J SYST SCI SYST ENG
Vol. 30, No. 6, December 2021, pp. 712–727
DOI: https://doi.org/10.1007/s11518-021-5509-1

ISSN: 1004-3756 (paper), 1861-9576 (online)
CN 11-2983/N

Quantifying the Research Diversification of Physicists

Jianlin Zhou,[a] Ying Fan[b]

[a]School of Economics and Management, China University of Geosciences (Beijing), Beijing 100083, China
jianlinzhou@cugb.edu.cn
[b]School of Systems Science, Beijing Normal University, Beijing 100875, China
yfan@bnu.edu.cn (✉)

Abstract. Scientists may shift research interests and span multiple research areas in their careers, reflecting the research diversification of scientists. Quantifying the scientists' research diversity can help to understand the research patterns of scientists. In this paper, we study the research diversification of scientists in Physics based on the Physics and Astronomy Classification Scheme (PACS) which can well reflect the research topics of physics papers. For each scientist, we first build a PACS codes co-occurrence network and reveal the research diversity by analyzing the connectivity and community structure of this network. Then we use diversity indicators to measure the research diversification of scientists and analyze the distribution of each indicator. Finally, we investigate the relationship between scientists' diversity indicators and their scientific impact using multiple regression analysis. The results show that the numbers of connected components of most PACS codes co-occurrence networks are less than 5, and some networks have significant community structures. The diversity indicators show the heterogeneity of the research diversity of physicists. We also find that some diversity indicators are weakly correlated with scientific impact indicators. Based on our findings, we suggest that physicists should focus on their main research fields and span multiple research fields over their entire careers which could promote their scientific impact.

Keywords: Research diversification, co-occurrence network, community structure, scientific impact

1. Introduction

Scientists often adopt different research strategies in their careers (Foster et al. 2015, Chakraborty et al. 2015), taking into account their abilities and the surrounding research environment. Some researchers have involved many research topics in their careers, and they like to work with others to solve some problems with interdisciplinary nature. Some scientists concentrate on several research topics in their careers. These research behaviors reflect the research diversity of scientists, which will influence the creation of scientific knowledge and may have an important impact on scientists' career development. In order to reveal the hidden research patterns of scientists and seek successful career strategies, quantifying the research diversification of scientists becomes more necessary and significative. By analyzing the research diversification of scientists, it can also make people understand the research process of scientists more clearly.

The increased availability of large-scale datasets has provided an opportunity to explore the behavioral patterns of scientists. A series of behaviors have been studied, including collaboration among scientists (Milojevic 2014, Li et al. 2020), submission patterns (Calcagno et al. 2012), the behavior of citing paper (Uzzi et al. 2013), career movements (Deville et al. 2014, Gomez et al. 2020), etc. Research diversity, as one of the important research patterns, has also been extensively studied. Currently, most studies analyze the research diversification of scientists mainly based on the research topics of the paper (Moschini et al. 2020, Deng and Xia 2020). Chakraborty et al. (2015) use the Microsoft Academic Search (MAS) Engine to categorize papers of computer science domain into 24 fields and analyze the diversity of a researcher's scientific career. They find that most scientists participate in various research fields or focus on very few fields, and highly cited scientists are involved in diverse fields over their entire career but concentrate on one or two fields at any given time. Abramo et al. conducted a series of studies on the research diversity of scientists by using a database of Italian professors in the sciences and using the Web of Science (WoS) subject category to identify the research topics or disciplines of a paper. They investigate the research diversification of scientists in different disciplines from three dimensions: extent of diversification, intensity of diversification, relatedness ratio (Abramo et al. 2017). They find that the above three aspects vary among disciplines. They also analyze the effects of gender, age, academic rank, and multidisciplinary collaborations on research diversification, and find that these factors play different roles in influencing the research diversity of scientists (Abramo et al. 2018ab). Jamali et al. (2020) use the Australian Fields of Research (FoR) codes assigned to journals to calculate the diversification of Australian professors' publications. They find that there is a correlation between the research diversity and impact, publication counts of scientists. In addition, Zeng et al. (2019) apply community analysis in the co-citing network of individual scientist to identify the research topics and investigate the research dynamics of scientists. In addition to exploring the research diversity of researchers, there are also many studies to quantify the interdisciplinarity or diversity of publications and investigate the relationship between the diversity and impact of scientific articles (Zeng et al. 2017). Herron et al. (2016) investigate the relationship between research diversification and scientific impact in Nanoscience at the national level. Zhang et al. (2021) also study the influence of interdisciplinarity on scientific impact of publications and find that interdisciplinarity has a positive effect on both citation and broader impact.

Previous studies are mainly based on the journal-level classification systems such as Web of Science (WoS) subject categories to study the research diversification of scientists. This type of classification system assigns one or more research fields (or topics) to a journal, but it does not directly assign research fields to the publication (Waltman and Van Eck 2012). In this case, the research fields of the publication are determined by the research fields to which the journal belongs, which makes the research fields of the publication not specific and detailed enough. But some disciplines have their own classification systems which are publication-level classification systems, such as the Medical Subject Headings (MeSH), the Physics and Astronomy Classification Scheme (PACS), the Mathematics Subject Classification (MSC), the Chemical Abstracts sections, and the Journal of Economic Literature (JEL) classification system. These classification systems have a more detailed subject categories and they directly assign individual publications to research fields. In order to ensure the accuracy and reliability of the empirical results, we think that these more detailed field classification systems are more suitable for studying the research diversity of scientists. PACS is a fine-grained level of subject classification in physics and it can identify fields and subfields of physics well (Smith 2019). Here we study the research diversification of physicists based on the PACS codes. Moreover, when analyzing the research diversification of scientists, past studies mainly calculate the number of research fields covered in the papers but ignore the relationship among the research fields of the papers. In this study, we consider the co-occurrence relationship of research fields and connect them by building a co-occurrence network of research fields.

