Interdisciplinarity vs. Unidisciplinarity: A Structural Comparison of Multi-generation Citations and References

Jiawei Xu[1], Chao Min[2], Win-bin Huang[1] and Yi Bu[1]

[1]*{imxujiawei, huangwb, buyi}@pku.edu.cn*
Department of Information Management, Peking University (China)

[2]*mc@nju.edu.cn*
School of Information Management, Nanjing University (China)

Abstract
This paper compares the structural characteristics between the citation/reference cascades (networks containing multi-generation citations and references) of interdisciplinary and unidisciplinary papers. By constructing multiple citation/reference cascades for each publication in the American Physical Society (APS) dataset, we found that: (1) for papers having a similar number of references, interdisciplinary papers tend to have a smaller and less 'radioactive' knowledge foundation than unidisciplinary papers, though the knowledge foundation of interdisciplinary papers is more heterogeneous; (2) for papers having a similar 'direct' scientific impact, unidisciplinary papers have a relatively greater 'indirect' impact, which indicates that the scientific impact of unidisciplinary papers is 'deeper' and more persistent; and, (3) compared to unidisciplinary papers, there exists at least a short time in which interdisciplinary papers trigger more follow-up discussions.

Introduction
Interdisciplinary research is found to facilitate knowledge integration and scientific innovation (Klein, 1990; Rafols et al., 2012). Yet, some researchers also hold the view that "the general superiority of interdisciplinarity over disciplinary knowledge" has not been fully proved (Jacobs & Frickel, 2009, p. 60). Within this debate, a considerable number of studies have focused on interdisciplinarity, many of which have paid special attention to the relationship between the interdisciplinarity of a single publication and its scientific impact. For instance, Larivière and Gingras (2010) found that publications in physics citing citation-intensive disciplines tend to obtain higher citation scores. Chen, Clément Arsenault and Larivière (2015) showed that the top 1% most-cited papers present higher levels of interdisciplinarity. These studies suggest to us that there might be a positive correlation between the level of interdisciplinarity of a publication and its 'direct' scientific impact, calculated by citation count-related indicators. Nonetheless, for two publications having a similar 'direct' impact, one might have a greater 'indirect' impact (quantified by using multi-generation citing publications) than the other. And the publication that triggers more follow-up discussions seems more invaluable.

To this end, in this paper, we adopt an existing citation network structure, namely citation cascade (Min, Sun & Ding, 2017; Min et al., 2021a) and reference cascade (Min et al., 2021b). For a given focal publication, its citation cascade contains its first-, second- (i.e., citing publications' citing publications), …, and *n*th generation citing publications, as well as their citing relationships. Symmetrically, its reference cascade includes its first-, second-, …, and *n*th generation references. Rousseau identified the concept of multi-generation citations and references dating back to 1987 (Rousseau, 1987), a concept which has inspired many further studies (Hu, Rousseau & Chen, 2011; Hu & Rousseau, 2016). Investigating citation cascades helps understand forward citation behaviour details and how a particular study inspires later research, regardless of whether this influence is direct or indirect. As for reference cascades, they include backward citation nuances in terms of knowledge foundations (e.g., scale, interlock-wise structure, etc.). This paper employs these two cascades and compares the 'indirect' scientific impact and knowledge foundation of interdisciplinary and unidisciplinary publications by adopting many structural-level measurements, such as structural virality, average clustering coefficient and network density.

Data and methods

Data

We adopt the American Physical Society (APS) dataset that covers 541,448 publications from 1893 to 2013 and the nearly 3 million citing relations among them. As a balance to guarantee a relatively long citation window and a sufficient number of publications, we particularly select all publications in 1996, 1997 and 1998 in our following empirical study.

All publications in or after 1975 are assigned one or more codes within the Physics and Astronomy Classification Scheme (PACS). These codes are adopted to identify fields in physics. There are several levels in the PACS. For simplicity, we employ the top-level codes (representing 10 fields) in our study to present the specialties of publications.

