Understanding Success through the Diversity of Collaborators and the Milestone of Career

Yi Bu

School of Informatics and Computing, Indiana University, Bloomington, IN., U.S.A.

Ying Ding

School of Informatics and Computing, Indiana University, Bloomington, IN., U.S.A.

School of Information Management, Wuhan University, Wuhan, Hubei, China

University Library, Tongji University, Shanghai, China

Jian Xu*

School of Information Management, Sun Yat-sen University, Guangzhou, Guangdong, China

Xingkun Liang

Department of Information Management, Peking University, Beijing, China

Gege Gao

School of Informatics and Computing, Indiana University, Bloomington, IN., U.S.A.

Yiming Zhao

Center for the Study of Information Resources, Wuhan University, Wuhan, Hubei, China

Correspondence concerning this article should be addressed to Dr. Jian Xu,
Affiliation and address: School of Information Management, Sun Yat-sen University, Guangzhou, Guangdong 510006, China. Email: xujianonline@gmail.com

Understanding Success through the Diversity of Collaborators and the Milestone of Career

Abstract: Scientific collaboration is vital to many fields, and it is common to see scholars seek out experienced researchers or experts in a domain with whom they can share knowledge, experience, and resources. To explore the diversity of research collaborations, this paper performs a temporal analysis on the scientific careers of researchers in the field of computer science. Specifically, we analyze collaborators using two indicators: the research topic diversity, measured by the Author-Conference-Topic model and cosine, and the impact diversity, measured by the normalized standard deviation of h-indices. We find that the collaborators of high-impact researchers tend to study diverse research topics and have diverse h-indices. Moreover, by setting Ph.D. graduation as an important milestone in researchers' careers, we examine several indicators related to scientific collaboration and their effects on a career. The results show that collaborating with authoritative authors plays an important role prior to a researcher's Ph.D. graduation, but working with non-authoritative authors carries more weight after Ph.D. graduation.

Keywords: scientific collaboration, topic diversity, impact diversity, Ph.D., team of team science, Exploratory Factor Analysis (EFA), scientific career, scholarly communication, scientometrics

INTRODUCTION

The number of collaborative studies in different scientific fields has been increasing for decades (Newman, 2004; de Solla Price and de Beaver, 1966). As pointed out in Bozeman and Boardman's (2014) "collaboration imperative", more than 90 percent of the publications in science, technology, and engineering feature more than one author.

Indeed, collaboration has become mandatory in many fields, where research success depends heavily on teamwork between scientists with mixed skillsets, such as between theoreticians, and those experienced with lab experiments. A team consisting of these scholars may have an advantage in terms of research innovation and success than each individual working alone (Bozeman and Boardman, 2014).

Scientific collaboration not only enhances the quality of research (Kumar and Ratnavelu, 2016), but also breeds innovation (Ding, 2011), making it crucial to a successful scientific career. The old saying—"standing on the shoulders of giants"—tells us that scholars should collaborate with authoritative authors (experienced researchers or expert in a domain) because they can provide profound knowledge, experience, and resources (Adegbola, 2013). Many scholars try to meet and work with authoritative authors in their domains in hopes that it might lead to a better scientific career; But due to advanced age or cost issues concerning their research, these experienced researchers, being often well-established and invested in their own study and status, are not necessarily anxious to try new ideas (Packalen and Bhattacharya, 2015), an attitude that can impede innovation (Chinchilla-Rodriguez, Ferligoj, Miguel, Kronegger, and de Moya-Anegon, 2012). This circumstance begs the question: Should scholars seek to collaborate with experienced researchers? If so, should they *only* focus on collaborating with experienced researchers? We use the h-index (Hirsch, 2005) to measure the impact of authors' collaborators and the normalized standard deviation (NSD) of their h-indices to measure the impact diversity of their collaborators. We find that although high-impact scholars often collaborate with experienced researchers, the impact of collaborators on these high-impact authors are diverse.

