https://doi.org/10.51408/issi2025_025

Knowledge Combination and Research Impact: A Comparison of Sources and Keywords Co-Citation

Hsiao Tsung-Ming[1], Tang Muh-Chyun[2]

[1]*162228@mail.tku.edu.tw*
Department of Information and Library Science, Tamkang University (Taiwan)

[2]*mctang@ntu.edu.tw*
Department of Library and Information Science, National Taiwan University (Taiwan)

Abstract

Combining knowledge from diverse origins has long been recognized as a key driver of innovation. Although many studies have examined how such combinations influence research impact, their findings remain inconsistent. One potential reason is the use of different units of measurement for novelty and conventionality. Using data from the DBLP Citation Network, this study compares two approaches—one based on sources and the other on keywords co-cited frequency—to measure research novelty. We found a low correlation between these two measures, suggesting that each captures distinct aspects of novelty. In line with Uzzi et al. (2013), it was found that papers exhibiting both high novelty and high conventionality (HNHC) are more likely to achieve high citation impact, especially when novelty is measured at the source level. Logit regression indicates that source-based HNHC is a strong predictor of highly-cited "hit" papers, though the keyword-based measure also contributes a smaller but statistically significant effect. These results highlight the importance of carefully selecting units of analysis when investigating the relationships between novelty, conventionality, and research impact.

Introduction

The synthesis of heterogeneous knowledge has long been recognized as a key driver of innovation. However, recent studies suggest that achieving high-impact research often requires balancing high novelty with strong conventionality (Uzzi et al., 2013). This is an intriguing development as previous studies of the relationships between novelty and research impact had often overlooked the need to situate novelty within conventional wisdom. Methodologically, research novelty is frequently measured by the rarity or unexpectedness of knowledge combined in a paper. A paper is considered novel if it synthesizes knowledge units that appear for the first time or occur rarely. Two types of knowledge units have been proposed to measure novelty: one based on the journals cited and the other on the keywords or subject headings used to index the paper. However, little research has examined the consistency of the novelty assessments produced by these two approaches. To address this gap, the present study compared novelty and conventionality measurements derived from source co-citation (journals) and keyword co-citation using DBLP, a large citation network dataset in the field of computer science. Additionally, the study evaluated how effectively combinations of novelty and conventionality, as measured by each approach, can identify highly cited papers. citation. A novel aspect of this research is the use of keyword co-citation, rather than keyword co-occurrence, as an indicator of a paper's novelty. Our findings reveal a slight correlation between journal-based

and keyword-based co-citation measures of novelty, suggesting they are capturing different aspects of novelty. Interestingly, the combination of high novelty and conventionality was associated with greater odds of producing a hit paper when cited journal is used as the basic knowledge unit.

Literature Review

Creating innovative ideas relies on combining knowledge from various sources, this is especially so when the combination is novel. In the last decades, various quantitative indicators have been proposed to measure novelty based on how rare and different the combination of knowledge unit is (e.g. Bornmann et al., 2019; Carayol et al., 2019;). While investigating the relationship between exploring new possibilities and exploiting established certainties in organizational learning, March (1991) found that the process of innovation relates to refining the existing technical combinations and creating new technical combinations. The combinatorial conjecture, "all creativity results from combinations of mental representations", was also evaluated by Thagard (2012) with great scientific discoveries and technological inventions. After surveying 200 invention examples, he supported this conjecture. As claimed by Kaplan and Vakili (2015), novel ideas require both a broad search for information and a process of recombining diverse knowledge. The recombination process plays a critical role in enhancing novelty and producing breakthrough-class papers (Bornmann et al., 2019).

Atypical and Conventional Combination

While novelty is often considered a necessary condition for innovation, it has been pointed out that novelty alone is not enough to drive impact. Uzzi and his colleagues (2013) argued that "balancing atypical knowledge with conventional knowledge may be critical to the link between innovativeness and impact" (p. 468). To determine the degree of how atypical or conventional an article's knowledge combinations are, they examined the sources of knowledge, namely the journals listed in its bibliography. Specifically, they built the journal co-citation networks by year with Web of Science (WoS) data and aggregated the frequency of journal pairing, two journals co-cited by articles. The atypical or conventional level of a combination was determined by comparing its observed frequency with the expected frequency, derived from a randomized simulation network that retains key features of its corresponding journal co-citation network. A journal pair was classified as an atypical combination if its observed frequency lowers than the expected frequency. Conversely, if a journal pair occurs frequently than expected, it is considered as conventional combination. The observed frequency of the journal pair was converted to a z-score to facilitate comparison.

For each paper, two summary statistics, novel and conventional values, were derived from the rank-ordering of the z-scores of all its journal pairings. As depicted in Figure 1, the novelty, left tail, was defined as the 10th percentile z-score, and the conventionality was defined as the median z-score. Novelty serves as a criterion for classifying papers into high or low novelty, and conventionality can be applied in a similar manner. Hence, papers can be categorized into one of four quadrants based

on their conventionality and novelty. Uzzi and his colleagues (2013) analyzed 17.9 million articles and 302 million articles references across all WoS disciplines from 1950 to 2000 and showed that articles properly balancing high levels of both novelty and conventionality have the highest potential of becoming high-impact publications. Based on the results, they argued that effectively embedding novel ideas into established traditions is the key drivers of scientific advancement.

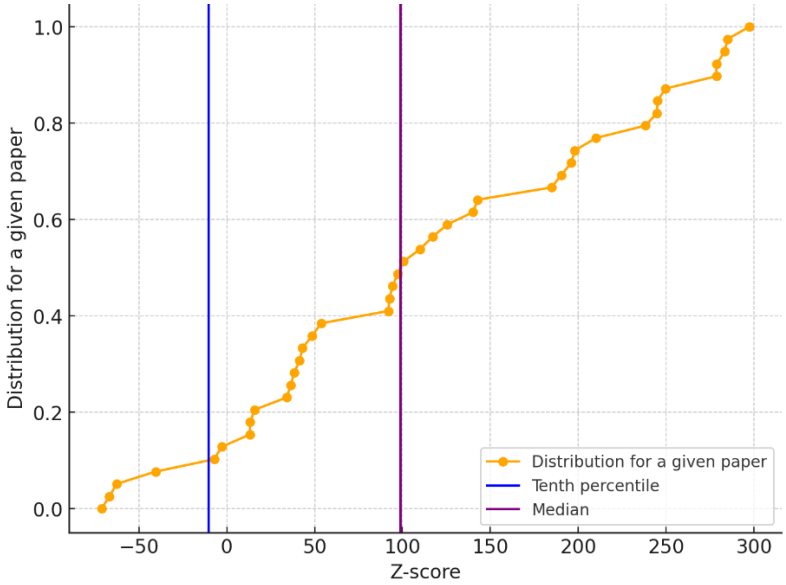


Figure 1. The distribution of z-scores of an article's journal pairings and how the novelty and conventionality of this article are determined. Redraw based on Uzzi et al. (2013).