Methods

Citation cascade and reference cascade

From the structural perspective, citation cascade and reference cascade are generated through citing relations. For a given focal publication, its citation cascade comprises its citing publications (first-generation citations), its citing publications' citing publications (second-generation citations), …, and *n*th-generation citations. A reference cascade is symmetric; that being said, the reference cascade of a given publication contains its references (first-generation references), its references' references (second-generation references), …, and *n*th-generation references. Thus, reference and citation cascades tend to be more informative than traditional bibliometric indicators, e.g., number of citations and its normalised variants (Bornmann & Daniel, 2008).

Previous scientometricians have focused on the conceptualisation and operationalisation of citation cascades (e.g., Min et al., 2017; Min et al., 2021a), described as "the constitution of a series of subsequent citing events initiated by a certain publication" (Min et al., 2021a, p. 110). Similarly, reference cascades contain a series of proceeding citing (referencing) events initiated by a certain publication.

We defined all publications in 1996, 1997 and 1998 as the 'initial publication set' $\boldsymbol{I}$. And, for every publication in $\boldsymbol{I}$, we created its citation cascade and reference cascade using all the citing relations in the APS dataset. We stipulate that, if the depth of a publication's reference or citation cascade is less than four, this publication, as well as its cascades, will be removed from our following empirical study. To this end, 29,310 publications remain.

Quantifying interdisciplinarity

As mentioned above, each publication has one or more PACS labels from the 10 labels. Following Porter and Chubin (1985), the references' PACS labels are used to measure how interdisciplinary a publication is. We characterise the degree of interdisciplinarity of one specific publication following Figure 1. As shown, we first examine the PACS labels of all references of i and adopt the Reverse Simpson Index (RSI_i) as an indicator:

$$RSI_i = 1 - \sum_{j=1}^{R} {p_j}^2$$

where R represents the number of fields publication i's references come from and p_j the proportion of field j among all fields. Per Simpson (1949), RSI is "a measure of the concentration of the classification" (p. 688). The greater the RSI is, the higher the degree of interdisciplinarity. In practice, we define the top 20% ranked publications in terms of RSI as interdisciplinary papers and the bottom 20% as unidisciplinary papers[1]. This results in 5837 and 5834 publications in each group, respectively[2].

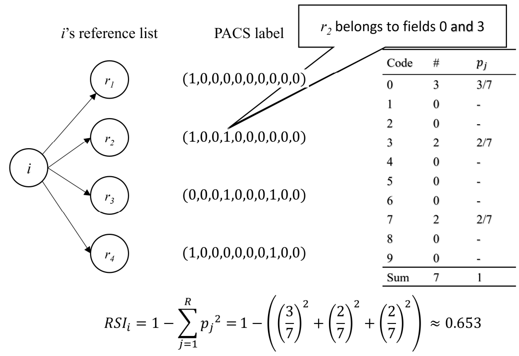

Figure 1. Illustration of quantifying interdisciplinarity of a publication *i*.

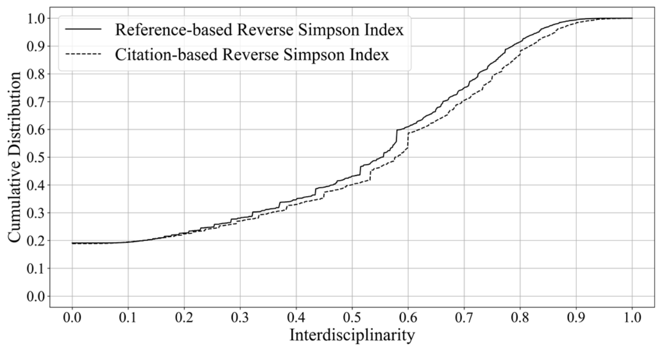

Figure 2. The cumulative distribution of the two interdisciplinarity indicators.

