STI2022
GRANADA | SEP 7+8+9

26th International Conference on Science and Technology Indicators
"From Global Indicators to Local Applications"

#STI2022GRX

Poster

STI 2022 Conference Proceedings
Proceedings of the 26th International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Proceeding Editors

Nicolas Robinson-Garcia
Daniel Torres-Salinas
Wenceslao Arroyo-Machado

OPEN ACCESS

Citation: Wang, H., Chen, H., Meng, F., & Bu, Y. (2022). Towards "rich clubs" in scientific publications: A preliminary exploration. In N. Robinson-Garcia, D. Torres-Salinas, & W. Arroyo-Machado (Eds.), *26th International Conference on Science and Technology Indicators*, STI 2022 (sti22117). https://doi.org/10.5281/zenodo.6951730

cc BY NC ND

Copyright: © 2022 the authors, © 2022 Faculty of Communication and Documentation, University of Granada, Spain. This is an open access article distributed under the terms of the Creative Commons Attribution License.

Collection: https://zenodo.org/communities/sti2022grx/

26th International Conference on Science and Technology Indicators | **STI** 2022
“From Global Indicators to Local Applications”
7-9 September 2022 | Granada, Spain
#STI22GRX

Towards "rich clubs" in scientific publications: A preliminary exploration[1]

Haoyang Wang*, Hongkan Chen*, Fan Meng* and Yi Bu*

**wanghaoyang@stu.pku.edu.cn; chenhongkan@pku.edu.cn; mengfan@pku.edu.cn; buyi@pku.edu.cn*
Department of Information Management, Peking University, Beijing 100871, China

Introduction

Complex network studies have researched core-periphery structures that focus on the interaction between nodes for quite a few years (Csermely, London, Wu, & Uzzi, 2013). From a topological perspective, “core” refers to a set of densely connected nodes in a network while “periphery” represents the nodes which lies at the fringe of the network and are sparsely connected with each other. Rich clubs, a vivid metaphor first introduced by Zhou and Mondragon (2004), is one of the ideas describing the core-periphery structure by the density of the connectiveness of the “rich” nodes.

The rich club phenomenon has been widely observed in many networks under various contexts covering biological, technological, and social world (Colizza, Flammini, Serrano, & Vespignani, 2006), among which science of science is an important research context (Szell & Sinatra, 2015). To measure the strength of rich clubs, Zhou and Mondragon (2004) first introduced the rich club coefficient to quantify the phenomenon. The rich club coefficient is defined as a function where, given a degree k, there is a corresponding coefficient measuring the density among those nodes whose value of degree is larger than k. The rich club coefficient has been defined in both undirected and direct networks in a similar way (Smilkov & Kocarev, 2010).

