Scientometrics manuscript No.
(will be inserted by the editor)

An efficient system to fund science: from proposal review to peer-to-peer distributions

Johan Bollen[1,3,4*], David Crandall[1,4], Damion Junk[1], Ying Ding[1,3], and Katy Börner[1,2,3,4]

the date of receipt and acceptance should be inserted later

[1]: School of Informatics and Computing, Indiana University
[2]: Department of Information and Library Science, Indiana University
[3]: Indiana University Network Institute, Indiana University
[4]: Center for Complex Network and Systems Research, Indiana University
∗: Corresponding Author.

Abstract This paper presents a novel model of science funding that exploits the wisdom of the scientific crowd. Each researcher receives an equal, unconditional part of all available science funding on a yearly basis, but is required to individually donate to other scientists a given fraction of all they receive. Science funding thus moves from one scientist to the next in such a way that scientists who receive many donations must also redistribute the most. As the funding circulates through the scientific community it is mathematically expected to converge on a funding distribution favored by the entire scientific community. This is achieved without any proposal submissions or reviews. The model furthermore funds scientists instead of projects, reducing much of the overhead and bias of the present grant peer review system. Model validation using large-scale citation data and funding records over the past 20 years show that the proposed model could yield funding distributions that are similar to those of the NSF and NIH, and the model could potentially be more fair and more equitable. We discuss possible extensions of this approach as well as science policy implications.

1 Introduction

Public agencies such as the U.S. National Science Foundation (NSF) and the National Institutes of Health (NIH) award tens of billions of dollars in science funding annually. How can this money be distributed as efficiently as possible to best promote scientific innovation and productivity?

During 2015 alone, the NSF conducted 231,000 proposal reviews to evaluate 49,600 proposals that directly funded 350,000 people (researchers, postdocs,

Address(es) of author(s) should be given

trainees, teachers, and students) [1]. Although considered the scientific gold standard [2], grant peer review requires significant overhead costs [3,4]. Herbert et al (2013) [5] estimate that Australian researchers alone spent five centuries worth of research time preparing proposals. Extrapolating for population size, the US grant review system may incur even greater overall overhead costs. This situation may be no less favorable at the individual or institutional levels where faculty spend significant amounts of valuable time preparing and submitting research proposals. According to at least one study, faculty report to spend 42% of their time attending to pre- and post-award administrative demands [6]. Principal Investigators (PI) and co-PIs may spend 116 hours and 55 hours per proposal respectively. As a result, the monetary value of any subsequent grant may be greatly diminished when taking into account all preparation, submission, reviewing, as well as professional and personal costs [7], even to the point that applying for grants could become, on balance, a poor investment for universities and faculty alike.

Peer review may furthermore be subject to biases, inconsistencies, and oversights that are difficult to remediate [8–20]. These issues have led some to propose less costly, and possibly more reliable and effective alternatives, such as the random distribution of funding [21] or person-directed funding that does not involve proposal preparation and review [22]. Proposals to reform funding systems have ranged from incremental changes to the peer review process including careful selection of reviewers [18] and post-hoc normalization of reviews [20], to more radical proposals such as opening the proposal review process to the entire online population [23] or removing human reviewers altogether by allocating funds equally, randomly, or through an objective performance measure [4].

Here we investigate a new class of funding models in which all scientists individually participate in the allocation of research funding. All participants receive an equal portion of all yearly funding, but they are then required to anonymously donate a fraction of everything that they have received to peers. Funding thus flows from one participant to the next, each acting as if he or she were a funding agency themselves. These distributed systems incorporate the opinions of the entire scientific community, but in a highly-structured framework that encourages fairness, robustness, and efficiency.

We use large-scale citation data (37 million articles, 770 million citations) as a proxy for how researchers might distribute their funds in a large-scale simulation of the proposed system. Model validation suggests that such a distributed system for science yields funding patterns similar to existing NIH and NSF distributions, but it may do so at much lower overhead while exhibiting a range of other desirable features. Our results indicate that self-correcting mechanisms in scientific peer evaluation can yield an efficient and fair distribution of funding. The proposed model can be applied in many other situations in which top-down or bottom-up allocation of public resources is either impractical or undesirable, e.g. public investments, distribution chains, and shared resource management.

