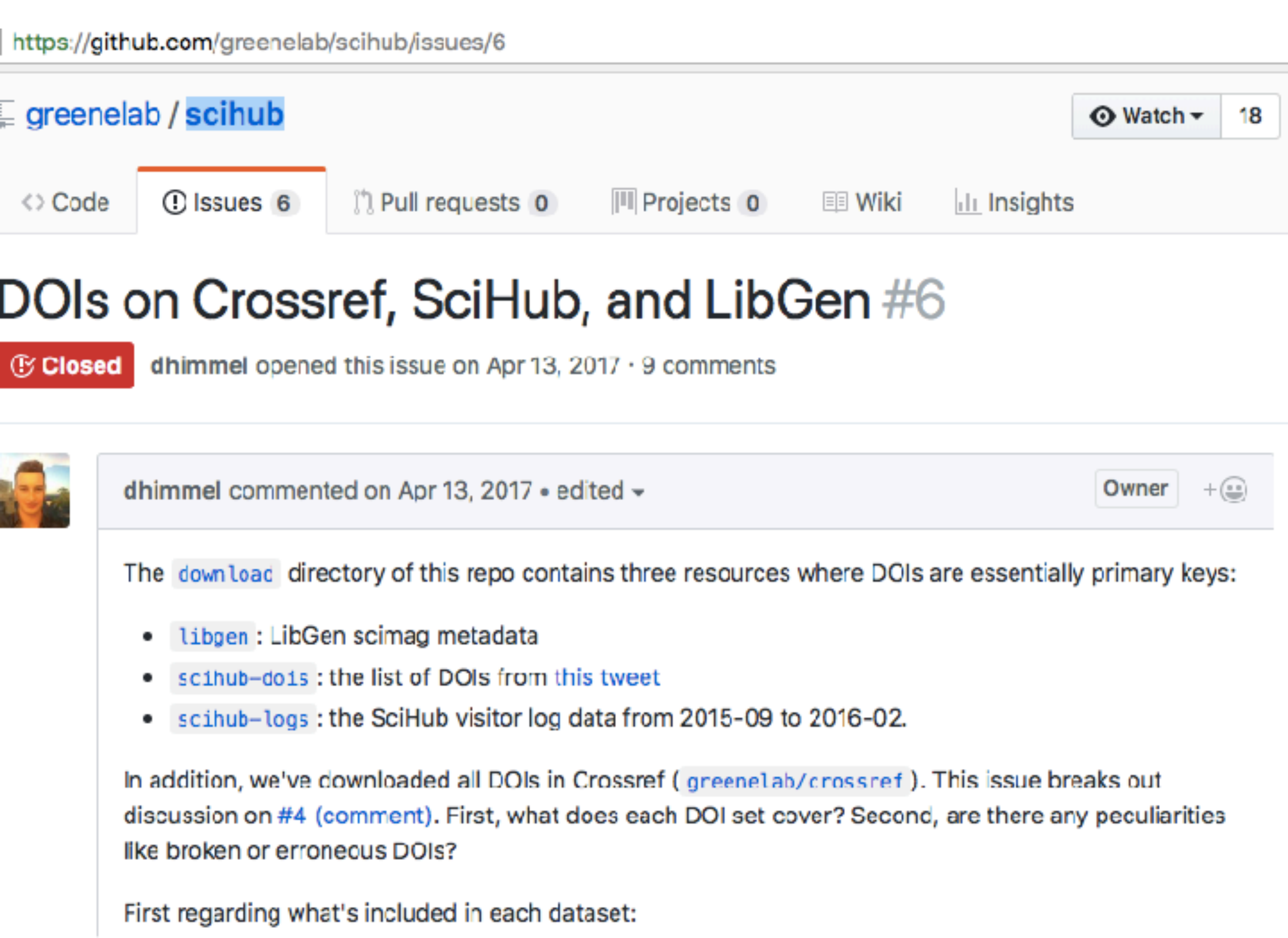
The overall data set includes three parts, which are described in the table below:

Data Set	#Paper	#File	Total Size	Date
Linking relations (matching)	64,639,608	1	1.6GB	2017-06-22
MAG papers	166,192,182	9	104GB	2017-06-09
AMiner papers	154,771,162	3	39GB	2017-03-22



The American Physical Society (APS) dataset consists of all papers published by journals of American Physical Society between 1893 and 2009 (Table S1). The dataset contains 463,348 papers, 4,710,547 citations, and 248,738 authors.

Web of Science (WOS) dataset contains all papers indexed by Thomson Reuters between 1955 and 2012. The dataset contains 37,553,657 papers, 672,321,250 citations, and 8,724,394 authors. The dataset offers comprehensive information for the study of credit allocation, containing papers from most research fields. This enables us to evaluate the robustness of our method by applying it to papers in different fields.



