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

Dataset: APS

Journal	#Papers	#Citations	Period
Physical Review (Series I)	1, 469	668	1893-1912
Physical Review	47, 941	590, 665	1913-1969
Physical Review A	53, 655	418, 196	1970-2009
Physical Review B	137, 999	1, 191, 515	1970-2009
Physical Review C	29, 935	202, 312	1970-2009
Physical Review D	56, 616	526, 930	1970-2009
Physical Review E	35, 944	154, 133	1993-2009
Physical Review Letters	95, 516	1, 507, 974	1958-2009
Review of Modern Physics	2, 926	115, 697	1929-2009
Physical Review Special Topics - Accelerators and Beams	1, 257	2, 457	1998-2009
Physical Review Special Topics - Physics Education Research	90	0	2005-2009
Total	463, 348	4, 710, 547	1893-2009

Parsing Issues

```
<articles>
<article doi="10.1103/PhysRevLett.1.11">
<journal jcode="PRL" short="Phys. Rev. Lett.">Physical Review Letters</journal>
<volume>1</volume>
<issue printdate="1958-07-01">1</issue>
<fpage>11</fpage>
<lpage>12</lpage>
<seqno>1</seqno>
<price></price><tocsec>Articles</tocsec>
<arttype type="article"></arttype><doi>10.1103/PhysRevLett.1.11</doi>
<title>Observation of Unpolarized A's Produced by 1.5-Bev n^{-} Interactions in Pb, Fe, and C</title>
<authgrp>
<author><givenname>Theodore</givenname><surname>Bowen</surname></author>
<author><givenname>Judson</givenname><surname>Hardy</surname><suffix>Jr.</suffix></author>
<author><givenname>George</givenname><middlename>T.</middlename><surname>Reynolds</surname></author>
<author><givenname>Guido</givenname><surname>Tagliaferri</surname></author>
<author><givenname>Albert</givenname><middlename>R.</middlename><surname>Werbrouck</surname></author>
<aff >Palmer Physical Laboratory, Princeton University, Princeton, New Jersey</aff>
</authgrp>
<authgrp>
<author><givenname>William</givenname><middlename>H.</middlename><surname>Moore</surname></author>
<aff >Brookhaven National Laboratory, Upton, New York</aff>
</authgrp>
<history>
<received date="1958-06-04"/>
</history>
<cpyrt>
<cpyrtdate date="1958" /><cpyrtholder>The American Physical Society</cpyrtholder>
</cpyrt>
</article>
<article doi="10.1103/PhysRevLett.1.12">
<journal jcode="PRL" short="Phys. Rev. Lett.">Physical Review Letters</journal>
```

Author Ambiguity

- **Same Name, Different Styles**

Scale-Free Networks: A Decade and Beyond

A.-L. Barabási

Science 325, 412-413 (2009)

Drug—target network

Muhammed A Yildirim, Kwang-Il Goh, Michael E Cusick, Albert-László Barabási & Marc Vidal

Remedy 1: combining the **Surname** with **Initial Letters** of the given and middle names

- **Different Person, Same Name**

Remedy 2:

http://www.researcherid.com/
associates each researcher with a unique id

https://github.com/SaschaSchweitzer/persons

Michael Jordan

Disorganized Authors' Names

PhysRevD.59.092002

Troconiz[11], S. Truitt[20], J. Tseng[19], N. Turini[27], T. Uchida[37], F. Ukegawa[26], J. Valls[32], S. C. van den Brink[15], S. Vejcik, III,[7], G. Velev[27], R. Vidal[7], R. Vilar[7,*], I. Volobouev,[18], D. Vucinic[19], R. G. Wagner[1], R. L. Wagner[7], J. Wahl[5], N. B. Wallace[27], A. M. Walsh[32], C. Wang[6], C. H. Wang[33], M. J. Wang[33], A.

PhysRevLett.55.430

P. Chaudhari, H.-U. Habermeier, and and S. Maekawa

PhysRevA.63.030302

R. T. Thew* and W. J. MunroSpecial Research Centre for Quantum Computer Technology, University of Queensland, Brisbane, Australia

PhysRevD.62.114028

Christian W. Bauer and Craig N. BurrellDepartment of Physics, University of Toronto, 60 St. George Street, Toronto, Ontario, Canada M5S 1A7

Noise in Authors' Names

PhysRevC.74.042201

V. Malafaia[1], M. T. Peña[1,2], Ch. Elster[3], and J. Adam, Jr.[4]

```
·<middlename>Adam,</middlename><surname>Jr.</surname></author>
```