2 Methods

In the proposed system, all scientists are given an equal and unconditional base amount of yearly funding. Each year, however, they are also required to distribute a given percentage of their funding to other scientists whom they feel would make

best use of the money (Fig. 1). Each scientist thus receives funds from two sources: the fixed based amount they receive unconditionally and the amounts they receive from other scientists. As scientists donate a fraction of their total funding to other scientists each year, funding moves from one scientist to the next. Everyone is guaranteed the base amount, but larger amounts will accumulate with scientists whom most scientists believe will make best use of the funding.

For example, suppose the base funding amount is set to $100,000. This roughly corresponds to the entire NSF budget in 2010 divided by the total number of senior researchers it funded [24]. If the required donation fraction is set to $F = 0.5$, i.e., 50%, scientist K receives her yearly base amount of $100,000. In addition she receives $200,000 from other scientists in 2012, bringing her total funding to $300,000. In 2013, K may spend half of that total, i.e. $150,000, on her own research program, but must donate the other half to other scientists for their use in 2014. Rather than painstakingly submitting project proposals, K and her colleagues only need to take a few minutes of their time each year to log into a centralized website and enter the names of the scientists they choose to donate to and how much each should receive.

More formally, suppose that a funding agency maintains an account for each of the N qualified scientists (chosen according to some criteria such as academic appointment status or recent research productivity), to whom we assign unique identifiers in $[1, N]$. Let $O^t_{i\to j}$ denote the amount of money that scientist i gave to scientist j in year t. The amount of funding A each scientist receives in year t is equal to the base funding B from the government plus the contributions from other scientists,

$$A^t_i = B + \sum_{j\in[1,N]} O^{t-1}_{j\to i}.$$

We require that every scientist gives away a fraction F of their funds each year,

$$\sum_{j\in[1,N]} O^t_{i\to j} = (F)A^t_i \qquad \forall i \in N,$$

while they are allowed to spend the remaining money on their research activities. Taken together, these two equations provide a "recursive" definition of the behavior of the overall funding system. A similar formulation has been used to rank webpages by transferring influence from one page to the next [25], as well as to rank scientific journals [26], and author "prestige" [27].

This simple, highly distributed process yields surprisingly sophisticated behavior at a global level. First, respected and highly-productive scientists are likely to receive a comparatively large number of donations. They must in turn distribute a fraction of this total to others; their high status among scientists thus affords them both greater funding and greater influence over how funding is distributed. Second, as the priorities and preferences of the scientific community change over time, the flow of funding will change accordingly. Rather than converging on a stationary distribution, the system will dynamically adjust funding levels as scientists collectively assess and re-assess each others' merits.

3 Results

How would funding decisions made by this system compare to the gold standard of peer review? No funding system will be optimal since research outcomes and impact are difficult to predict in advance [28]. At the very least, one might hope that the outcome of the proposed system would match those of existing funding systems.

To investigate whether this minimal criterion might be satisfied, we conducted a large-scale, agent-based simulation to test how the proposed funding system might operate in practice. For this simulation, we used citations as a proxy for how each scientist might distribute funds in the proposed system. We extracted 37 million academic papers and their 770 million references from Thomson-Reuters' 1992 to 2010 Web of Science (WoS) database. We converted this data into a citation graph by matching each reference with a paper in the database by comparing year, volume, page number, and journal name, while allowing for some variation in journal names due to abbreviations. About 70 percent of all references could be matched back to a manuscript within the WoS data itself.

From the matching 37 million papers, we retrieved 4,195,734 unique author names; we took the 867,872 names who had authored at least one paper per year in any consecutive five years of the period 2000–2010 to be in the set of qualified scientists for the purpose of our study. For each pair of authors, we determined the number of times one had cited the other in each year of our citation data (1992–2010). We also retrieved NIH and NSF funding records for the same period, a data set which provided 347,364 grant amounts for 109,919 unique scientists [29].

We then ran our simulation beginning in the year 2000, in which everyone was given a fixed budget of $B = \$100,000$. We simulated the system by assuming that all scientists would distribute their funding in proportion to their citations over the prior five years. For example, if a scientist cited paper A three times and paper B two times over the last five years, then she would distribute three-fifths of her budget equally among the authors of A, and two-fifths amongst the authors of B.

Importantly, we stress that we are merely using citation data as a proxy for how scientists *might* distribute their funding for purposes of simulation, and are *not* proposing that actual funding decisions be made on the basis of citation analysis. Of course, scientists cite papers for a variety of reasons, not all of which indicate positive endorsement or influence on their work [30]. Nevertheless, although this proxy is an imperfect prediction of scientists' local funding distribution decisions, it permits a large-scale simulation that provides an initial indication of how the overall system may operate in practice.

