NETWORK NEUROSCIENCE

an open access journal

CrossMark click for updates

Citation: Betzel, R. F., Fukushima, M.,w 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

The MIT Press

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], Djamil Lakhdar-Hamina[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: (co-citation analysis, bibliometrics, 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 we 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; Garfield, 1955; Newman, 2001; Patience, Patience, Blais, & Bertrand, 2017; Shi, Leskovec, & McFarland, 2010). 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, specializations, and new ideas in science (Boyack & Klavans, 2010; Marshakova-Shaikevich, 1973; Small, 1973; Zuckerman, 2018).

(Uzzi, Mukherjee, Stringer, & Jones, 2013) used a novel approach for co-citation analysis using, as input, 17.9 million articles and their cited references, from the Web of Science (WoS) to characterize a subset of highly cited articles with respect to both novel and conventional combinations of prior research. In this study, the frequency with which references were calculated and reduced to journal pairs frequencies (observed co-citation frequencies). Expected values were generated from Monte Carlo simulations under a random graph model. Observed frequencies were then normalized (shifted and scaled) to averaged expected values from ten simulations and termed as *z-scores*, which were interpreted as indicating conventional thought if positive and novel thought if negative. Consequently, every article was associated with multiple z-scores corresponding to co-cited journal pairs in its refernces. For each article, positional statistics of z-scores were calculated to set thresholds for a binary (high/low) classification of conventionality (C) using the median and novelty (N) using the tenth percentile of z-scores within an article. Thus, HCLN would denote high conventionality and low novelty with all four combinations being possible. Uzzi et al. (2013) observed that HCHN articles were twice as likely to be highly cited, suggesting that novel combinations of ideas flavoring a body of conventional thought were a 'nearly universal' feature for impact.

Key to Uzzi *et al.*, is their random graph model and its underlying assumptions. The citation switching algorithm used to generate a null model elegantly preserves the number of publications, the number of references in each publication, the year of publication of both publications and references, and the disciplinary composition of the references cited in these publications. However, random substitution of the references used to generate expected values is equiprobable with respect to disciplinary origin and citation count. The model permits substitution of a reference in quantum physics with equal probability by a reference in the related quantum chemistry field or in some entirely different field, such as classical literature, evolutionary biology, or anthropology. Such substitutions poorly model the disciplinary nature of scientific endeavor and citation behavior (Garfield, 1979; Klavans & Boyack, 2017; Moed, 2010; Wallace, Lariviere, & Gingras, 2012). In addition, under this random model, a reference cited over 100 times in a given year is selected with the same probability as a reference cited only once, which appears inconsistent with the power law or lognormal citation distributions described in the literature (Perline, 2005; Stringer, Sales-Pardo, & Amaral, 2010). Accordingly, model mis-specification is likely to arise from the expected value calculations used to generate z-scores and characterize articles in terms of conventionality and novelty when disciplinary subsets of the WoS are analyzed in the context of expected values generated across all of WoS.

A follow-up study by (Boyack & Klavans, 2014) explored the impact of discipline and journal effects on the definition of conventionality and novelty. While their study had some methodological differences in the use of Scopus data rather than WoS data, a smaller dataset, and a χ^2 calculation rather than Monte Carlo simulations to generate expected values of journal pairs, Boyack and Klavans also reported the trend that HCHN is more probable in highly cited papers. However, they note that "only 64.4% of 243 WoS subject categories" in the Uzzi *et al.* 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, they noted that three multidisciplinary journals accounted for 9.4% of all atypical combinations, suggesting strong effects from both disciplines and journals that were not reported by Uzzi et al.

Despite different methods used to generate expected values, both of these key preceding studies (*vide supra*) measured co-citation frequencies across the scientific literature without disciplinary constraints and subsequently analyzed disciplinary subsets. In extending their prior work, we hypothesized that normalization using expected values drawn from datasets focused in disciplinary areas, rather than all of WoS, would reduce model mis-specification. Consequently, we used keyword searches of the scientific literature to construct citation networks themed around academic disciplines: *applied physics, immunology, and metabolism*. Within these disciplinary frameworks, we calculated observed and expected co-citation frequencies using a refined random graph model and an efficient Monte Carlo simulation algorithm.

