Supplementary Information for "Co-citations In Context"

James Bradley, Sitaram Devarakonda, Avon Davey, Siyu Liu, Dmitriy Korobskiy, Tandy Warnow, George Chacko

Corresponding author: George Chacko, E-mail: netelabs@nete.com

Supporting Information Text

This document provides supplemental data sets for an article that is under review. Where data sets are larger than a single page, a link to a Github site is provided. All Python, R, SQL, Spark, and Bash code are archived on our Github site. https://github.com/NETESOLUTIONS/ERNIE/tree/master/P2_studies

Descriptive Statistics and Statistical Tests. Our study considered a wide range of data sets and parameter settings in its analysis. Data sets from applied physics, immunology, metabolism, and the Web of Science for the years 1985, 1995, 2005 were analyzed. Following Uzzi et al. (1)), our analysis was carried out over a considerable range of parameters: hits were defined as the 1%, 2%, 5%, and 10% most-cited articles and novelty was defined at the 1st and 10th percentile of articles' citation journal-pair z-score distributions. Our main paper focused its comments on the circumstance where hits were defined as the 1% most-cited articles and novelty as defined by the 10th percentile of citation journal-pair z-scores because this is the most stringent definitions of these two classification dimensions: this most exclusive definitions, arguably, might be the most appropriate ones. This document reports, more comprehensively, on all our results.

The main paper documents a significant difference in the percentage of articles being classified as highly novel across data sets. A factor in the varying characterization of articles across reference data set is that the z-scores of journal pairs vary with the reference data set. Figure S1 shows, for example, that 28.6% of journal-pairs' z-scores differ in their sign depending on whether the reference data set is the 1995 Immunology data set or the 1995 Web of Science data set. Whether an article is deemed to be highly novel depends on the distribution of its journal pair z-scores and, specifically, on whether the 1st or 10th percentile is negative. Having such a significant percentage of journal pairs whose signs change across data sets can cause any particular article to be classified as highly novel for one data set but not in another.

The main paper qualitatively summarizes our observations of hit rates, which are defined for each of the four categories as the percentage of publications in the category that are hit articles (e.g., articles cited in the top 1% of all articles in the network): the hit rate for each category might vary significantly from the overall percentage of articles considered to be hits, as we demonstrated. The qualitative observations in the main paper are supplemented here in Table S1 and Table S2, which display our results in their entirety: these data are also shown graphically in Figures S2 through S5 for the semantically themed data sets (immunology, metabolism, and applied physics) as well as for the Web of Science, respectively. The main paper specifically references hit rates for the most stringent characterization of hits and novelty where hits are defined by the top 1% most-cited articles and novelty by the 10th percentile of z-score distribution for 1995 data sets. In that circumstance, articles in the WoS network had the highest hit rates in the HNHC category, but the top 1% of papers in the immunology network had the highest hit rates in the LNHC category in all but one case where it was virtually tied with the hit rate in the HNHC category. Results for metabolism are similar to immunology, but applied physics had the highest hit rate most often in the HNLC category. The category with the highest hit rate, therefore, varies with the context, that is, the local reference data set. As documented in Table S1 and Table S2, the categories demonstrating the highest hit rate remains consistent within each data set. These observations give rise to our comment in the main paper about the WoS and the themed networks having the highest hit rates in different categories, that the patterns persist over the years implying a systematic difference.

We evaluated the statistical significance of the categorical hit rates by testing the distribution of hit articles across the four categories using the null hypotheses that hits were distributed randomly across the four categories in proportion to the number of articles in each category. The alternate hypothesis is that the hit rates vary significantly among the categories. The null hypothesis was rejected at $p < 0.001$ using a Chi-Square Goodness of Fit test for the data referenced in the main paper in all but one case where a valid test was not possible for the immunology data due to insufficient data. Similarly, the null hypothesis was rejected with $p < 0.001$ in all valid tests and for all years as shown in the comprehensive results in Table S3. We conclude that it is unlikely that hits are distributed randomly among categories and that, instead, hit rates vary among the categories. The observed hit rate statistics demonstrate that the manner in which hit rates differ among the categories is consistent within each data set.

We also tested the explanatory power of each framework dimension individually by categorizing data, first, as LN or HN and, conversely, as LC or HC, with the null hypothesis that hits are distributed between LN and HN (LC and HC) in proportion to the number of articles in those categories. The results of those tests, as shown in Tables S4 and S5 for novelty defined at the 10th and 1st percentiles, respectively, confirmed that the association of novelty and conventionality with impactful research varies across data sets with both HN and HC having a strong association in the WoS data set, HC having the foremost association in the immunology and metabolism data, and HN having the foremost association for the applied physics data. The same trends hold across three different time points (1985, 1995, and 2005), showing that the differences between WoS, metabolism, and immunology are not transient.

Work Flow. Relevant folders at this level are

- cocitation_analysis- code for statistical analysis of MCMC data
- imm_metabolism- data sets generated by high level search terms
- workflow_pnas.001.jpeg- graphic of our principal MCMC workflow (jpeg)
- workflow_pnas.pdf- graphic of our principal MCMC workflow (pdf)
- workflow_png.001.png- graphic of our principal MCMC workflow (png)
- Permutation_Testing- assorted production and analytical scripts in SQL, R, Python, Bash

Discussion of algorithmic approach in Uzzi et al (2013) vs the one we used. The MCMC algorithmic approach in Uzzi et al (2013), DOI: 10.1126/science.1240474 for citation switching involves building three dicts containing publications, references, and year of publication information, and using them as lookup tables for various operations. In plain language, an iteration process selects publication in turn. Then each reference in said publication is replaced by a random selection from the *set* of eligible references published in the same year. If the potential replacement candidate is not the same as the reference to be replaced then a replacement is made. If it is the same, then up to 20 tries are made to find a non-self replacement. This process occurs for all the references in the set of publications being analyzed. Thus, reference a in three publications A,B,C could be replaced by references [b,c,d] or [b,b,d] but not [a,b,c]. Secondly, a reciprocal switch is made with a publication that cites the replacement. Thus, if publication A cites reference a published in year X then a is substituted with reference b also published in year X and and a randomly selected publication, say publication B that cites b, will have b replaced with a. See 'satyam_mukherjee_mcmc.py' kindly provided by the authors of this paper DOI: 10.1126/science.1240474.

Our approach is roughly similar. References are first grouped by year of publication and then the sample function in R is used on the *multi-set* of potential replacements to permute all references in a single step. A check is then run to see if the permutation process has created any duplicate references within each publication. Those publications with duplicate references are then deleted (typically <= 0.2%). See 'permute_script.R'.

A key difference is that the pool of replacement candidates is the *set* in one case and the *multi-set* in the other. Every substitution in the first approach is independent for instances of the same reference. Using the *multi-set* accounts for existing citation frequency when selecting possible replacements. Thus a publication in year X that has accumulated 10,000 citations is more likely to be selected than a publication that is cited only once. Reference a in publications A,B,C could be replaced by references [a,b,c]. This process is very fast in comparison and we have scaled it up even further by porting it to the Spark environment. In a recent comparison of publications in WoS in year 1985 (39,1860 pubs and 5,588,861 total references), ten simulations using the satyam_mukherjee_mcmc.py code took roughly 22 hours per simulation on a 32 Gb CentOS VM. Run times using our approach on comparable hardware amounted to roughly 30 seconds per shuffle plus an overhead of 3-4 hours to consolidate the data. We also performed 1,000 simulations in 60 hrs using a small Spark cluster for the much large 2005 data set (886,648 publications and 19,036,324 total references)

For input data we selected all publications of type 'Article' in WoS for a given year. Articles are then filtered to those that have at least two references in them. Further, only those references that have complete records in the Web of Science Core Collection are considered. This eliminates those that have cryptic references to other data sources or are just placeholders. Publications and references are mapped to their respective journals using ISSNs as identifiers. Where a reference has more than one ISSN, the most popular one is assigned to ensure that each reference is associated only with one journal.