<u>Matching</u>

To reveal the differences regarding interdisciplinarity and unidisciplinarity, we set up two groups of publications. For each unidisciplinary publication, we try to find a 'similar' publication in the interdisciplinary group. To this end, we employ a one-to-one matching and stipulate that one publication cannot be matched with multiple publications. When implementing this matching process, we follow the below rules:

(1) This pair of publications have at least one same PACS field;
(2) For a given unidisciplinary publication, we only keep a certain interdisciplinary publication that has >90% field similarity[3];
(3) Compared with the higher cited publication (regardless of which groups), the lower cited publication has at least 90% as many citations, i.e., their numbers of citations have a ≤10% difference;
(4) Compared with the higher citing publication (regardless of which groups), the lower citing publication has at least 90% as many as references;
(5) This pair of publications have the same document type; and
(6) This pair of papers were published in the same journal in the same year.

Finally, we obtained 822 pairs of publications. The matched groups (822*2=1644 publications) are hereafter abbreviated as interdisciplinarity and unidisciplinary groups.

---

[1] According to Figure 2, there are ~20% publications with an *RSI* of 0.0. We therefore set the bottom 20% as unidisciplinary publications. Correspondingly, the top 20% are defined as interdisciplinary ones.

[2] Because there might be more than one publication exactly at the 20% threshold, we have to include all of these publications in practice. This is why there are minor differences regarding the number of publications in the two groups.

[3] For each publication, we construct a 10-dimension vector that contains its PACS code; each dimension represents one of the 10 fields under physics. If a publication has a certain field label, this dimension will be marked as one; otherwise, zero. We compare two publications' field similarity by calculating the cosine similarity of the two vectors.

Measurements

For each publication in the unidisciplinarity or interdisciplinarity group, we construct three citation cascades with different generations, i.e., citation cascade with the first two, three and four generations, as well as three reference cascades with different generations, i.e., reference cascade with the first two, three and four generations. We calculate the below indicators particularly:

- Structural popularity (SP): For citation cascades, SP represents the range to which knowledge is diffused; in terms of reference cascades, it indicates the scale of knowledge foundation. We utilise the number of nodes in a cascade as a measurement.
- Structural virality (SV): SV measures the knowledge diffusion depth of the publication. We use the average depth of the publication's citation cascade to represent it.

$$SV = \frac{1}{|T|-1} \sum_{v \in T, v \neq i} d(v, i),$$

T is the cascade initiated by publication i, $|T|$ is the number of nodes in T and $d(v, i)$ is the shortest path between v and i.

- Average clustering coefficient (ACC): ACC reflects the goodness of clustering in a network. Specifically:

$$ACC = \frac{1}{|V|} \sum_{v_k} C(v_k)$$

where V is the node set and $C(v_k)$ is the clustering coefficient of node v_k:

$$C(v_k) = \frac{number\ of\ closed\ triads\ connected\ to\ v_k}{number\ of\ triples\ of\ vertices\ centered\ on\ v_k}$$

- Network density (ND): ND describes the proportion of the potential connections in a network that are actual connections.

$$ND = \frac{2|E|}{|V|(|V|-1)}$$

where E is the edge set. The higher the ND is, the more connections the network has.

Results and Discussion

To statistically test whether the interdisciplinary and unidisciplinary publications have significant differences regarding the aforementioned indicators, we employ the Wilcoxon signed-rank test. As shown in Table 1, in terms of reference cascades, papers from the two groups have shown significant differences in SP and ACC when we select the first three and four generations in the cascade. As for citation cascades, we observe a significant difference for SV and ND for the first two and three generations. We did not find any significance in the first four generations for citation cascades, partly because of, at least conceptually, the insufficient topical similarity between a focal publication and its fourth-generation citing publications.

Table 1. Wilcoxon signed-rank test results.