Our analysis of these semantically-themed citation networks challenges the assertion of universality. We demonstrate reduced model misspecification and observe that while HC remains highly correlated with hit articles in the immunology and metabolism datasets, HN is not. In addition, while HN is highly correlated with hit articles in applied physics, HC is not. Thus, disciplinary networks are different from each other and the full WoS network. Beyond variation across disciplines and the case for contextual analysis, we found that the categories demonstrating the highest percentage of hits are not robust with respect to varying parameters that define hit articles and the threshold for highly novel citation patterns.

these remaining sentences require some work The varying citation patterns when sampling from a disciplinary-focused dataset versus a broader dataset due, was a cause of journal pair z-scores changing signs in 28.6% of instances: the effect of any one of these journal pairs on an article's novelty or conventionality is contradictory between broad and, more narrow disciplinary datasets. We contend that the interpretation due to a disciplinary dataset, which preserved citation patterns, is more appropriate and an improvement on current methods.

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 (Keserci, Davey, Pico, Korobskiy, & Chacko, 2018). 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 from ERNIE. The total number of publications in this dataset was just over 25 million publications (25,134,073), which were then stratified by year of publication. For each of these years, we further restricted analysis to publications 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, Table 1, we analyzed the two boundary years (1985 and 2005) and the mid-point (1995) performing 1,000 simulations for each dataset. We also calculated the number of times each of these articles was cited in the first 8 years since publication.

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 (Table 2). Pubmed IDs (pmids) returned were matched to WoS IDs (wos_ids) and used to retrieve relevant articles. For the applied physics dataset, we directly searched subject labels in WoS for 'applied physics'. We also examined publications in the five major research areas in the Web of Science; life sciences & biomedicine, physical sciences, technology, social sciences, arts & humanities using the extended subcategory classification of 153 sub-groups to categorize disciplinary composition of cited references in the datasets we studied.

Monte Carlo simulations, normalization of observed frequencies, annotations, and 'hit' papers. Analysis was performed on publications from 1985, 1995, and 2005. Building upon prior work Uzzi et al. (2013), all $\binom{n}{2}$ reference pairs were generated for each publication, where

Table 1. Summary of base WoS Analytical Dataset.The number of publications of type Article (UP), unique references (UR), total references (TR) 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. Only publications of type Article with at least two references, and references with complete publication data were counted.

year	UP	UR	TR	TR/UR
1985	391,860	2,266,584	5,588,861	2.47
1986	402,309	2,316,451	5,708,796	2.46
1987	412,936	2,427,347	5,998,513	2.47
1988	426,001	2,545,647	6,354,917	2.50
1989	443,144	2,673,092	6,749,319	2.52
1990	458,768	2,827,517	7,209,413	2.55
1991	477,712	2,977,784	7,729,776	2.60
1992	492,181	3,134,109	8,188,940	2.61
1993	504,488	3,278,102	8,676,583	2.65
1994	523,660	3,458,072	9,255,748	2.68
1995	537,160	3,680,616	9,875,421	2.68
1996	663,110	4,144,581	11,641,286	2.81
1997	677,077	4,340,733	12,135,104	2.80
1998	693,531	4,573,584	12,728,629	2.78
1999	709,827	4,784,024	13,280,828	2.78
2000	721,926	5,008,842	13,810,746	2.76
2001	727,816	5,203,078	14,261,189	2.74
2002	747,287	5,464,045	15,001,390	2.75
2003	786,284	5,773,756	16,024,652	2.78
2004	826,834	6,095,594	17,167,347	2.82
2005	886,648	6,615,824	19,036,324	2.88

Table 2. Disciplinary Datasets. PubMed and WoS were searched for articles using search terms, 'immunology', 'metabolism', and 'applied physics'. Counts of 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

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 as identifiers. 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 fewer than two references were discarded. Journal pair frequencies were summed across the dataset to create observed frequencies (F_{obs}).

We developed a performant citation switching algorithm, *runtime enhanced permuting citation switcher (repcs)* that randomly permuted citations grouped by year of publication to switch citations while preserving the year of publication for both articles and references, the number of publications and the number of references in each dataset and the disciplinary composition of the references in each dataset. We used the citation switching algorithm of Uzzi et al. (2013) referred to as *umsj* (Boyack & Klavans, 2014), as a benchmark for the repcs algorithm using code kindly provided by the authors. A comparative analysis showed that while 10 simulations of the WoS 1985 dataset (391,860 publications) completed in 2,186 hours using the umsj algorithm, it completed in less than one hour using the repcs algorithm implemented on a Spark cluster. Using either the repcs or umsj algorithms for 10 simulations resulted in expected value coverage of 75% of the observed journal pair frequencies. Subsequent rounds of development and testing resulted in 99% coverage of observed journal pairs when 1,000 simulations were conducted. The runtime advantage was significant enough that we chose to use the repcs algorithm in our study and generated expected values averaged from 1,000 simulations for improved coverage for every dataset we analyzed.