Source-based Approach

The method proposed by Uzzi et al. (2013) is based on how the sources of references are co-cited by papers. This source-based approach is applied by several studies, and one of them is Boyack & Klavans (2014). They used Scopus data of articles published from 2001 to 2010 to replicate the findings of Uzzi et al. (2013) and further explore the disciplinary effects on the relations between paper's atypical and conventional combinations and its probability of being highly-cited papers. Instead of deciding the expected frequency of the source combination with simulated citation network, K50 was used to determine the novel/conventional degrees of a journal pair was determined by K50, a method examined by Klavans & Boyack (2004). While affirming the findings of Uzzi and his colleagues, Boyack & Klavans (2014) highlighted the potential mediating effects of disciplines and publication venues. According to their findings, the relationship was less evident when identifying the top 5 percent of highly-cited papers within individual disciplines, compared to findings across all disciplines (Uzzi et al., 2013). Further investigation of the top 20 highly-cited journals revealed that leading physics journal typically exhibited high conventionality and low novelty. Conversely, top biomedical journals combined

high novelty with high conventionality. Meanwhile, multidisciplinary journals like *Nature* and *Science* exhibited high novelty but low conventionality.

Source-based approach is used by two later studies, Lee et al. (2015) and Wang et al. (2017). Lee et al. (2015) analyzed 9,428 WoS-indexed publications, covering publication years 2001 to 2006, to examine how team size, field diversity, and task diversity influence creativity. Building on Uzzi et al. (2013), they defined the commonness of a journal pairing in a specific year as the fraction of its observed frequency to its expected frequency. In their research, the expected frequency is calculated as the total number of all journal pairings multiplied by the joint probability of the co-occurrence of the two journals. Instead of using co-occurrence of the two journals directly, Wang et al. (2017) proposed that the similarity between two journals can be defined as the cosine similarity of their corresponding row vectors, extracted from the journal co-citation matrix, as shown in Figure 2. The higher the similarity of a journal pairing, the lower its novelty. Hence, a paper's novelty is measured by the sum of the differences in all its journal pairings.

	J_1	J_2	J_3	J_4	J_5	…
J_1	/	0	3	0	5	…
J_2	0	/	6	2	3	…
J_3	3	6	/	5	4	…
J_4	0	2	5	/	0	…
J_5	5	3	4	0	/	…
…	…	…	…	…	…	/

Figure 2. The journal co-citation matrix. The number in each cell represents the frequency that two journals are co-cited. Redraw based on Wang et al. (2017).

The source-based approaches proposed by studies reviewed above are examined by Bornmann et al. (2019) and Fontana et al. (2020). Bornmann et al. (2019) used the human recommendations of papers from F100Prime, a post-publication peer review system, to evaluate the validity of two novelty metrics proposed by Lee et al. (2015) and Wang et al. (2017), referred to as novelty score U and novelty score W, respectively. While novelty U followed the unexpected combinations in Uzzi et al. (2013), novelty W counted the number of novel combinations in the references. In addition, they introduced a novelty score K determined by comparing the new keywords to the existing ones in a specific subject category. According to their research findings, novelty score U agreed with their formulated expectations mostly, while novelty score W lacked convergent validity with the FMs' assessments. The logistic regression result revealed that as novelty score K increased, the likelihood of an article being included in F1000 Prime decreased. In another research, Fontana et al. (2020) analyzed the novelty indicators proposed by Uzzi et al. (2013) and Wang et al. (2017) with 230,854 articles published on 8 journals of the American Physical Society. Notably, they used the domain-specific subject classification, instead of journal, as the basic unit based on which the novelty and interdisciplinarity measures were calculated. Infrequent co-occurrences of subject headings were considered

more novel. They showed that novelty score W lacks ability to tell novel and non-novel articles and that novelty score U correlated well with interdisciplinarity.
In summary, a series of studies measure the novelty of a publication based on its references. These studies propose that the novelty level of an article depends on the rarity of its co-citation relationships. The studies reviewed above consider co-citation relationships at the level of sources where the references are published. Instead of focusing on sources, some studies explore how research articles combine topics. The following section will review these related studies.

Topic-based Approach

An alternative approach to defining novelty is by analyzing how articles combine topics. The novelty score K used by Bornmann et al. (2019) is based on comparing the new keywords to existing ones in a specific domain. Besides keywords, some studies use controlled vocabularies like chemical annotations or MeSH to assess novelty. Foster et al. (2015) proposed that "five strategies available to a scientist facing a network of known scientific relationship: *jump*, *new consolidation*, *new bridge*, *repeat consolidation*, or *repeat bridge*" (p. 881). Figure 3 illustrates the five strategies, and these strategies were further divided into two classes: innovation (*jump*, *new consolidation*, and *new bridge*) and tradition (*repeat consolidation* and *repeat*). Traditional strategies involve scientists delving deeper into established knowledge entities and relationships, while innovative strategies involve introducing novel ones. Their study used chemical annotations, extracted from abstracts in the MEDLINE collection, as nodes in the knowledge network, with edges representing the co-occurrence of chemical entities within an abstract. According to their findings, while innovative work has higher impact potential, its rewards do not compensate for the risk of non-publication.

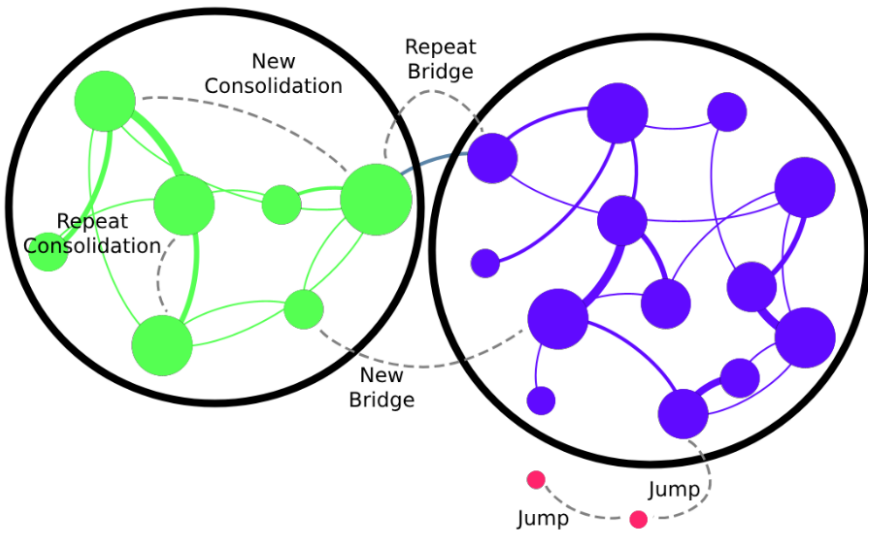


Figure 3. The five strategies in a knowledge network. Bridge connects knowledge entities of two domains, consolidation links knowledge entities with the same domain. Redraw based on Foster et al. (2015).

