NETWORK NEUROSCIENCE

an open access journal

RESEARCH ARTICLE

# Co-citations in context: Disciplinary heterogeneity is relevant

**James Bradley**[1], **Sitaram Devarakonda**[2], **Avon Davey**[2], **Dmitriy Korobskiy**[2], **Siyu Liu**[2], **Tandy Warnow**[3] **and George Chacko**[2]

[1]Raymond A. Mason School of Business, College of William and Mary, Williamsburg, VA, USA

[2]Netelabs, NET ESolutions Corporation, McLean, VA 22102, USA

[3]Department of Computer Science, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA

**Keywords:** (bibliometrics, co-citation analysis, random graphs )

## ABSTRACT

Citation analysis of the scientific literature has been used to study and define disciplinary boundaries, to trace the dissemination of knowledge, and to estimate impact. Co-citation, the frequency with which pairs of publications are cited, provides insight into how documents relate to each other and across fields. Co-citation analysis has been used to characterize combinations of prior work as conventional or innovative and to derive features of highly cited publications. Given the organization of science into disciplines, a key question is the sensitivity of such analyses to frame of reference. Our study examines this question using semantically-themed citation networks. We observe that trends reported to be true across the scientific literature do not hold for focused citation networks, and conclude that co-citation analysis requires a contextual perspective.

## INTRODUCTION

Citation and network analysis of scientific literature reveals information on semantic relationships between publications, collaboration between scientists, and the practice of citation itself (de Solla Price, 1965; Newman, 2001; **?**; **?**; **?**). Co-citation, the frequency with which two documents are cited together in other documents provides additional insights, including the identification of semantically related documents, fields, idea, and specialization in science (**??**).

In an intriguing study, co-citation analysis of 17.9 million articles and their cited references from the Web of Science (WoS) was used to document novel combinations of prior research and to characterize highly cited papers (**?**). In this study, pairwise combinations of the references in each article were generated to create a multiset of journal-pairs. Observed frequencies of these journal pairs across the dataset were computed and normalized to expected values generated by Monte Carlo simulations that shuffled cited references in the context of a random graph model. The resultant normalized, journal pair frequencies were termed *z-scores* (Materials and Methods). Thus, every article was associated with multiple z-scores. For each article, positional statistics of these z-scores were calculated to describe conventionality;(high conventionality (HC) if the median z-score for an article was greater than the median of median z-scores of all articles and low conventionality for the inverse (LC). Similarly with novelty; high novelty (HN) if the tenth percentile of z-scores for an

Citation: Betzel, R. F., Fukushima, M., He, Ye, Zuo, Xi-Nian, Sporns, O. (2016) Dynamic fluctuations coincide with periods of high and low modularity in resting-state functional brain networks Network Neuroscience, 1

DOI:
http://dx.doi.org/10.1162/NETN-00001

Supporting Information:
http://dx.doi.org/10.7910/DVN/PQ6ILM

Received: 20 October 2016
Accepted: 7 November 2016
Published: 26 January 2016

Competing Interests: The authors have declared that no competing interests exist.

Corresponding Author:
George Chacko
netelabs@nete.com

Handling Editor:
Xi-Nian Zuo

Copyright: © 2019
Massachusetts Institute of Technology
Published under a Creative Commons Attribution 4.0

The MIT Press

DRAFT

article was less than zero and low novelty (LN) for the inverse. Accordingly, each article was labeled with respect to conventionality and novelty, e.g, HCLN, with all four combinations being possible. The finding of considerable interest was that HCHN articles were twice as likely to be found in highly cited papers (?) suggesting that novel ideas flavoring a base of conventional thought were a recipe for impact. The authors examined subsets of publications grouped by disciplined and asserted universality of this feature.