Averaging the result of 1,000 simulations for each dataset studied, z-scores were then calculated for each journal-pair using the formula $(F_{obs} - F_{exp})/\sigma$ where F_{obs} is the observed frequency, F_{exp} is the averaged simulated frequency, and σ is the standard deviation of the simulated frequencies. 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. Positional statistics of z-scores were calculated for each publication, which was then labeled according to conventionality and novelty: (i) HC if the median z-score exceeded the median of median z-scores for all publications and LC if the median z-score was equal to or less than the median of median z-scores for all publications, and (ii) HN if the tenth percentile of z-scores for a publication was less than zero, and LN if the tenth percentile of z-scores for a publication was greater than zero.

To consider the relationship between citation impact, conventionality, and novelty we calculated percentiles for the number of accumulated citations in the first 8 years since publication for each article we studied and stratified. We also investigated multiple definitions of hit articles with hits defined as the 1%, 2%, 5%, and 10% top-cited articles. Also, mention different thresholds, 1st and 10th, for novelty threshold.

RESULTS

Model Misspecification and the Attributes of Disciplinary Context

We invoke model mis-specification arising from irrelevant references as a concern when disciplinary subsets are analyzed using a random graph model that generates expected co-citation frequencies from the full body of scientific literature. To address this issue, we first assembled three exemplar datasets using keyword searches for the terms 'applied physics', 'immunology', and 'metabolism' (Materials and Methods). Using the WoS extended classificatation of 153 subject areas, we then analyzed the disciplinary composition of the references in these three datasets into the five research areas in the Web of Science (life sciences and biomedicine, physical sciences, social sciences, technology, and arts and humanities) for the years 1985, 1995, and 2005. We observed that 84.2% of the references in the applied physics dataset were classified as belonging to the WoS physical sciences research area. Similarly, 93.7% and 92.1% of the references in the immunology and metabolism datasets belonged to the life sciences and biomedicine research area, indicating strongly that references are more likely belong to the same research area as the publications that cite them.

A symptom of model misspecification is that random sampling of citations associated with an article do not match with the distribution of citations across the disciplines observed in articles from that discipline. Wallace et al. (2012) found that, often, a majority of an article's citations are from the specialty of the article, although that percentage varied among disciplines in the eight specialties they investigated: from approximately 39% to 89% for 2006. We also analyzed the consequences of citation shuffling within a disciplinary set or all of the Web of Science. References in publications belonging to these three datasets were classified using the WoS Extended Subjects classification of 153 subject areas and summarized as a frequency distribution. A single shuffle of the references in these datasets or all the references in the corresponding WoS year slice was performed, using either the rpcs or umsj algorithms, after which subject frequencies were computed again. The fold change in subject frequencies of references before and after shuffling was calculated for these groups using all 153 Subject categories and are summarized as boxplots in Fig 1. For example, the applied physics dataset contained 1 reference labeled Genetics and Heredity . This proportion was preserved in the intra-network shuffle but increased 1496 times when the WoS background was used as a source of substitution during Monte Carlo simulation with the repcs algorithm. Similarly, with the metabolism dataset.which contained one reference labeled Philosophy that was increased to 661 after a single repcs shuffle. The data show convincingly that the disciplinary composition of references in a network is preserved when citation shuffling is conducted within it but is significantly altered when the full WoS network is used as a source of substitution. A second inference is that the choice of algorithm is not relevant beyond runtime advantage but that the background network used for the shuffle is critical.

We predicted that model misspecification as measured by the Kullback-Leibler (K-L) Divergence (Kullback & Leibler, 1951) between observed and simulated frequencies in a disciplinary network would be less than the divergence for observed and simulated frequencies in the WoS superset. The results in Table 3 indicate that for the set of journals common to both a disciplinary network and the WoS superset, simulations under our model consistently have a lower K-L divergence, roughly half, compared to simulations that draw from the WoS superset (and its attendant substitutions that are ectopic with respect to field and discipline).

We used the criteria of Uzzi to calculate normalized journal pair frequencies (z-scores) and classify publications according to conventionality and novelty. We observe that z-score calculations are sensitive to the background network (disciplinary or WoS). Figure 2 shows that the z-scores for the same journal pair can be positive (negative) when computed with respect to one data set but be negative (positive) for another data set. The journal-pair z-scores in Figure 2 have consistent signs for both Immunology and WoS data sets in 71.4% of the instances and different signs for 28.6% of the journal pairs. When a journal-pair z-score is negative with respect to one dataset and positive with respect to another dataset, articles citing that pair are more likely to be deemed novel in the first instance and less likely to be deemed novel in the second instance. It seems that any journal pair should either be indicative of novelty or not, but it should never have contradictory implications that depend on the reference dataset. We view, therefore, these contradictions as symptomatic of inappropriately sampling from a broad dataset and ignoring observed citation patterns. Figure 2 reflects that the WoS data set has approximately 44,000 fewer negative z-scores than does the immunology data set, which contributes to its significantly lower percentage of high-novelty articles. *The plots I made of the cumulative z-score distributions for Immunology and WoS also could be used to demonstrate this phenomenon, although the scatter plot does this more clearly, and it is more striking. The scatter plot also encodes more data, that is, the matched z-scores for each journal pair.*

Further, we stratified publications in these datasets by percentiles using the number of citations accumulated in the first 8 years since publication and according to conventionality and novelty in a second dimension Fig 3 and note that [INSERT WHATEVER WE NOTE].