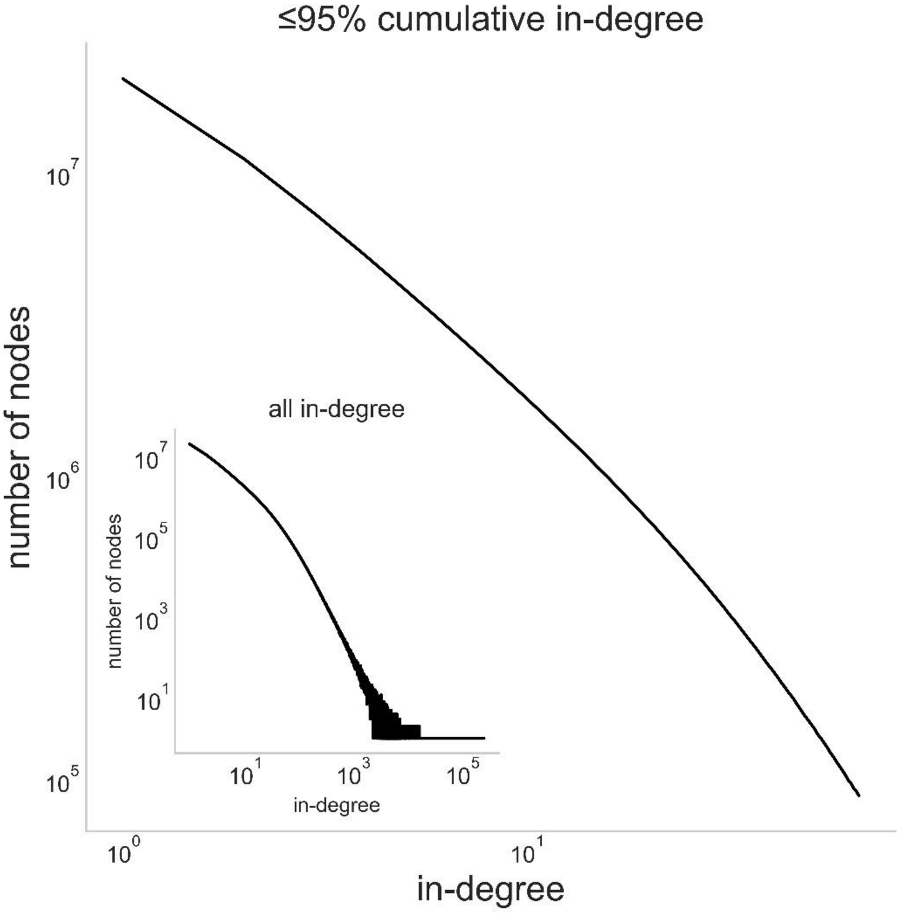
Scientometricians have established many types of networks in their research, e.g., citation, coauthorship, bibliographic coupling, co-citation, topic, and co-words networks (Yan & Ding, 2012). Among these, citation networks are one of the most frequently adopted networks as they offer a good performance in research evaluation, mapping knowledge domains, and scientific and technical information mining (Waltman, 2016). Topological analyses on citation networks may thus supply various benefits (Shibata, Kajikawa, & Matsushima, 2007) and provide methodological and political implications. In this paper, we present a preliminary quantitative study upon the rich club phenomenon in publication citation networks. We first introduce the dataset used and define the way quantifying rich clubs. We then perform temporal, field-wise analyses on the rich club phenomenon.

[1] This work was supported by the Youth Project of Humanities and Social Sciences of the Ministry of Education of China (No: 21YJC870001) and the 7th Young Elite Scientists Sponsorship Program by the China Association for Science and Technology (No: 2021QNRC001).

Data

The Microsoft Academic Graph (MAG) is adopted as the empirical dataset in this paper (Wang et al., 2020). MAG has been widely utilized in various scientometric studies (Huang, Lu, Liu, Cheng, & Bu, 2022) for its broad coverage, open access, and the well-designed field classification system (Paszcza, 2016). The MAG copy we adopt in the current study ranges from 1800 to February 2020. Based on all publications in this copy, we establish a citation network where publications A cites B would derive an edge from A to B. That being said, a node (publication) with a great value of in-degree indicates that the publication has received many citations. As we are targeting potential rich-club phenomenon in citations, we first explore the in-degree distribution (i.e., the distribution of the number of citations) where both the horizontal and the vertical axes are manipulated in a logarithmic scale as shown in Figure 1. In Figure 1, the relationship between in-degree and the number of nodes almost fits a straight curve. This indicates that the in-degree distribution of MAG follows a typical power-law distribution, echoing many extant studies in complex networks (Eom & Fortunato, 2011; Huang, Bu, Ding, & Lu, 2021).

Figure 1. Distribution of the in-degree (publications' citation counts) in the MAG publication citation network (inner). For a better visualization, we only show those whose in-degree cumulated percentage is lower than 95% (outer).



MAG has a hierarchical, six-level field classification system where each publication is assigned to each of the six levels. The six levels are labelled as L0 (the top level), L1, L2, L3, L4, and L5 (the bottom level). In this study, for simplicity, we adopt the L0-level fields (19 in total) to investigate field variance regarding the potential rich club coefficients.

In-degree rich club coefficient

To quantify the degree of rich club, if any, we employ the in-degree rich club coefficient ($\Phi(\mathrm{k})$) to obtain the rich club strength of a citation network:

$$\Phi(\mathrm{k}) = \frac{E_k}{N_{>k}(N_{>k}-1)} \tag{1}$$

where $N_{>k}$ refers to the number of nodes with in-degree higher than k and E_k indicates the number of citing relationships among all nodes within $N_{>k}$ scope. Using Eq. (1), we derive the in-degree rich club coefficient for each value of in-degree from zero to the maximum in-degree. From Eq. (1), we know that $\Phi(\mathrm{k})$ defines the connection density among those nodes (publications) whose in-degree is greater than a certain value (a.k.a., k), indicating the closeness among "rich people".