In this paper, we propose to quantify the research diversification of scientists considering the number of topics and the co-occurrence relationship among research topics. Here, we apply network analysis and diversity indicators to quantify the research diversity of physicists. We first build a PACS codes co-occurrence network for individual scientist based on empirical data and then analyze the network's structural characteristics. The results show that the PACS codes co-occurrence networks may contain multiple connected components, in which any two nodes are connected by paths, and significant community structure. Then we calculate several common diversity indicators for each scientist and display their distribution. We finally investigate the relation between the research diversification and the impact of scientists. We find that some diversity indicators show a weak positive correlation with the scientific impact of scientists.

The main contributions of this paper are the following two points. Firstly, this study gives a more detailed and comprehensive insight into the research diversification of physicists by analyzing PACS codes co-occurrence networks and calculating diversity indicators. Secondly, the study reveals the relationship between research diversity and scientific impact of physicists, which can guide researchers to formulate future career plans.

The paper proceeds as follows: Section 2 introduces the empirical data and illustrates our method used to analyze the research diversification of physicists in this study. Section 3 presents the empirical results. Section 4 presents our comments on the paper. Finally, we conclude this paper in Section 5.

2. Methodology

This paper mainly analyzes the diversification of physicists from two dimensions: the relevance of research topics, and the diversity in the type and number of research topics. In terms of topic relevance, we construct a PACS codes co-occurrence network for individual scientist considering the co-occurrence relationship between research topics which means that two research topics appear in the same paper. We reveal the research diversity of individual physicists by analyzing connectivity and community structure of their PACS codes co-occurrence networks. In terms of the diversity in the type and number of research topics, we adopt four common indicators to measure the research diversity of individual physicists: number of different PACS codes, ratio of papers involved in main topics, Simpson diversity index, and Shannon entropy. Our analysis method is universal. As long as the research topics of each paper can be accurately identified, readers can use our method to analyze the research diversity of scientists. The detailed description of our method is presented in the following subsections.

2.1 Dataset

In this paper, we use the dataset provided by American Physical Society (APS), which involves nine representative physics journals: Physical Review A, B, C, D, E, Letters, Series I & II, Special Topics, and Reviews of Modern Physics. The data contains over 450,000 papers, ranging from year 1893 to year 2010. We can get the information of each paper from the dataset, such as title, author names, affiliations, printed time, received time, references, PACS codes, and so on. To remove the influence of author name ambiguity on the analytical results, we use the author name dataset which Sinatra et al. (2016) have conducted a comprehensive disambiguation process in the APS data, and 236,884 authors are found in this dataset. We determine our research objects mainly based on the following two considerations. Firstly, the research interest or direction of physicists is not stable enough at an early stage of their career, which makes the quantitative results of the research diversification of physicists at this stage unreliable. Secondly, to eliminate authors that leave research at an early stage of their career, in this paper we limit our analysis to scientists that (i) have at least 10 publications, (ii) their career time is no less than 5 years. Moreover, their published papers all contain PACS codes. In the end, a total of 11,020 scientists in the APS dataset are identified as our research objects.

The Physics and Astronomy Classification Scheme (PACS) was introduced into Physics in 1975, and its function was to organize subject indexes (Smith 2019). The basic form of a PACS code is "XX.YY.ZZ", where "XXYY" are all integers with values from 0 to 9, "ZZ" are alphanumerical (Radicchi and Castellano 2011). PACS code itself has a hierarchical structure, and it is divided into four levels of research topics, which can reflect the subfields of physics well. The first element "X" represents the first level of topics, the first two elements "XX" represents the second level of topics, the first four elements "XX.YY" represents the third level of topics, and the entire six-digit PACS code represents the fourth level of topics. From the first level of topics to the fourth level of topics, the number of topics in each level is increasing, and the corresponding research scope is getting narrower. For example, in the PACS code "89.75.Fb", the first digit "8" denotes "Interdisciplinary physics and related areas of science and technology", "89" represents "Other areas of applied and interdisciplinary physics", "89.75" denotes "Complex systems" and "89.75.Fb" denotes "Structures and organization in complex systems". In this study, we will make use of the first four digits of the PACS codes to represent the topics of publications, because this level of topics could represent the subfields of physics well and keep relatively stable. In addition, 328,210 papers contain the PACS codes in the dataset we used. Among these papers, 9.71% of the papers contain only one PACS code, 22.8% of the papers contain two PACS codes, 38.18% of the papers have three PACS codes, and 29.21% of the papers contain four PACS codes.

2.2 Network Construction and Community Detection

In this paper, we construct for each individual scientist a PACS codes co-occurrence network to reveal the research diversification of scientists. The nodes in the co-occurrence network are the PACS nodes involved in the all published papers of a scientist, and two nodes are linked if they appear together in one or more papers written by this scientist (Pan et al. 2012). Considering that two PACS codes may appear in multiple papers and some articles contain only one PACS code, so the PACS codes co-occurrence network we built is an undirected, weighted, and self-looped network. The weight of the link between the nodes i and j can be defined as

$$\omega_{ij} = \sum_{\alpha} \frac{1}{n_\alpha - 1} \tag{1}$$

where α is the paper in which PACS codes i and j appear, and n_α is the number of different PACS codes in the article α. If node i has a self-loop, it shows that some papers of individual scientist only have the PACS code i and the self-loop weight ω_{ii} of node i is equal to the number of such articles above. In this PACS codes co-occurrence network, it can be observed that the strength of node i, $s_i = \sum_j \omega_{ij}$, is exactly equal to the number of papers which contain the PACS code i. For the PACS codes co-occurrence network of a scientist, it can be a disconnected graph, because the PACS codes in some papers are entirely different from those in other papers of the scientist, which results in multiple connected components.