DOIs on Crossref, SciHub, and LibGen #6

dhimmel opened this issue on Apr 13, 2017 · 9 comments

dhimmel commented on Apr 13, 2017 • edited

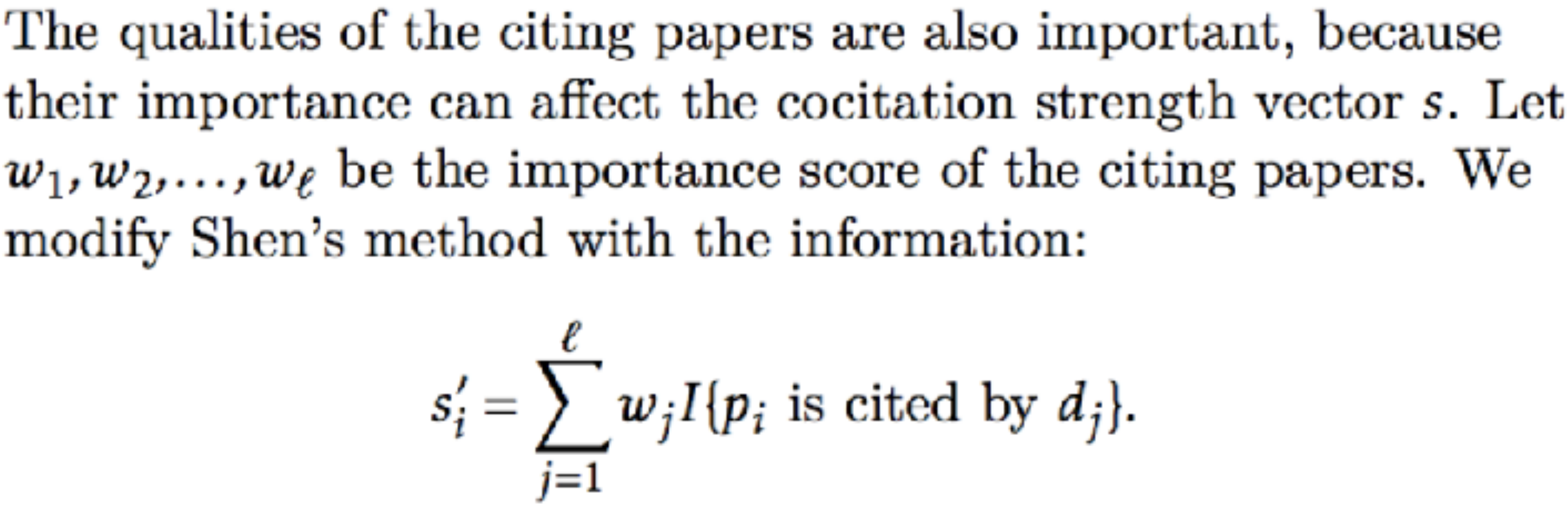
The `download` directory of this repo contains three resources where DOIs are essentially primary keys:

- `libgen`: LibGen scimag metadata
- `scihub-dois`: the list of DOIs from this tweet
- `scihub-logs`: the SciHub visitor log data from 2015-09 to 2016-02.

In addition, we've downloaded all DOIs in Crossref (`greenelab/crossref`). This issue breaks out discussion on #4 (comment). First, what does each DOI set cover? Second, are there any peculiarities like broken or erroneous DOIs?

First regarding what's included in each dataset:

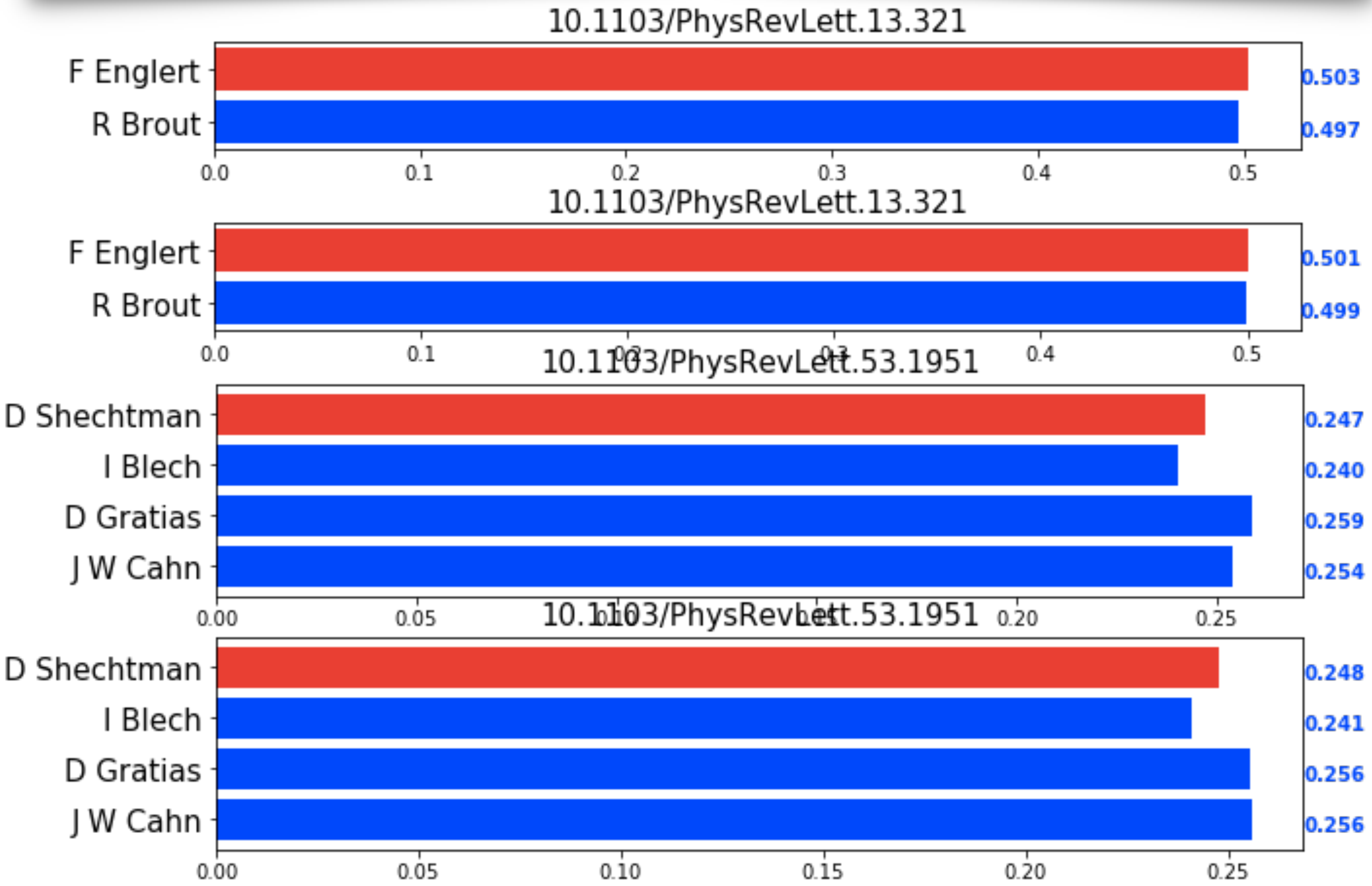
Proposed Method

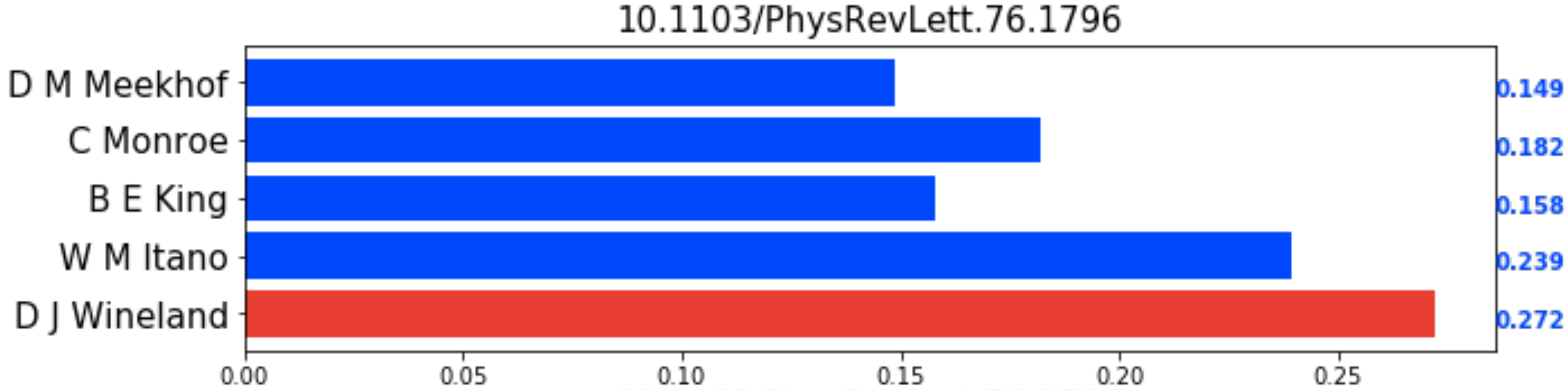


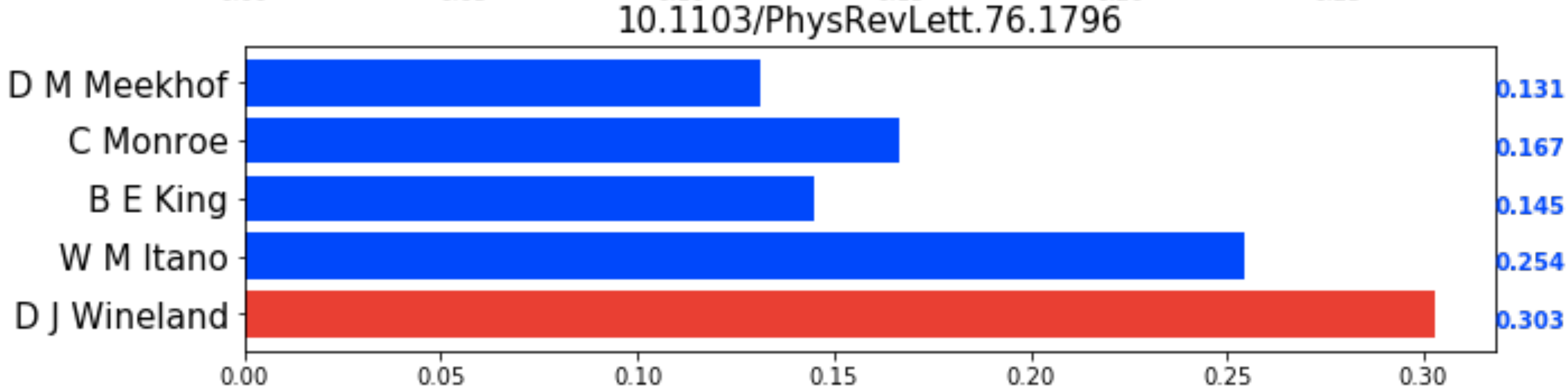
The qualities of the citing papers are also important, because their importance can affect the cocitation strength vector s. Let $w_1, w_2, \ldots, w_\ell$ be the importance score of the citing papers. We modify Shen's method with the information:

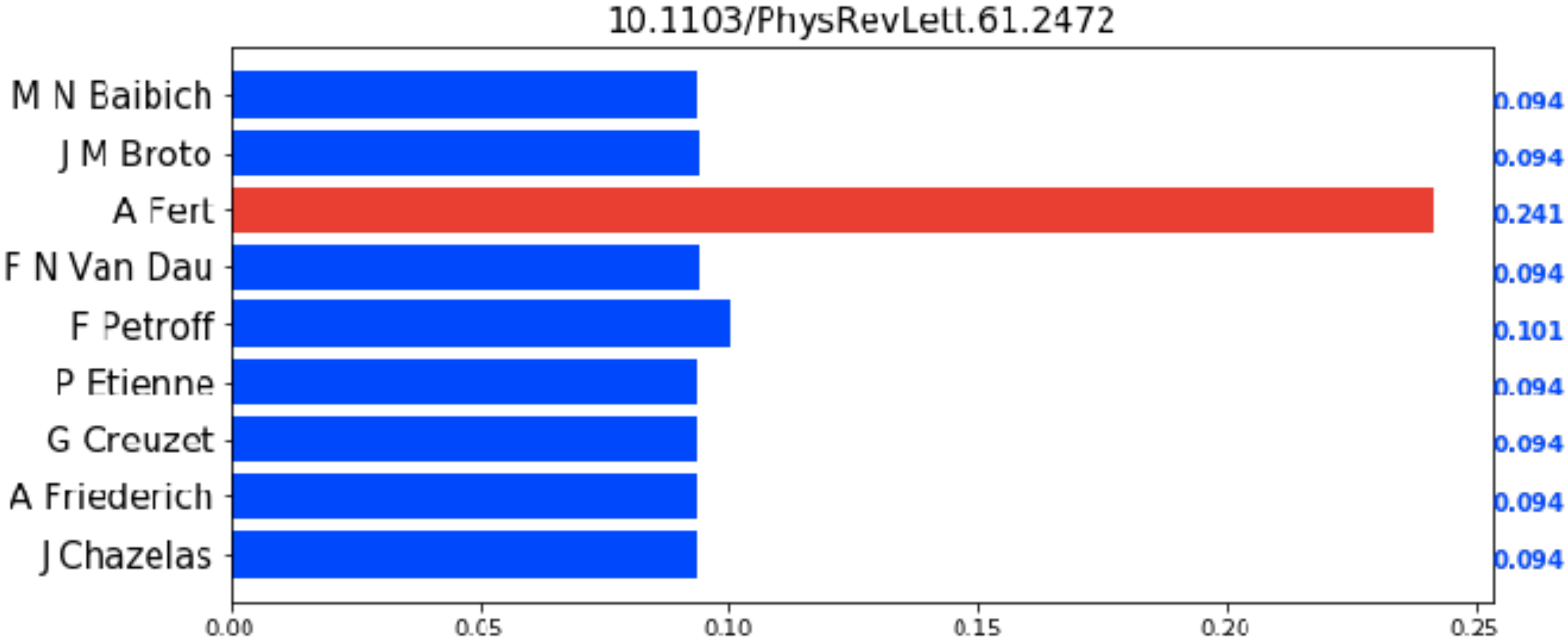
$$s'_i = \sum_{j=1}^{\ell} w_j I\{p_i \text{ is cited by } d_j\}.$$