Key to the Uzzi study, however, is the random graph approach and the mechanism of citation shuffling employed. Random substitutions are made under conditions that preserve the number of publications in each year of the dataset as well as the number of references cited by each of these publications. However, the approach inadequately accounts for observed citation behavior (????) since (i) random substitutions are selected with equal probability irrespective of disciplinary origin (ii) replacements are drawn from the set of eligible references rather than the multi-set that accounts for the frequency with which a reference is cited. In illustration, a musicology reference from can replace a quantum physics reference during the simulation process. Similarly, a reference cited over 100 times in a given year is selected with the same probability as a reference cited only once. Accordingly, the expected value calculations generated by simulations and used in normalizing observed journal-pair-frequencies can be reasonably questioned on grounds of model mis-specification.

A follow-up study using a smaller dataset of 12 million articles, also suggests that Uzzi and colleagues do not fully account for disciplinary and journal effects (?). While the two studies are conceptually similar, the authors of the latter study used Scopus data and the K50 measurement, a derivative of the $\chi^2$ method, rather than Monte Carlo simulations for normalizing observed to expected journal-pair frequencies. Boyack and Klavans note that "only 64.4% of 243 WoS subject categories" in the Uzzi study met the criterion of having the highest probability of hit papers in the HCHN category. Further, they observed that journals vary widely in terms of size and influence and that 20 journals accounted for nearly 15% of co-citations in their measurements. Lastly, that three multidisciplinary journals accounted for 9.4% of all atypical combinations, suggesting strong effects from both disciplines and journals.

Despite different methods used to generate expected values, expected journal-pair frequencies were generated without disciplinary constraints in both these prior studies (??), after which subsets of publications were examined for disciplinary effects. To avoid the problem of discipline-insensitive substitutions, we chose instead to first construct semantically related sets of documents (disciplinary networks) and then measure observed and expected frequencies within these networks. Accordingly, we used keyword searches of the scientific literature to create three citation networks themed around broad search terms. For each of these, we calculated expected values using an efficient simulation algorithm that randomly permuted references grouped by year of publication in a random graph model that accounted for existing citation frequencies in these networks. We hypothesized that our approach would reduce model misspecification and better simulate citation practice, in turn leading to different and more relevant conclusions. Our study on these semantically-themed citation networks reveals significantly different patterns of conventionality and novelty between citation networks and disciplines that challenges the conclusion of universality in HCHN articles have a greater probability of being represented in highly cited publications. Instead, we conclude that conventional thought as commonly defined in our study, Uzzi, and Boyack is more likely to drive higher citations.. *these are just filler conclusions that must be carefully rewritten and added to.*

## MATERIALS AND METHODS

*Bibliographic data* We have previously developed ERNIE, an open source knowledge platform into which we parse the Web of Science (WoS) Core Collection **?**. WoS data stored in ERNIE spans the period 1900-2019 and consists of over 72 million publications. For this study, we generated an analytical dataset from years 1985 to 2005. The total number of publications in this dataset was just over 25 million publications (25,134,073). For each of these years, we further restricted analysis to those of type Article. Since WoS data also contains incomplete references or references that point at other indexes, we also considered only those references for which there were complete records (Table 1). For example, WoS data for year 2005 contained 1,753,174 publications, which after restricting to type Article and considering only those references described above resulted in 916,573 publications, 6,095,594 unique references (set of references), and 17,167,347 total references (multiset of references). Given consistent trends in the data, we analyzed the two boundary years (1985 and 2005) and the mid-point (1995).

*Disciplinary datasets* We constructed three disciplinary datasets based on keyword searches. (i) immunology (ii) metabolism (iii) applied physics. For the first two, rooted in biomedical research, we searched Pubmed for the term 'immunology' or 'metabolism' in the years 1985, 1995, and 2005. Pubmed IDs (pmids) returned were converted to WoS IDs (wos_ids). For the applied physics dataset, we directly searched article level labels in WoS for 'applied physics'.