The PACS codes co-occurrence network exhibits the research topics in which the scientist works and the relationship among these topics. To explore whether the research topics of a scientist show the clustering phenomenon, we will perform community detection analysis based on the topology of the PACS codes co-occurrence network. Currently, there are so many methods to detect communities based on the network structure such as modularity optimization, statistical inference, spectral methods (Fortunato and Hric 2016). In this study, we use a fast algorithm proposed by Newman (2004) to optimize weighted modularity for detecting communities. The common form of weighted modularity (Rubinov and Sporns 2011) is defined as follows:

$$Q = \frac{1}{2W}\sum_{ij}\left(\omega_{ij} - \frac{s_i s_j}{2W}\right)\delta\left(C_i, C_j\right) \qquad (2)$$

where ω_{ij} is the link weight between nodes i and j, $W = \frac{1}{2}\sum_{ij}\omega_{ij}$, $s_i = \sum_j \omega_{ij}$, C_i and C_j indicate the communities to which node i and j belong, $\delta\left(C_i, C_j\right) = 1$ if $C_i = C_j$ and 0 otherwise. This fast detection algorithm is an agglomerative hierarchical clustering method. It first treats each node as a community, and then repeatedly joins communities together in pairs, which can optimize the weighted modularity, until all the nodes are merged into a community. At each merge step, one can get a partition and calculate the weighted modularity. Finally, we choose the partition which has the largest weighted modularity value as the optimal partition results.

2.3 Indicators of Diversity Measures

Considering the number and type of PACS codes into which all publications of a scientist are involved, we use the following four indicators to measure the research diversity of scientists. The first indicator is the number of different PACS codes in the publications of a scientist, which describes the number of topics covered by a scientist. The more PACS codes a scientist is involved in, the more diverse his or her research activity is. The second indicator is the proportion of papers containing the main research topics of a scientist. When the value of this index is large, it indicates that the output of the scientist is mainly concentrated on the main research topics, and this scientist's research activity tends to be specific. And when the value of this index is small, it shows that the scientist's research activity tends to diversify. In this study, we define the main research topics of a scientist as those PACS codes that cover the number of papers in all publications of the scientist exceeds the average number of papers covered by each PACS code.

The other two indicators are the Simpson diversity index (SDI) (Simpson 1949) and Shannon entropy (SE) (Shannon 1948), which can be used to measure scientists' research diversity (Chakraborty et al. 2015) . The Simpson diversity index of a scientist can be calculated by the following formula:

$$SDI = 1 - \sum_{i=1}^{m} p_i^2 \qquad (3)$$

where m is the number of different PACS codes covered in the all published papers of the scientist. p_i is the ratio of the scientist's number of papers containing the PACS code i to the total number of papers written by the scientist and $\sum_{i=1}^{m} p_i = 1$. If a publication α contains PACS code i and has n_α^i PACS codes, we assume that only $\frac{1}{n_\alpha^i}$ paper covers PACS code i. So p_i can be defined as:

$$p_i = \frac{\sum_\alpha \frac{1}{n_\alpha^i}}{N} \qquad (4)$$

where N is the total number of publications of the scientist, n_α^i is the number of different PACS codes in the article α which covers PACS code i. The Shannon entropy of the scientist can also be calculated based on m and p_i, and its calculation formula is as follows:

$$SE = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad (5)$$

The range of SDI and SE are $[0, 1 - 1/m]$ and $[0, \log_2 m]$, respectively. When the scientist only involves one PACS code, The SDI and SE both get the minimum value of 0. For a given m, when the scientist publishes the same

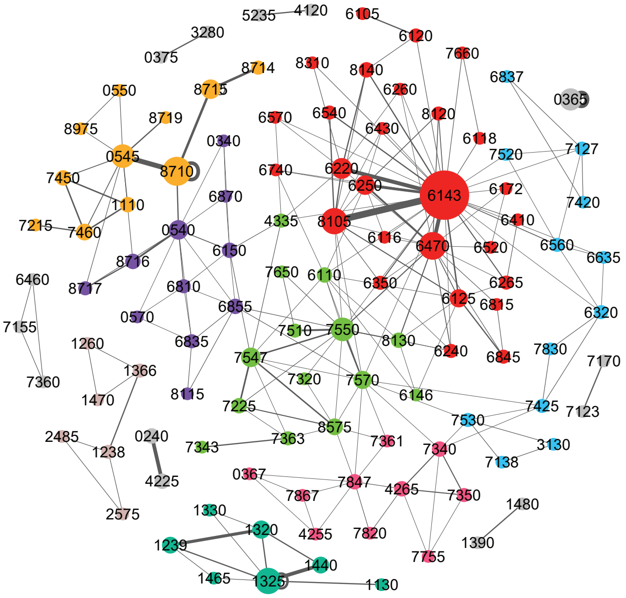

Figure 1 The Visualization of the PACS Codes Co-Occurrence Network of a Certain Physicist [1]

number of papers for each PACS code, the SDI and SE get the maximum value. For the scientists with the same number of PACS codes, the closer the scientist is to the maximum values of SEI and SD, the more diverse his or her scientific research is.

3. Results

3.1 The Connectivity and Community Structure of PACS Codes Co-occurrence Networks

In this article, 11,020 PACS codes co-occurrence networks have been established. We first give a visualization of the PACS codes co-occurrence network of a certain physicist in Figure 1. We can find that the PACS codes co-occurrence network of this physicist is a disconnected graph and has an obvious community structure. We observe that many networks are not connected graphs. The connectivity of the PACS codes co-occurrence network can reflect the correlation among scientist's research topics to some extent. If there are multiple connected components in this network, it may indicate that the scientist performs research from multiple unrelated directions. We first investigate the distribution of the number of connected components in those networks we built, and the result is shown in Figure 2 (A). One can see that nearly 70% of networks have only one or two connected components, and less than 10% of networks have over four connected components. The giant component of PACS codes co-occurrence network usually represents a maximum number of interrelated research topics which the scientist have studied. For each network, we calculate the ratio of the size of the giant component to the entire network, and the corresponding distribution is as shown in Figure 2 (B). It can be found that there are 90% giant components whose size is more than half of the size of their corresponding networks.