We addressed this consideration by analyzing disciplinary subsets of the scientific literature, thereby restricting random selection of references to only those references in the disciplinary network being studied. *What about z-scores?* We have conjectured that the approach of Uzzi et al. for generating journal-pair z-scores is misspecified in its sampling from a broad dataset and its disregard for the frequency with which journal pairs are citedOur principal consideration was to restrict model misspecification arising from disciplinarily irrelevant references. We addressed this consideration by analyzing disciplinary subsets of the scientific literature, thereby restricting random selection of references to only those references in the disciplinary network being studied.As observed in the Introduction, Monte Carlo simulations that use references from all publications do not account for observed citation practice (sentence needs to be made more eloquent)...

This paragraph needs work, but I wanted to create a placeholder for the introduction to this subsection. It uses some of the passages form the former Chacko subsection. . Like Uzzi et al., we also used a Monte Carlo approach to simulate under a random graph model, although our principal consideration was to restrict model misspecification arising from disciplinarily irrelevant references. We addressed this consideration by analyzing disciplinary subsets of the scientific literature, thereby restricting random selection of references to only those references in the disciplinary network being studied. In this subsection we demonstrate the effects of model misspecification or Uzzi et al.'s approach and the effectiveness of using disciplinary datasets in resolving the concomitant issues with the former.

Note also in Figure 2 that many z-scores values are significantly different between the WoS and Immunology datasets, although the density of points near the origin in Figure 2 indicates that some journal pairs change signs between the datasets while their magnitudes are not significantly different. While no metric is without downside, this observation points to a weakness of measuring novelty based simply on the sign of z-scores where no distinction is made between a small negative value and a large negative value, but a negative value of small magnitude is viewed as being significantly different than a small positive value.

The variation in z-scores with respect to the reference datasets causes the percentages of articles denoted as highly conventional and highly novel to be significantly different, as demonstrated in Figure 3. *Insert text about Figure 3.*

Novelty and Conventionality as Determinants of Impact in Disciplinary Contexts

Figure 4, Panels (a) and (b), compares hit rates for the four categories among the Immunology, Applied Physics, and WoS datasets for 1995: the hit rate is defined as the number of hit articles in each category divided by the number of articles in the category. We evaluated the statistical significance of the categorical hit rates using multiple methods, some of which we describe here. Our first test was based on the null hypotheses that hits were distributed randomly among the four categories with uniform probability in proportion to the number of articles in each category. Using a Chi-Square Goodness of Fit test, rejecting the null hypothesis in favor of the alternate hypothesis supports a non-uniform dispersion of hits: that is, some of the four categories are individually associated with higher than expected, or lower than expected hit rates. The null hypothesis was rejected at a $p < 0.001$ in all cases