Many studies have pointed out that transdisciplinary scientific collaboration can allow researchers to exchange innovative ideas and methods in various fields (Bridle, Vrieling, Cardillo, Araya, and Hinojosa, 2013; Xu, Ding, and Malic, 2015a; Ding and

Stirling, 2016). An interconnected world needs new approaches to intellectual inquiry that can challenge common disciplinary and institutional boundaries (Davoudi and Pendlebury, 2010). On the contrary, others have illustrated some of the drawbacks of cross-disciplinary study, such as being more time-consuming (Schaltegger, Beckmann, and Hansen, 2013). Other obstacles include attitude, communication (e.g. jargon, intellectual turf, leadership, facilitating interactions), and academic and professional barriers (e.g. publications and professional organizations, funding, peer review, and career development and training) (Institute of Medicine, 2000). These studies, however, generally focus on the discipline or domain level. While there can be many different topics within a discipline or domain, the relations between these topics in terms of researchers' impact have not been deeply explored in studies of collaborator diversity. In this paper, we analyze the collaborators' research diversity at the topic level, using the Author-Conference-Topic (ACT) model (Tang, Jin, and Zhang, 2008a) to extract the authors' topic distributions and cosine to measure the diversity. The results of our empirical studies show that high-impact authors have a tendency to collaborate with those having different research topics. In other words, the research topics of high-impact scholars' collaborators are more diverse compared with the collaborators of those authors having lower impact.

Scientific collaboration can be distinct depending on the stage of researchers' scientific careers. When researchers pursue a Ph.D., for example, they may be eager to collaborate with top scholars, i.e. authoritative authors (AAs), in their domains to gain valuable experience and enhance their careers. After graduation and while working in universities or other research institutes, researchers may spend more time with their students or postdocs to accomplish projects or produce publications. As their research interests evolve over time, they may also collaborate with new researchers (Baker and Pifer, 2011). Identifying all of these changes in scientific collaboration could be relevant or important to career trajectory because they may

provide temporal clues for a successful career. In this paper, we analyze high-impact researchers in the field of computer science before and after their Ph.D. graduation—an important milestone of their scientific careers—to identify important indicators to their success at different stages in their careers (Costas, Nane, and Larivière, 2015). We find that these high-impact authors prefer to collaborate with many researchers, and also have extensive experience working with authoritative authors before their Ph.D. graduation, but they generally regard collaborations with non-authoritative authors as an important focus of their studies after receiving their doctoral degrees.

This paper is outlined as follows. Related work is discussed from the perspective of scientific collaborators' research topic diversity, collaborators' impact diversity, and scientific collaboration among different stages of researchers' careers. The dataset and methodology used in this paper are then described. Results are discussed and compared with existing related studies. Finally, conclusions and suggestions for future work are offered.

RELATED WORK

Collaborators' Research Topic Similarity/Diversity in Scientific Collaboration

Many researchers have demonstrated the relation between authors' research interests and their scientific collaboration. Applying qualitative methods such as interview and observation, Kraut, Egido, and Galegher (1988) found that scientists' sharing of similar research interests encourages ongoing collaboration. However, limited by the development of quantitative algorithms for extracting the researchers' research topics/interests, such research has long stayed at a qualitative level. Newman (2004) used network science theories and methodologies to analyze coauthorship networks, and proposed that different disciplines feature distinct distributions of collaborator