Indicator	SP	SV	ACC	ND
Reference Cascade (first 2 generations) Sig.	0.149	-	0.240	0.379
Reference Cascade (first 3 generations) Sig.	**0.027***	-	**0.000*****	0.652
Reference Cascade (first 4 generations) Sig.	**0.005****	-	**0.000*****	0.406
Citation Cascade (first 2 generations) Sig.	0.108	**0.007****	0.269	**0.048***
Citation Cascade (first 3 generations) Sig.	0.105	**0.036***	0.076	0.069
Citation Cascade (first 4 generations) Sig.	0.208	0.117	0.198	0.139

Notes. SP=Structural popularity; **SV**=Structural virality; **ACC**=Average clustering coefficient; **ND**=Network density. $^{*}p$<0.05, $^{**}p$<0.01, $^{***}p$<0.001. All significant results under the 0.05 threshold are **bolded**.

We are particularly interested in the indicators that show statistically significant differences. In this research-in-progress paper, we present four of them in Figure 3, representing SP and ACC for reference cascades (sub-figures A and B, respectively) and SV and ND for citation cascades (sub-figures C and D, respectively). Figure 3(A) indicates the cumulative distribution of reference cascade SP (first four generations). We can see that, when SP<400, the unidisciplinarity and interdisciplinarity groups distribute closely and that there are more interdisciplinarity papers whose SPs are between 400 and 1100. Yet, when SP is greater, unidisciplinary papers dominate. As SP of reference cascades represents the knowledge foundation of the focal publication, Figure 3(A) reveals that interdisciplinary publications tend to have a smaller amount of knowledge foundation than unidisciplinary publications.

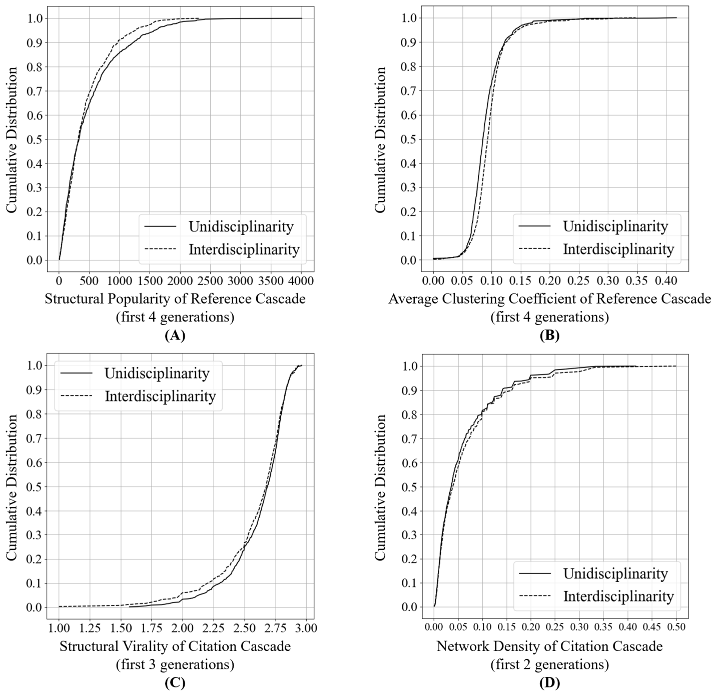

Figure 3. The cumulative distribution of the four structural indicators.

Figure 3(B) shows the cumulative distribution of reference cascade ACC (first four generations), and it demonstrates that an interdisciplinary paper's reference cascade is more similar to an interlocked network, which reveals that an interdisciplinary paper's knowledge foundation is more 'interlocked', though interdisciplinarity reflects the diversity of references.

As to citation cascade, Figure 3(C) presents the cumulative distribution of citation cascade SV (first three generations), in which we find that there are more interdisciplinary publications with a lower SV value. As SV is quantified by the average 'distance' between the focal publication and other publications in the citation cascade, this demonstrates that interdisciplinary publications have a more 'direct' or 'shallow' impact while unidisciplinary publications have a more 'indirect' or 'deep' impact.