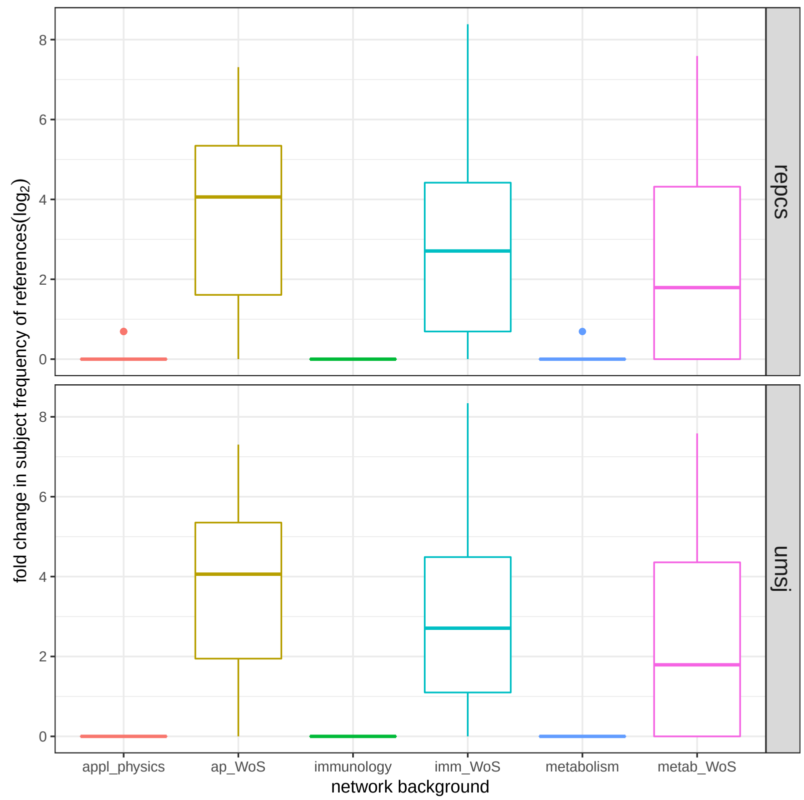

Figure 1. Intra-network citation shuffling preserves the disciplinary composition of references. Publications of type Article belonging to the three disciplinary networks (i) Applied Physics (ii) Immunology (iii) Metabolism were subject to a single shuffle of all their cited references using either the cited references in these networks as a source of random substitutions (bg_local) or references from all articles in WoS (bg_allWoS). Citation shuffling was performed using either our algorithm (repcs) or that of Uzzi (umsj). The disciplinary composition of cited references before and after shuffling was measured as frequencies for each of 153 sub-disciplines (from the extended subject classification in WoS) and expressed as fold difference between citation counts grouped by subject for original (o) and shuffled (s) references using the formula (fold_difference = $ifelse(o > s, o/s, s/o)$) and rounded to the nearest integer. A fold difference of 1 indicates that citation shuffling did not alter disciplinary composition. Data are shown for articles published in 1985. All eight boxplots are generated from 153 observations. Null values were set to 1. Note y-axis: log_2 scale.

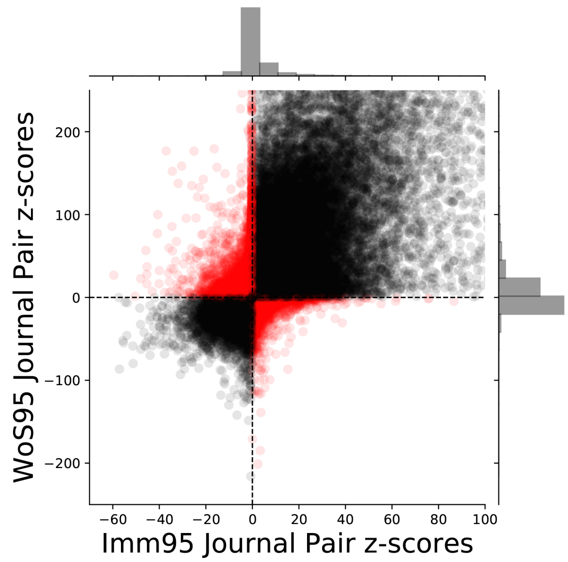

Figure 2. Journal pair z-scores vary with reference dataset. Scatter plot points (319,005) indicate journal pair z-scores for the 1995 Immunology dataset along the x-axis and the 1995 Web of Science dataset on the y-axis. Black indicates journal pairs whose z-scores have the same sign when computed for both reference datasets while red points indicate the 28.6% of journal pairs whose z-scores change sign across datasets. Regions with deeper hues indicate higher point densities and the histograms show the marginal distributions for each dataset separately.

in Figure 4, with the exception of the immunology and applied datasets where hit articles are designated as the top 1% of articles: valid tests were not possible in those instances due to too few expected hits. The null hypothesis was rejected with $p < 0.001$ for all valid tests for all parameter settings, all datasets, and all years: hypotheses tests were valid in 73 of 96 instances. We conclude that it is likely that the distribution of hits among categories is not uniform but, rather, hit rates vary among the categories in all datasets. (Should we insert a table with all statistical results?)

We computed hit rates for the WoS dataset, which mirrored Uzzi et al.'s results whereby the largest hit rates were for the HNHC category, despite our methodological improvement of sampling citations in proportion to their frequency. We found contrary results in the 1995 immunology dataset where articles in the LNHC category often had the greatest hit rates, as reflected in Table 4 and Figure 4. Across all year's data, all datasets, and all parameter settings, the highest hit rates in the immunology datasets were sometimes in the LNHC category and sometimes in the HNHC category. The metabolism hit rates reflected this same pattern. The greatest hit rates for the applied physics data were often in the HNLC category as reflected in Table 4 and Figure 4, and otherwise in the HNHC category. We conclude that Uzzi et al.'s finding of high hit rates in the HNHC category does not hold generally for disciplinary-based datasets and that novel citation patterns are not always indicative of impactful research, as was the case with immunology. Furthermore, the categories displaying the greatest hit rate vary with parameter settings and with the year. The lack of stable results across parameter settings suggests that parameters must be selected judiciously.