PhysRevD.51.R949

F. Abe, M. G. Albrow@f, D. Amidei, J. Antos, C. Anway-Wiese@f, G. Apollinari, H. Areti@f, M. Atac@f, P. Auchincloss, F. Azfar, P. Azzi, N. Bacchetta, W. Badgett, M. W. Bailey, J. Bao, P. de Barbaro, A. Barbaro-Galtieri, V. E. Barnes, B. A. Barnett, P. Bartalini, G. Bauer, T. Baumann@f, F. Bedeschi, S. Behrends@f, S. Belforte, G. Bellettini, J. Bellinger, D. Benjamin, J. Benlloch, J. Bensinger@f, D. Benton, A. Beretvas@f, J. P. Berge@f, S. Bertolucci@f, A. Bhatti, K. Biery, M. Binkley@f, F. Bird, D.

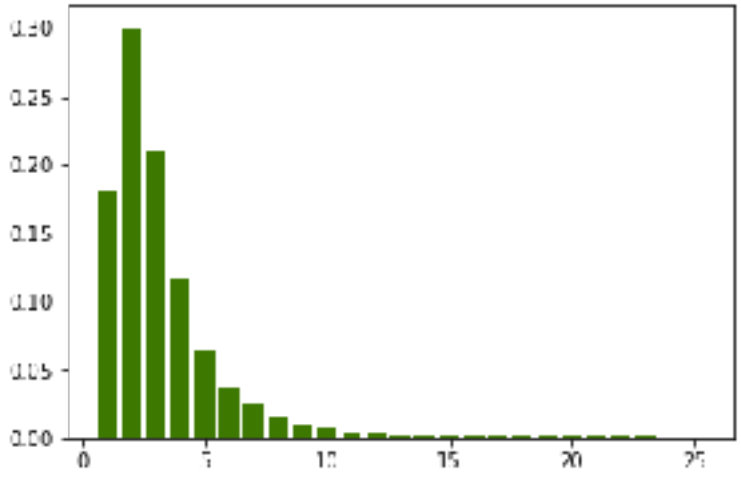
PhysRevB.64.094510

L. Pintschovius and and W. Reichardt

Data Structures

Articles

462,145



	id	doi	journal	numauth
40	40	10.1103/PhysRev.10.101	PR	2
41	41	10.1103/PhysRev.10.116	PR	2
42	42	10.1103/PhysRev.10.129	PR	1
43	43	10.1103/PhysRev.10.140	PR	1
44	44	10.1103/PhysRev.10.156	PR	1
45	45	10.1103/PhysRev.10.166	PR	1
46	46	10.1103/PhysRev.10.171	PR	1
47	47	10.1103/PhysRev.10.217	PR	1
48	48	10.1103/PhysRev.10.226	PR	1
49	49	10.1103/PhysRev.10.244	PR	1

Data Structures

Authors

244,256

	id	given	middle	surname	name
40	120529	J.	A.	Gracey	J A Gracey
41	81554	Lance	NaN	Horng	L Horng
42	135465	Patanjali	NaN	Kambhampati	P Kambhampati
43	203085	Irina	NaN	Sushko	I Sushko
44	44465	I.	NaN	Reinhard	I Reinhard
45	202726	K.	Z.	Nóbrega	K Z Nóbrega
46	178834	Thomas	NaN	Strobel	T Strobel
47	90709	M.	P.	Srivastava	M P Srivastava
48	202360	Roman	E.	Limberger	R E Limberger
49	20811	T.	NaN	Masuda	T Masuda

Data Structures

Authorship

	article	author
40	33	34
41	34	35
42	35	36
43	36	37
44	36	38
45	37	39
46	38	40
47	39	41
48	40	42
49	40	43

Citation Network

	citing_doi	cited_doi
40	10.1103/PhysRevLett.67.407	10.1103/PhysRevLett.67.406
41	10.1103/PhysRevE.53.5130	10.1103/PhysRevLett.67.406
42	10.1103/PhysRevE.70.061405	10.1103/PhysRevE.59.603
43	10.1103/PhysRev.51.306	10.1103/PhysRev.49.317
44	10.1103/PhysRevLett.74.4899	10.1103/PhysRevB.49.4003
45	10.1103/PhysRevB.51.1370	10.1103/PhysRevB.49.4003
46	10.1103/PhysRevB.53.2523	10.1103/PhysRevB.49.4003
47	10.1103/PhysRevB.63.214406	10.1103/PhysRevB.49.4003
48	10.1103/PhysRevB.55.5404	10.1103/PhysRevB.49.4003
49	10.1103/PhysRevB.50.3779	10.1103/PhysRevB.49.4003