For n <= 1000 simulations on disciplinary networks (immunology, metabolism, applied physics) permute_script.R is used to generate n files each with shuffled references. Typically, run times are less than 2 min per simulation on a 32 Gb CentOS box with 8 vCPUs in MS Azure. The permutation_testing_script.sh shell script is then run, which calls four Python scripts in turn.

1. observed_frequency.py: generates journal pair frequencies for the year slice of the WoS or disciplinary data set being analyzed.

2. background_frequency.py: generates journal pair frequencies for the background model implemented using permute_script.R

3. journal_count.py: joins all permuted files generated by background_frequency and calculates mean,std and z-scores.

4. Table_generator.py: final output file which contains all publications, reference pairs along with observed frequency and z-scores.

Thus, the workflow is (i) generate year slice (input data) (ii) generate background models by shuffling references (iii) calculate journal pair frequencies (iv) consolidate observed and simulated frequencies into a single table and calculate z-scores.

This process tends to slow down with large data set such as WoS in 2005 with 886,000 publications and 5.8 million journal pairs. Consequently, the entire process has been ported to Spark and provisioning a cluster, copying source data from the ERNIE PostgreSQL database over to Spark, conducting in-memory calculations, and copying a final table back to PostgreSQL has been automated (see Spark folder). Comparative performance data has been generated and will be posted soon.

Tables and Figures.

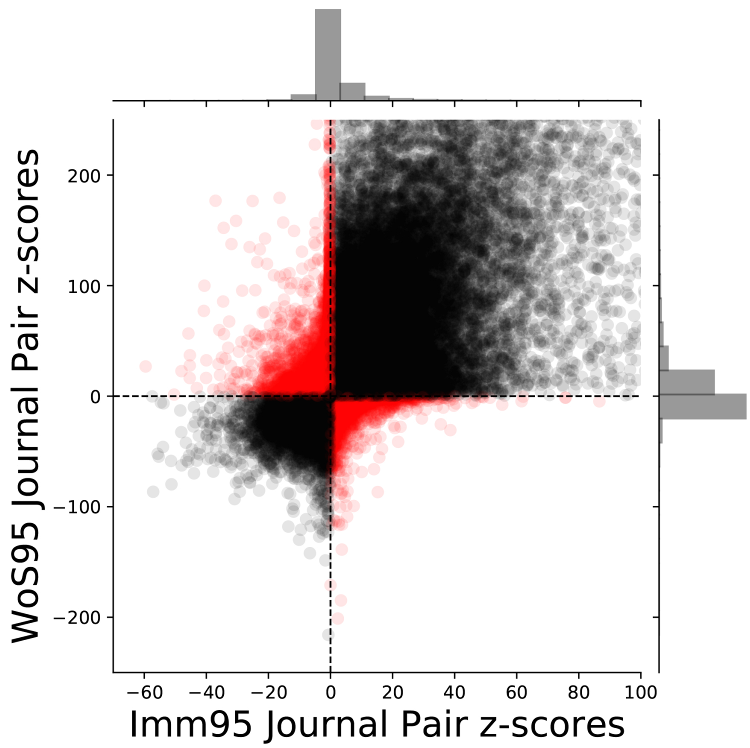


Fig. S1. Effect of Context on Journal-Pair z-scores. Journal pair z-scores (319,005) for the 1995 Immunology data set with two background networks (local network on the x-axis and the 1995 Web of Science 1995 on the y-axis). Black indicates journal pairs whose z-scores have the same sign for both background data sets and red points indicate the 28.6% of journal pairs whose z-scores change sign. Regions with deeper hues indicate higher point densities and histograms show the marginal data set distributions. The significant percentage of journal pairs that change sign can cause an article to be assessed as novel with respect to one background data set while it is not novel with regard to another data set.

Table S1. Hit Rates By Category: Novelty Defined at 10th Percentile

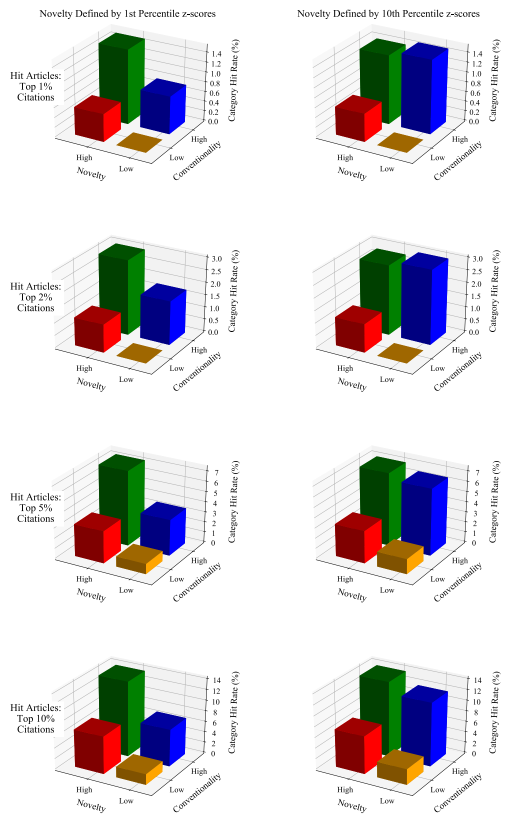
Data Set	Year	Highly Cited Min. Percentile	LNLC	LNHC	HNLC	HNHC
Immunology	1985	1%	0.0	**1.5**	0.6	**1.4**
Immunology	1985	2%	0.0	**3.0**	1.2	**2.8**
Immunology	1985	5%	1.5	6.5	3.2	**7.1**
Immunology	1985	10%	3.0	12.0	7.1	**14.0**
Immunology	1995	1%	0.0	**1.9**	0.5	1.4
Immunology	1995	2%	0.0	**3.4**	1.1	2.8
Immunology	1995	5%	0.6	**7.2**	3.2	6.8
Immunology	1995	10%	1.7	**12.8**	7.6	**12.9**
Immunology	2005	1%	0.0	**1.5**	0.6	**1.3**
Immunology	2005	2%	0.0	**2.7**	1.3	**2.7**
Immunology	2005	5%	0.0	6.0	3.6	**6.7**
Immunology	2005	10%	2.0	10.8	8.1	**12.9**
Metabolism	1985	1%	0.1	**1.5**	0.5	**1.5**
Metabolism	1985	2%	0.3	**2.7**	1.3	**2.9**
Metabolism	1985	5%	0.7	6.6	3.4	**7.0**
Metabolism	1985	10%	2.3	12.3	7.2	**13.8**
Metabolism	1995	1%	0.1	**1.7**	0.6	**1.4**
Metabolism	1995	2%	0.3	**3.1**	1.2	**2.8**
Metabolism	1995	5%	0.7	**7.0**	3.4	**6.7**
Metabolism	1995	10%	1.9	**13.0**	7.4	**13.3**
Metabolism	2005	1%	0.6	**1.3**	0.7	**1.3**
Metabolism	2005	2%	1.0	**2.5**	1.5	**2.6**
Metabolism	2005	5%	2.3	6.0	3.9	**6.6**
Metabolism	2005	10%	4.1	11.4	8.1	**12.6**
Applied Physics	1985	1%	0.0	0.5	**1.2**	**1.2**
Applied Physics	1985	2%	0.9	0.9	**2.5**	**2.4**
Applied Physics	1985	5%	2.8	3.0	5.5	**6.5**
Applied Physics	1985	10%	5.2	6.7	10.6	**13.2**
Applied Physics	1995	1%	0.2	0.7	**1.2**	**1.0**
Applied Physics	1995	2%	0.2	1.3	**2.5**	2.1
Applied Physics	1995	5%	0.9	3.4	**6.0**	5.2
Applied Physics	1995	10%	4.7	7.9	**12.3**	10.9
Applied Physics	2005	1%	0.8	0.6	**1.1**	**1.3**
Applied Physics	2005	2%	1.1	1.3	**2.1**	**2.3**
Applied Physics	2005	5%	1.6	3.3	5.4	**5.9**
Applied Physics	2005	10%	3.9	7.5	10.7	**11.4**
Web of Science	1985	1%	0.4	1.2	1.0	**1.6**
Web of Science	1985	2%	0.9	2.4	2.0	**3.3**
Web of Science	1985	5%	2.6	5.9	5.1	**8.4**
Web of Science	1985	10%	5.8	11.4	10.4	**15.8**
Web of Science	1995	1%	0.4	1.3	0.9	**1.7**
Web of Science	1995	2%	0.9	2.4	1.9	**3.3**
Web of Science	1995	5%	2.5	6.0	5.0	**8.0**
Web of Science	1995	10%	5.6	11.5	10.4	**15.6**
Web of Science	2005	1%	0.4	1.2	1.0	**1.7**
Web of Science	2005	2%	0.9	2.3	2.0	**3.4**
Web of Science	2005	5%	2.5	5.7	5.3	**8.1**
Web of Science	2005	10%	5.6	11.2	10.8	**15.0**