Considering that the studies of a scientist are likely to focus on several main research areas during his or her career, that is, the PACS codes involved in his or her published papers

[1] This network is an undirected, weighted, and self-looped network, where nodes represent PACS codes involved in the papers the physicist published, and the link between two nodes shows that these two PACS codes have ever appeared in a paper. The node colors represent different communities, and the node labels show the first four digits of the PACS codes. The node size is proportional to the node's strength, and the link's width is proportional to the link weight.

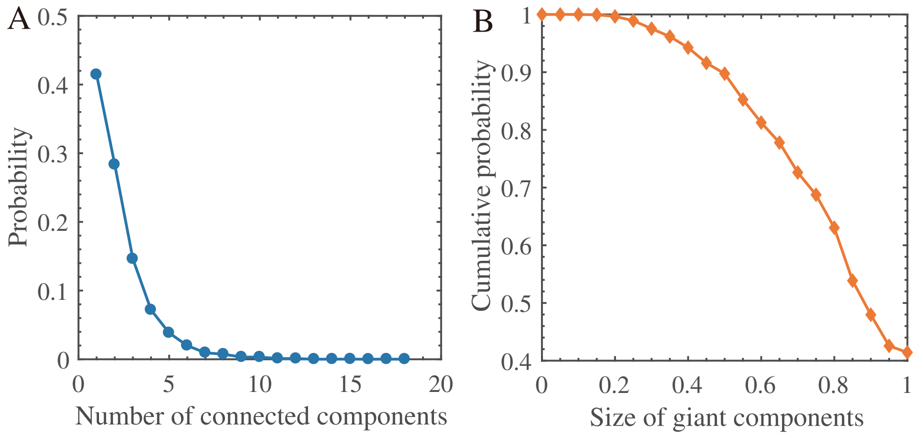

Figure 2 (Color Online) The Basic Statistical Properties of the PACS Codes Co-Occurrence Networks [1]

[1] (A) The distribution of the number of connected components in these networks. (B) The cumulative distribution of the size of giant components in these networks.

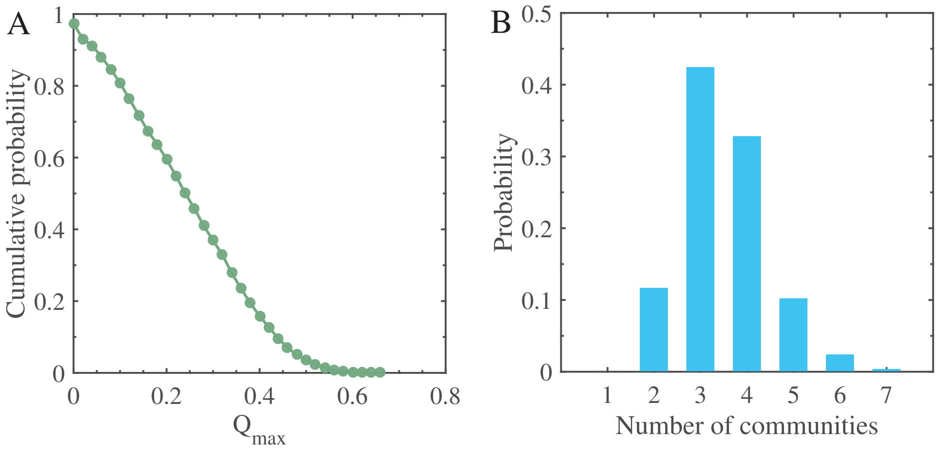

Figure 3 (Color Online) The Community Structure of the PACS Codes Co-Occurrence Networks Which Contain Only One Connected Component[1]

[1] (A) The cumulative distribution of maximum modularity values of these PACS codes co-occurrence networks. (B) The distribution of the number of communities in these networks.

tend to cluster into communities. To verify whether there exists a community structure in the PACS codes co-occurrence network, in this paper, we apply the fast detection algorithm proposed by Newman to those networks which contain only one connected component. The distribution of maximum modularity Q values is shown in Figure 3 (A). In practice, if the maximum Q value of a real network is higher than 0.3, there will exist a significant community structure in a network (Chen and Redner 2010). We find that the Q values of 36.88% of 4566 networks are higher than 0.3, which shows that there are meaningful communities in the PACS codes co-occurrence networks of some scientists. We also calculate the distribution of the number of communities for the networks whose maximum Q values are higher than 0.3 in Figure 3 (B). One can see that over 75% of these networks have three or four communities, and less than 5% of these networks have over five communities.

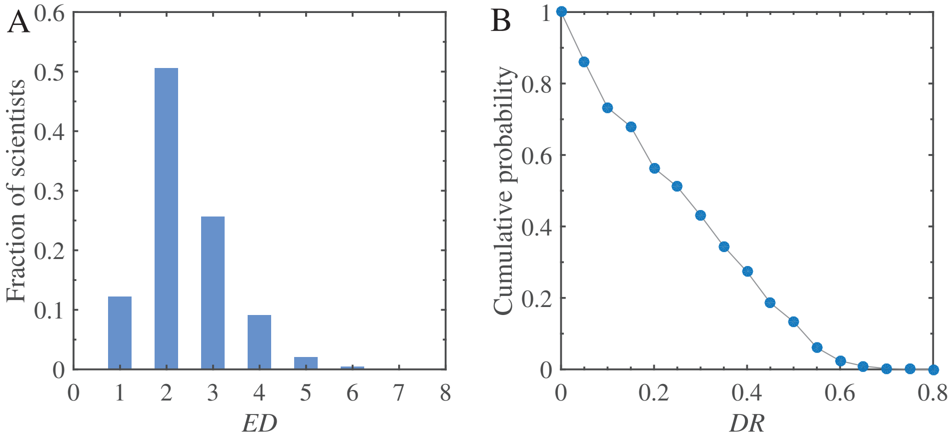

Figure 4 (Color Online) The Distributions of *ED* and *DR* [1]

[1] (A) The distribution of *ED*. (B) The cumulative distribution of *DR*.