Data Structures

Citation Network

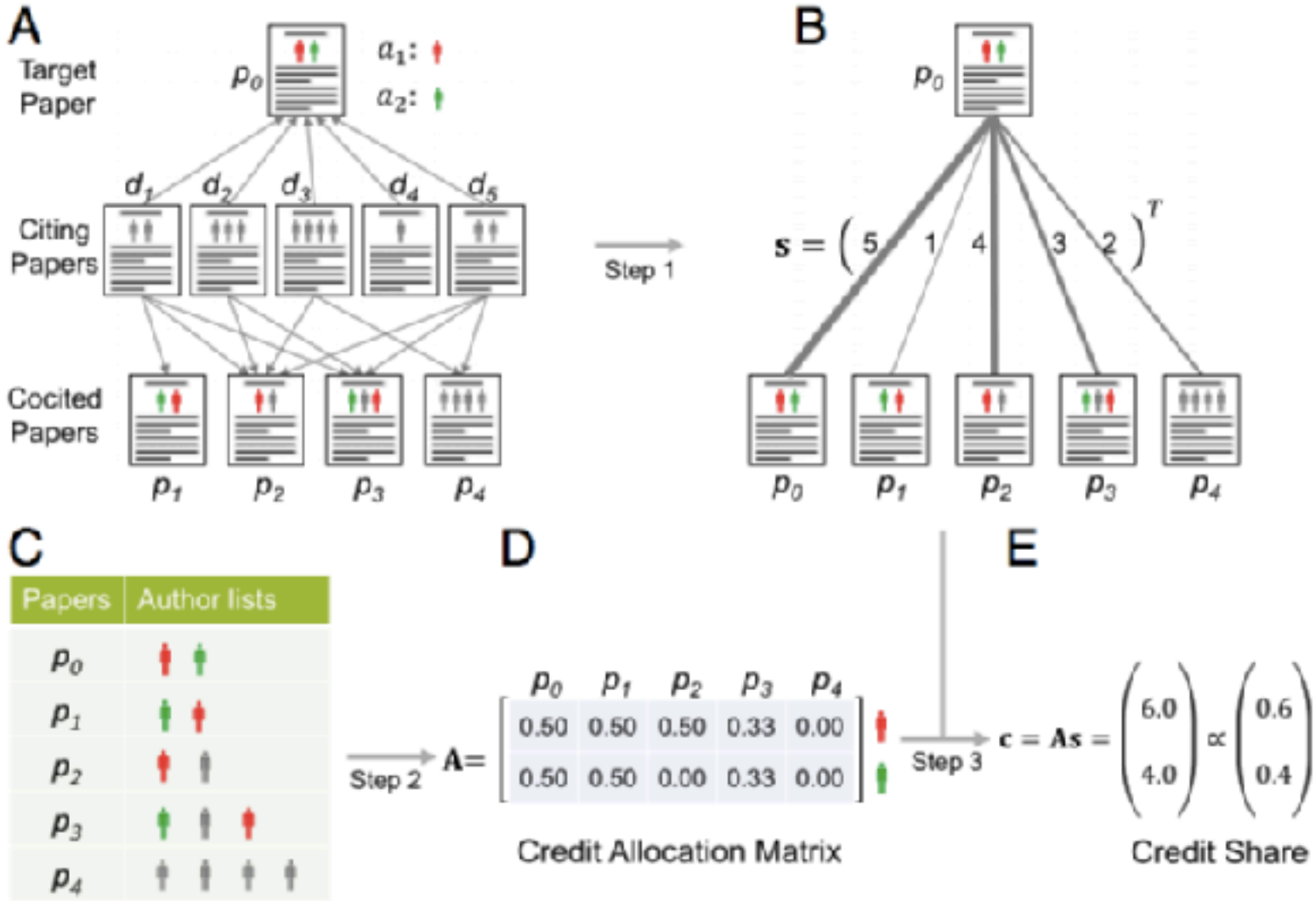
	citing_doi	cited_doi
40	10.1103/PhysRevLett.67.407	10.1103/PhysRevLett.67.406
41	10.1103/PhysRevE.53.5130	10.1103/PhysRevLett.67.406
42	10.1103/PhysRevE.70.061405	10.1103/PhysRevE.59.603
43	10.1103/PhysRev.51.306	10.1103/PhysRev.49.317
44	10.1103/PhysRevLett.74.4899	10.1103/PhysRevB.49.4003
45	10.1103/PhysRevB.51.1370	10.1103/PhysRevB.49.4003
46	10.1103/PhysRevB.53.2523	10.1103/PhysRevB.49.4003
47	10.1103/PhysRevB.63.214406	10.1103/PhysRevB.49.4003
48	10.1103/PhysRevB.55.5404	10.1103/PhysRevB.49.4003
49	10.1103/PhysRevB.50.3779	10.1103/PhysRevB.49.4003

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



Shen's 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.