The hit rate is the percentage of publications in the referenced category that are in the top 1%, 2%, 5%, or 10% of papers according to citation count (see column 3) for novel articles defined as those with the 10th percentile z-score being negative. The z-scores are computed using the local network. The category with the highest percentile is boldfaced (the second highest is also boldfaced if within 0.3% and greater than the overall percentage of articles considered to be hits).

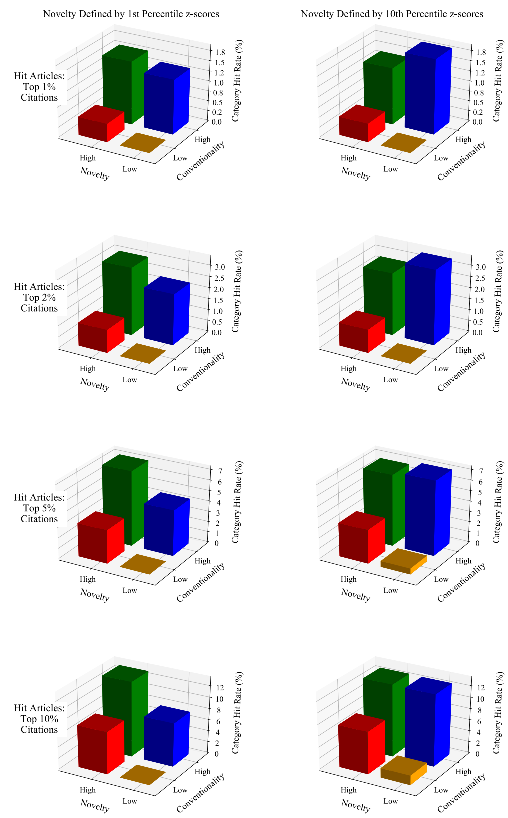
Table S2. Hit Rates By Category: Novelty Defined at 1st Percentile

Data Set	Year	Highly Cited Min. Percentile	LNLC	LNHC	HNLC	HNHC
Immunology	1985	1%	0.0	0.8	0.6	**1.5**
Immunology	1985	2%	0.0	1.8	1.1	**3.0**
Immunology	1985	5%	1.0	3.5	3.2	**7.3**
Immunology	1985	10%	2.0	7.0	7.1	**14.1**
Immunology	1995	1%	0.0	**1.4**	0.5	**1.6**
Immunology	1995	2%	0.0	2.3	1.0	**3.0**
Immunology	1995	5%	0.0	4.4	3.2	**7.2**
Immunology	1995	10%	0.0	7.8	7.5	**13.3**
Immunology	2005	1%	0.0	0.6	0.6	**1.5**
Immunology	2005	2%	0.0	1.1	1.3	**2.8**
Immunology	2005	5%	0.0	3.0	3.5	**6.8**
Immunology	2005	10%	1.9	6.4	8.0	**12.8**
Metabolism	1985	1%	0.0	1.0	0.5	**1.6**
Metabolism	1985	2%	0.0	2.0	1.2	**3.0**
Metabolism	1985	5%	0.0	4.6	3.4	**7.2**
Metabolism	1985	10%	0.6	8.6	7.1	**13.9**
Metabolism	1995	1%	0.0	1.1	0.5	**1.5**
Metabolism	1995	2%	0.2	1.9	1.1	**3.0**
Metabolism	1995	5%	0.2	4.4	3.3	**7.1**
Metabolism	1995	10%	0.4	8.6	7.3	**13.7**
Metabolism	2005	1%	0.2	0.9	0.7	**1.3**
Metabolism	2005	2%	0.7	1.9	1.5	**2.7**
Metabolism	2005	5%	2.2	4.3	3.8	**6.6**
Metabolism	2005	10%	3.3	8.0	8.0	**12.6**
Applied Physics	1985	1%	0.0	0.4	**1.2**	**1.1**
Applied Physics	1985	2%	1.2	0.5	**2.4**	**2.3**
Applied Physics	1985	5%	3.5	2.3	5.4	**6.1**
Applied Physics	1985	10%	5.8	5.5	10.5	**12.5**
Applied Physics	1995	1%	0.3	0.5	**1.2**	**1.0**
Applied Physics	1995	2%	0.3	1.1	**2.4**	2.0
Applied Physics	1995	5%	1.3	2.7	**5.9**	5.2
Applied Physics	1995	10%	4.3	6.7	**12.2**	10.8
Applied Physics	2005	1%	0.5	0.2	**1.1**	**1.2**
Applied Physics	2005	2%	0.8	0.4	**2.1**	**2.4**
Applied Physics	2005	5%	1.0	1.4	5.3	**5.9**
Applied Physics	2005	10%	2.8	4.1	10.6	**11.6**
Web of Science	1985	1%	0.1	0.7	0.9	**2.2**
Web of Science	1985	2%	0.3	1.5	1.9	**4.2**
Web of Science	1985	5%	1.0	3.9	4.9	**10.0**
Web of Science	1985	10%	2.6	8.1	10.1	**18.4**
Web of Science	1995	1%	0.1	0.8	0.8	**1.9**
Web of Science	1995	2%	0.3	1.6	1.7	**3.7**
Web of Science	1995	5%	1.0	4.0	4.7	**9.1**
Web of Science	1995	10%	2.4	8.0	9.9	**17.3**
Web of Science	2005	1%	0.2	0.8	0.9	**1.7**
Web of Science	2005	2%	0.4	1.6	1.9	**3.3**
Web of Science	2005	5%	1.2	3.9	4.9	**8.2**
Web of Science	2005	10%	2.8	7.9	10.1	**15.5**

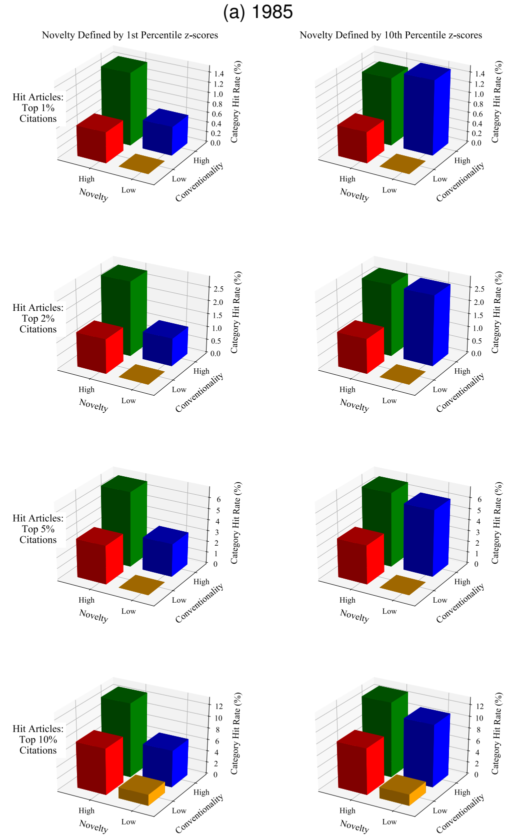
The hit rate is the percentage of publications in the referenced category that are in the top 1%, 2%, 5%, or 10% of papers according to citation count (see column 3) for novel articles defined as those with the 1st percentile z-score being negative. The z-scores are computed using the local network. The category with the highest percentile is boldfaced (the second highest is also boldfaced if within 0.3% and greater than the overall percentage of articles considered to be hits).