*Normalization of observed and expected values* We analyzed Building upon the work of Uzzi and colleagues, $nC2$ reference pairs were generated for each publication where $n$ is the number of cited references in the publication. These reference pairs were then mapped to the journals they were published in using ISSN numbers creating journal pairs. Where multiple ISSN numbers exist for a journal, the most frequently used one in the WoS was assigned to the journal. In addition, publications containing less than two references were discarded. Journal pair frequencies were summed up across the dataset to create observed frequencies$(F_{obs})$. In contrast to the preceding study (?), we we generated 1,000 rather than 10 null models for each dataset by randomly shuffling references while preserving the number of publications, the number of references in each publication, and the frequency with which these references were cited within the year of interest. Expected values $(F_{exp})$ were generated by averaging the result of 1,000 simulations. z-scores were calculated for each journal-pair using the formula $(F_{obs} - F_{exp})/\sigma$ where $\sigma$ is the standard deviation of the frequencies generated by simulation. As a result of these calculations, each publication becomes associated with a set of z-scores corresponding to the journal pairs derived from pairwise combinations of its cited references and positional statistics (quantiles) of z-scores were calculated for each publication. Publications were also labeled according to conventionality and novelty. (i) HC if the median z-score exceeded the median of median z-scores for all publications. LC if the median z-score was equal to or less than the median of median z-scores for all publications (ii) HN if the tenth percentile of z-scores for a publication was less than zero. LN if the tenth percentile of z-scores for a publication was greater than zero.

## RESULTS

(i) disciplinary networks are different (ii) background makes a difference (iii) (iv)

## DISCUSSION

## ACKNOWLEDGMENTS

**Table 1.** Summary of WoS Analytical Dataset. UP: unique publications, UR: unique references, TR: total references. The number of publications, unique references, total references and the ratio of total references to unique references increases monotonically with each year indicating that both the number of documents and citation activity increase over time. Data for reference years is flanked by horizontal lines and shown in boldface. Only publications of type Article and references with complete WoS records are included in these counts.

| Year | UP | UR | TR | TR/UR |
|---|---:|---:|---:|---:|
| **1985** | **418495** | **2281297** | **5615496** | **2.46** |
| 1986 | 402309 | 2316451 | 5708796 | 2.46 |
| 1987 | 412936 | 2427347 | 5998513 | 2.47 |
| 1988 | 426001 | 2545647 | 6354917 | 2.50 |
| 1989 | 443144 | 2673092 | 6749319 | 2.52 |
| 1990 | 458768 | 2827517 | 7209413 | 2.55 |
| 1991 | 477712 | 2977784 | 7729776 | 2.60 |
| 1992 | 492181 | 3134109 | 8188940 | 2.61 |
| 1993 | 504488 | 3278102 | 8676583 | 2.65 |
| 1994 | 523660 | 3458072 | 9255748 | 2.68 |
| **1995** | **559685** | **3692575** | **9897946** | **2.68** |
| 1996 | 663110 | 4144581 | 11641286 | 2.81 |
| 1997 | 677077 | 4340733 | 12135104 | 2.80 |
| 1998 | 693531 | 4573584 | 12728629 | 2.78 |
| 1999 | 709827 | 4784024 | 13280828 | 2.78 |
| 2000 | 721926 | 5008842 | 13810746 | 2.76 |
| 2001 | 727816 | 5203078 | 14261189 | 2.74 |
| 2002 | 747287 | 5464045 | 15001390 | 2.75 |
| 2003 | 786284 | 5773756 | 16024652 | 2.78 |
| 2004 | 826834 | 6095594 | 17167347 | 2.82 |
| **2005** | **916573** | **6629595** | **19066249** | **2.88** |

**Table 2.** Disciplinary Datasets. PubMed and WoS were searched for articles using search terms, ‘immunology’, ‘metabolism’, and ‘applied physics’. Counts of retrieved publications are shown for each of the three years analyzed.

| Year | Immunology | Metabolism | Applied Physics |
|---|---:|---:|---:|
| 1985 | 21,606 | 78,998 | 10,298 |
| 1995 | 29,320 | 121,247 | 21,012 |
| 2005 | 37,296 | 200,052 | 35,600 |