Nonetheless, citation cascade ND (first two generations) displays the opposite pattern – we observe more interdisciplinary publications that have a greater value of ND (Figure 3(D)). Figure 3(D) reveals that there are more connections between the publications in the citation cascade of an interdisciplinary paper, and that interdisciplinary research may intrigue researchers and results in more discussions.

Conclusions

In this paper, we implement a structural comparison between the reference/citation cascades of unidisciplinary and interdisciplinary papers. With a sophisticated matching process, our study suggests that interdisciplinary papers tend to have a smaller knowledge foundation and a more 'shallow' or 'direct' impact on the follow-up research. Nonetheless, interdisciplinary papers contribute to the interactions between researchers. We speculate that interdisciplinary research may be relatively pioneering and there are fewer prior works for researchers to build on. Due to this characteristic of interdisciplinary research, policy makers should even more strongly encourage interdisciplinary research and devote more resources to it. Also, more future works are needed to explore the features and mechanisms of interdisciplinary research.

There are some limitations in our study. First, we do not set up any restriction when constructing the cascades, e.g., publication year of forward citations. Future work may consider involving a temporal dimension as a filter and set up different thresholds to examine differences. Second, since our experiment is implemented only in the discipline of physics, we cannot generalise the conclusions to other domains. In the future, we are going to apply the current empirical study in a more comprehensive, cross-discipline bibliographic dataset. Finally, only a few constraints are taken into account in our matching experiment, which may influence the reliability of the experiment. Many other factors could be controlled in our future study, such as the number of co-authors in a publication and their affiliations.

Acknowledgments

We acknowledge the two anonymous referees as well as the organisers for their useful comments. We also acknowledge the American Physical Society (APS) for publicly share the dataset for research.

References

Bornmann, L. & Daniel, H. D. (2008). What do citation counts measure? a review of studies on citing behaviour. *Journal of Documentation, 64*(1), 45-80.

Chen, S., Arsenault, C. & Larivière, V. (2015). Are top-cited papers more interdisciplinary? *Journal of Informetrics, 9*(4), 1034–1046.

Hu, X. & Rousseau, R. (2016). Scientific influence is not always visible: the phenomenon of under-cited influential publications. *Journal of Informetrics, 10*(4), 1079-1091.

Hu, X., Rousseau, R. & Chen, J. (2011). On the definition of forward and backward citation generations. *Journal of Informetrics, 5*(1), 27–36.

Jacobs, J. A. & Frickel, S. (2009). Interdisciplinarity: A critical assessment. *Annual Review of Sociology*, 35, 43–65. http://dx.doi.org/10.1146/annurev-soc-070308-115954

Klein, J.T. (1990). *Interdisciplinarity: History, theory, and practice.* Detroit, MI: Wayne State University Press.

Larivière, V. & Gingras, Y. (2010). On the relationship between interdisciplinarity and scientific impact. *Journal of the American Society for Information Science and Technology, 61*(1), 126-131.

Min, C., Chen, Q., Yan, E., Bu, Y. & Sun, J. (2021a). Citation cascade and the evolution of topic relevance. *Journal of the Association for Information Science & Technology*, *72*(1), 110-127.

Min, C., Xu, J., Han, T. & Bu, Y. (2021b). References of References: How Far is the Knowledge Ancestry. https://arxiv.org/abs/2101.08577

Min, C., Sun, J. & Ding, Y. (2017). Quantifying the evolution of citation cascades. *Proceedings of the Association for Information Science and Technology, 54*(1),761-763.

Porter, A.L. & Chubin, D.E. (1985). An indicator of cross-disciplinary research. *Scientometrics, 8*(3–4), 161–176

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P. & Stirling, A. (2012). How journal rankings can suppress interdisciplinary research: A comparison between innovation studies and business & management. *Research Policy, 41*(7), 1262–1282.

Rousseau, R. (1987). The Gozinto theorem: Using citations to determine influence on a scientific publication. *Scientometrics*, 11, 217–229.

Simpson, E. H. (1949). Measurement of diversity. *Nature*, 163, 688.