(a) 1985



(b) 1995



(c) 2005

Fig. S2. Categorical Hit Rates for the Immunology Data Sets for 1985, 1995, 2005. The immunology data sets contain 16,651, 21,917, and 28,063 data points for 1985, 1995, and 2005, respectively. Results show that for the more stringent definition of novelty (at the 10th percentile of journal-pair z-scores) that the highest hit rates in immunology are most often in the LNHC category.

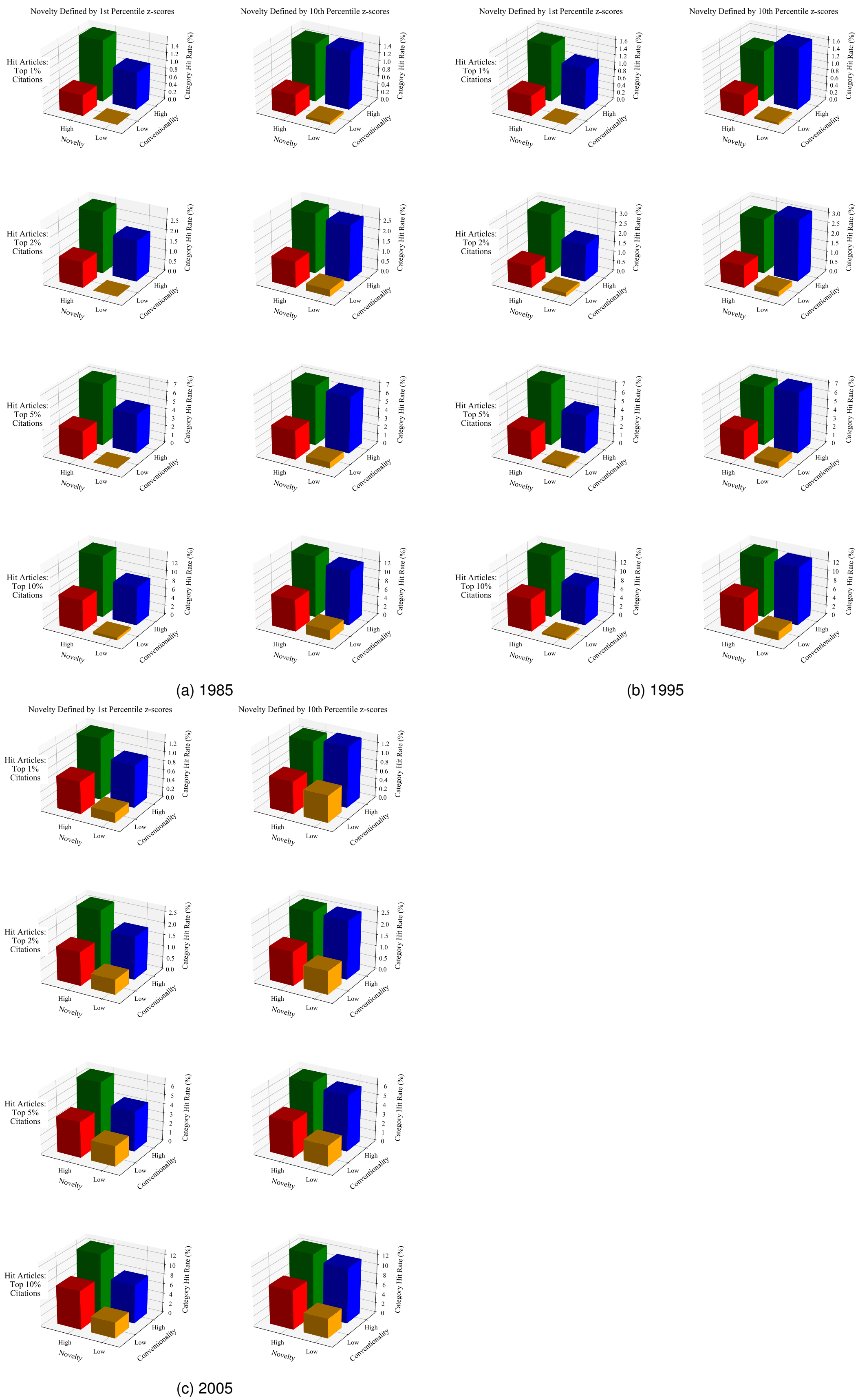


Fig. S3. Categorical Hit Rates for the Metabolism Data Sets for 1985, 1995, 2005. The number of data points are 62,992, 97,405, and 157,839 data points for 1985, 1995, and 2005, respectively. Results show that for the more stringent definition of novelty that the highest hit rates in immunology are most often in the LNHC category.