The results of our simulation suggest that the resulting funding distribution would be heavy-tailed (Fig. 2a), and similar in shape to the actual funding distribution of NSF and NIH for the period 2000-2010 if $F \simeq 0.5$. As expected, the redistribution fraction F controls the shape of the distribution, with low values creating a nearly uniform distribution (less redistribution) and high values creating a highly biased distribution (more redistribution). This suggests that the value of F could be changed on the basis of policy objectives to control the level of funding inequality.

Finally, we used a very conservative (and simple) heuristic to match the author names from our simulation results to those listed in the actual NSF and NIH funding records: we simply normalized all names to consist of a last name, a

first name, and middle initials, and then required exact matches between the two sets. This conservative heuristic yielded 65,610 matching scientist names. For each scientist we compared their actual NSF and NIH funding for 2000–2010 to the amount of funding predicted by simulation of the proposed system, and we found that the two were correlated with Pearson $R = 0.268$ and Spearman $\rho = 0.300$ (Fig. 2b).

4 Discussion and conclusion

These results suggest that our proposed system would lead to funding distributions that are highly similar in shape and individual level of funding to those of the NIH and NSF, if scientists are compelled to redistribute 50 percent of their funding each year – but at a fraction of the time and cost of the current peer-review system.

We note that the ability to mimic or reproduce the shape of the existing funding distribution is not a *conditio sine qua non*. It is certainly possible, and perhaps even likely given the underlying mathematics, that the proposed system yields funding distributions that are dissimilar from those obtained by the grant peer review system, and could potentially be more valid, equitable, and supportive of scientific innovation. This, however, cannot be determined from our simulation, which used citation behavior as a proxy for actual funding decisions.

As shown in Fig. 2, our results indicate very high levels of funding inequality in the present distribution of NIH and NSF funding, indicating that few individuals (or projects) receive very high amounts of funding while most receive rather low amounts of funding. Determining whether this situation promotes scientific innovation or not falls outside the scope of this work, but the proposed system provides a straightforward mechanism to create higher or lower levels of funding equality. At high levels of redistribution, the system becomes more strongly focused on merit as scientists retain less of their base amount and become more strongly dependent on donations from other scientists. This may lead to increasing levels of funding inequality. At low levels of redistribution, scientists retain more of the base amount and receive fewer donations. In other words, the system becomes less meritorious and more egalitarian, leading to higher levels of funding equality. The redistribution factor (F) thus provides policy makers with powerful new leverage to render the distribution of science funding more equitable (less redistribution) or more meritorious (more redistribution) than is presently the case.

The proposed framework would fund *people* [31] instead of funding *projects*. Changes to the present grant peer review system have not significantly reduced the average age of first time grant recipients [32], but the proposed system supports all scientists equally regardless of age or career development stage. Since every scientist receives an unconditional base amount every year, early career scientists could focus on building their research programs rather than spending valuable research time and resources to acquire funding [33].

In general, our system would significantly reduce the amount of time scientists spend preparing and submitting proposals, freeing more time for scientific discovery and innovation. It also introduces incentives for a more open communication of scientific results and research plans. To receive donations, scientists would have to actively communicate the value of their work to the larger scientific community and to the public. A strong commitment to clear communication, open science,

and transparency might prove to be more effective than pursuing traditional tokens of academic merit, e.g., publications and citations. Conferences, workshops, journals, and publishers may have to change their *modus operandi* accordingly.

Of course, funding agencies and governments may still wish to play a directive role, e.g., to encourage advances in certain areas of national interest or to foster diversity. This could be included in the outlined system in a number of straightforward ways. Most simply, traditional peer-reviewed, project-based funding could be continued in parallel, using the proposed system to distribute only a portion of public science funding. Traditional avenues may also be needed to fund projects that develop or rely on large scientific tools and infrastructure. Alternatively, funding agencies could vary the base funding rate B across different constituencies, allowing them to temporarily inject more money into certain disciplines or certain underrepresented groups. The system could also include explicit temporal dampening to prevent large funding changes from year to year, e.g. by allowing scientists to save or accumulate portions of their funding.