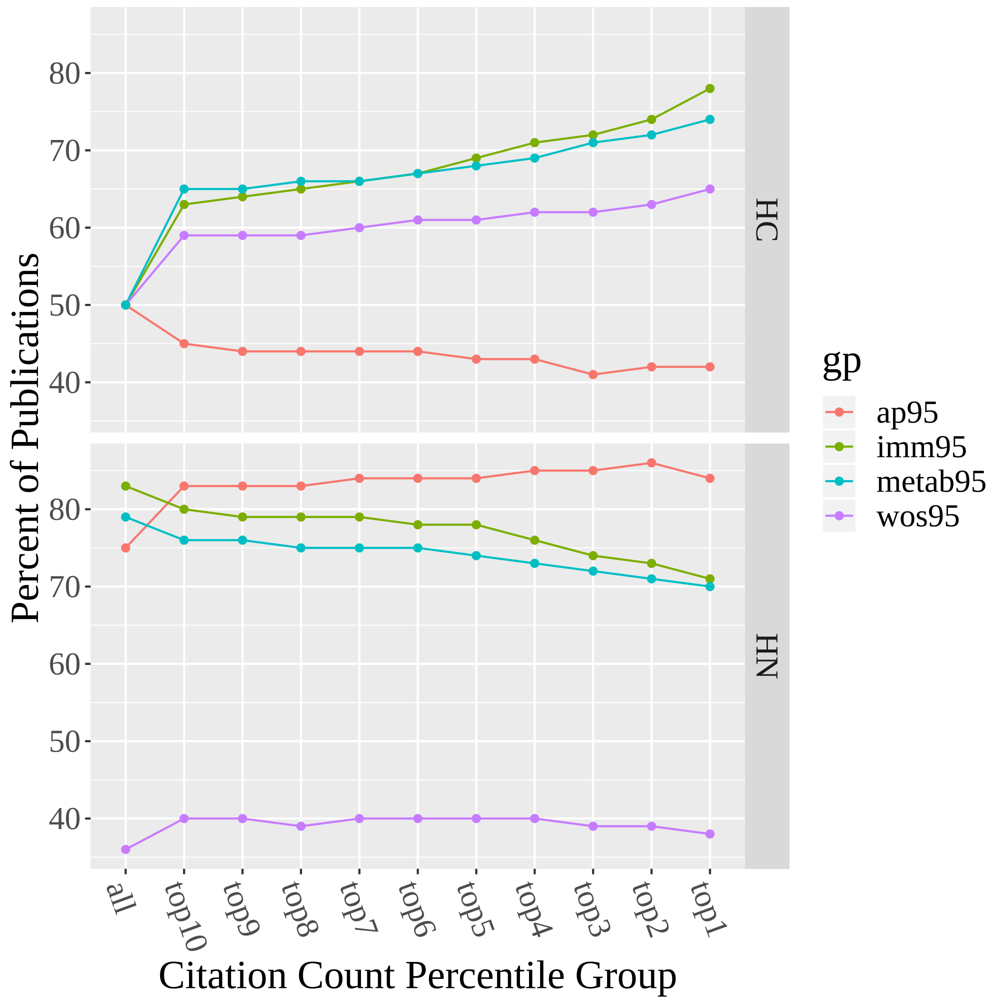

Figure 3. Effect of Research Discipline, Background Network, and Citation Count on Conventionality and Novelty. Data are shown for the applied physics (18,305), immunology (21,917), metabolism (97,405) and WoS (476,288) networks for 1995. The number of publications in each network is shown in parentheses. Citation counts shown are cumulative over the first 8 years since publication. X-axis: publications were classified into percentile groups based on citation counts (e.g., Top 1 indicates those publications in the top 1%). Y axis: The percent of applications in each group that are high conventionality (HC) and high novelty (HN). The z-scores are computed for each disciplinary network based on the selected background network; thus, *imm* denotes the immunology network with immunology z-scores and *imm_wos* denotes the immunology network with z-scores from WoS z-scores. The figure shows striking differences between the WOS network compared to the metabolism and immunology networks: across all networks, the percentage of high conventionality (HC) publications increases with citation counts, while the percentage of high novelty (HN) publications decreases with citation counts for the biological networks but not for WOS.

