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.

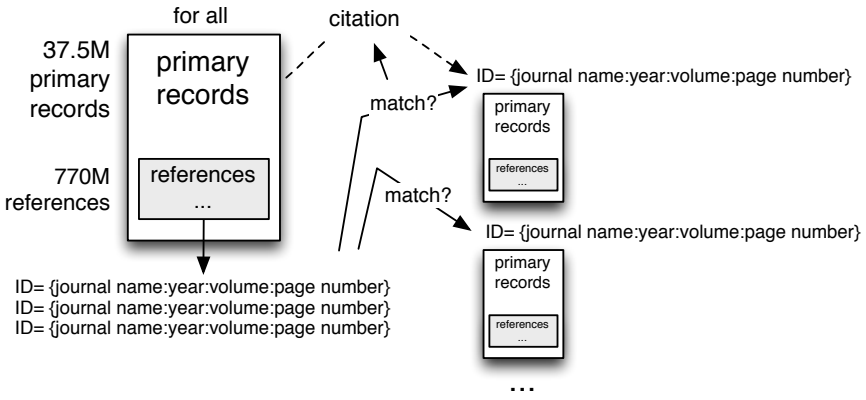

Fig. 3: Matching of metadata-based identifiers generated for (1) references to those generated for (2) primary records to determine citation relations between the two primary records involved.

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.

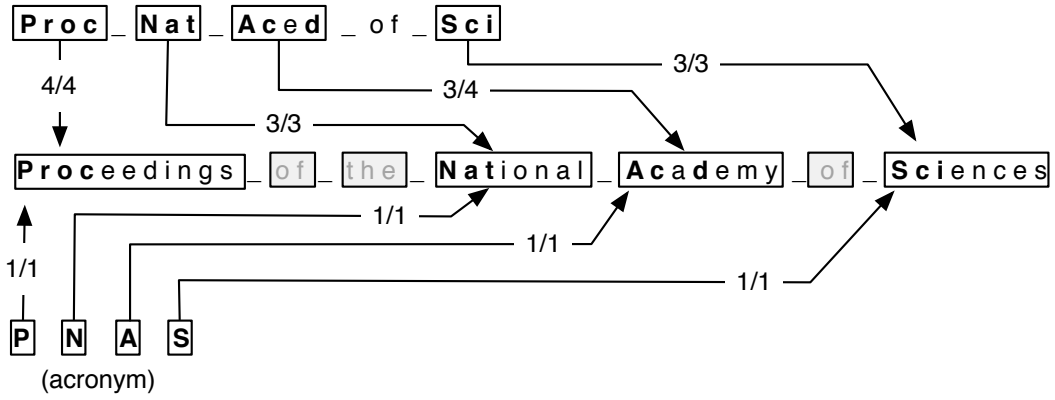

Fig. 4: Matching abbreviated journal title variants by means of longest common substring matching, expanded to handle acronyms.

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.