Instead of using chemical entities, Boudreau et al. (2016) and Ruan et al. (2023) utilized MeSH terms to measure the novelty. To measure novelty, Boudreau et al. (2016) utilized MeSH term combinations and analyzed their appearance in the entire existing related literature. They proposed that novelty was determined by the proportion of term combinations in a given proposal that had not appeared before. The novelty was expressed as a percentile, ranging from 1% (least novel) to 100% (most novel). Ruan et al. (2023) also utilized MeSH term as the unit to determine topic combination novelty. Their design followed Uzzi et al. (2013) and Lee et al. (2015) and adopted "the proportion of the observed and expected frequency of a combination of MeSH terms to denote the *commonness* of the MeSH pair" (p. 5). Carayol et al. (2019) proposed a novelty measure based on the pairwise author keyword co-occurrence in papers indexed in Web of Science in a given year and field. The less common a pair of keywords cooccur, the higher its novelty. It was found that higher novelty was more likely to be observed in larger teams, especially those spanning across institutional and geographic boundaries. Importantly, the correlation between novelty measured through pairwise keyword co-occurrence and journal co-citation was found to be small. Furthermore, pairwise keyword novelty was positively associated with higher citation counts within a three-year citation window, and papers with high novelty had greater odds of becoming "hit papers."

Our review showed increasing efforts in measuring the novelty of a paper, and exploring the relationship between novelty and impact. It is still unclear that, whether novelty along, or the combination of both novelty and conventionality is more conducive to higher impact. Furthermore, as different studies used different knowledge unit for the base of combination, it is difficult to assess how consistent the results are. The potential inconsistence between these using source vs. topic as the base of measuring knowledge combination poses a great challenge to clarify the relationship between knowledge combination and research impact. The purpose of this study is therefore to compare two types of methods, the source-based and topic-based approaches, in measuring an article's novelty and conventionality, and to explore the relationships between the resulting novelty/convention combination and research impact. Specifically, our research questions are as follows:

1. When determining the novelty and conventionality of a given paper, do the source-based and topic-based approaches produce consistent results? This can be further tested by:

a. Are the novel/conventional rank-orders revealed by the two approaches aligned with one another?"

b. When categorizing a paper as high or low in novelty/conventionality, are the classifications from the two approaches consistent?

2. Following Uzzi et al. (2013), do papers integrating high novelty and conventionality (HNHC) resulting in higher odds of being highly cited? And if so,

3. When identifying HNHC, which approach (source vs. keyword-based) yields a higher probability of highly cited papers?

4. Does combining the two approaches offer greater advantage in revealing the relationship between HNHC and high impact?

Research Design

This study uses the DBLP dataset, which comprises over 7 million research articles in computer science. The version, DBLP-Citation-network v13, utilized in this study is maintained by Tang et al. (2008) and has been widely used in various types of research, such as developing recommendation systems (Huang et al., 2024; Kanwal & Amjad, 2024) and predicting scholar impact (Zhang & Wu, 2024). The raw dataset, formatted in JSON, comprises 5,354,306 publications and 48,277,950 citation relations. We extracted publication metadata and citation relations from the dataset. Following the extraction process, publications published prior to 2000 and after 2015 were excluded. Given that both source-based and topic-based approaches were included in this study, any publication with five or fewer references or keywords was excluded to avoid possible bias. After these procedures, 1,725,037 publications were included in this study.

By utilizing the citation relationships in this dataset, the yearly article co-citation networks were built. The source-based approach included in this study was a slightly modified version of the method employed by Boyack and Klavans (2014). Therefore, the article co-citation networks were transferred into source co-citation networks (SCCN) with the procedures reported in the supplementary materials of Uzzi et al. (2013b). The keyword co-citation networks (KCCN) were constructed in similar ways. The source and keywords were the paper venue ID and keywords extracted from the DBLP dataset. Specifically, we used the 'venue.id' field provided in DBLP v13 as the source ID, which indicated the venue in which an article was published. For keywords, we employed the author-provided keywords included in the dataset. The novelty and conventionality for a source pairing or a keyword pairing were determined by K50, a method used for measuring the relatedness of two entities (Boyack & Klavans, 2014; Klavans & Boyack, 2006). For a pair of entities *i* and *j*, their K50 value was calculated using the following formula.

$$K50_{i,j} = K50_{j,i} = max\left[\frac{(F_{i,j} - E_{i,j})}{\sqrt{S_i S_j}}, \frac{(F_{j,i} - E_{j,i})}{\sqrt{S_i S_j}}\right]$$

$$E_{i,j} = S_i S_j / (SS - S_i)$$

$$SS = \sum_{i=1}^{n} S_i$$

$$S_i = \sum_{j=1}^{n} F_{i,j}$$

$F_{i,j}$ denotes the observed frequency with which entities i and j co-occur in the reference documents of a specific year, and $E_{i,j}$ represents the expected frequency. Therefore, any publication included in this study had two distributions of K50, based on its source pairings and keyword pairings. The median of a K50 distribution was the conventionality of a publication. For novelty, we referred to Boyack and Klavans

(2014) and defined the 5th percentile of the K50 distribution as novelty. In addition, alternative thresholds: the 1st and 10th percentile of the K50 distribution were also tested as supplementary novelty metrics to evaluate the robustness of our findings. Each publication in this research features two sets of novelty/conventionality values, calculated separately from SCCN and KCCN.

The publications were classified into four categories based on their novelty/conventionality and their comparison to the novelty/conventionality scores of all publications in the same year. Our study adopts a relative criterion for identifying high/low novelty. High/low novelty was determined by the 40th percentile (PR40) of all novelty scores, and high/low conventionality was determined by the median of all conventionality scores. The four categories are:

- High novelty & high conventionality (HNHC): The novelty value is less than PR40; the conventionality value is higher than the median.
- High novelty & low conventionality (HNLC): The novelty value is less than PR40; the conventionality value lowerthan the median.
- Low novelty & high conventionality (LNHC): The novelty value is higher than PR40; the conventionality value is higher than the median.
- Low novelty & low conventionality (LNLC): The novelty value is higher than PR40; the conventionality value lower than median.

Note that a lower novelty value indicates that the source/keyword pair occurs less frequently, which suggests a novel combination. Therefore, publications with lower novelty values are classified as high novelty.