In practice, the system will require Conflict of Interest rules similar to the ones that presently keep grant peer-review fair and unbiased. For example, scientists might be prevented from funding advisors, advisees, close collaborators, coauthors, or researchers at their own institution. The interface of an online funding platform might automatically preclude such donations. Donations must furthermore be kept confidential in order to prevent groups of people from colluding to affect the global funding distribution. Such collusion might however be easily detectable since large-scale, objective, but anonymous donation data will be available to analysts and policy-makers. In fact, being able to detect "gaming" may be another major advantage over the current system.

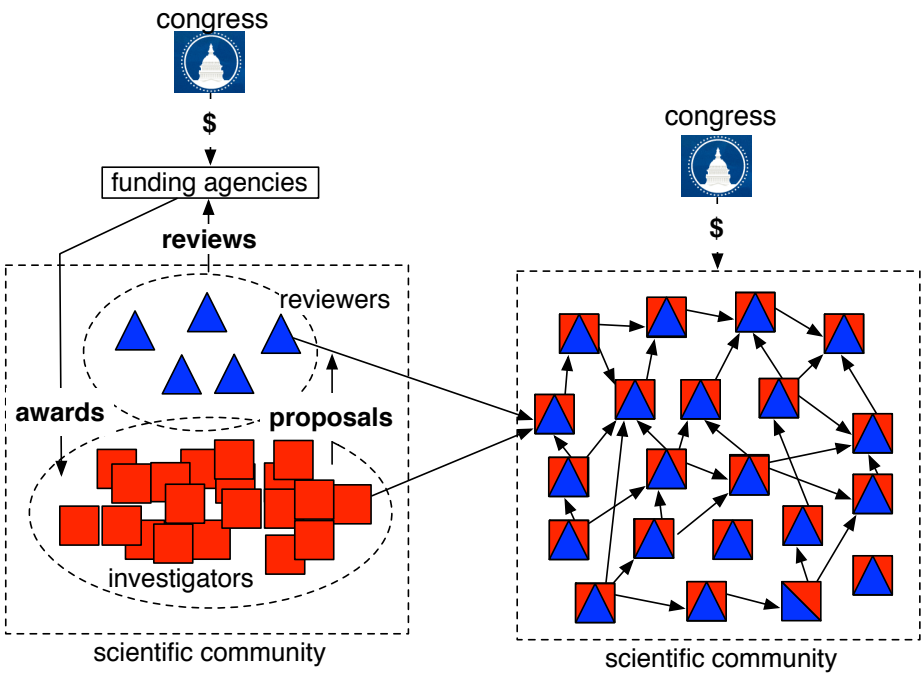
In summary, peer-review of funding proposals has served science well for decades, but funding agencies may want to consider alternative approaches to public science funding that build on theoretical advances and leverage modern technology and big data analytics. The system proposed here requires a fraction of the costs associated with peer review. The potential savings of financial as well as human resources could be used to better identify targets of opportunity, to translate scientific results into products and jobs, and to help communicate scientific and technological advances to the public and to policy makers.