numbers for scientists in biology, physics, and mathematics. Different from previous studies in which domain- or discipline-level information is applied as the measurement of authors' research interests (Newman, 2004; Milojević, 2010), Ding (2011) mined topic-level information into a coauthorship network analysis using the Author-Conference-Topic (ACT) model proposed by Tang et al. (2008a). She found that in the information retrieval field, productive authors prefer to collaborate with those with whom they share similar research interests. By employing Exponential Random Graph Models (ERGMs) and the ACT model, and including all-impact authors in the field of information retrieval instead of only the productive authors per Ding (2011), Zhang, Bu, and Ding (2016) concluded that research topic similarity does not necessarily affect which scholars one author will collaborate with. Besides the field of information retrieval, research interest similarity/diversity of collaborators has also been studied in library and information science (Huang and Chang, 2011), health sciences (Lee, McDonald, Anderson, and Tarczy-Hornoch, 2009), the social sciences (Bredereck et al., 2014), and cognitive science (Derry, Schunn, and Gernbacher, 2014). Other similar studies exploring the relationships between authors' research interests and their scientific collaboration include those of Huang, Zhuang, Li, and Giles (2008), Bird et al. (2009), and Sie, Drachsler, Bitter-Rijpkema, and Sloep (2012).

Although collaborating with researchers who do not share the same interests may be more time-consuming, error-prone, and resource-intensive (Schaltegger et al., 2013; Xu et al., 2015a), several studies claimed that diverse research topics of collaborators offer many potential benefits (Stokols, Harvey, Gress, Fugua, and Phillips, 2005; Wickson, Carew, and Russell, 2006; Kessel and Rosenfield, 2008; Pohl, 2007; Adegbola, 2013). For example, Pohl (2005), who studied collaborators' topic diversity in environmental research using qualitative interviews, argued that most thinking about collaborators with diverse research topics takes place at the level of program

management and problem solving. Xu et al. (2015a) posited that in problem-oriented fields, transdisciplinarity—which essentially requires some levels of research topic diversity of collaborators—is regarded as a means to solve complex research questions through a broader exchange of ideas, theoretical approaches, and best practices (Bridle et al., 2013). Most of these studies, however, were either qualitative in their approach or ignore the correlations between researchers' success (impact) and research interest diversity of their collaborators. To effectively address this gap, this paper explores the relationships between the impact of the researchers as well as the degree of their collaborators' research interest diversity to quantitatively measure the collaborators' research interests at a topic level.

Collaborators' Impact Similarity/Diversity in Scientific Collaboration

Research focusing on the similarity and diversity of researchers' impact and their scientific collaborations has been explored in two main branches of study. The first branch mainly demonstrates the relation between collaborators' productivity and collaboration, where de Solla Price and de Beaver (1966), for example, investigated the collaboration of informal publications (mostly article preprints) between members in health-related domains. They found a correlation between the authors' number of publications and their number of collaborative articles. Zuckerman (1967) indicated that the higher an author's number of scientific collaborations is, the more papers he/she publishes. Pravdić and Oluić-Vuković (1986) explored the relations between scientific output and collaboration in the field of chemistry and found that authors' productivity is dependent to a large degree on the frequency of their collaborations. Ebadi and Schiffauerova (2015) investigated the roles of researchers occupying important positions in the collaboration network and discovered that highly productive researchers not only have many collaborators but also perform an essential role in connecting other researchers in a network.

The second main branch of studies illustrates the relation between collaborators' number of citations and their collaborations. For example, Thurman and Birkinshaw (2006) found that the number of citations is significantly correlated with the number of coauthors for the six top journal articles in the field of medicine. Leimu and Koricheva (2005), however, did not find a significant positive correlation between the influence of collaborations and the impact of the resulting work in the field of ecology. Similarly, Ding (2011) argued that authors with a large number of citations do not generally coauthor with each other in the field of information retrieval. Different from previous research that has largely focused on the article level, Zhang et al. (2016)'s analysis at the author level found that the number of citations one author has received does not influence other authors' collaboration preference for him/her. Similar studies concerning the similarity/diversity of researchers' impact and their scientific collaborations include those of Pao (1982), Lee and Bozeman (2005), and Freeman and Huang (2014).