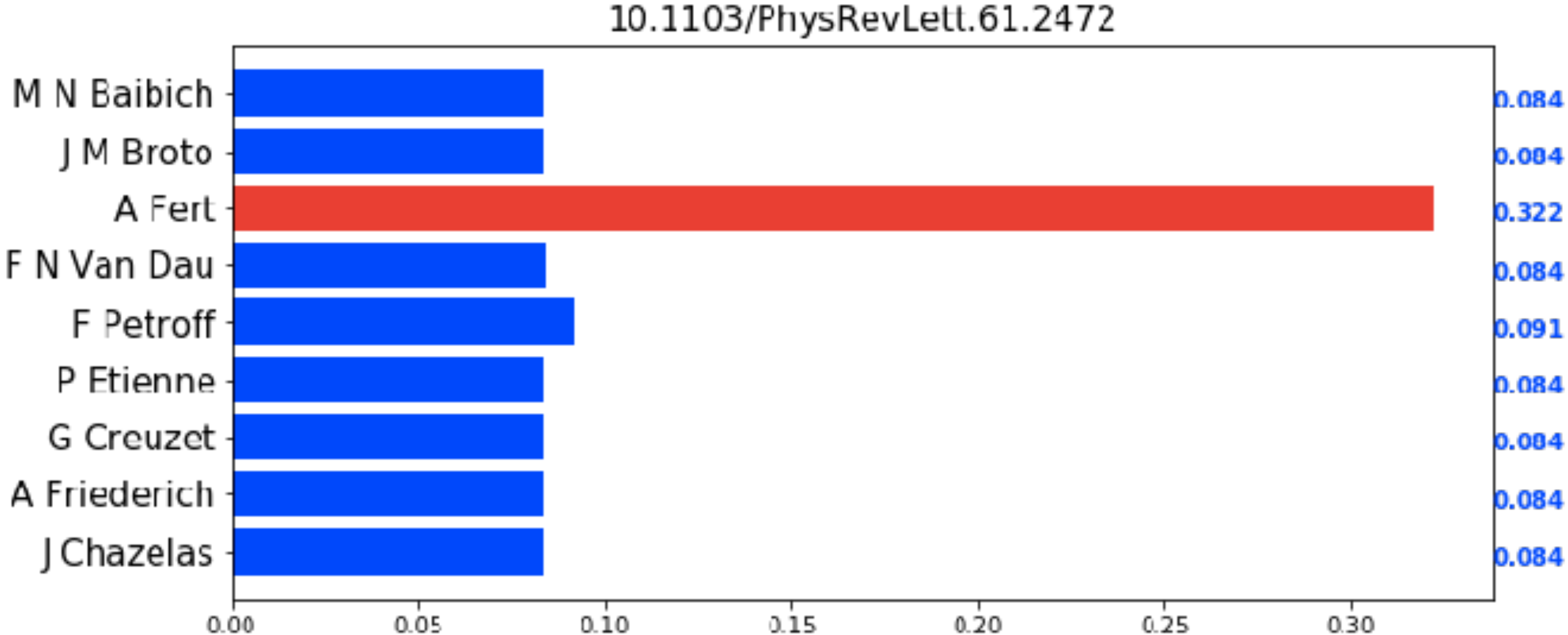
SB v.s Modification

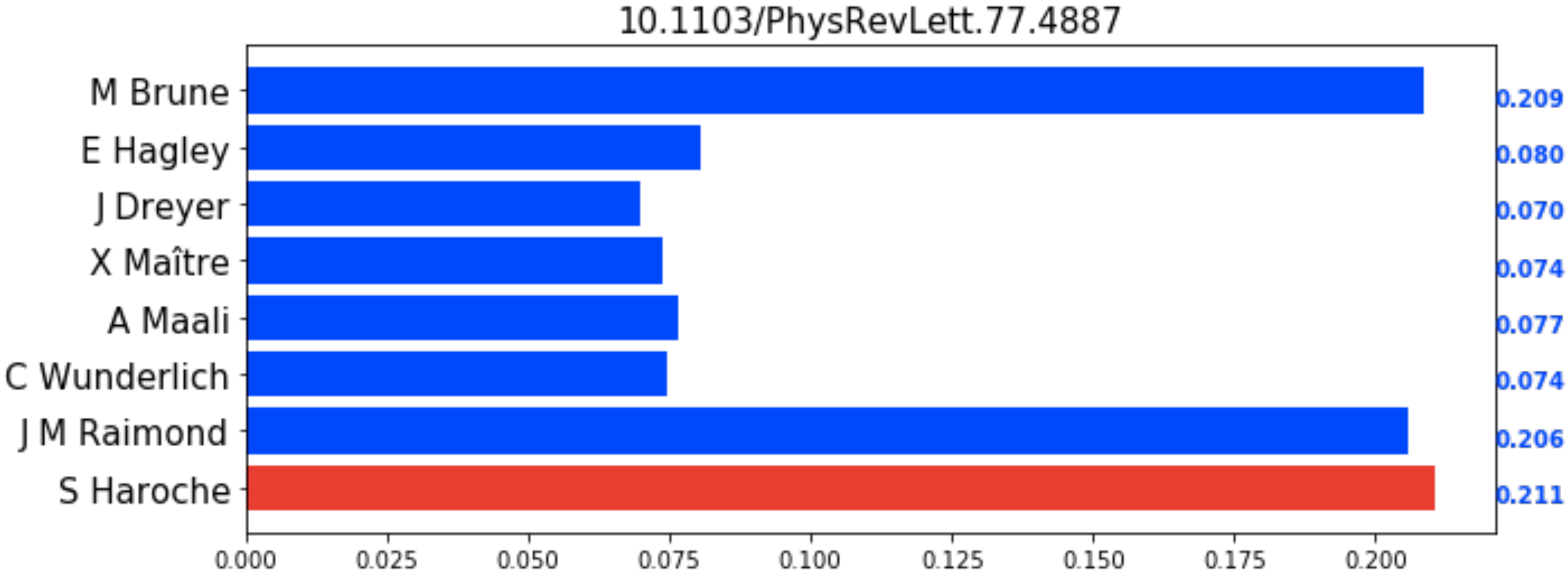


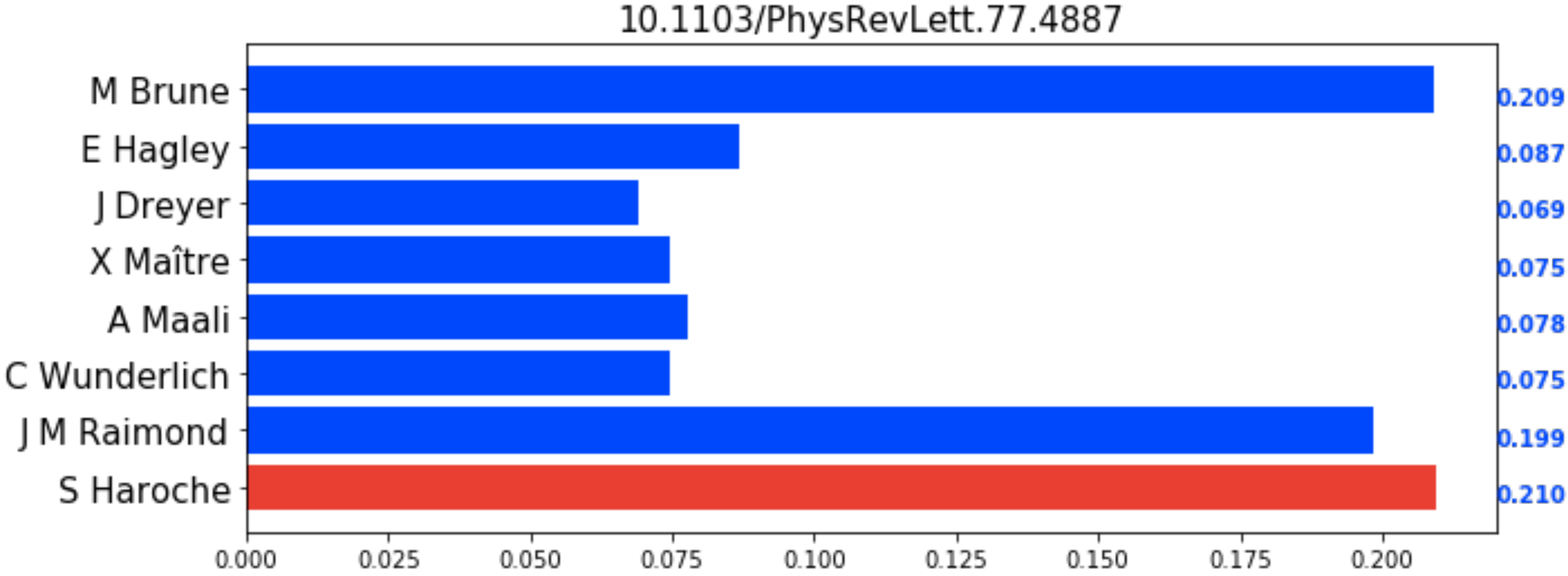

10.1103/PhysRevLett.76.1796
D M Meekhof
C Monroe
B E King
W M Itano
D J Wineland
0.149
0.182
0.158
0.239
0.272
0.00
0.05
0.10
0.15
0.20
0.25