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:

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

How to compute the importance scores?

Importance Scores

1. Number of citations
2. Sum up the credits of each coauthors
3. ...

Next Steps

1. Produce a relative ranking of the authors based on their cumulative credits earned from all articles
2. Produce a relative ranking of the articles
3. Analyze the primary factors (e.g. the number of co-citation articles) with the largest impact on the credit allocation (remove nodes or edges to observe the differences)
4. Evaluate the improvement over Shen's method: increasing the credit gap between the Nobel winners & other coauthors.
5. Visualization of the authorship bipartite network and the citation network

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.

https://github.com/greenelab/scihub/issues/6

greenelab / scihub

Watch 18

Code | Issues 6 | Pull requests 0 | Projects 0 | Wiki | Insights

DOIs on Crossref, SciHub, and LibGen #6

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

dhimmel commented on Apr 13, 2017 • edited ▾ Owner +😀

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.

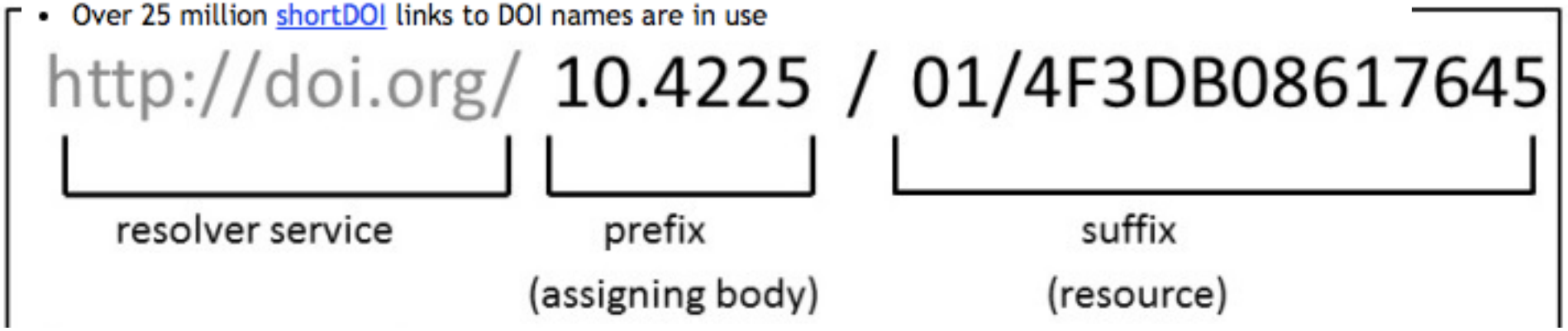
In addition, we've downloaded all DOIs in Crossref (`greenelab/crossref`). This issue breaks out discussion on #4 (comment). First, what does each DOI set cover? Second, are there any peculiarities like broken or erroneous DOIs?

First regarding what's included in each dataset:

DOI

A **digital object identifier (DOI)** is a character string used to uniquely identify an object such as an electronic document. Metadata about the object is stored in association with the DOI name and this metadata may include a location, such as a URL, where the object can be found. The DOI for a document remains fixed over the lifetime of the document, whereas its location and other metadata may change.

- Foundation launched to develop system in 1998. First applications launched 2000
- Currently used by well over 5,000 assigners, e.g., publishers, science data centres, movie studios, etc.
- Approximately 148 million DOI names assigned to date
- Over 22,000 DOI name prefixes within the DOI System
- Over 5 billion DOI resolutions per year
- DOI names are assigned by multiple RAs worldwide
- Over 25 million shortDOI links to DOI names are in use



This is the web site of the International DOI Foundation (IDF), a not-for-profit membership organization that is the governance and management body for the federation of Registration Agencies providing Digital Object Identifier (DOI) services and registration, and is the registration authority for the ISO standard (ISO 26324) for the DOI system. The DOI system provides a technical and social infrastructure for the registration and use of persistent interoperable identifiers, called DOIs, for use on digital networks.

Next Step

Design a dynamically updated allocation algorithm to propagate a fixed amount of credits on the authorship bipartite network and the citation network.

Citations only is not accurate:

Bibliometrics: Is your most cited work your best?

John P. A. Ioannidis, Kevin W. Boyack, Henry Small, Aaron A. Sorensen & Richard Klavans