Results and Discussion

After preprocessing procedures detailed in the research design, a total of 1,725,037 publications were included in this study. The yearly distribution of these publications is presented in Table 1. The number of included publications increases steadily from 32,364 in 2000 to 198,275 in 2015. Tables 2 and 3 provide the statistics for SCCN and KCCN from 2000 to 2015, respectively. The number of sources ranges from 8,372 in 2000 to 26,124 in 2015. Similarly, the number of source pairings grows from 509,928 in 2000 to 4,642,351 in 2015. Overall, the number of sources and source pairings increase three- and ninefold, respectively, during this period. In the same year, the network scale of KCCN is larger than SCCN. Between 2000 and 2015, the number of keywords grows from 45,642 to 101,837, while the number of keyword pairings expands from 21,958,870 to 124,796,940. These figures indicate two- and sixfold increases, respectively.

Table 1. Number of Included Publications.

Year	*Articles*	*Year*	*Articles*	*Year*	*Articles*	*Year*	*Articles*
2000	32,364	2004	63,913	2008	113,347	2012	156,086
2001	35,611	2005	77,090	2009	125,375	2013	172,782
2002	41,209	2006	93,273	2010	134,322	2014	186,702
2003	49,909	2007	102,392	2011	142,387	2015	198,275

Table 2. Yearly statistics of SCCN.

Year	*Source*	*Pairs*	*Year*	*Source*	*Pairs*
2000	8,372	509,928	2008	17,096	1,931,129
2001	9,084	597,745	2009	18,465	2,201,253
2002	9,846	698,292	2010	19,771	2,521,501
2003	10,838	807,826	2011	20,982	2,823,376
2004	11,920	984,009	2012	22,117	3,100,438
2005	13,151	1,232,116	2013	23,277	3,606,916
2006	14,541	1,427,682	2014	24,539	4,022,984
2007	15,671	1,721,678	2015	26,124	4,642,351

Table 3. Yearly statistics of KCCN.

Year	*Keywords*	*Pairs*	*Year*	*Keywords*	*Pairs*
2000	45,642	21,958,870	2008	77,207	63,226,895
2001	48,515	24,176,947	2009	80,913	69,289,002
2002	51,126	27,834,411	2010	84,500	77,334,472
2003	55,244	30,885,836	2011	87.938	83,561,294
2004	59,401	35,953,067	2012	91,750	91,903,280
2005	64,005	43,137,967	2013	95,690	103,920,458
2006	68,679	49,626,908	2014	98,927	112,976,940
2007	72,955	56,844,272	2015	101,837	124,797,226

Rank-Order Similarity and Classification Consistency: Source vs. Keyword Approaches

This study utilized Spearman's rank correlation to investigate whether source-based and keyword-based approaches evaluate the publications' novelty/conventionality consistently. The results are reported in Figure 4. Novelty (1), Novelty (5), and Novelty (10) represent the results based on utilizing the 1st, 5th, and 10th percentile of the K50 distribution as measures of a publication's novelty. The Spearman rank correlation coefficients range from 0.1 to 0.3, suggesting a weak positive correlation. When Novelty (1) is excluded, the coefficients drop below 0.25. The findings suggest that the novelty/conventionality rank orders from the two approaches are weakly related.

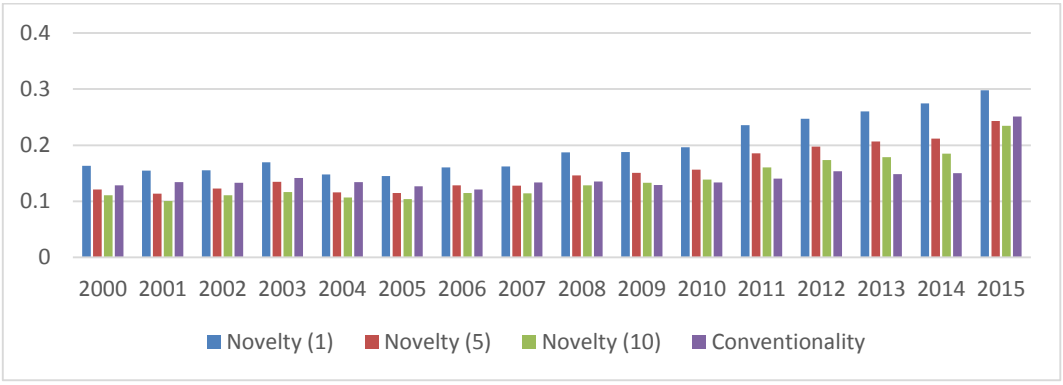


Figure 4. The Spearman rank correlation between two approaches.

We examine whether a publication is classified into the same category by the two approaches. For example, if the source-based approach classifies a publication as high novelty, does the keyword-based approach do the same? Cohen's Kappa, a statistical measure for evaluating agreement between two classifiers, was used to assess the consistency of categorization results between the two approaches. Cohen's Kappa measures the difference between observed agreement and expected agreement by chance. It ranges from -1 to 1, with 1 signifying perfect agreement and 0 indicating agreement equivalent to random chance. Table 4 reported the details. The results indicate that the degree of consistency between the two approaches is only marginally better than what is expected by chance. While consistency has improved over time and increased significantly since 2008, the correlation between the two approaches remains low. We currently suspect that this phenomenon may be attributable to the growth in data size. However, further research is needed to fully address this issue.

Table 4. Classification consistency of binary classes (high/low).

	Novelty (1)	*Novelty (5)*	*Novelty (10)*	*Conventionality*
2000	0.09	0.07	0.06	0.09
2001	0.08	0.07	0.05	0.10
2002	0.09	0.07	0.06	0.09
2003	0.10	0.08	0.06	0.09
2004	0.09	0.07	0.06	0.09
2005	0.09	0.06	0.06	0.08
2006	0.10	0.08	0.08	0.08
2007	0.09	0.08	0.07	0.09
2008	0.12	0.10	0.08	0.09
2009	0.12	0.11	0.09	0.08
2010	0.13	0.11	0.09	0.08
2011	0.16	0.13	0.11	0.09
2012	0.18	0.15	0.12	0.10
2013	0.19	0.16	0.12	0.10
2014	0.19	0.16	0.12	0.10
2015	0.19	0.17	0.16	0.20

Note. Novelty (1) refers to the results of examining the classification consistency of the novelty type derived from two approaches based on the publication's Novelty (1). The same applies to the other notations.

By combining novelty and conventionality values, each approach classifies a publication into one of four possible categories: NHNC, NHLC, LNHC, and LNLC. We further examine the classification consistency of four categories with Cohen's Kappa. Similarly, the degree of consistency between two approaches is weak. Table 5 reports the details. The result indicates that two approaches may evaluate the publication's novelty/conventionality from different perspectives and classify the same publication into various categories.