Null model

Colizza et al. (2006) have empirically observed that, even in a randomized Erdős–Rényi or Barabási-Albert network, the rich club coefficient increases when the degree increases. To this end, we need a null model to demonstrate expected in-degree rich club coefficient by randomly reshuffling edges in the real citation network and keeping the same in- and out-degree of nodes in the network. The random network "provides information about the overall rich-club ordering in the network with respect to an ideally uncorrelated graph" (Colizza et al., 2006, p. 112). The comparison between the real and the random networks differentiates the level of rich club phenomenon from the data itself or from the real citation relationship in the network. For the whole MAG citation data, we keep the year of citing paper in a citation relationship unchanged to randomize the citation network. For instance, suppose papers A cites B, and papers C cites D. If A and C were published in the same year, we then switch the citing relationship; that is to say, in the null model, A links to D, and B links to C.

For the randomization process of different fields, we first extract citation sub-networks for each field where all edges in this sub-network belong to the citation within the same field. This extraction follows the rule that both the citing and cited papers belong to the target field. Such randomization process repeats 19 times and we finally derive 19 random sub-networks for each field.

The definition of null models further defines a null-model-based in-degree rich club coefficient. In another word, in a random citation network, we can calculate its in-degree rich club coefficient. Then we delineate the **divided in-degree rich club coefficient** that equals the in-degree rich club coefficient in the real citation network divided by the in-degree rich club coefficient in the corresponding random network when the in-degree is same:

$$\Phi_{divided}(\mathrm{k}) = \frac{\frac{E_k}{N_{>k}(N_{>k}-1)}}{\frac{E_k^{'}}{N_{>k}^{'}(N_{>k}^{'}-1)}} \tag{2}$$

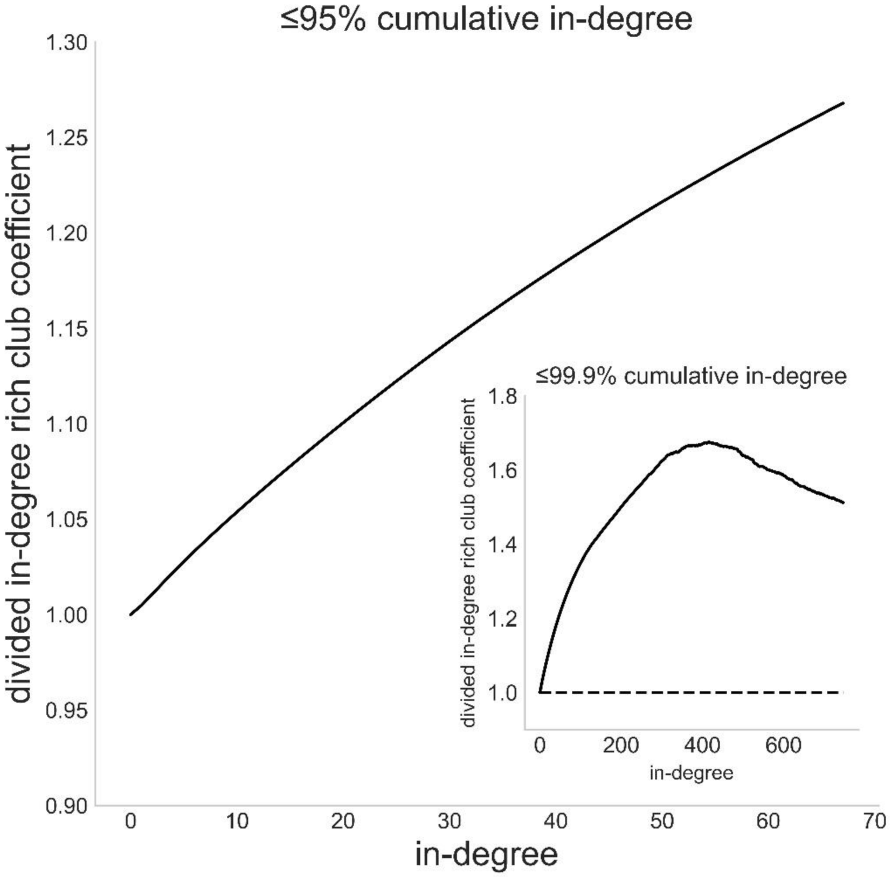
where $N_{>k}$ and $N_{>k}^{'}$ refer to the number of nodes with in-degree higher than k in the real and the random networks, respectively. E_k and $E_k^{'}$ indicate the number of citing relationships among all nodes within $N_{>k}$ and $N_{>k}^{'}$ scope, respectively.