10.1103/PhysRevLett.76.1796
D M Meekhof
C Monroe
B E King
W M Itano
D J Wineland
0.131
0.167
0.145
0.254
0.303
0.00
0.05
0.10
0.15
0.20
0.25
0.30

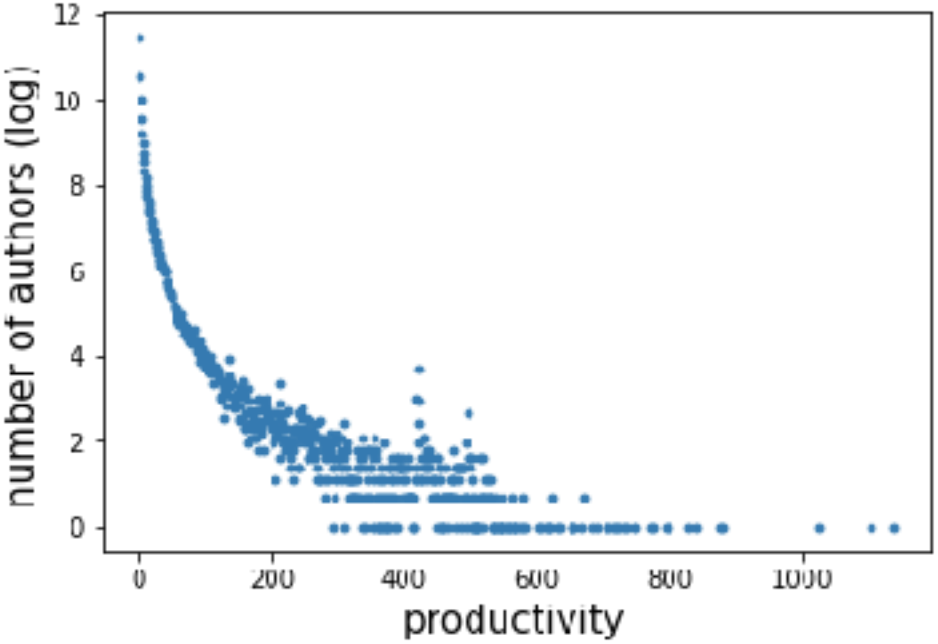

10.1103/PhysRevLett.61.2472
M N Baibich
J M Broto
A Fert
F N Van Dau
F Petroff
P Etienne
G Creuzet
A Friederich
J Chazelas
0.094
0.094
0.241
0.094
0.101
0.094
0.094
0.094
0.094
0.00
0.05
0.10
0.15
0.20
0.25

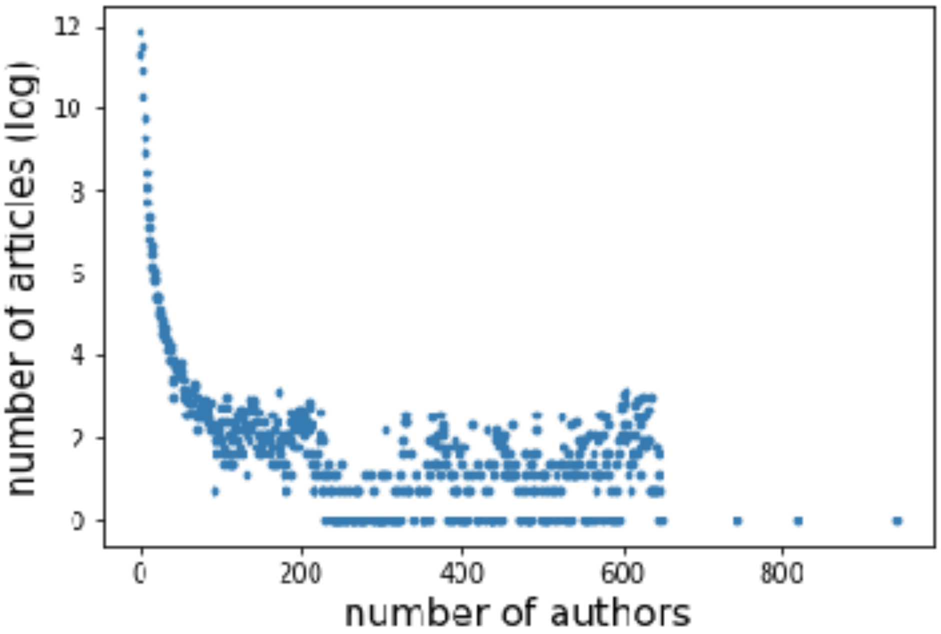

10.1103/PhysRevLett.61.2472
M N Baibich
J M Broto
A Fert
F N Van Dau
F Petroff
P Etienne
G Creuzet
A Friederich
J Chazelas
0.084
0.084
0.322
0.084
0.091
0.084
0.084
0.084
0.084
0.00
0.05
0.10
0.15
0.20
0.25
0.30


10.1103/PhysRevLett.77.4887
M Brune
E Hagley
J Dreyer
X Maître
A Maali
C Wunderlich
J M Raimond
S Haroche
0.209
0.080
0.070
0.074
0.077
0.074
0.206
0.211
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200