Table 5. Classification consistency of multiple classes.

	Novelty (1)	*Novelty (5)*	*Novelty (10)*
2000	0.07	0.07	0.06
2001	0.07	0.07	0.06
2002	0.07	0.07	0.06
2003	0.08	0.07	0.06
2004	0.07	0.07	0.06
2005	0.07	0.06	0.06
2006	0.08	0.07	0.07
2007	0.07	0.07	0.06
2008	0.08	0.08	0.07
2009	0.08	0.08	0.07
2010	0.08	0.08	0.07
2011	0.10	0.09	0.08
2012	0.11	0.10	0.09
2013	0.11	0.10	0.09
2014	0.11	0.10	0.09
2015	0.14	0.13	0.13

Note. Novelty (1) refers to the results of examining the classification consistency of the four classes determined by two approaches based on the publication's Novelty (1) and conventionality value. The same applies to the other notations.

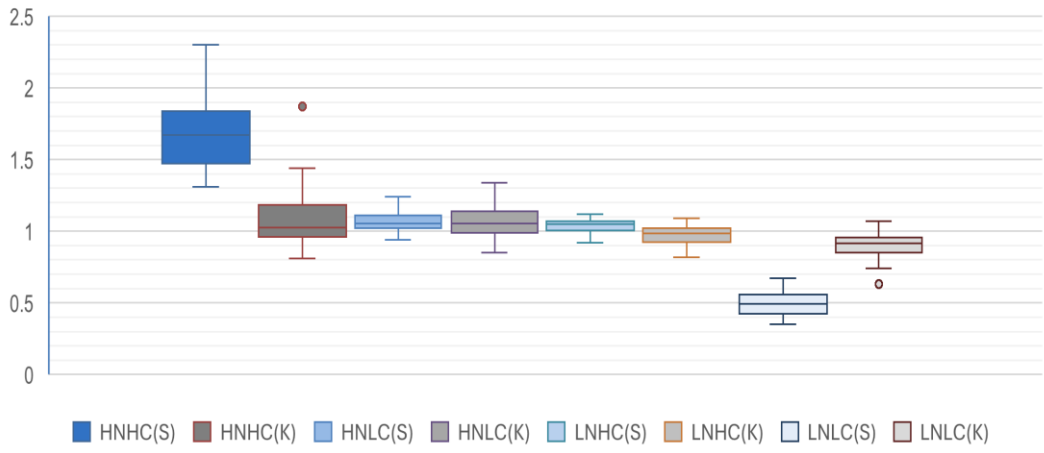
Ability in Identifying Highly Cited Research Publications

To better understand the differences between the two approaches, we analyze the probability that publications categorized into different groups was highly cited. Specifically, we compared the probabilities of being highly cited across the four categories (HNHC, HNLC, LNHC, LNLC) within each approach and investigated whether the probabilities of being highly cited in corresponding categories differ between the source-based and keyword-based approaches. Highly cited publications are defined as those with citation counts in the top 5% for a given year. To ensure the robustness of our findings, we also examine results using alternative thresholds, defining highly cited publications as those in the top 1% and 10%, respectively.

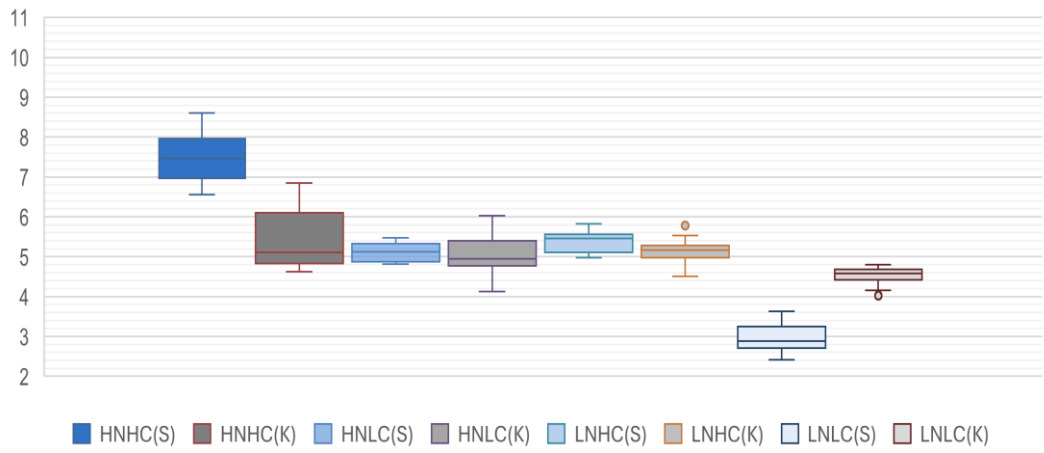
Figure 5 presents the yearly probabilities distribution of being highly cited, with novelty defined as the 5th percentile (PR5) and conventionality as the median of a publication's K50 distribution. We used the notation S and K to denote the combinatory unit of knowledge source and keywords, respectively. Thus, HNHC(S) denotes the category based on the source-based approach, while HNHC(K) denotes the category based on the keyword-based approach. When analyzing the probabilities of being highly cited for groups formed using the source-based approach, the results suggest that in both approaches HNHC has the highest

probability of being highly cited, though the differences is much more salient when journals are used as the basic unit of knowledge.

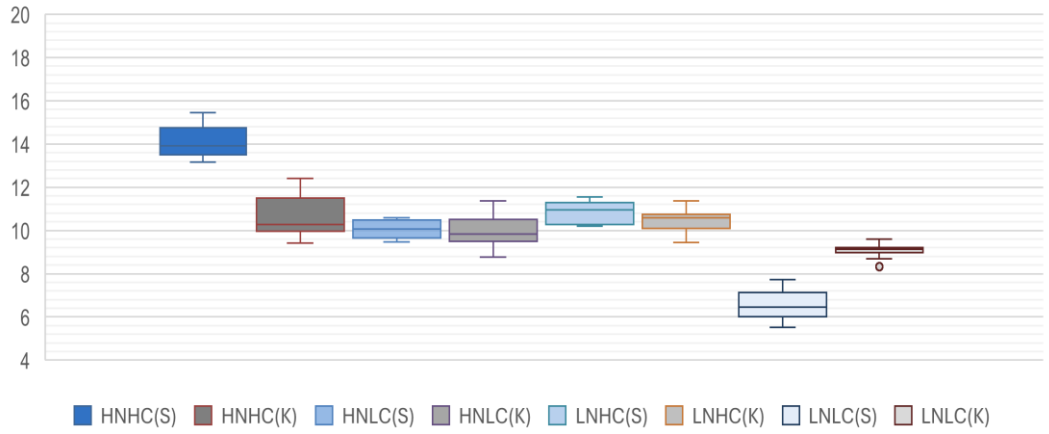
HNHC(S) consistently exhibits nearly double the probability of being highly cited compared to HNLC(S) and LNHC(S). This difference grows to approximately 3–4 times when compared with LNLC(S). However, this pattern is less apparent when examining groups formed using the keyword-based approach. The probabilities of being highly cited for HNHC(K) is only slightly higher than the rest, with only LNLC(K) showing the lowest probability of producing highly cited papers. This pattern remained robust across various thresholds used to define highly cited papers (see Appendices I and II).