3.2 Research Diversity of Scientists

We first use two indicators proposed by Abramo et al. (2017) to measure the research diversification of physicists from the perspective of journals' subject categories. The first indicator is the extent of diversification (ED), which is calculated by the number of subject categories of journals covered in the published papers of individual scientist. The second indicator is the diversification ratio (DR), which is calculated by the proportion of papers belonging to non-dominant subject categories. The dominant subject category for a scientist means the most recurrent subject category involved in his or her published papers. The value of *ED* is not less than 1, and it is also an integer. The value of *DR* can vary between 0 and 1. The higher the values of *ED* and *DR*, the greater the diversity of scientist's research activity. Based on the Web of Science subject categories of APS journals shown in Table 1, we calculate *ED* and *DR* values for each physicist and then display the distributions of *ED* and *DR* in Figure 4. We can observe that the maximum value of *ED* is 7 and the minimum value of *ED* is 1. There are about 50% of scientists whose *ED* values are 2 and no more than 5% of scientists whose *ED* values are over 4. For the *DR* values of scientists, we can see that the cumulative probability of *DR* decreases approximately linearly with the increase of *DR*, and there are very few scientists whose *DR* values exceed 0.6. In addition, the Spearman's correlation coefficient between *ED* and *DR* is 0.712, which means there is a high correlation between *ED* and *DR*. That is to say, if a scientist has a relatively large *ED* value, then his or her *DR* value will also be relatively large. This analysis method can reveal the research diversity of scientists to some extent. However, there are few journals in APS dataset, which cannot reflect the research topics of scientists in detail, making the analysis method based on the subject categories of journals limited.

PACS was developed by the American Institute of Physics (AIP) and used to identify fields and sub-fields of physics since 1970s. It is a more fine-grained level of categorization and works at the level of individual publications. Since PACS numbers are attributed to papers by authors themselves, this ensures that we can distinguish the research fields of different APS papers based on their PACS numbers in most cases. We further make use of diversity indicators based on PACS codes to analyze the research diversity of scientists. For each scientist, we first calculate their diversity indicators and perform descriptive statistical analysis on

Table 1 The Web of Science (WoS) Subject Categories of APS Journals

Journals	WoS Subject Categories
Physical Review A	Physics, Atomic, Molecular & Chemical; Optics
Physical Review B	Materials Science, Multidisciplinary; Physics, Applied; Physics, Condensed Matter
Physical Review C	Physics, Nuclear
Physical Review D	Astronomy & Astrophysics; Physics, Particles & Fields
Physical Review E	Physics, Fluids & Plasmas; Physics, Mathematical
Physical Review Letters	Physics, Multidisciplinary
Physical Review Special Topics-Accelerators and Beams	Physics, Nuclear; Physics, Particles & Fields
Physical Review Special Topics-Physics Education Research	Education & Educational Research; Education, Scientific Disciplines
Reviews of Modern Physics	Physics, Multidisciplinary

Table 2 Descriptive Statistics of Diversity Indicators

Indicators	Minimum	Maximum	Mean	First quartile	Median	Third quartile	Variance
Number of different PACS codes	1	112	17.519	11	16	22	82.394
Ratio of papers involved in main topics	0	1	0.907	0.857	0.923	1	0.010
Simpson diversity index	0	0.986	0.846	0.811	0.875	0.913	0.011
Shannon entropy	0	6.408	3.368	2.856	3.435	3.927	0.640

each indicator. The basic statistical results are shown in Table 2. We can find that the average of the number of different PACS codes is larger than its median, while the other three indicators are just the opposite. Then we also analyze the distribution characteristics of different diversity indicators. The cumulative distribution of total number of PACS codes is presented in Figure 5 (A). It can be seen that the distribution of the number of PACS codes is a heavy-tailed distribution, which indicates that only a few scientists participate in a large number of research topics. There are nearly 70% of scientists whose number of different PACS codes is less than 20, and no more than 3% of scientists have over 40 PACS codes. Next, we identify the main PACS codes of each scientist, which can reflect the main research topics of scientists. The distribution of the ratio of papers involved in the main PACS codes in Figure 5 (B) shows that more than 80% of scientists whose proportion of papers containing the main PACS codes is higher than 0.8. This indicates that many scientists have invested much effort into