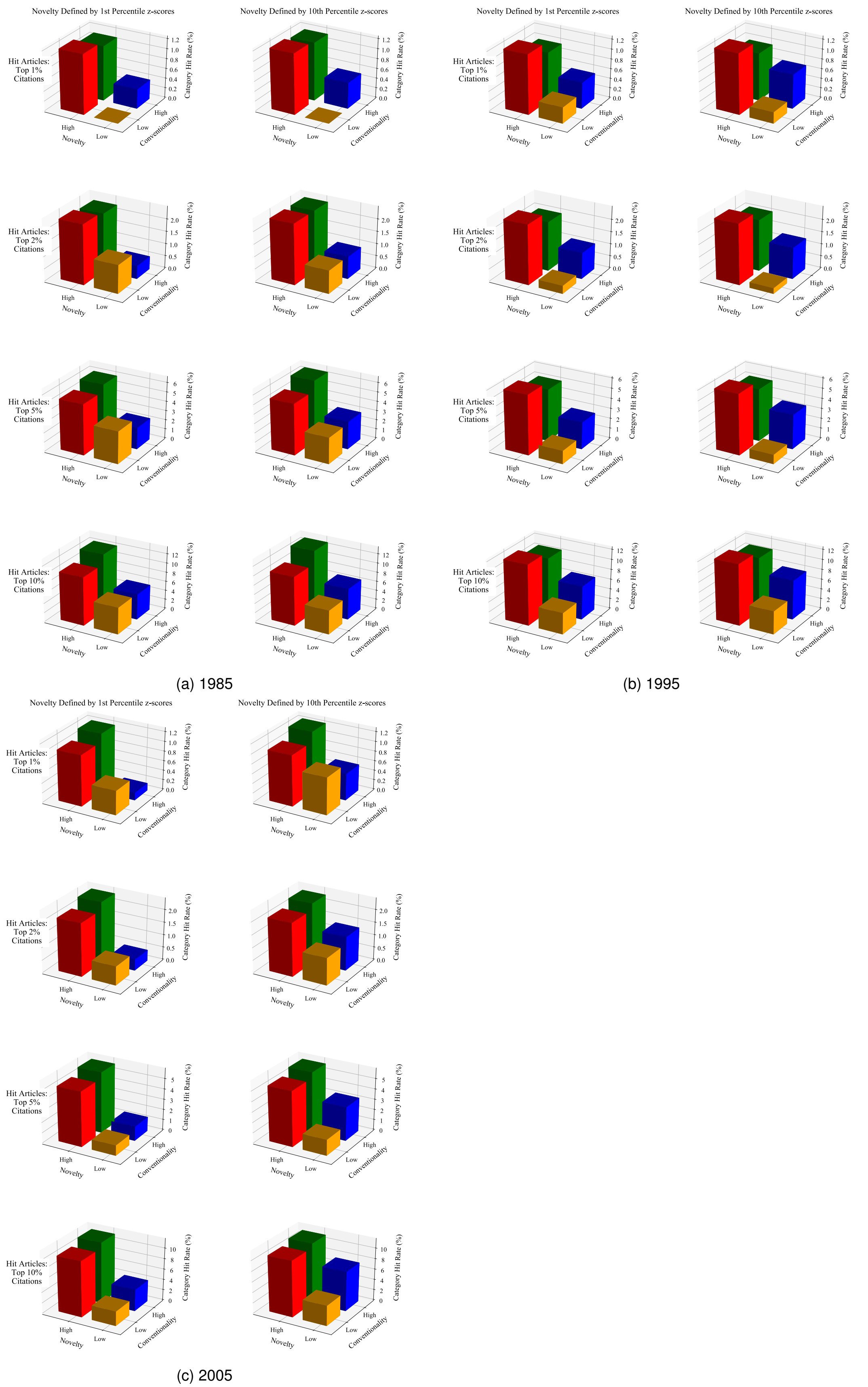
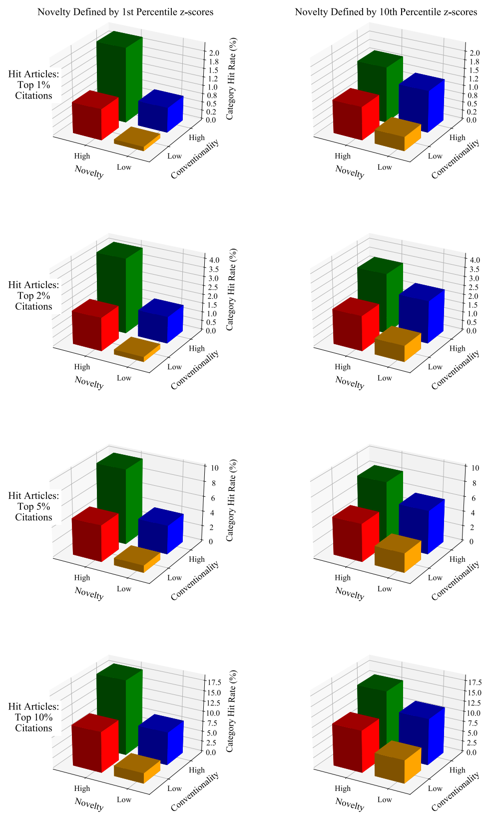
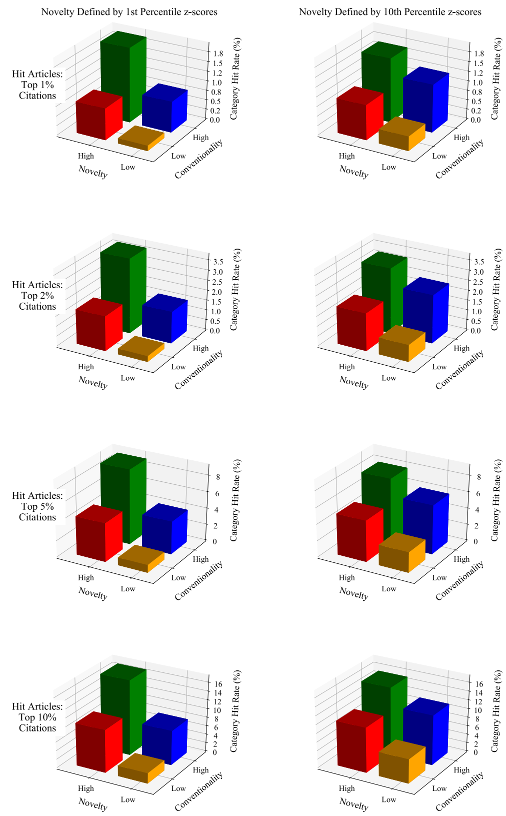


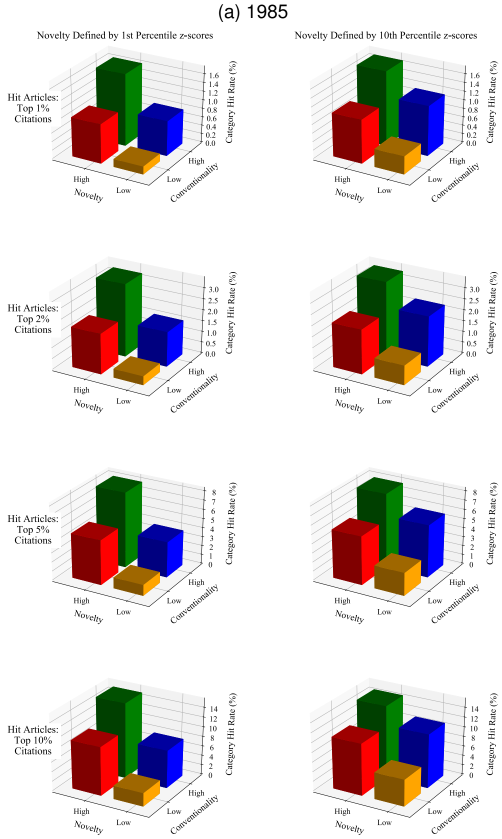
Fig. S4. Categorical Hit Rates for the Applied Physics Data Sets for 1985, 1995, 2005. The number of data points are 8,413, 18,305, and 32976 for 1985, 1995, and 2005, respectively. Results show the highest hit rates are often in the HNLC category.



(a) 1985



(b) 1995



(c) 2005

Fig. S5. Categorical Hit Rates for the Web of Science Data Sets for 1985, 1995, 2005. The number of data points are 327,756, 476,288, and 812,939 for 1985, 1995, and 2005, respectively. Results show the highest hit rates in the HNHC category consistent with Uzzi et al. (1).

Table S3. Statistical Significance of Deviation from a Random Distribution of Hits

Data Set	Year	Highly Cited Min. Percentile	p value Novelty Def.: 1%	p value Novelty Def.: 10%
Immunology	1985	1%	† 0.000	† 0.000
Immunology	1985	2%	† 0.000	† 0.000
Immunology	1985	5%	† 0.000	**0.000**
Immunology	1985	10%	**0.000**	**0.000**
Immunology	1995	1%	† 0.000	† 0.000
Immunology	1995	2%	† 0.000	† 0.000
Immunology	1995	5%	† 0.000	**0.000**
Immunology	1995	10%	**0.000**	**0.000**
Immunology	2005	1%	† 0.000	† 0.000
Immunology	2005	2%	† 0.000	† 0.000
Immunology	2005	5%	† 0.000	**0.000**
Immunology	2005	10%	**0.000**	**0.000**
Metabolism	1985	1%	**0.000**	**0.000**
Metabolism	1985	2%	**0.000**	**0.000**
Metabolism	1985	5%	**0.000**	**0.000**
Metabolism	1985	10%	**0.000**	**0.000**
Metabolism	1995	1%	**0.000**	**0.000**
Metabolism	1995	2%	**0.000**	**0.000**
Metabolism	1995	5%	**0.000**	**0.000**
Metabolism	1995	10%	**0.000**	**0.000**
Metabolism	2005	1%	† 0.000	**0.000**
Metabolism	2005	2%	**0.000**	**0.000**
Metabolism	2005	5%	**0.000**	**0.000**
Metabolism	2005	10%	**0.000**	**0.000**
Applied Physics	1985	1%	† 0.027	† 0.025
Applied Physics	1985	2%	† 0.000	† 0.000
Applied Physics	1985	5%	**0.000**	**0.000**
Applied Physics	1985	10%	**0.000**	**0.000**
Applied Physics	1995	1%	† 0.010	† 0.013
Applied Physics	1995	2%	**0.000**	**0.000**
Applied Physics	1995	5%	**0.000**	**0.000**
Applied Physics	1995	10%	**0.000**	**0.000**
Applied Physics	2005	1%	† 0.000	**0.000**
Applied Physics	2005	2%	**0.000**	**0.000**
Applied Physics	2005	5%	**0.000**	**0.000**
Applied Physics	2005	10%	**0.000**	**0.000**
Web of Science	1985	1%	**0.000**	**0.000**
Web of Science	1985	2%	**0.000**	**0.000**
Web of Science	1985	5%	**0.000**	**0.000**
Web of Science	1985	10%	**0.000**	**0.000**
Web of Science	1995	1%	**0.000**	**0.000**
Web of Science	1995	2%	**0.000**	**0.000**
Web of Science	1995	5%	**0.000**	**0.000**
Web of Science	1995	10%	**0.000**	**0.000**
Web of Science	2005	1%	**0.000**	**0.000**
Web of Science	2005	2%	**0.000**	**0.000**
Web of Science	2005	5%	**0.000**	**0.000**
Web of Science	2005	10%	**0.000**	**0.000**