We deeply thank Kevin Boyack and Dick Klavans for critically constructive discussions. We thank the authors of Uzzi et al. (2013) for sharing their Python simulation code. Research and development reported in this publication was partially supported by Federal funds from the National Institute on Drug Abuse, National Institutes of Health, US Department of Health and Human Services, under Contract Nos. HHSN271201700053C (N43DA-17-1216) and HHSN271201800040C (N44DA-18-1216). The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. TW receives funding from the Grainger Foundation. All the code used in this study is freely available from a Github site. Access to bibliographic data requires a license from Clarivate Analytics, which had no role in funding, experimental design, review of results, and conclusions presented.

## AUTHOR CONTRIBUTIONS

This study was designed by GD, JB, SD, and TW. Simulations and analysis were performed by AD, GC, JB, and SD. Infrastructure, and workflows used to generate data used in this study were developed by AD, DK, SL, SD, and GC. All authors reviewed and commented on the manuscript, which was written by GC, JB, and TW.

## REFERENCES

Boyack, K., & Klavans, R. (2014). Atypical combinations are confounded by disciplinary effects. In *International conference on science and technology indicators* (pp. 49–58). Leiden, Netherlands: CWTS-Leiden University.

de Solla Price, D. J. (1965). Networks of Scientific Papers. *Science*, *149*(3683), 510–515. Retrieved 2019-02-03, from http://www.sciencemag.org/cgi/doi/10.1126/science.149.3683.510 doi: 10.1126/science.149.3683.510

Garfield, E. (1955). Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science*, *122*(3159), 108–111. Retrieved 2019-05-16, from https://science.sciencemag.org/content/122/3159/108 doi: 10.1126/science.122.3159.108

Garfield, E. (1979). *Citation Indexing-Its Theory and Application in Science, Technology, and Humanities* . John Wiley and Sons, ISI Press.

Keserci, S., Davey, A., Pico, A. R., Korobskiy, D., & Chacko, G. (2018). ERNIE: A data platform for research assessment. *bioRxiv*. (https://www.biorxiv.org/content/early/2018/07/19/371955) doi: 10.1101/371955

Klavans, R., & Boyack, K. W. (2017). Research portfolio analysis and topic prominence. *Journal of Informetrics*, *11*(4), 1158–1174. Retrieved 2019-05-16, from http://www.sciencedirect.com/science/article/pii/S1751157717302110 doi: 10.1016/j.joi.2017.10.002

Marshakova-Shaikevich, I. (1973). System of document connections based on references. *Nauch-Techn.Inform, Ser.2*, *6*(4), 3-8. Retrieved 2019-02-03, from http://doi.wiley.com/10.1002/asi.4630240406 doi: 10.1002/asi.4630240406

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of informetrics*, *4*(3), 265–277.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, *98*(2), 404–409. Retrieved 2019-03-22, from https://www.pnas.org/content/98/2/404 doi: 10.1073/pnas.98.2.404

Patience, G. S., Patience, C. A., Blais, B., & Bertrand, F. (2017). Citation analysis of scientific categories. *Heliyon*, *3*(5), e00300.

Shi, X., Leskovec, J., & McFarland, D. A. (2010). Citing for high impact. In *Proceedings of the 10th annual joint conference on digital libraries* (pp. 49–58). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1816123.1816131 doi: 10.1145/1816123.1816131

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265–269. Retrieved 2019-02-03, from http://doi.wiley.com/10.1002/asi.4630240406 doi: 10.1002/asi.4630240406

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science (New York, N.Y.)*, *342*(6157), 468–472. doi: 10.1126/science.1240474

Wallace, M. L., Lariviere, V., & Gingras, Y. (2012). A Small World of Citations? The Influence of Collaboration Networks on Citation Practices. *PLOS One*, *7*, e33339. Retrieved 2019-05-16, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0033339 doi: 10.1371/journal.pone.0033339