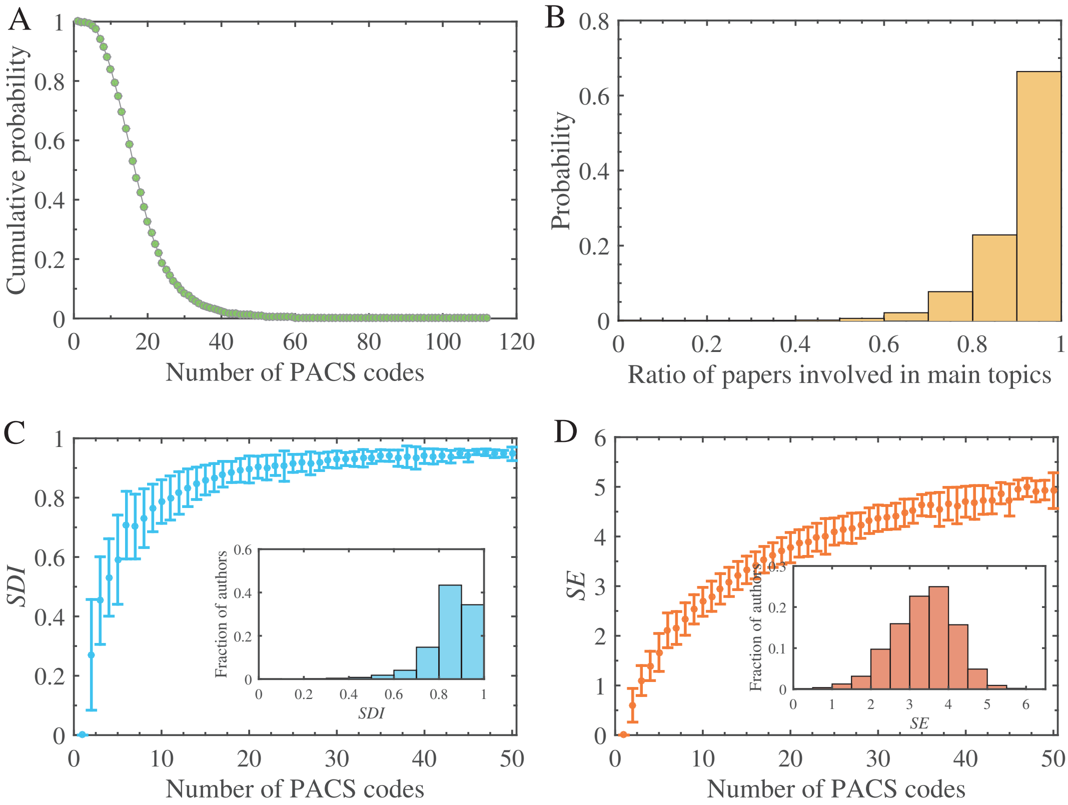

Figure 5 (Color Online) The Distributions of Diversity Indicators of Scientists[1]

their main research topics. Since the calculation of Simpson diversity index and Shannon entropy is related to the number of the scientist's PACS codes, we show the relation between these two indicators and the number of PACS codes by plotting the average indicator of the researchers with the same number of PACS codes in Fig. 5 (C) and (D). One can see that the average *SDI* first quickly increases with the number of PACS codes and then slowly increases. When the number of PACS codes is less than 10, the variance of *SDI* values of scientists with a different number of PACS codes is relatively large. We also find that the average *SE* increases with the number of PACS codes. We also investigate the distributions of the *SDI* and *SE* of all scientists. One can observe that more than 75% of scientists whose Simpson diversity index values lie in 0.8-1 and about 80% of the scientists whose values of Shannon entropy are mainly concentrated in 2.5-4.5. Here, we also provide some criteria to judge whether the *SDI* value or *SE* value of the scientist is a "big" value, as shown below: 1) His or her diversity index value is greater than the average index value of all scientists; 2) His or her diversity index value is greater than the average index value of scientists with the same number of PACS codes. Finally, we investigate the Spearman rank correlation between different diversity indicators, and the results are shown in Table 3. One can observe that the diversity indicators based on PACS codes are weakly correlated with the diversity indicators

[1] (A) The cumulative distribution of the number of PACS codes. (B) The distribution of the ratio of papers involved in main topics. (C) shows the average *SDI* with the same number of PACS codes and the distribution of *SDI*. The error bars in this figure represent standard deviations. (D) shows the average *SE* with the same number of PACS codes and the distribution of *SE*. The error bars in this figure represent standard deviations.

Table 3 Spearman's Correlation Coefficients between Diversity Indicators

	Extent of diversification	Diversification ratio	Number of different PACS codes	Ratio of papers involved in main topics	Simpson diversity index	Shannon entropy
Extent of diversification	1.000	0.712**	0.350**	–0.237**	0.254**	0.301**
Diversification ratio	0.712**	1.000	0.271**	–0.128**	0.224**	0.255**
Number of different PACS codes	0.350**	0.271**	1.000	-0.255**	0.838**	0.929**
Ratio of papers involved in main topics	–0.237**	–0.128**	–0.255**	1.000	–0.296**	–0.304**
Simpson diversity index	0.254**	0.224**	0.838**	–0.296**	1.000	0.974**
Shannon entropy	0.301**	0.255**	0.929**	–0.304**	0.974**	1.000

**. Correlation is significant at the 0.01 level (2-tailed).

based on journals' subject categories. Among the diversity indicators based on PACS codes, there is a strong positive correlation between the number of different PACS codes, Simpson diversity index, and Shannon entropy. We also find that the ratio of papers involved in main topics shows a negative and weak correlation with other diversity indicators.

3.3 Correlation between Research Diversification and Impact of Scientists

After calculating the diversity indicators of scientists, we try to explore the relationship between research diversification and the scientific impact of scientists. Many factors will affect the relationship between research diversification and the scientific impact of scientists, such as the number of papers published, the career length, and the country in which the scientists work. For example, senior scientists who have produced more papers will on average have higher h-indexes and cover more research topics. Authors in countries with big scientific communities may also publish more papers, receive higher citations, and participate in more research topics. In this paper, we use two indicators: h-index and the average number of citations per publication to represent the scientific impact of a scientist. We first calculate the correlation between the diversity indicators and scientific impact indicators, and their Spearman correlation matrix is shown in Table 4. The calculated Spearman's correlation coefficients show that the four diversity indicators are weakly correlated with the scientific impact indicators. Next, we perform a multiple regression analysis further to verify the relationship between diversity indicators and

Table 4 Spearman's Correlation Coefficients between Diversity Indicators and Scientific Impact

	Number of different PACS codes	Ratio of papers involved in main topics	Simpson diversity index	Shannon entropy
h-index	0.314**	0.092**	0.090**	0.155**
average number of citations per publication	0.029**	0.063**	–0.022*	–0.010

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

impact indicators. The dependent variable is the h-index or the average number of citations per publication, and the explanatory variables are four diversity indicators, the number of papers published, the career length, and the country in which the scientists work. Since the country in which the scientists work is a categorical variable, we use dummy variables to indicate the country where the scientists are located when performing regression analysis. We use stepwise multiple linear regression to select significant variables. The regression results are shown in Table 5 and Table 6. From Table 5, we can see that *SDI* and *SE* do not appear in the regression equation, which indicates that there is no significant relationship between these two diversity indicators and h-index. The number of different PACS codes and the ratio of papers involved in main topics show positive effects on the h-index. Their standardized regression coefficients are 0.042 and 0.049, respectively. When we explore the effect of diversity indicators on the average number of citations per publication, among the four diversity indicators, only ratio of papers involved in main topics is retained in the regression equation, and it also exerts a positive influence on the average number of citations per publication, which its standardized regression coefficient is 0.038. In conclusion, based on regression analysis, we can see that there is a weak positive correlation between some diversity indicators and scientific impact indicators of scientists.