In Eq. (2), if the value of the divided in-degree rich club coefficient is greater than one, the strength of rich club phenomenon in the real network is greater than that in the random network. In other words, the value of the divided in-degree rich club coefficient quantifies the level of the real network compared with the random network.

Overview

We plot the distribution of divided in-degree rich club coefficient for the whole MAG citing network in Figure 2. The cumulative 95% nodes are shown in the outer sub-figure while the cumulative 99.9% are in the inner sub-figure for a better visualization. As shown in both inner and outer sub-figure of Figure 2, the value is always greater than one, demonstrating the existence of the rich club coefficient in a publication citation network. Besides, the divided in-degree rich club coefficient raises with the increase of in-degree in the outer sub-figure, indicating that the rich club phenomenon becomes more recognizable when zooming into highly cited publication sets (the “richer communities”). Moreover, when 99.9% publications are involved, we observe that the divided in-degree rich club coefficient keeps increasing and only slightly decreases after the optimal point.

Figure 2. How the divided in-degree rich club coefficient changes with the increase of in-degree: (outer) remaining only nodes with 95% or less cumulative in-degree; (inner) remaining nodes with 99.9% or less cumulative in-degree. The dotted line in the inner sub-figure is a reference line indicating an equivalent value of in-degree rich club coefficients in the real and the random networks (i.e., divided in-degree rich club coefficient = 1.0).

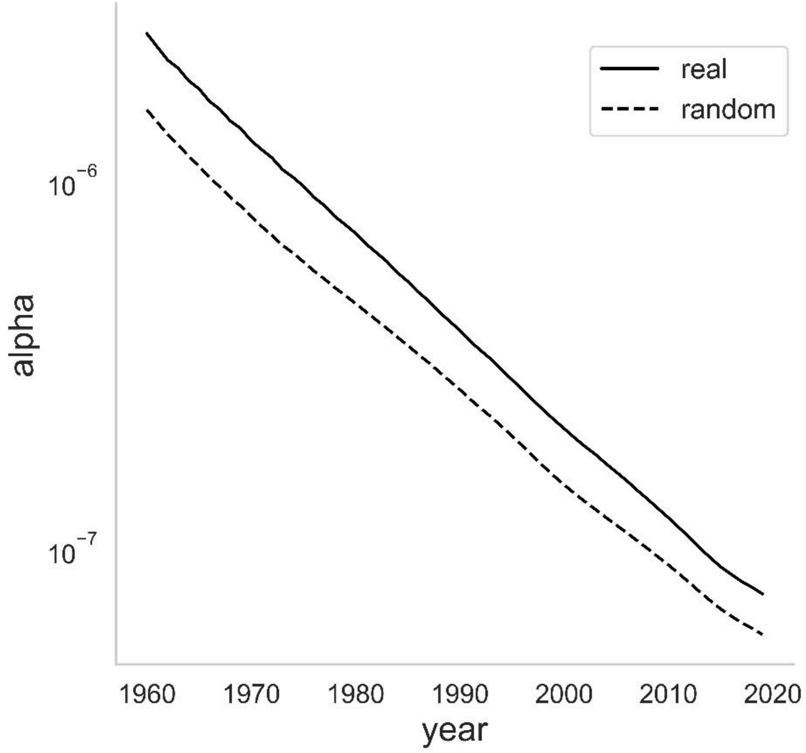


Quantifying temporal patterns of rich club

To quantify temporal patterns of the rich club phenomenon over years, we define **the Alpha index** as an effective indicator. We first build up the annual publication citation network where all nodes (publications) were published in or before a certain year and edges represent citing relations among these publications; hence, such a network is essentially temporally cumulative. In each annual network, we plot a curve like Figure 2 where the horizontal axis shows in-degree and the vertical axis indicates the divided in-degree rich-club coefficient. We then apply an Ordinary Least Square (OLS) method on these data points to derive the linear slope of these points. Alpha is defined as the OLS coefficient of the linear slope of this annual network. This would generate an alpha with each publication year. We then duplicate this process of establishing annual network for each year.