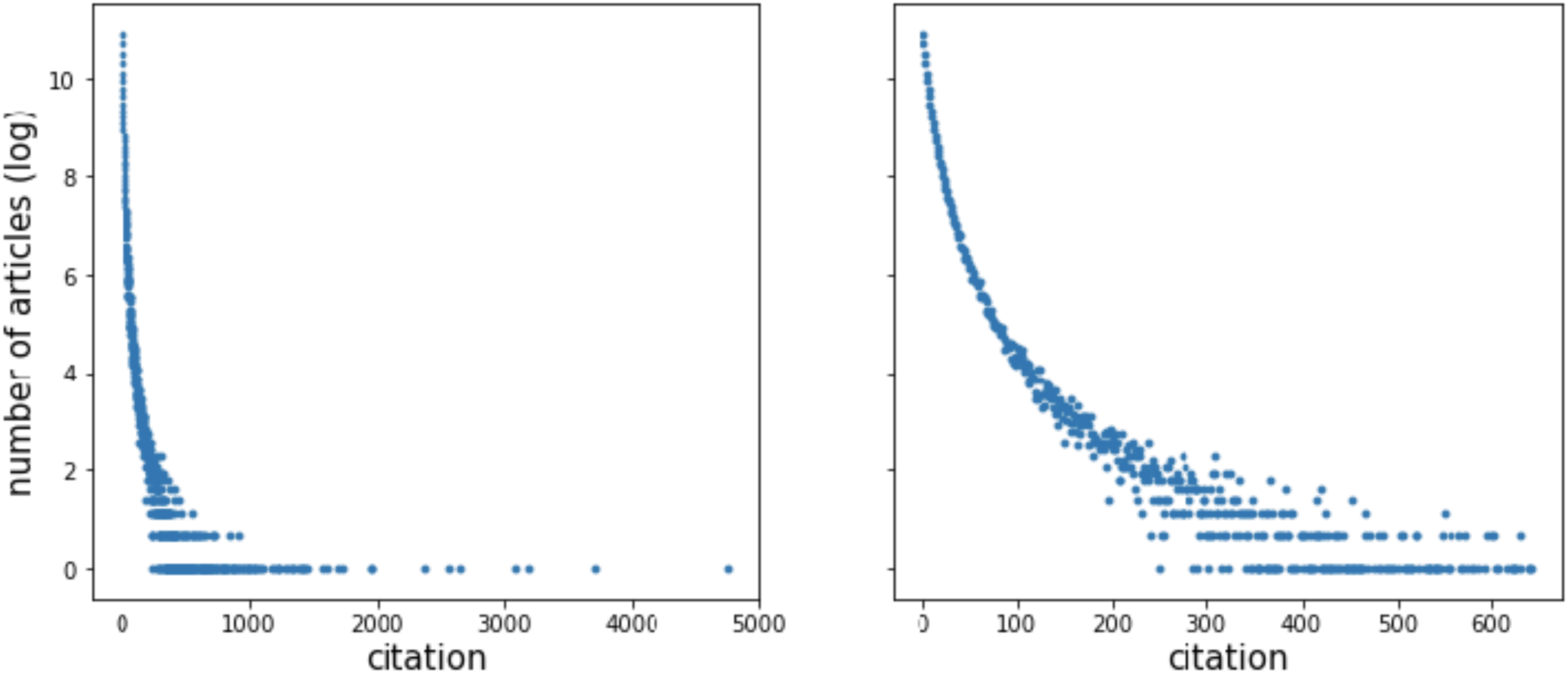

10.1103/PhysRevLett.77.4887
M Brune
E Hagley
J Dreyer
X Maître
A Maali
C Wunderlich
J M Raimond
S Haroche
0.209
0.087
0.069
0.075
0.078
0.075
0.199
0.210
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200