4. Discussion

This paper analyzes the research diversification of physicists and explores the relationship between this diversity and the impact of physicists. The results show that the PACS codes co-occurrence networks of some physicists have obvious community structures, which means that the research of the scientist is concentrated in several main directions. A previous study (Zeng et al. 2019) also shows the similar results at paper level and they find that the co-citing network of papers of a scientist exhibits a clear community structure where each major community represents a research topic. In addition, one of important findings by analyzing the relationship between diversity indicators and the impact of physicists is that the ratio of papers involved in main topics is weakly and positively correlated with the h-index and the average number of citations per publication, which implies that focusing on the main research topics can increase the impact of scientists. Jamali et al. (2020) also find that the ratio of papers involved in main topics has a positive but weak correlation with the number of publications and 10% most frequently cited paper. They also find that the number of different topics per publication is negatively correlated with output and citation impact (total citations, 10% most frequently cited papers, and total ci-

Table 5 The Regression Results for the h-index as the Dependent Variable

R	R^2	Adjusted R^2	F	Sig.	
0.731	0.534	0.533	572.964	0.000	
Explanatory variable	Unstandardized coefficient	Std. error	Standardized coefficient	t	Sig.
Number of different PACS codes	0.014	0.003	0.042	4.811	0.000
Ratio of papers involved in main topics	1.580	0.223	0.049	7.101	0.000
Simpson diversity index	-	-	-	-	-
Shannon entropy	-	-	-	-	-
Number of papers published	0.221	0.003	0.717	86.996	0.000
Career length	–0.092	0.004	–0.147	–20.855	0.000

Table 6 The Regression Results for the Average Number of Citations Per Publication as the Dependent Variable

R	R^2	Adjusted R^2	F	Sig.	
0.318	0.101	0.099	58.943	0.000	
Explanatory variable	Unstandardized coefficient	Std. error	Standardized coefficient	t	Sig.
Number of different PACS codes	-	-	-	-	-
Ratio of papers involved in main topics	1.953	0.478	0.038	4.088	0.000
Simpson diversity index	-	-	-	-	-
Shannon entropy	-	-	-	-	-
Number of papers published	0.080	0.005	0.163	17.090	0.000
Career length	–0.262	0.009	–0.264	–27.658	0.000

tations received in the first 3 years after publications) of scientists. But in our study we find that the number of different PACS codes is positively correlated with the h-index. We think that this difference is caused by the empirical data used and the definition of the impact indicator of scientists.

The Physics and Astronomy Classification Scheme (PACS) is a hierarchical subject classification scheme used to classify and categorize the journal articles in physics and astronomy. Since there may be more than one PACS code in a Physical Review paper, considering the co-occurrence relationship among PACS codes, we build PACS codes co-occurrence networks to study the research diversification of scientists. Some disciplines use subject classification codes directly to distinguish research topics or fields of papers. We can use similar analysis methods in this study to explore the research diversity of scientists in other disciplines. Take the field of economics as an example, we first collect all economic papers published by an economist. Based on the Journal of Economic Literature (JEL) codes contained in these papers, we build a JEL codes co-occurrence network considering the co-occurrence relationship between JEL codes and calculate the diversity indicators. We reveal the research diversity of economists by analyzing the structural characteristics of the co-occurrence network and analyzing the value of diversity indicators.

The present study can be helpful to guide the reality. Firstly, our empirical investigation reveals the research diversification of physicists, which enables researchers to understand the behavioral patterns of physicists, and further help physics beginners to make future career plans. Secondly, the present study shows that spanning more PACS codes or increasing the ratio of papers involved in main topics can promote the scientific impact of physicists. Therefore, in order to have a successful career, we suggest that physicists should focus on the main research fields while spanning multiple research fields. Thirdly, physicists can use our methods to analyze their own research diversification at different career stages and they can decide to adjust their future career plans according to their own circumstances.

This study has several limitations: (i) For any author, we do not distinguish his or her contribution to each paper. In this case, when an author as a collaborator participates in multiple papers, the research fields involved in these papers may not be his or her real research fields. (ii) We only analyze those physicists whose published papers all contain PACS codes. But in reality, not all APS papers contain PACS codes. We need to design a reasonable automated algorithm to extract the research fields of the publications and then quantify the research diversity of scientists.

This paper provides a new perspective for analyzing the research diversification of physicists and reveals many behavioral characteristics of physicists. The above analysis is mainly based on the static PACS codes co-occurrence networks. However, the PACS codes co-occurrence network of each scientist is dynamically changing. In other words, the research interests or topics of scientists change over time. At present, few studies have investigated the macroscopic and microscopic dynamics of research-interest or topic evolution (Jia et al. 2017, Zeng et al. 2019). In the future, we will explore the evolution mechanism of individual PACS codes co-occurrence network and explore different research strategies adopted by scientists at different career stages. Finally, we will also try to reveal the effect of these different research strategies on the successful career of scientists.

5. Conclusions

Scientists often involve multiple research topics in their scientific careers. To quantify the research diversification of scientists, many studies have proposed various diversity indicators based on journals' subject categories. In this paper, we analyze the research diversification of physicists based on the PACS codes. We mainly investigate the research diversification of scientists by analyzing the co-occurrence network structure and calculating the diversity indicators. The network structure analysis shows that some PACS codes co-occurrence networks have obvious community structures, which these networks are mainly divided into three or four communities. It indicates that many scientists do not randomly choose research topics, but they focus on several research directions in their careers. We also observe that most scientists are involved in no more than 25 PACS codes, and only about 40% of the networks is a connected graph. The correlation analysis between diversity indicators shows that the diversity indicators based on the PACS codes are weakly correlated with the diversity indicators based on the subject categories of journals, which indicates that there are significant differences between the two types of diversity indicators. By investigating the relationship between research diversification and the scientific impact of scientists using correlation analysis and regression analysis, the results show that research diversification can affect the impact of scientists. The number of different PACS codes is weakly correlated with the *h*-index, and the ratio of papers involved in main topics is weakly correlated with the *h*-index and the average number of citations per publication.