What does alpha connotate? Alpha means that when in-degree is increased by one, to what extent the in-degree rich club coefficient increases. From this point of view, alpha defines the margin increasing rate of the in-degree rich club coefficient. Figure 3 shows the temporal pattern of alpha where we observe that, in both real and random networks, the alpha declines with time but that the real network's alpha is always greater than the random one. Again, this supports our finding in Figure 2 in that the rich club phenomenon in citation networks exists.

Figure 3. The changes of alpha over time both in real network and random network. In the figure, we only involve those whose cumulated in-degree is smaller than 95% published between 1960 and 2019[2].



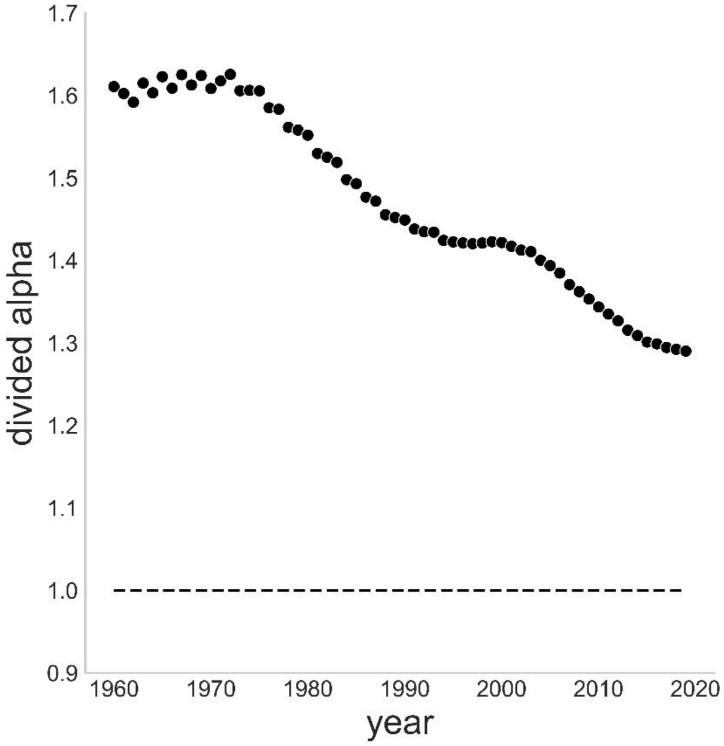
The decreasing trend of alpha also makes sense that is attributable to how we process the data. By definition, the citation network of one specific year is the subset of the citation network of the following year. Thus, the citation network grows larger and, of course, contains the original relation in previous years' networks. Rich clubs in previous years may not be densely connected to rich clubs in nowadays, partly because of the knowledge obsolescence (Gou, Meng, Chinchilla-Rodríguez, & Bu, 2021) and myopic referencing behaviour (Pan, Petersen, Pammolli, & Fortunato, 2018). The rich clubs of different periods, therefore, tend to separate from each other albeit the number of papers rapidly increases over years.

Furthermore, following the same idea in terms of comparing the in-degree rich club coefficient in real and random networks, we define the **divided alpha index** with the original alpha index in a real citation network divided by the alpha calculated from the random network for each year. The temporal pattern of the divided alpha is shown in Figure 4. Similar to Figure 2, the

[2] The reason for choosing 1960 as the start year for our analysis is that, before 1960, the number of citation pairs is very small compared to the whole MAG and the cumulated citation pairs only reach 0.5% in 1960. As for why we exclude all 2020 publications, that is due to the incomplete records of 2020 publications in our MAG copy.

dotted, vertical line (divided alpha = 1.0) is a reference line indicating an equivalent value of alpha in the real and the random networks.