References

1. National Science Foundation. Fy2015 agency financial report. NSF report nsf16002 (2015). URL http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf16002
2. T. Jefferson, Nature web debate (2006). DOI doi:10.1038/nature05031
3. F. Southwick, The Scientist (2012)
4. R. Roy, Science, Technology, & Human Values **10**, 73 (1985)
5. D.L. Herbert, A.G. Barnett, N. Graves, Nature **495**(314) (2013). DOI doi:10.1038/495314d
6. F.S.C. of the Federal Demonstration Partnership, A profile of federal grant administrative burden among federal demonstration partnership faculty. Tech. rep. (2007). URL http://www.iscintelligence.com/archivos_subidos/usfacultyburden_5.pdf
7. D.L. Herbert, J. Coveney, P. Clarke, N. Graves, A.G. Barnett, BMJ Open **4**(3) (2014). DOI 10.1136/bmjopen-2013-004462
8. T. Gura, Nature **416**(6878), 258 (2002)
9. S. Suresh, Nature **490**, 337 (2012). DOI 10.1038/490337a
10. S. Wessely, The Lancet **352**, 301 (1998)
11. D.V. Cicchetti, Behavioral and Brain Sciences **14** (1991)
12. G.S. S. Cole, J.R. Cole, Science **214**, 881 (1981)
13. P. Azoulay, J.S.G. Zivin, G. Manso. NIH peer review: Challenges and avenues for reform. National Bureau of Economic Research Working Paper 18116 (2012)
14. D. Horrobin, The Lancet **348**, 1293 (1996)
15. Editorial, Nature **493**(7434) (2013)
16. J.M. Ioannidis, J.P.A. Nicholson, Nature **492**, 34 (2012). DOI doi:10.1038/492034a
17. R. Smith, Journal of the Royal Society of Medicine **99**, 178 (2006)
18. H.W. Marsh, U.W. Jayasinghe, N.W. Bond, American Psychologist **63**, 160 (2008)
19. L. Bornmann, R. Mutz, H.D. Daniel, Journal of Informetrics **1**, 226 (2007)
20. V.E. Johnson, Proceedings of the National Academy of Sciences **105**, 11076 (2008)
21. N. Geard, J. Noble, in *3rd World Congress on Social Simulation* (2010)
22. P. Azoulay, J.S.G. Zivin, G. Manso, Incentives and creativity: Evidence from the academic life sciences. Working Paper 15466, National Bureau of Economic Research (2009). DOI 10.3386/w15466. URL http://www.nber.org/papers/w15466
23. Editor, Nature **468**, 1002 (2010)
24. National Science Foundation. Fiscal Year 2010 Budget Request to Congress (2009)
25. S. Brin, L. Page, Computer Networks and ISDN Systems **30**, 107 (1998)
26. J. Bollen, M.A. Rodriguez, H. Van de Sompel, Scientometrics **69**(3), 669 (2006)
27. Y. Ding, E. Yan, A. Frazho, J. Caverlee, Journal of the American Society for Information Science and Technology **60**(11), 2229 (2009)
28. N. Myhrvold, Science **282**, 621 (1998)
29. G.L. Rowe, S. Burgoon, J. Burgoon, W. Ke, K. Börner, in *Proceedings of the 11th International Conference on Scientometrics and Informetrics* (2007), pp. 457–462
30. M.H. MacRoberts, B.R. MacRoberts, Journal of the American Society for Information Science **40**(5) (1989)
31. N. Gilbert, Nature (online) **462**(145), 145 (2009). DOI doi:10.1038/462145a
32. J. Kaiser, Science **322**(5903), 834 (2008)
33. Editor, Scientific American **April 25** (2011)

Acknowledgements (1) The authors acknowledge the generous support of the National Science Foundation under grant SBE #0914939, the National Institutes of Health under grants #P01AG039347 and #U01GM098959, and the Andrew W. Mellon Foundation. We also thank the Los Alamos National Laboratory Research Library, the LANL Digital Library Prototyping and Research Team, Thomson-Reuters, and the Cyberinfrastructure for Network Science Center at Indiana University for furnishing the data employed in this analysis. The authors thank Marten Scheffer (Wageningen University) for his extensive feedback on our work and his support of *in vivo* implementations.
(2) The authors declare that they have no competing financial interests.
(3) Correspondence and requests for materials should be addressed to Johan Bollen (email: jbollen@indiana.edu).

Figures and tables



Existing funding system **Proposed funding system**

Fig. 1: Illustrations of existing (left) and the proposed (right) funding systems, with reviewers marked with triangles and investigators marked by squares. In most current funding models like those used by NSF and NIH, investigators write proposals in response to solicitations from funding agencies, these proposals are reviewed by small panels, and funding agencies use these reviews to help make funding decisions, providing awards to some investigators. In the proposed system, all scientists are *both* investigators and reviewers: every scientist receives a fixed amount of funding from the government and other scientists but is required to redistribute some fraction of it to other investigators.

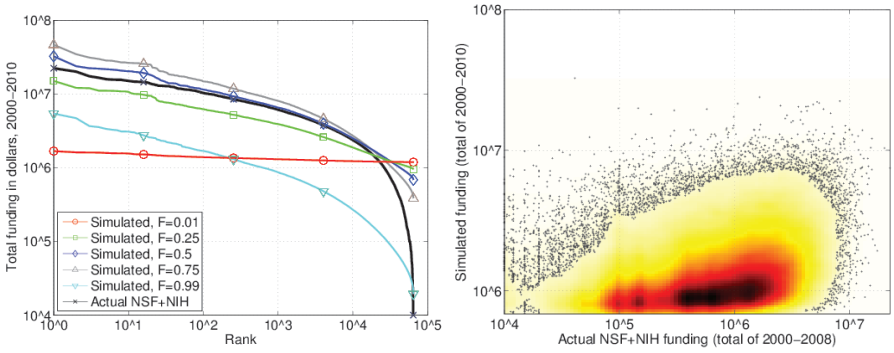


Fig. 2: Results of the distributed funding system simulation for 2000-2010. *(a):* The general shape of the funding distribution is similar to that of actual historical NSF and NIH funding distribution. The shape of the distribution can be controlled

by adjusting F, the fraction of funds that scientists must give away each year. *(b):* On a per-scientist basis, simulated funding from our system (with F=0.5) is correlated with actual NSF and NIH funding (Pearson $R = 0.2683$ and Spearman $\rho = 0.2999$).