These are hypothesis test data for the null hypothesis that hits are distributed among the categories randomly in proportion to the number of articles in each category using a Chi Square Goodness of Fit Test for novel articles defined as those with the 10th percentile z-score being negative and the 1st percentile z-score being negative. Rejecting the null hypothesis supports the alternate hypothesis that hit rates vary among the categories. The p values indicate the significance of the difference between the observed number of hits and the expected number of hits given by the random null model. † denotes that the Chi Square Goodness of Fit Test is not valid because the expected number of hits in at least one category was less than the required five hits. This was caused by the combination of a small number of articles in a category due to a low overall hit rate (1% or 2%) and a definition of novelty (1%) that resulted in few articles being defined as being of low novelty. Results that are significant at the 0.05 level are shown in bold font and those significant at the 0.10 level are shown in italics.

Table S4. Explanatory Power of Novelty and Conventionality: Novelty Defined at 10th Percentile

Data Set	Year	Highly Cited Min. Percentile	Cumulative Probabilities Conventionality Low	Conventionality High	Novelty Low	Novelty High
Immunology	1985	1%	1.000	**0.000**	**0.020**	0.866
Immunology	1985	2%	1.000	**0.000**	**0.002**	0.940
Immunology	1985	5%	1.000	**0.000**	**0.002**	0.924
Immunology	1985	10%	1.000	**0.000**	**0.017**	0.853
Immunology	1995	1%	1.000	**0.000**	**0.000**	0.985
Immunology	1995	2%	1.000	**0.000**	**0.000**	0.991
Immunology	1995	5%	1.000	**0.000**	**0.000**	0.988
Immunology	1995	10%	1.000	**0.000**	**0.000**	0.972
Immunology	2005	1%	1.000	**0.000**	**0.005**	0.882
Immunology	2005	2%	1.000	**0.000**	**0.007**	0.862
Immunology	2005	5%	1.000	**0.000**	**0.012**	0.837
Immunology	2005	10%	1.000	**0.000**	0.265	0.609
Metabolism	1985	1%	1.000	**0.000**	**0.000**	0.991
Metabolism	1985	2%	1.000	**0.000**	**0.000**	0.986
Metabolism	1985	5%	1.000	**0.000**	**0.000**	0.999
Metabolism	1985	10%	1.000	**0.000**	**0.000**	0.997
Metabolism	1995	1%	1.000	**0.000**	**0.000**	1.000
Metabolism	1995	2%	1.000	**0.000**	**0.000**	1.000
Metabolism	1995	5%	1.000	**0.000**	**0.000**	1.000
Metabolism	1995	10%	1.000	**0.000**	**0.000**	1.000
Metabolism	2005	1%	1.000	**0.000**	**0.000**	0.997
Metabolism	2005	2%	1.000	**0.000**	**0.000**	0.998
Metabolism	2005	5%	1.000	**0.000**	**0.000**	0.998
Metabolism	2005	10%	1.000	**0.000**	**0.000**	0.996
Applied Physics	1985	1%	0.177	0.860	0.998	0.071
Applied Physics	1985	2%	0.066	0.952	1.000	**0.013**
Applied Physics	1985	5%	0.200	0.818	1.000	**0.003**
Applied Physics	1985	10%	0.390	0.625	1.000	**0.000**
Applied Physics	1995	1%	0.062	0.955	0.996	0.090
Applied Physics	1995	2%	**0.018**	0.988	1.000	**0.011**
Applied Physics	1995	5%	**0.002**	0.999	1.000	**0.001**
Applied Physics	1995	10%	**0.000**	1.000	1.000	**0.000**
Applied Physics	2005	1%	0.319	0.706	1.000	*0.028*
Applied Physics	2005	2%	0.272	0.748	1.000	**0.015**
Applied Physics	2005	5%	0.117	0.897	1.000	**0.000**
Applied Physics	2005	10%	0.102	0.909	1.000	**0.000**
Web of Science	1985	1%	1.000	**0.000**	0.986	**0.002**
Web of Science	1985	2%	1.000	**0.000**	1.000	**0.000**
Web of Science	1985	5%	1.000	**0.000**	1.000	**0.000**
Web of Science	1985	10%	1.000	**0.000**	1.000	**0.000**
Web of Science	1995	1%	1.000	**0.000**	0.969	**0.007**
Web of Science	1995	2%	1.000	**0.000**	1.000	**0.000**
Web of Science	1995	5%	1.000	**0.000**	1.000	**0.000**
Web of Science	1995	10%	1.000	**0.000**	1.000	**0.000**
Web of Science	2005	1%	1.000	**0.000**	1.000	**0.000**
Web of Science	2005	2%	1.000	**0.000**	1.000	**0.000**
Web of Science	2005	5%	1.000	**0.000**	1.000	**0.000**
Web of Science	2005	10%	1.000	**0.000**	1.000	**0.000**

This table lists p-values in the form of cumulative right-hand tail probabilities for the observed number of hits in the Low Novelty, High Novelty, Low Conventionality, and High Conventionality categories under the sampling distribution generated by the null hypothesis of a random distribution of hit articles in proportion to the number of articles in each of the categories. A small p-value, therefore, indicates a number of hits that exceeds the expected number. Results that indicate statistically significant numbers of hits in excess of the expected number at the 0.05 level using a two-tailed test are highlighted in bold font, and those significant at the 0.10 level are italicized. These data are for the circumstances where novel citation patterns are defined by whether an article's 10th percentile z-score is negative.