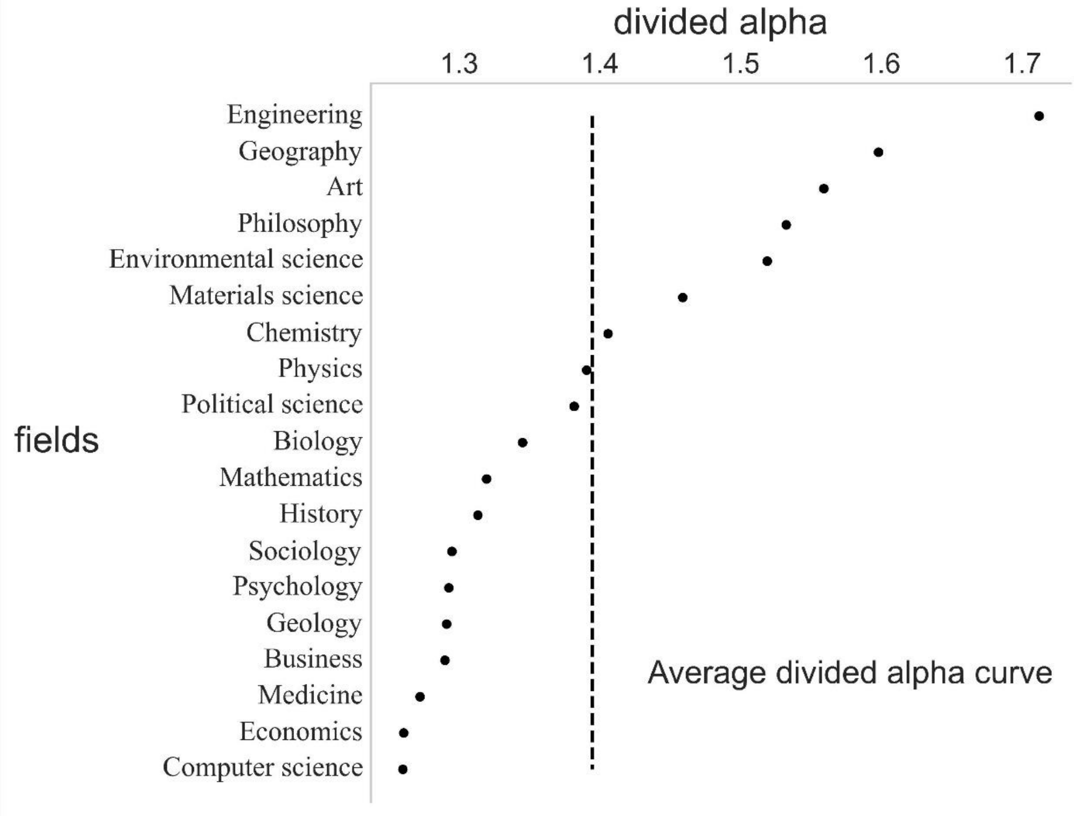
Figure 4. The changes of divided alpha over time.



From Figure 4, we see that the value of divided alpha is always greater than one, which is, again, consistent with our previous observation of the existence of rich club phenomenon. Meanwhile, the divided alpha emerges an overall downward trend, indicating that compared with random citation networks (baseline), the marginal effect intensity of rich club phenomenon is overall decreasing over years.

We repeat the same analytical strategy to observe the divided alpha in each field in Figure 5. The figure shows that the value of divided alpha is greater than one in all fields. We also see that Engineering tends to have the greatest value of divided alpha and that Computer science features the lowest. These illustrate that Engineering has the highest level of the marginal effect intensity of rich club phenomenon while Computer Science has the lowest level of such marginal effect. The low value of Computer Science may result from the tendency that computer scientists pursue the latest, state-of-art technologies and algorithms that makes the domain updates in a rapid manner. This characteristic may further link to the citing behavior in computer science and partly explain its low divided alpha. Moreover, Chemistry, Physics, and Political science have the medium level of the marginal effect as the points representing these fields are quite close to the average effect of the marginal level in all fields.

Figure 5. The divided alpha in each field. Average divided alpha in all fields is marked in the dotted curve.



Conclusions

This paper explores the rich club phenomenon in publication citation networks using MAG. We notice that the whole citation network shows a strong rich club phenomenon as divided in-degree rich club coefficient raises with the increase of in-degree of publications. Considering the temporal factor, we conclude that the divided alpha is decreasing as time approaches which indicates that the level of the rich club phenomenon is decreasing. Field-level analyses show various rich-club phenomena among disciplines. We clearly observe that Engineering, the level of rich club phenomenon is the highest while Computer science the lowest.