Supplemental materials

Overview

The simulation of our proposed funding system (which we call FundRank) was based on the assumption that we could use authors' citation behavior as a proxy of their potential donation behavior. In other words, we assumed that we could estimate *whom people would donate funding to* based on *whom they frequently cited in the recent past*.

To determine author citation behavior we created an author-to-author citation network from article citation data as follows:

1. Extract an article-to-article citation network from 20 years of Thomson-Reuters' Web of Science (WoS) reference data;
2. Extract the authors from each of the articles in our article citation network;
3. Aggregate article-to-article citations into author-to-author citations;
4. Created an author citation network for each year of the mentioned 20 years of WoS data.

Data

This analysis is based on Web of Science (WoS) citation data that was kindly made available to our project by Thomson-Reuters, by way of the Los Alamos National Laboratory (LANL) Research Library (RL), where it was pre-processed by the Digital Library Prototyping and Research Team of the LANL RL (please see acknowledgements).

Our WoS data spanned 20 years (1990 to 2010) and offered bibliographic data for a total of 37.5 million publications. Each primary bibliographic record in the data corresponded to one unique scholarly publication for which the record provided a unique identifier, the publication date, issue, volume, keywords, page range, journal title, article title, volume, and the record's bibliography (list of references).

The references indicate which articles are cited by the primary record, but consisted of a summarized citation that only contained a single author, year, page number, journal title, and volume. Not all references contained values for each of the mentioned fields, and no unique identifiers were provided. A total of about 770 million reference records were contained within the 37.5 million primary bibliographic records in our data.

The primary records and their references should in principle map to the same set of articles, establishing a citation relation between the primary bibliographic record and the articles it references. Given the differences in formatting and the lack of bibliographic information in the references, our set of bibliographic records was thus initially separated into two separate types of data: (1) 37.5 million primary records, and (2) a total of about 770 million summary reference entries contained by the former.

Reference metadata matching for article citation network creation

To establish an article citation network over the 37.5 million primary records in our data, it was necessary to determine which of the 770 million million reference entries mapped back to any of the 37.5 primary bibliographic records. In other words, we had to determine whether an article A cited another article B by determining whether any of A's references matched the bibliographic information of article B. The reference data contained only abbreviated journal titles and abbreviated author names (typically only the first author), so we needed to match

this information to the more detailed bibliographic data provided in the primary records in the WoS data. To do this, we assigned a metadata-based identifier to each primary record,

$$\text{ID} = (\text{journal name}, \text{journal volume}, \text{page number}, \text{year of publication}),$$

which we expected to provide a reasonably unique identifier since it consisted of bibliographic information that was 1) available for both primary records and references, and 2) well-defined, unambiguous numerical information.

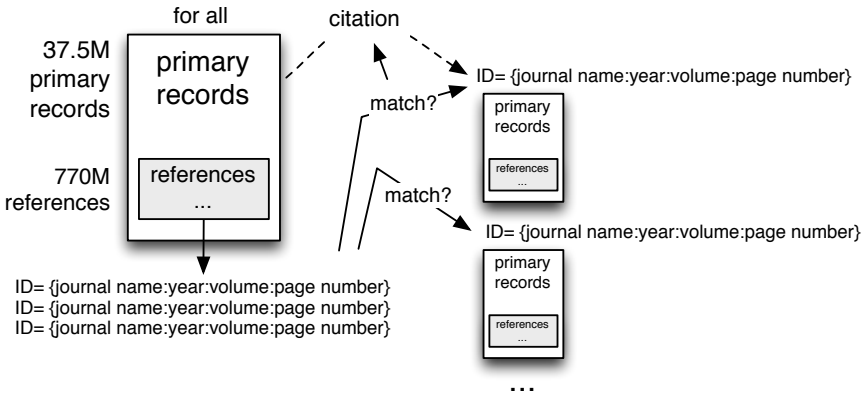


Fig. 3: Matching of metadata-based identifiers generated for (1) references to those generated for (2) primary records to determine citation relations between the two primary records involved.