Table S5. Explanatory Power of Novelty and Conventionality: Novelty Defined at 1st Percentile

Data Set	Year	Highly Cited Min. Percentile	Cumulative Probabilities: Conventionality Low	Conventionality High	Novelty Low	Novelty High
Immunology	1985	1%	1.000	**0.000**	0.874	0.414
Immunology	1985	2%	1.000	**0.000**	0.850	0.416
Immunology	1985	5%	1.000	**0.000**	0.997	0.267
Immunology	1985	10%	1.000	**0.000**	1.000	0.168
Immunology	1995	1%	1.000	**0.000**	0.240	0.583
Immunology	1995	2%	1.000	**0.000**	0.416	0.532
Immunology	1995	5%	1.000	**0.000**	0.939	0.379
Immunology	1995	10%	1.000	**0.000**	1.000	0.245
Immunology	2005	1%	1.000	**0.000**	0.961	0.391
Immunology	2005	2%	1.000	**0.000**	0.996	0.325
Immunology	2005	5%	1.000	**0.000**	1.000	0.260
Immunology	2005	10%	1.000	**0.000**	1.000	0.195
Metabolism	1985	1%	1.000	**0.000**	0.865	0.386
Metabolism	1985	2%	1.000	**0.000**	0.896	0.365
Metabolism	1985	5%	1.000	**0.000**	0.999	0.194
Metabolism	1985	10%	1.000	**0.000**	1.000	0.059
Metabolism	1995	1%	1.000	**0.000**	0.521	0.504
Metabolism	1995	2%	1.000	**0.000**	0.937	0.367
Metabolism	1995	5%	1.000	**0.000**	1.000	0.199
Metabolism	1995	10%	1.000	**0.000**	1.000	0.084
Metabolism	2005	1%	1.000	**0.000**	0.841	0.413
Metabolism	2005	2%	1.000	**0.000**	0.931	0.367
Metabolism	2005	5%	1.000	**0.000**	1.000	0.183
Metabolism	2005	10%	1.000	**0.000**	1.000	*0.042*
Applied Physics	1985	1%	0.177	0.860	0.999	0.105
Applied Physics	1985	2%	0.066	0.952	1.000	**0.023**
Applied Physics	1985	5%	0.200	0.818	1.000	**0.009**
Applied Physics	1985	10%	0.390	0.625	1.000	**0.001**
Applied Physics	1995	1%	0.062	0.955	0.999	0.113
Applied Physics	1995	2%	**0.018**	0.988	1.000	*0.036*
Applied Physics	1995	5%	**0.002**	0.999	1.000	**0.002**
Applied Physics	1995	10%	**0.000**	1.000	1.000	**0.000**
Applied Physics	2005	1%	0.319	0.706	1.000	**0.023**
Applied Physics	2005	2%	0.272	0.748	1.000	**0.004**
Applied Physics	2005	5%	0.117	0.897	1.000	**0.000**
Applied Physics	2005	10%	0.102	0.909	1.000	**0.000**
Web of Science	1985	1%	1.000	**0.000**	1.000	**0.000**
Web of Science	1985	2%	1.000	**0.000**	1.000	**0.000**
Web of Science	1985	5%	1.000	**0.000**	1.000	**0.000**
Web of Science	1985	10%	1.000	**0.000**	1.000	**0.000**
Web of Science	1995	1%	1.000	**0.000**	1.000	**0.000**
Web of Science	1995	2%	1.000	**0.000**	1.000	**0.000**
Web of Science	1995	5%	1.000	**0.000**	1.000	**0.000**
Web of Science	1995	10%	1.000	**0.000**	1.000	**0.000**
Web of Science	2005	1%	1.000	**0.000**	1.000	**0.000**
Web of Science	2005	2%	1.000	**0.000**	1.000	**0.000**
Web of Science	2005	5%	1.000	**0.000**	1.000	**0.000**
Web of Science	2005	10%	1.000	**0.000**	1.000	**0.000**

This table lists p-values in the form of cumulative right-hand tail probabilities where novel citation patterns are defined by whether an article's 1st percentile z-score is negative for the observed number of hits in the Low Novelty, High Novelty, Low Conventionality, and High Conventionality categories under the sampling distribution generated by the null hypothesis of a random distribution of hit articles in proportion to the number of articles in each of the categories. A small p-value, therefore, indicates a number of hits that exceeds the expected number. Results that indicate statistically significant numbers of hits in excess of the expected number at the 0.05 level using a two-tailed test are highlighted in bold font, and those significant at the 0.10 level are italicized. The z-scores are computed using the local network.

Table S6. Counts of Publications and References used in this study

	Year	unique publications	unique references	total references
1	1985	391860	2266584	5588861
2	1986	402309	2316451	5708796
3	1987	412936	2427347	5998513
4	1988	426001	2545647	6354917
5	1989	443144	2673092	6749319
6	1990	458768	2827517	7209413
7	1991	477712	2977784	7729776
8	1992	492181	3134109	8188940
9	1993	504488	3278102	8676583
10	1994	523660	3458072	9255748
11	1995	537160	3680616	9875421
12	1996	663110	4144581	11641286
13	1997	677077	4340733	12135104
14	1998	693531	4573584	12728629
15	1999	709827	4784024	13280828
16	2000	721926	5008842	13810746
17	2001	727816	5203078	14261189
18	2002	747287	5464045	15001390
19	2003	786284	5773756	16024652
20	2004	826834	6095594	17167347
21	2005	886648	6615824	19036324

These data were generated using the uz_ds.sql script in (2). Only publications of type Article and references for which complete records exist in the Web of Science Core Collection are included. The data set consists of 138.6 million unique references that are cited 398.9 million times by 19.7 million publications spanning 21 years (1985-2005)

Table S7. Kullback-Leibler Distances between simulated and observed frequencies.

	data set	D1	D2	D3
1	ap_2005	1.13	0.95	1.04
2	wos_2005 (ap)	1.80	2.35	2.08
3	imm_2005	0.82	0.73	0.78
4	wos_2005 (imm)	1.71	1.92	1.81
5	metab_2005	1.25	1.19	1.22
6	wos_2005 (metab)	2.12	2.60	2.36
7	ap_1995	0.89	0.86	0.87
8	wos_1995 (ap)	1.82	2.37	2.10
9	imm_1995	0.85	0.78	0.81
10	wos_1995 (imm)	1.56	1.70	1.63
11	metab_1995	1.10	1.07	1.22
12	wos_1995 (metab)	1.91	2.33	2.12
13	ap_1985	1.22	1.21	1.22
14	wos_1985 (ap)	1.96	2.37	2.17
15	imm_1985	0.82	0.75	0.79
16	wos_1985 (imm)	1.56	1.68	1.62
17	metab_1985	1.15	1.11	1.13
18	wos_1985 (metab)	1.91	2.24	2.08

The Kullback-Leibler (K-L) Divergence was calculated for simulated (mean value from 1000 simulations) and observed journal pair frequencies for the set of journal pairs common to a disciplinary network and the WoS network, e.g., ap_2005 and wos_2005 (ap). Journal pair frequencies were converted to a probability distribution and rhe *seewave* package in R and K-L_distance_1985.R, K-L_distance_1995.R, and K-L_distance_2005.R scripts were used. D1 and D2 are asymmetric distances, D3 is the symmetric K-L distance.