Acknowledgments

This work has been supported by the MOE (Ministry of Education in China) Liberal Arts and Social Sciences Foundation under Grant No. 20YJC870015, the Fundamental Research Funds for the Central Universities under Grant No. 2652019010, and the National Natural Science Foundation of China under Grant No. 61573065. The authors sincerely thank the referees for their help to improve the quality of this paper.

References

Abramo G, D'Angelo C A, Di Costa F (2017). Specialization versus diversification in research activities: The extent, intensity and relatedness of field diversification by individual scientists. *Scientometrics* 112(3): 1403-1418.

Abramo G, D'Angelo C A, Di Costa F (2018). The effects of gender, age and academic rank on research diversification. *Scientometrics* 114(2): 373-387.

Abramo G, D'Angelo C A, Di Costa F (2018). The effect of multidisciplinary collaborations on research diversification. *Scientometrics* 116(1): 423-433.

Calcagno V, Demoinet E, Gollner K, Guidi L, Ruths D, de Mazancourt C (2012). Flows of research manuscripts among scientific journals reveal hidden submission patterns. *Science* 338(6110): 1065-1069.

Chakraborty T, Tammana V, Ganguly N, Mukherjee A (2015). Understanding and modeling diverse scientific careers of researchers. *Journal of Informetrics* 9(1): 69-78.

Chen P, Redner S (2010). Community structure of the physical review citation network. *Journal of Informetrics* 4(3): 278-290.

Deng S, Xia S (2020). Mapping the interdisciplinarity in information behavior research: a quantitative study using diversity measure and co-occurrence analysis. *Scientometrics* 124(1): 489-513.

Deville P, Wang D, Sinatra R, Song C, Blondel V D, Barabási A L (2014). Career on the move: Geography, stratification, and scientific impact. *Scientific Reports* 4: 4770.

Fortunato S, Hric D (2016). Community detection in networks: A user guide. *Physics Reports* 659: 1-44.

Foster J G, Rzhetsky A, Evans J A (2015). Tradition and innovation in scientists' research strategies. *American Sociological Review* 80(5): 875-908.

Gomez C J, Herman A C, Parigi P (2020). Moving more, but closer: Mapping the growing regionalization of global scientific mobility using ORCID. *Journal of Informetrics* 14(3): 101044.

Herron P, Mehta A, Cao C, Lenoir T (2016). Research diversification and impact: The case of national nanoscience development. *Scientometrics* 109(2): 629-659.

Jamali H R, Abbasi A, Bornmann L (2020). Research diversification and its relationship with publication counts and impact: A case study based on Australian professors. *Journal of Information Science* 46(1): 131-144.

Jia T, Wang D, Szymanski B K (2017). Quantifying patterns of research-interest evolution. *Nature Human Behaviour* 1(4): 0078.

Li J, Yin Y, Fortunato S, Wang D (2020). Scientific elite revisited: Patterns of productivity, collaboration, authorship and impact. *Journal of the Royal Society Interface* 17(165): 20200135.

Milojević S (2014). Principles of scientific research team formation and evolution. *Proceedings of the National Academy of Sciences of the United States of America* 111(11): 3984-3989.

Moschini U, Fenialdi E, Daraio C, Ruocco G, Molinari E (2020). A comparison of three multidisciplinarity indices based on the diversity of Scopus subject areas of authors' documents, their bibliography and their citing papers. *Scientometrics* 125(2): 1145-1158.

Newman M E (2004). Fast algorithm for detecting community structure in networks. *Physical Review E* 69(6): 066133.

Pan R K, Sinha S, Kaski K, Saramäki J (2012). The evolution of interdisciplinarity in physics research. *Scientific Reports* 2: 551.

Radicchi F, Castellano C (2011). Rescaling citations of publications in physics. *Physical Review E* 83(4): 046116.

Rubinov M, Sporns O (2011). Weight-conserving characterization of complex functional brain networks. *NeuroImage* 56(4): 2068-2079.

Shannon C E (1948). A mathematical theory of communication. *Bell System Technical Journal* 27(3): 379-423.

Simpson E H (1949). Measurement of diversity. *Nature* 163(4148): 688-688.

Sinatra R, Wang D, Deville P, Song C, Barabási A L (2016). Quantifying the evolution of individual scientific impact. *Science* 354(6312): aaf5239.

Smith A (2019). From PACS to PhySH. *Nature Reviews Physics* 1(1): 8-11.

Uzzi B, Mukherjee S, Stringer M, Jones B (2013). Atypical combinations and scientific impact. *Science* 342(6157): 468-472.

Waltman L, Van Eck N J (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology* 63(12): 2378-2392.

Zhang L, Sun B, Jiang L, Huang Y (2021). On the relationship between interdisciplinarity and impact: Distinct effects on academic and broader impact. *Research Evaluation* rvab007.

Zeng A, Shen Z, Zhou J, Fan Y, Di Z, Wang Y, Stanley H E, Havlin S. (2019). Increasing trend of scientists to switch between topics. *Nature Communications* 10(1): 3439.

Zeng A, Shen Z, Zhou J, Wu J, Fan Y, Wang Y, Stanley H E (2017). The science of science: From the perspective of complex systems. *Physics Reports* 714: 1-73.

Jianlin Zhou is currently a postdoc with the School of Economics and Management, China University of Geosciences (Beijing). He received the Ph.D. degree from Beijing Normal University, in 2019. His research interests include complexity science, complex network, science of science, and social science computing.

Ying Fan is a professor of the School of Systems Science, Beijing Normal University, China. She received her Ph.D. degree in systems science from Beijing Normal University. She is the Deputy Secretary-General and the Executive Director of the China Systems Engineering Society. Her research interests include the self-organization theory, nonlinear dynamics, and complex networks.