As shown in Fig. 3, we then attempted to match the article identifier generated for all 37.5 million primary bibliographic records to those of all 770 million references. Each metadata identifier between reference vs. primary record match was taken to indicate a citation relation between the matching primary records.

In doing this matching, we allowed for imperfect, partial matches on some elements of the metadata identifier to account for errors and typos and references:

– Page numbers: The page number could be either within the range indicated by the master, e.g. “1540” matches “1539-1560,” or an exact text match of the page entry for the primary and reference identifier, e.g “13a.”
– Journal titles: Due to significant variation of journal names, for example differing and inconsistent abbreviations, partial matches were allowed for journal titles.

For this latter item, we defined a heuristic algorithm to detect matching journal titles across various spelling and abbreviation standards, as shown in Fig. 4.

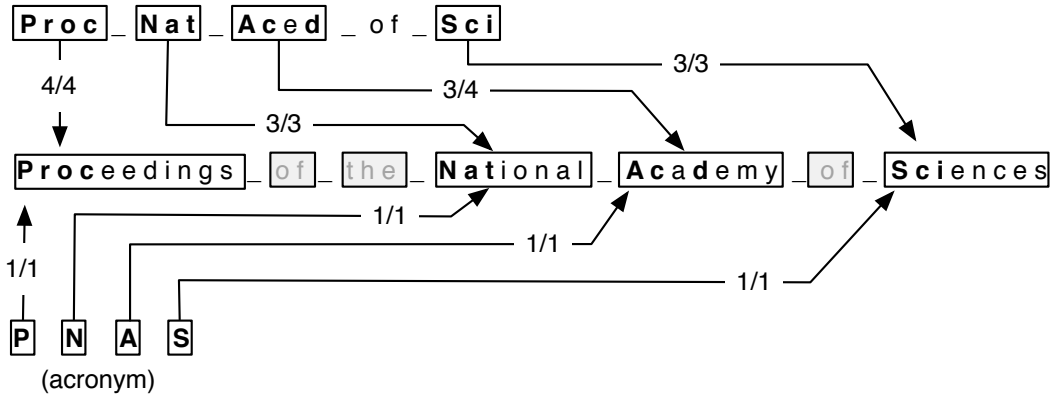


Fig. 4: Matching abbreviated journal title variants by means of longest common substring matching, expanded to handle acronyms.

First, numbers, symbols, and stop words were removed, and repeated spaces were reduced to a single space. All characters were converted to lower-case. Second, the resulting titles were split into individual terms, which were scanned character by character from left to right to compute the degree of overlap between the individual space-delimited words of each title. Then, for each pair of terms across the titles, we calculated the fraction of the characters in the shortest term that matched the characters of the longest term without interruption from the left to the right over the length of the shortest term.

For example, the two titles "Proc Nat Acad Sci" and "proceedings of the national academy of sciences" would first have numbers, symbols, and stop words removed, after which each pair of terms in the resulting titles would be compared character by character to determine their degree of overlap. That is, *"proc"* would be compared to *"proceedings"*, *"nat"* to *"national"*, etc. All four characters of "acad" match the first four characters of "academy", and therefore the two terms were considered a perfect match. The average of the ratios across the entire titles produced a "match ratio" which could be used to assess the degree to which they referred to the same journal.

We included an additional heuristic to handle journal title acronyms. If a title was less than six characters in length and contained no spaces it was considered an acronym. In our matching system, each letter of the acronym was considered as an individual word to be matched against possible targets in the longer title. This allows for comparison such as "PNAS" vs. "Proc. Nat. Acad. Sci", and "PNAS" vs. "Proceedings of the National Academy of Sciences" to result in positive matches.

Finally, a reference record was considered a match when the (year of publication, journal volume, page number) tuple matched exactly, and the journal title match score exceeded a configurable threshold; we used 90 percent in this paper.

This procedure for matching reference identifiers to primary record identifiers allowed us to connect nearly 70 percent of all references to a primary record, thereby achieving an article to article citation network with wide coverage across the entire set of 37.5 million primary records and 20 years of our WoS data.

Author to Author Edge Lists

In the 37 million papers extracted from the WoS data, we found 4,195,734 unique author names. In principle it is trivial to derive an author-to-author network

from the established article-to-article citation network by simply looking up the author names of the primary records, and collating citation numbers across all publication records of the individual authors. Because we were interested activity over a five year sliding window, only citations within that window are considered when building our author-to-author network.