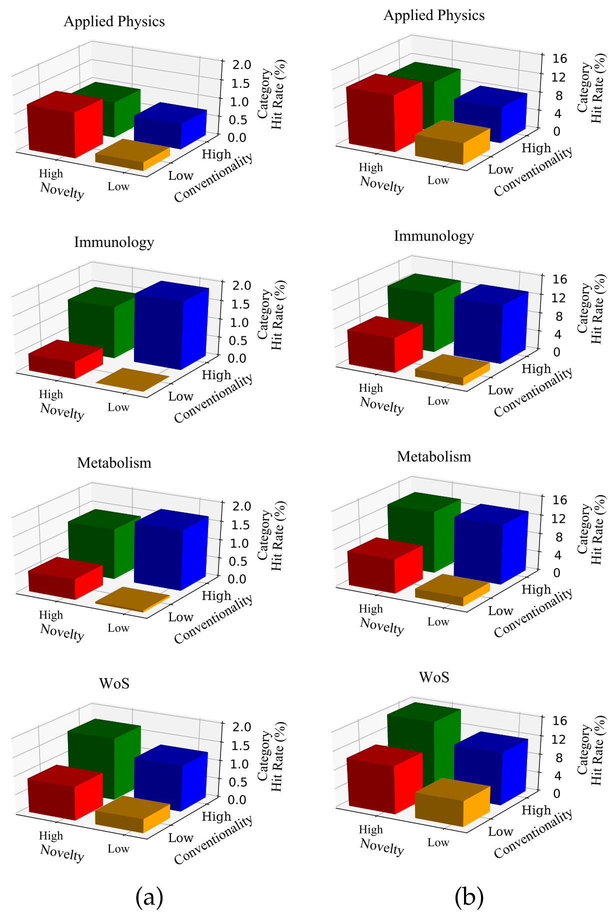

Figure 4. Effect of Context on Journal-Pair z-scores and Categorical Hit Rates: Immunology, Applied Physics, and WoS for 1995. Panels (a) and (b) show hit rates for the LNLC, LNHC, HNLC, and HNHC categories for the immunology, applied physics, and WoS datasets when hit articles are defined as the top 1% and top 10% of articles, respectively. Novelty in both panels is defined at the 10th percentile of articles' z-score distributions. The number of data points in the immunology, applied physics, and WoS data sets are 21,917, 18,305, and 476,288, respectively. The results for the WoS data set mirrored previous results from Uzzi et al. Uzzi et al. (2013) where the highest hit rate was for the HNHC category. The highest hit rates for the immunology dataset, in contrast, are in the LNHC category. The LNHC category often had the highest hit rate for the immunology dataset for various parameter settings with the HCHN category having the highest hit rate in other cases: metabolism results are similar. The applied physics dataset shows further contrast as the hit rate for it was highest in the HNLC category. The HNLC category demonstrated the highest hit rate often for applied physics: otherwise the highest hit rate was for HNHC articles.

Table 3. Measuring Model Misspecification. For the set of journal pairs in common between a disciplinary network and the full WoS dataset, Kullback-Leibler (K-L) divergences between empirical and simulated journal pair frequencies were computed for the years 1985, 1995, and 2005 for the three disciplinary datasets (applied_physics, immunology, and metabolism) using either the disciplinary network as background or the WoS superset (all_wos) to generate the null model (Background). K-L divergence was calculated using the R seewave package with a base (logarithm) of 2. The ratio between the K-L divergence for disciplinary networks versus the full WoS ranges from 1.96 to 2.77 and is greater than 2.0 for eight out of nine cases, strongly suggesting that simulations that constrain substitutions to the given disciplinary network better model the observed data.

	Disciplinary Network	Year	Background	K-L Divergence	Ratio
1	appl_physics	1985	appl_physics	1.21	
2		1985	all_wos	2.37	1.96
3		1995	appl_physics	0.86	
4		1995	all_wos	2.37	2.77
5		2005	appl_physics	0.95	
6		2005	all_wos	2.35	2.47
7	immunology	1985	immunology	0.75	
8		1985	all_wos	1.68	2.24
9		1995	immunology	0.78	
10		1995	all_wos	1.70	2.19
11		2005	immunology	0.73	
12		2005	all_wos	1.92	2.63
13	metabolism	1985	metabolism	1.11	
14		1985	all_wos	2.24	2.02
15		1995	metabolism	1.07	
16		1995	all_wos	2.33	2.17
17		2005	metabolism	1.19	
18		2005	all_wos	2.60	2.18