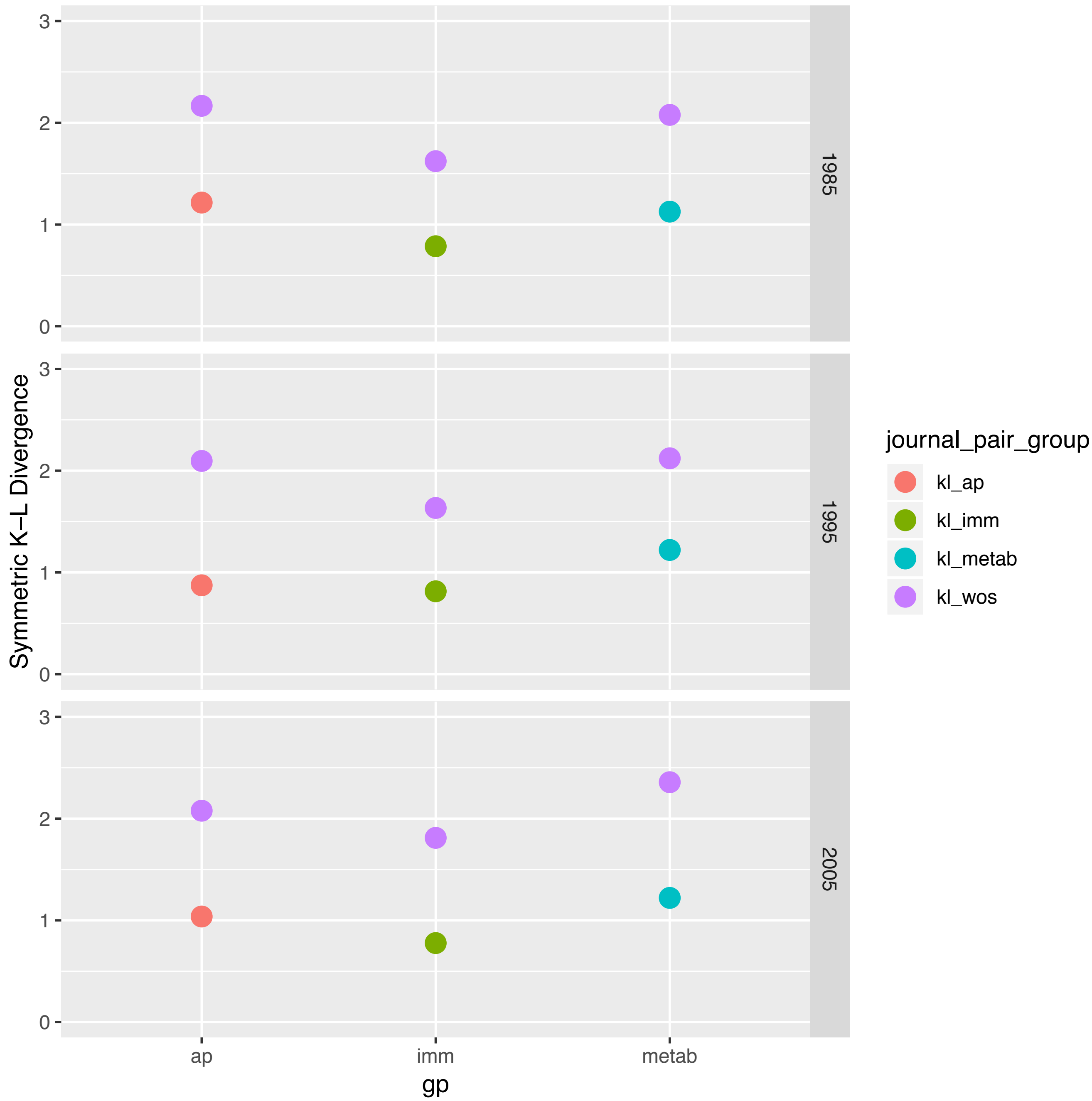


Fig. S6. K-L Divergence demonstrates improved model fits to observed journal pair frequencies for disciplinary networks compared to WoS. For the same journal pairs. the K-L divergence for the WoS network is consistently greater than that for the disciplinary network. from Table S3 are plotted for the years 1985, 1995, and 2005. kl_app, kl_imm, kl_metab, and kl_WoS refer to the symmetrical Kullback-Leibler divergence for mean simulated frequencies and observed frequencies in each of these data sets. The plot is faceted by year.

Table S8. Comparison of Citation Switching Algorithms

	data set	Wos 1985	WoS 1985	WoS 1985	WoS 1995	WoS 2005
1	Input publications	391,860	391,860	391,860	537,160	886,648
2	Input journals	8,075	8,075	8,075	10,983	15,203
3	Observed input journal pairs	1,277,349	1,277,349	1,277,349	2,373,226	5,847,432
4	Simulated journal pairs	961,487	959765	1,200,403	2,288,225	5,835,794
5	Journal pair coverage	75.27%	75.14%	93.97%	96.41%	99.80%
6	Min z-score	-132.71	-148.14	-104.50	-215.96	-273.708
7	Q1 z-score	-2.131	-2.151	-1.43	-1.49	-1.56
8	Median z-score	-0.536	-0.54	-0.24	-0.25	0.555
9	Q3 z-score	3.333	3.365	4.29	4.15	2.423
10	Max z-score	16598.534	22015.891	12,028.55	12,662.15	6,152.57
11	Environment	CentOS 7.4	Spark 2.3	Spark 2.3	Spark 2.3	Spark 2.3
12	Number of simulations	10	10	1000	1000	1000
13	Run time	2186h (22 hr /sim)	$<$ 1 hr	$<$ 50h	50h	60h
14	Algorithm	(1)	(2)	(2)	(2)	(2)

The citation switching algorithm of Uzzi et al. (2013) (1) has been implemented in Python. Ten simulations of the WoS 1985 data set were executed on a 32 Gb, 8 vCPU CentOS 7.4 virtual machine to generate the data in Col 2. Each simulation took roughly 22 hours to complete. Scaling up the experiment, 10 or 1000 simulations of our modifications (2) of this algorithm were executed on a 4-node Apache Spark cluster for the WoS 1985 data set. 1000 simulations completed in less than 50 hours (roughly 3 minute/simulation). These simulations also produced greater journal pair coverage. Performance was compared to 10 or 1000 simulations of our modifications of this algorithm (2) on a 9 node Spark cluster. 1000 simulations completed in less than 50 hours with greater journal pair coverage.

Table S9. Profile of Disciplinary Networks

	data set	Input Publications	Journal Pairs	Min	Q1	Median	Q3	Max
1	ap1985	10298	34,267	-23.05	-0.94	-0.21	3.03	1490.42
2	ap1995	21012	60,340	-45.36	-0.97	-0.24	2.86	646.03
3	ap2005	35600	199,928	-47.76	-0.80	-0.20	3.53	2158.47
4	imm85	17942	159,107	-48.33	-1.09	-0.27	2.49	934.63
5	imm95	22759	319,855	-59.56	-1.10	-0.28	2.37	1507.61
6	imm2005	28539	751,950	-74.54	-0.99	-0.30	1.84	2560.51
7	metab1985	67342	431,993	-97.00	-1.46	-0.34	2.34	4193.49
8	metab1995	100350	865,406	-132.85	-1.56	-0.37	2.16	3998.44
9	metab2005	159910	2,349,005	-127.81	-1.60	-0.41	1.83	3472.77

Data shown represent the results of 1000 simulations for the applied physics (ap), immunology (imm), and metabolism (metab disciplinary network. Summary statistics for z-scores are provided as well as the number of publications in each data set that were the input to the simulation process.

Table S10. Comparison of top 5% cited publications vs all publications in applied physics (ap) immunology (imm), metabolism (metab), and WoS data sets

	year	category	ap_5	ap_all	imm_5	imm_all	metab_5	metab_all	wos_5	wos_all
1	1985	HCHN	34	25	29	33	24	29	6	6
2	1985	HCLN	15	25	21	17	26	21	44	44
3	1985	LCHN	50	47	48	49	48	48	32	29
4	1985	LCLN	1	3	2	1	2	2	18	21
5	1995	HCHN	35	27	29	34	26	31	6	7
6	1995	HCLN	15	23	21	16	24	19	44	43
7	1995	LCHN	50	48	48	49	48	48	33	29
8	1995	LCLN	0	2	2	1	2	2	17	21
9	2005	HCHN	31	29	36	36	30	30	8	7
10	2005	HCLN	19	20	14	14	20	20	42	43
11	2005	LCHN	48	49	50	49	49	49	30	27
12	2005	LCLN	2	2	0	1	1	1	20	23

Numbers shown are percent of publications in each group. Data are shown for reference years 1985, 1995, and 2005

References

1. Uzzi B, Mukherjee S, Stringer M, Jones B (2013) Atypical combinations and scientific impact. *Science (New York, N.Y.)* 342(6157):468–472.
2. Korobskiy D, Davey A, Liu S, Devarakonda S, Chacko G (2019) Enhanced Research Network Informatics Environment (ERNIE), (NET ESolutions Corporation), Github repository. https://github.com/NETESOLUTIONS/ERNIE/tree/master/P2_studies/Permutation_Testing.