Unfortunately, in practice this procedure is difficult because there is not a one-to-one mapping between author names and scientist names, because (a) different scientists may have the same name, (b) some scientists publish under different names (most notably inconsistently using middle names or middle initials, or using nicknames), and (c) there are typos and other errors in the WoS data.

We first attempted to collapse together duplicate names for the same author. How best to do this depends on the name in question; for a very unique last name, for example, it is sensible to collapse more aggressively than for common names. We used the following heuristic to do this. We partition the list of approximately 4 million names into groups of mostly-similar names, i.e. where the last name and first initial are the same, and proceeded to apply the following procedure to each group:

1. If there's only a single name in the group, then we're done.
2. If there are multiple names, then look at each of them in sequence. For each name X:
   (a) test whether X is a "subname" of exactly one other name in the group, i.e. an abbreviated form of the same name. For example:

   `Lastname, D J` is a subname of `Lastname, Daniel J`
   `Lastname, Daniel` is a subname of `Lastname, D J`
   `Lastname, D` is a subname of `Lastname, D J`
   `Lastname, Dan` is a subname of `Lastname, Daniel`

   If so, then merge those two names together.
   (b) if X is a subname of multiple other names in the group (a set Y of names), then look at all of the names in X and Y. If they are all "mutually compatible" with one another, meaning that they all could refer to the same person, then merge them all together. If not, then do nothing because there is an inherent ambiguity that can't be resolved without additional information.

This procedure amounts to a rather aggressive approach to collapsing names, but it stops whenever ambiguity arises. For example, if the set of author names includes "Lastname, David", "Lastname, D", "Lastname David J", and "Lastname D J", then they are all collapsed into a single author. However if there are also some additional names like "Lastname, Daniel", "Lastname, Dan", "Lastname, Donald", "Lastname, D X", and "Lastname, Donald X", then we end up with the following collapsed equivalence classes:

```
1: {"Lastname, David", "Lastname, David J", "Lastname, D J"}
2: {"Lastname, Daniel", "Lastname Dan"}
3: {"Lastname, Donald", "Lastname D X", "Lastname Donald X"}
4: {"Lastname, D"}
```

The last one becomes a singleton author set because we simply can not resolve the ambiguity without more information.

FundRank simulation

From the resulting list of unique scientists, we filtered out people who had not authored at least one paper per year in any five years of the period 2000-2010. The remaining 867,872 are the group of authors for whom we conduct our FundRank simulation (our Scientists).

The FundRank simulation is carried out as follows. On Jan 1 of each year, each Scientist receives $100,000 as their equal share of the total amount of available funding. On Dec 31 of each year, all Scientists must donate a fraction F of their funding to others, distributed according to the number of citations that points from their papers written that year to other authors, with the restrictions that (a) Scientists cannot contribute money to themselves (even if they cited their own paper) and (b) papers that are more than five years old do not count.

In other words, if an author cited n papers in a given year, and each one had an average of m authors per paper, then this author splits his contributions across the mn (not necessarily distinct) scientists. If a person didn't write any papers in a given year, or didn't cite any papers, they simply distribute their money uniformly across the entire community of Scientists.

Correlations with NSF/NIH funding

We received NIH and NSF funding data from the Cyberinfrastructure for Network Science Center at the School of Library and Information Science at Indiana University. The NIH data lists all details pertinent to 451,188 grants that were made to Principal Investigators (PIs) from January 1990 through the end of 2011. For each grant, the dataset includes the PI name, the award date, PI institution, award amount in US Dollars, and the grant's subject keywords. The NSF data lists all details pertinent to 198,698 grants awarded from 1990-2011: PI name, award date, PI institution, award amount, and NSF program. Both datasets only list the number of co-PIs but do not include their names, so all comparisons are performed for the set of PIs listed only.

Many grants are quite small, and intended to fund small workshops and teaching needs instead of research projects. We attempted to remove these by filtering out any awards of less than $2,000 USD. Similarly, a few awards are unusually large, and correspond to multi-institution grants for major equipment development (e.g. building telescopes) that would be outside the scope of FundRank. We thus also filtered out single awards greater than $2 million USD as well.

We used a very conservative (and simple) heuristic to match up PI names between the NSF and NIH datasets and the author names from our FundRank simulation results: we simply normalized all names to consist of a last name and first and middle initials, and then required *exact* matches between the two sets. This conservative heuristic yielded 65,610 matching author/PI names, which were then used to perform correlations.