We also tested the explanatory power of each framework dimension by classifying articles as Low Novelty (LN) or High Novelty (HN) and, separately, as having Low Conventionality (LC) or High Conventionality (HC). We tested the null hypothesis that hits are distributed between LN and HN (LC and HC) in proportion to the total number of articles assigned to those categories. That null hypothesis was rejected for the WoS data along both dimensions. Consistent with previous analysis, hit articles were overrepresented in the HC category in every instance of WoS data at a $p < 0.001$ and hit articles were overrepresented in the HN category at a $p < 0.001$ in all but two cases. The p-values, in those cases, were 0.002 and 0.007. Hits in the immunology and metabolism data were overrepresented in the HC category with the same statistical significance as for WoS. The relationship of novelty with hits in the immunology and metabolism data differed dramatically from the WoS, however, with statistically significant findings of hit articles being sometimes overrepresented in the LN category, and sometimes being underrepresented. Of the 12 tests for applied physics, the statistical significance supporting a positive relationship between hit articles and HN were all $p < 0.10$, and 10 of 12 were $p < 0.05$. These tests also indicated strong support for the relationship between LN and hit articles in applied physics in a limited number of tests, with $p < 0.10$ in 5 of 12 instances and $p < 0.05$ in 3 of 12 instances. These results suggest that (1) both conventionality and novelty are strongly related to hits in the WoS, (2) the conventionality dimension is strongly related with hits in immunology

Table 4. Hit Rates by Category. The last four columns indicate the proportion of publications that are hits for each respective category.

Data Set	Hits as % of Articles	Novelty Percentile	LNLC	LNHC	HNLC	HNHC
Imm95	1%	10%	0.000	0.019	0.005	0.014
Imm95	10%	10%	0.017	0.128	0.076	0.129
Metab95	1%	10%	0.001	0.017	0.006	0.014
Metab95	10%	10%	0.019	0.130	0.074	0.133
AP95	1%	10%	0.002	0.007	0.012	0.010
AP95	10%	10%	0.047	0.079	0.123	0.109
WoS95	1%	10%	0.004	0.013	0.009	0.017
WoS95	10%	10%	0.056	0.115	0.104	0.156

and metabolism and novelty is not, (3) novelty is more strongly related with hits in applied physics than is conventionality. More generally, we find that the dimensions most strongly related with hit articles vary across disciplines and between disciplinary and broad data sets.

DISCUSSION

We conclude thathigh hit rates in the HNHC category does not hold generally for all datasets and that conventionality flavored with novelty is not always indicative of impactful research. The z-scores that change sign across datasets is either contradictory or an acceptable variation due to the difference in reference sets. We contend that it is the former because the inappropriate substitution of citations with those from disciplines that are implausible and because the observed citation frequencies are ignored. That fewer z-scores are negative in the WoS dataset relative to the immunology dataset may be directly due to the uniform sampling of references whereby many resulting journal pairs are never observed in the literature and so that the expected frequencies of those observed pairs have lower expected values, thus biasing their z-scores upward. In addition, treating citation as a set versus a multiset means that a random model will sample popular citations downward, thus further increasing their z-scores.*need confirmation from Jim here*.

Contrary to Uzzi et al. we found the category displaying the highest hit rate to be sensitive to the experimental parameter settings, which included the percentage of articles deemed to be hits and the percentile of articles' z-score distributions that delineated between articles of low novelty and high novelty. Faced with this lack of robustness, we are confronted with the necessity of defining defendable parameters, if possible, that determine novelty and conventionalty. More significantly, we should question whether the simple approach of defining novelty and conventionality as is done this article and other research is sufficient or whether we should expect such variation in the factors associated with impact as we have observed.

ACKNOWLEDGMENTS

We are grateful to Kevin Boyack and Dick Klavans for constructively critical 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 the bibliographic data analyzed in this study 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 GC, 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. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, *61*(12), 2389–2404. Retrieved 2019-01-05, from https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21419 doi: 10.1002/asi.21419

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* (1st ed.). The address: John Wiley and Sons, ISI Press. (An optional note)

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

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. Retrieved 2019-05-22, from https://projecteuclid.org/euclid.aoms/1177729694 doi: 10.1214/aoms/1177729694

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.

Perline, R. (2005). Strong, Weak and False Inverse Power Laws. *Statistical Science*, *20*(1), 68–88. Retrieved 2019-05-22, from https://www.jstor.org/stable/20061161

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

Stringer, M. J., Sales-Pardo, M., & Amaral, L. A. N. (2010). Statistical validation of a global model for the distribution of the ultimate number of citations accrued by papers published in a scientific journal. *Journal of the American Society for Information Science and Technology*, *61*(7), 1377–1385. Retrieved 2019-05-22, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3158611/ doi: 10.1002/asi.21335

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

Zuckerman, H. (2018). The sociology of science and the garfield effect: Happy accidents, unanticipated developments and unexploited potentials. *Frontiers in Research Metrics and Analytics*, *3*, 20. Retrieved from https://www.frontiersin.org/article/10.3389/frma.2018.00020 doi: 10.3389/frma.2018.00020