(a) Top 1% cited articles as highly-cited articles



(b) Top 5% cited articles as highly-cited articles

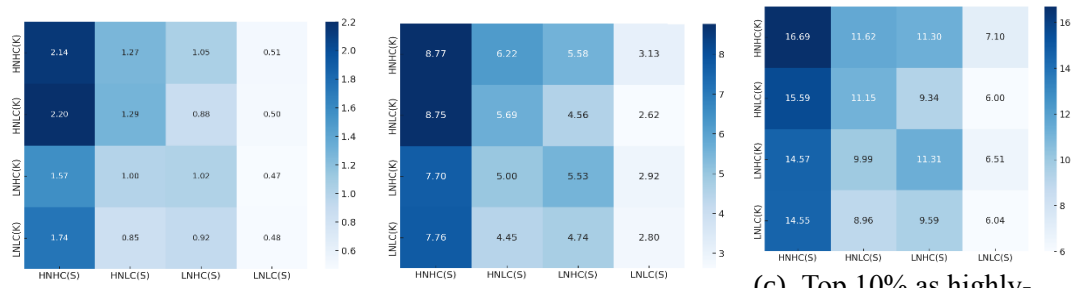


(c) Top 10% cited articles as highly-cited articles

Figure 5. Probabilities of being top highly cited across groups: Novelty (5). Y-axis represents the probability.

Analyzing the Effects of Combining Two Classification Approaches

Given HNHC category yielded the highest probabilities of identifying highly cited papers when either journal or keyword-based approaches was used, though to different extent, journal-based approach is a much stronger predictor. Using a keyword-based approach to measure novelty and conventionality, the difference in the probability of being among the top 5% most-cited articles across the four categories is less than 1% on average. However, while using a journal-based approach, the difference is more the 4%. And as shown earlier, the correlation between these two approaches is low. We explore whether combining both source-based and keyword-based approach enhances the relationship between HNHC and citation impact. In other words, does keyword-based HNHC provide extra explanatory power over and above source-base HNHC is predicting highly cited papers. As shown in Figure 6, if a publication is classified as HNHC by the source-based approach and topic-based approach, its probability of being highly cited is slightly higher.



(a) Top 1% as highly-cited (b) Top 5% as highly-cited (c) Top 10% as highly-cited

Figure 6. Highly-Cited Probabilities for Combined Classifications from Source-Based and Topic-Based Approaches.

To further test our hypothesis, we conducted a logistic regression analysis using highly cited papers as the dependent variable. The source-based and keyword-based HNHC categories served as the two key predictors, allowing us to assess the explanatory power of each classification approach. Specifically, two dummy variables were created to indicate whether an article was categorized as HNHC by the source-based approach and the keyword-based approach, respectively.

As shown in Table 6, As when all predictor variables are set to zero, the model estimates a log-odds of -3.0329, which corresponds to a predicated probability of about 4.6%, closed to our definition for highly-cited articles. Both predictors were significant, despite great differences in their coefficient. If an article is categorized into HNHC by source-based approach, its probability of being highly-cited articles increases to 8.26%, a 79.9% relative increase in the likelihood of the event. Similarly, the probability rises to 5.13%, a smaller 11.6% increase, when an article is categorized into HNHC by topic-based approach. If both approaches classify an article into HNHC, the event probability reaches 9.18%, a nearly 99.8% increase from the baseline.

Table 6. The odds of being highly-cited papers when identified as HNHC by two approaches.

Variable	*Coefficient*	*Stand error*	*Percentage change in odds*
NHNC(S)	0.62***	0.009	79.9
NHNC(K)	0.12***	0.012	11.6
Constant	-3.03***	0.004	

Note. *** <.001

Conclusion

While combination of heterogenous knowledge has long been recognized as a great source for innovation. It remains unclear that novelty along, or the combination of both novelty and conventionality is able to yield high impact research. Novelty has been shown to be associated with frontier research projects (Boudreau et al., 2016), higher research impact (Carayol et al., 2019), and seminal works in scientometrics (Tahamtan & Bornmann, 2018). On the other hand, studies also suggest that novelty alone is not enough, that it is also important to situate the novel ideas in established wisdom, therefore the importance of combining novelty and conventionality (Uzzi et al., 2013; Boyack & Klavans, 2014). One possible explanation for such inconsistency is the use of different knowledge units used when measuring novelty. The first step to clarify the relationship between knowledge combination and impact is therefor to examine how consistent when source based and topic based measurement of the construct of novelty/conventionality. Using DBLP dataset in the domain of computer sciences, we set out to compare how results from topic vs. source based knowledge unit are consistent with each other. The results show that the correlation between the two method is low, suggesting they are capturing different aspect of novelty. Furthermore, Consistent with the original research by Uzzi et al. (2013), we found that a paper combining high novelty and conventionality increases its likelihood of becoming highly cited within a given year. However, this

relationship between high novelty and conventionality and the likelihood of a hit paper was observed was much more salient when novelty and conventionality were calculated using journal co-citation data, and less so when keyword co-citation data was used.

These findings indicate that the source-based approach may better be able highlight the advantages of integrating high novelty and high conventionality, as demonstrated by the increased likelihood of being highly cited. Several limitations need to be noted, one of which is the lack of vocabulary control of author assigned keywords. While it is argued that author keywords, compared by control vocabulary such as MeSH, offers a great granularity therefore more precise representation of the topics (Carayol et al., 2019). Yet, without the benefit of vocabulary control, the actually cooccurrence frequence of topics is likely to highly underestimated because of morphological and semantic variations of the topics. And it is difficult to assess the extent of this underestimation and how this might impact the measurement of novelty. One possible solution to this dilemma is to utilize automatically-assigned topics such as SciVal topics used in Scopus and micro citation topics used in Web of Science. It should also be noted that, instead of keyword co-occurrence is the focal paper, as commonly done in previous research, we have adopted a novel approach of using keywords co-occurrence appearing in the cited references by the focal paper, which resulting a much greater set of keyword co-occurrence pairs. Future studies need to be done, to examine the consistency of these two approaches—using focal-paper keywords versus cited-reference keywords—in measuring topic-based novelty.