Among the limitation for the current research, when performing field-level analyses, we only include citing relationships where citing and cited papers belong to exactly the same field. This process discards among-field citation relationships that leads to an incomplete analysis. Interdisciplinary rich club phenomenon is thus neglected in the current analyses.

We observe that the divided alpha declines as the time goes by, but the mechanisms behind remains to be unclear. In the future, we are going to explore the reason for the decreasing divided alpha, among which might be the rise of the total number of papers. Meanwhile, we broadly compare the level of rich club phenomenon in 19 MAG L0 fields but fail to dive into details—for instance, different patterns of rich club phenomenon in various domains, as well as potential interpretations are missing.

Another important future work related to the current paper comes from the aggregation level. We are quite curious about the rich club phenomenon after grouping the citation record into different scholars (authors). Rich clubs, if any, in an author citation network would be quite intriguing as they shed lights on many aspects that offers important political and pedagogical implications for making better science.

Acknowledgements

We thank all members in the Knowledge Discovery (KD) lab at the Department of Information Management, Peking University for their constructive comments and fruitful discussions on this study. This work was supported by the high-performance computing platform of Peking University.

References

Colizza, V., Flammini, A., Serrano, M. A., & Vespignani, A. (2006). Detecting rich-club ordering in complex networks. *Nature Physics*, *2*(2), 110–115.

Csermely, P., London, A., Wu, L.-Y., & Uzzi, B. (2013). Structure and dynamics of core-periphery networks. *Journal of Complex Networks*, *1*(2), 93–123.

Eom, Y.-H., & Fortunato, S. (2011). Characterizing and Modeling Citation Dynamics. *PLOS ONE*, *6*(9), e24926.

Gou, Z., Meng, F., Chinchilla-Rodríguez, Z., & Bu, Y. (2021). Revisiting the Obsolescence Process of Individual Scientific Publications: Operationalisation and a Preliminary Cross-discipline Exploration. In Proceedings of the 18th International Conference on Scientometrics & Informetrics (ISSI 2021) (pp. 477–488), July 12-15, 2021, KU Leuven, Belgium.

Huang, Y., Bu, Y., Ding, Y., & Lu, W. (2021). Partitioning highly, medium and lowly cited publications. *Journal of Information Science*, *47*(5), 609–614. SAGE Publications Ltd.

Huang, Y., Lu, W., Liu, J., Cheng, Q., & Bu, Y. (2022). Towards transdisciplinary impact of scientific publications: A longitudinal, comprehensive, and large-scale analysis on Microsoft Academic Graph. *Information Processing & Management*, *59*(2), 102859.

Pan, R. K., Petersen, A. M., Pammolli, F., & Fortunato, S. (2018). The memory of science: Inflation, myopia, and the knowledge network. *Journal of Informetrics*, *12*(3), 656–678.

Paszcza, B. (2016, November 22). *Comparison of Microsoft Academic (Graph) with Web of Science, Scopus and Google Scholar* (masters). University of Southampton. Retrieved April 25, 2022, from https://eprints.soton.ac.uk/408647/

Shibata, N., Kajikawa, Y., & Matsushima, K. (2007). Topological analysis of citation networks to discover the future core articles. *Journal of the American Society for Information Science and Technology*, *58*(6), 872–882.

Smilkov, D., & Kocarev, L. (2010). Rich-club and page-club coefficients for directed graphs. *Physica A: Statistical Mechanics and its Applications*, *389*(11), 2290–2299.

Szell, M., & Sinatra, R. (2015). Research funding goes to rich clubs. *Proceedings of the National Academy of Sciences*, *112*(48), 14749–14750.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, *10*(2), 365–391.

Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y., & Kanakia, A. (2020). Microsoft Academic Graph: When experts are not enough. *Quantitative Science Studies*, *1*(1), 396–413.

Yan, E., & Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *Journal of the American Society for Information Science and Technology*, *63*(7), 1313–1326.

Zhou, S., & Mondragon, R. J. (2004). The rich-club phenomenon in the Internet topology. *IEEE Communications Letters*, *8*(3), 180–182. Presented at the IEEE Communications Letters.