Power Law in Publication

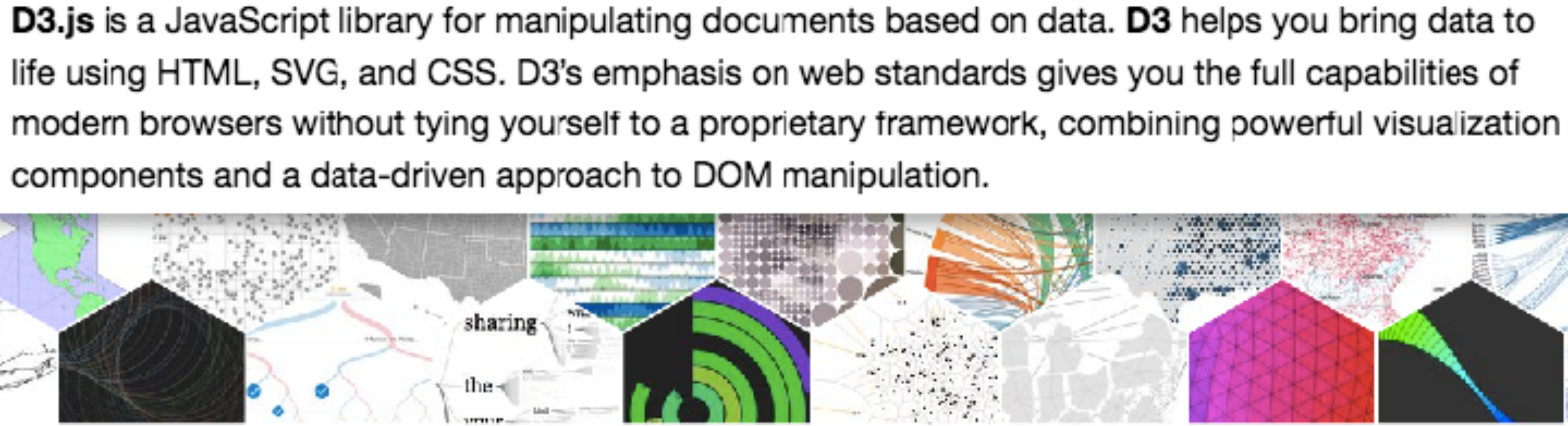




Power Law in Publication



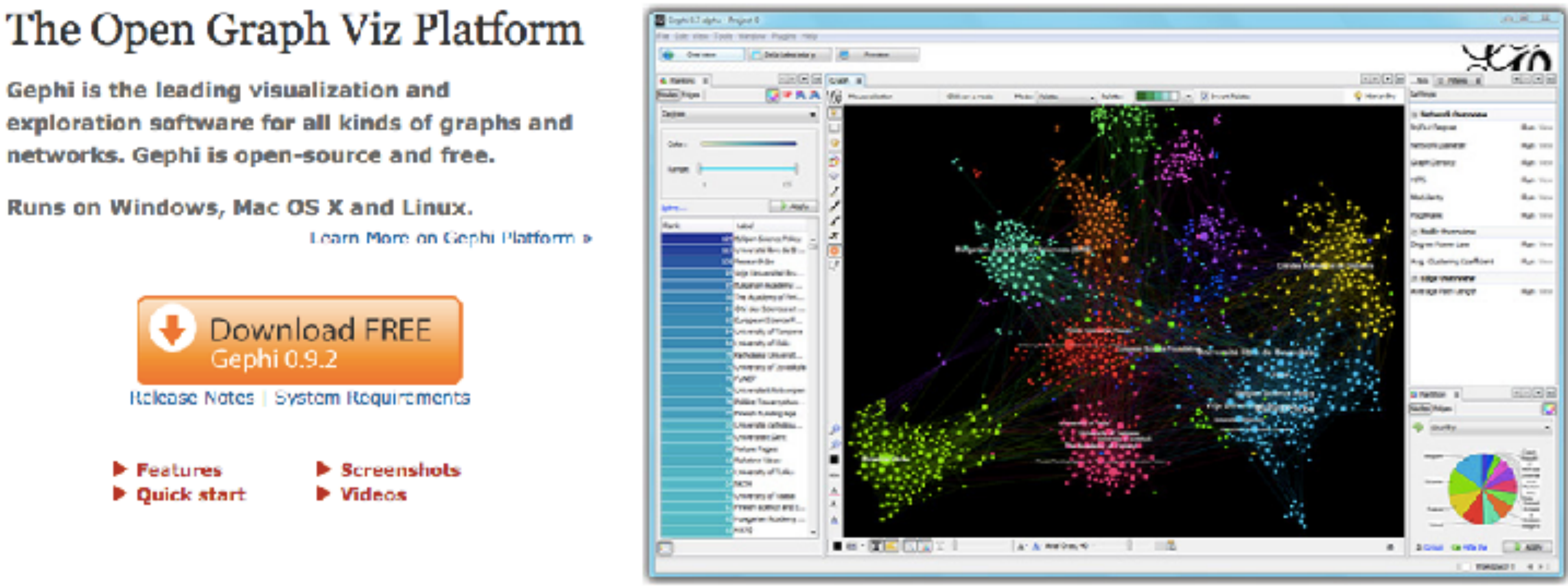
Visualization Issues



D3.js is a JavaScript library for manipulating documents based on data. **D3** helps you bring data to life using HTML, SVG, and CSS. D3's emphasis on web standards gives you the full capabilities of modern browsers without tying yourself to a proprietary framework, combining powerful visualization components and a data-driven approach to DOM manipulation.

Problem: SVG DOM elements eat too much of RAM and ultimately cause the browser to crash.

Visualization Issues



The Open Graph Viz Platform

Gephi is the leading visualization and exploration software for all kinds of graphs and networks. Gephi is open-source and free.

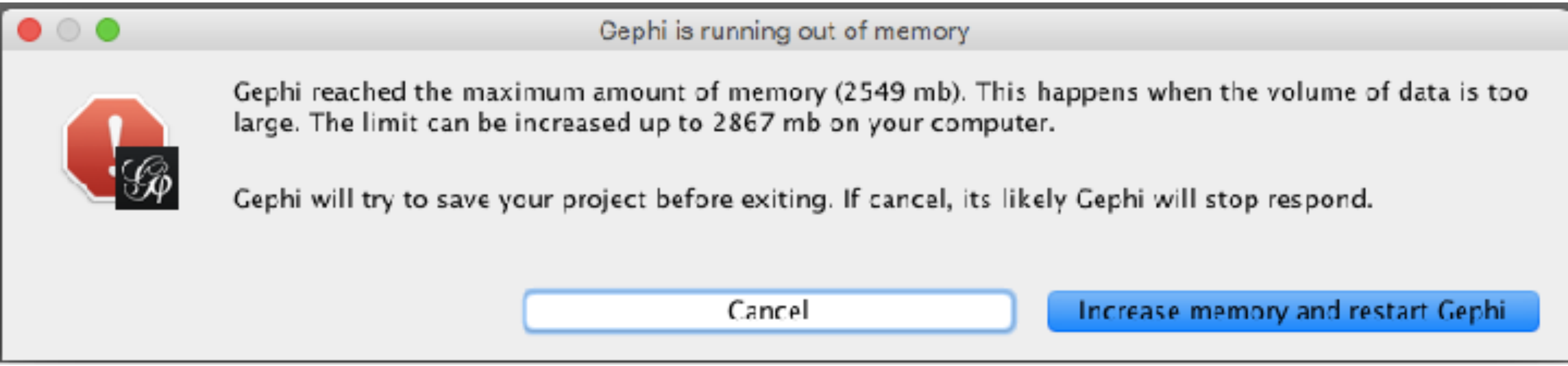
Runs on Windows, Mac OS X and Linux.

Learn More on Gephi Platform »

Download FREE Gephi 0.9.2

Release Notes | System Requirements

- Features
- Quick start
- Screenshots
- Videos



Gephi is running out of memory

Gephi reached the maximum amount of memory (2549 mb). This happens when the volume of data is too large. The limit can be increased up to 2867 mb on your computer.

Gephi will try to save your project before exiting. If cancel, its likely Gephi will stop respond.

Next Steps

1. Visualization of the authorship bipartite network and the citation network
2. Evaluate the improvement over Shen's method: increasing the credit gap between the Nobel winners & other coauthors.