Acknowledgments

This work was financially supported by the Universities and Colleges Humanities and Social Sciences Benchmarking Project (Grant no. 113L9A001), Ministry of Education in Taiwan, and National Science and Technology Council, R.O.C. (Taiwan) under the grant numbers 113-2410-H-032-031-. The authors are also grateful to anonymous reviewers for their comments and suggestions.

References

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., & Riedl, C. (2016). Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Science, 62*(10), 2765–2783. https://doi.org/10.1287/mnsc.2015.2285

Bornmann, L., Tekles, A., Zhang, H. H., & Ye, F. Y. (2019). Do we measure novelty when we analyze unusual combinations of cited references? A validation study of bibliometric novelty indicators based on F1000Prime data. *Journal of Informetrics, 13*(1), 100979. https://doi.org/10.1016/j.joi.2019.100979

Boyack, K., & Klavans, R. (2014). Atypical combinations are confounded by disciplinary effects. In *Proceedings of the 19th International Conference on Science and Technology Indicators*. Leiden, The Netherlands.

Carayol, N., Agenor, L., & Oscar, L. (2019). The right job and the job right: Novelty, impact and journal stratification in science. *Impact and Journal Stratification in Science* (March 5, 2019).

Fontana, M., Iori, M., Montobbio, F., & Sinatra, R. (2020). New and atypical combinations: An assessment of novelty and interdisciplinarity. *Research Policy, 49*(7), 104063. https://doi.org/10.1016/j.respol.2020.104063

Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists' research strategies. *American Sociological Review, 80*(5), 875–908. https://doi.org/10.1177/0003122415601618

Huang, Z., Tang, D., Zhao, R., & others. (2024). A scientific paper recommendation method using the time decay heterogeneous graph. *Scientometrics, 129*, 1589–1613. https://doi.org/10.1007/s11192-024-04933-4

Kanwal, T., & Amjad, T. (2024). Research paper recommendation system based on multiple features from citation network. *Scientometrics, 129*, 5493–5531. https://doi.org/10.1007/s11192-024-05109-w

Kaplan, S., & Vakili, K. (2015). The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal, 36*(10), 1435–1457. https://doi.org/10.1002/smj.2294

Klavans, R., & Boyack, K. W. (2006). Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology, 57*(2), 251–263.

Lee, Y.-N., Walsh, J. P., & Wang, J. (2015). Creativity in scientific teams: Unpacking novelty and impact. *Research Policy, 44*(4), 684–697. https://doi.org/10.1016/j.respol.2014.10.007

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science, 2*(1), 71–87. https://doi.org/10.1287/orsc.2.1.71

Ruan, X., Ao, W., Lyu, D., Cheng, Y., & Li, J. (2023). Effect of the topic-combination novelty on the disruption and impact of scientific articles: Evidence from PubMed. *Journal of Information Science.* https://doi.org/10.1177/01655515231161133

Tahamtan, I., & Bornmann, L. (2018). Creativity in science and the link to cited references: Is the creative potential of papers reflected in their cited references? *Journal of Informetrics, 12*(3), 906–930. https://doi.org/10.1016/j.joi.2018.08.001

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). ArnetMiner: Extraction and mining of academic social networks. *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008)*, 990–998. https://doi.org/10.1145/1401890.1402008

Thagard, P. (2012). Creative combination of representations: Scientific discovery and technological invention. In R. Proctor & E. J. Capaldi (Eds.), *Psychology of science* (pp. 389–405). Oxford: Oxford University Press.

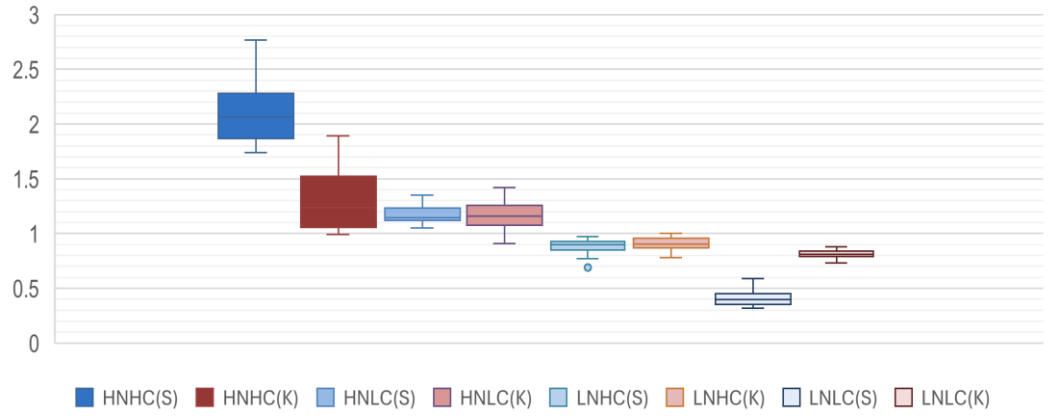
Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science, 342*, 468-472. https://doi.org/10.1126/science.1240474

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013b). Supplementary materials for atypical combinations and scientific impact. *Science, 342*, 468. https://doi.org/10.1126/science.1240474

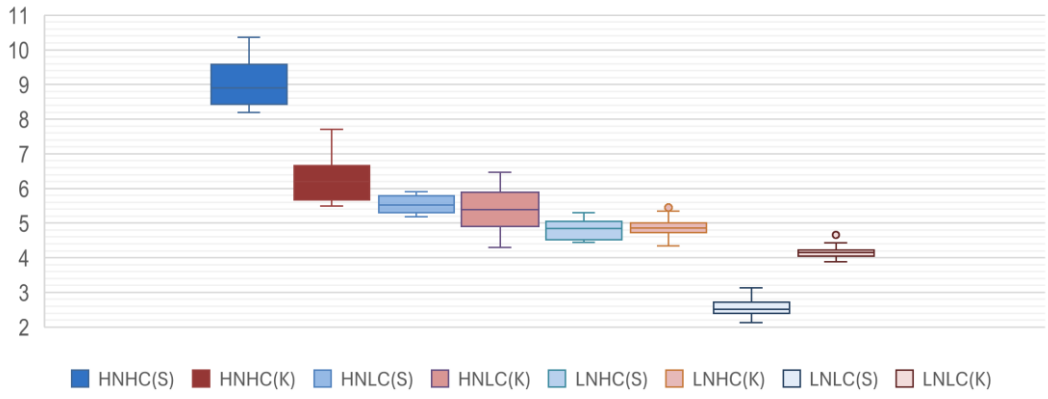
Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy, 46*(8), 1416–1436. https://doi.org/10.1016/j.respol.2017.06.006

Zhang, F., & Wu, S. (2024). Predicting citation impact of academic papers across research areas using multiple models and early citations. *Scientometrics, 129*, 4137–4166. https://doi.org/10.1007/s11192-024-05086-0

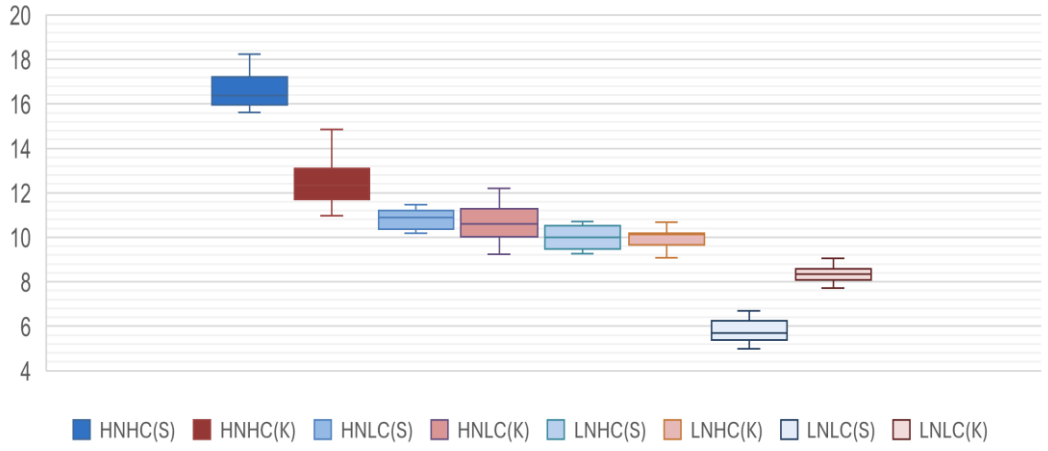
Appendix I



(a) Top 1% cited articles as highly-cited articles



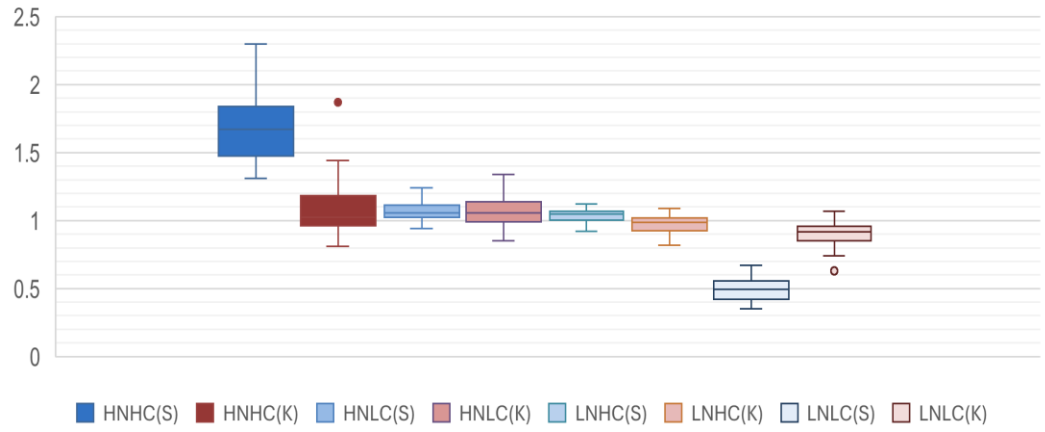
(b) Top 5% cited articles as highly-cited articles



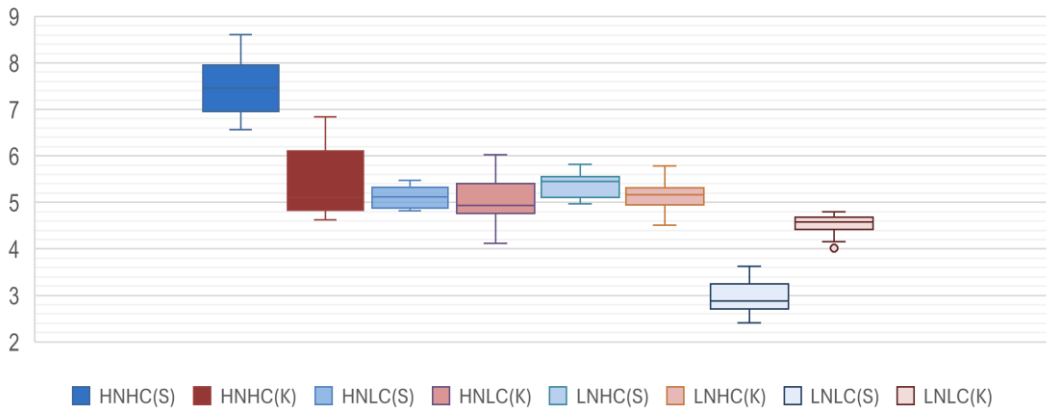
(c) Top 10% cited articles as highly-cited articles

Figure A1. Probabilities of being top highly cited across groups: Novelty (1). Y-axis represents the probability.

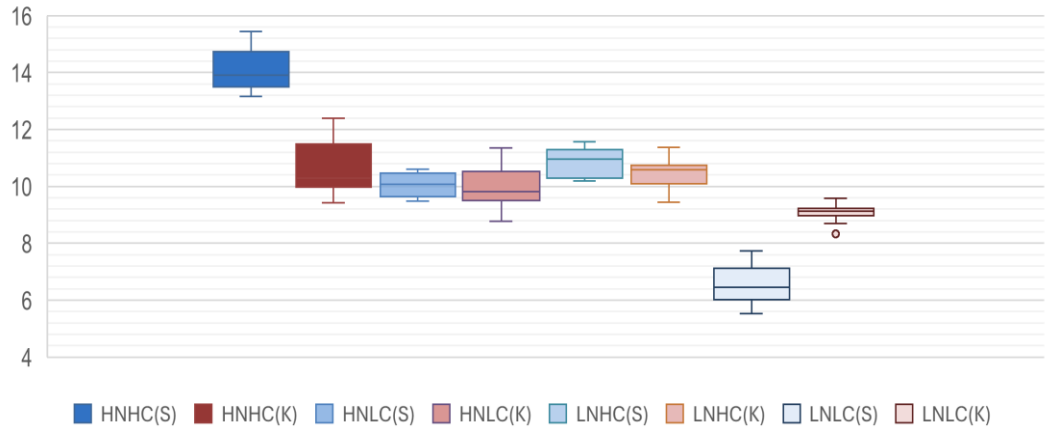
Appendix II



(a) Top 1% cited articles as highly-cited articles



(b) Top 5% cited articles as highly-cited articles



(c) Top 10% cited articles as highly-cited articles

Figure A2. Probabilities of being top highly cited across groups: Novelty (10). Y-axis represents the probability.