Some scholars have identified the relation between researchers' scientific impact and their collaborators' impact diversity. Adegbola (2013) argued that multi-ethnic, diverse scholars working collaboratively can benefit each other through creating innovation and actions that reduce research disparities, a process that also provides the potentially profound knowledge and experience in working with giants in a domain. These scholars can thus grow as rising stars by collaborating with high-impact authors, pushing them into better domain development (Kram and Isabella, 1985; Quatman and Chelladurai, 2008; Adegbola, 2010; Amjad et al., 2017). Similar studies focusing on the relationships between researchers' scientific impact and their collaborators' impact diversity include those of Dannerfer, Uhlenberg, Foner, and Abeles (2005), and Mccaughrean, Zinnecker, Andersen, Meeus, and Lodieu (2002).

Scientific Collaboration among Different Stages of Scientific Career

At the beginning of a career, scientific collaboration typically involves advisor-advisee relationships. At this early stage in their careers, researchers are usually doctoral students who take responsibility for the predominant research load while their mentors provide guidance (Kumar and Ratnavelu, 2016) in this mentoring relationship (Hart, 2000). Muschallik and Pull (2016) pointed out that such relationships help increase the productivity of advisees. Under this stage, if junior researchers can build new contacts they might obtain more opportunities to establish collaborations and work on new projects (Wang et al., 2010). Although several of these studies concluded that collaboration with a senior researcher is helpful for the career of the junior scholar, Packalen and Bhattacharya (2015) argued that senior researchers, especially older experienced researchers, are seldom open to investigating new ideas.

Specifically, when researchers become older or more established, their scientific collaborations are indeed subject to change. Hamermesh (2015) observed different research styles of older researchers compared with those of the younger researchers, which might be attributed to different “interpersonal relationships” as well as skills (Krapf, 2015). Kumar and Ratnavelu (2016) found that researchers who have spent more than ten years in their current affiliations had a smaller proportion of scientific collaborations then those who did not.

Studies exploring scientific collaboration during different stages of scientific careers include those of Holgate (2012), who proposed that young scholars should identify the “key people” in a collaboration and develop lasting relationships with them, especially with authoritative authors in a domain at the beginning of one’s research career. Using examples from French scientific leaders from 1799 to 1830, de Beaver and Rosen (1979) illustrated that collaborative associations with elite members in

France had facilitated the visibility of the young scientists. Luukkonen, Persson, and Sivertsen (1992) showed the significance of international scientific collaboration for both junior and senior scholars in terms of cognitive, social, historical, geopolitical, and economic factors. Loannidis, Boyack, and Klavans (2014) analyzed uninterrupted and continuous presence" (UCP)—the phenomenon of maintaining a continuous stream of publications—for researchers in the entire Scopus database and argued that different collaborators in distinct stages of careers may affect researchers' UCP. Yet few studies have addressed the changes in scientific collaboration as well as how to benefit from different scientific collaborations during different stages of researchers' careers. This paper divides the careers of the researchers in the field of computer science into two parts, before and after Ph.D. graduation, and identifies important scientific-collaboration-related indicators to their success in each career stage.

METHODOLOGY

Data

The dataset used in this article is derived from the AMiner platform (Tang et al., 2008b) which contains scientific publications in the field of computer science within the Association for Computer Machinery (ACM) from 1936 to 2014. It contains 2,092,356 papers, 1,207,061 unique authors, and 8,024,869 citation relationships. For author name disambiguation, we apply Tang, Fong, Wang, and Zhang (2012)'s algorithm, in which a unified probabilistic framework is proposed employing Markov Random Fields (Kindermann and Snell, 1980) and the number of actual authors are estimated. Both content-based information and structure-based information are considered as features with corresponding weights, where two parameter estimation steps are included to disambiguate the authors' name, that of estimating the weights of feature functions and assigning papers to different authors.

Figure 1 displays the relation between the logarithm of the authors' h-indices and the logarithm of the number of authors with the same h-index, where they show an approximate power law distribution. We select as high-impact authors those with h-indices greater than or equal to ten, in order to better analyze high-impact authors and identify how their careers developed before and after Ph.D. graduation. Finally, 8,621 authors are selected to form the author set in the analysis of "*Scientific Collaboration before and after Ph.D. Graduation*" section.

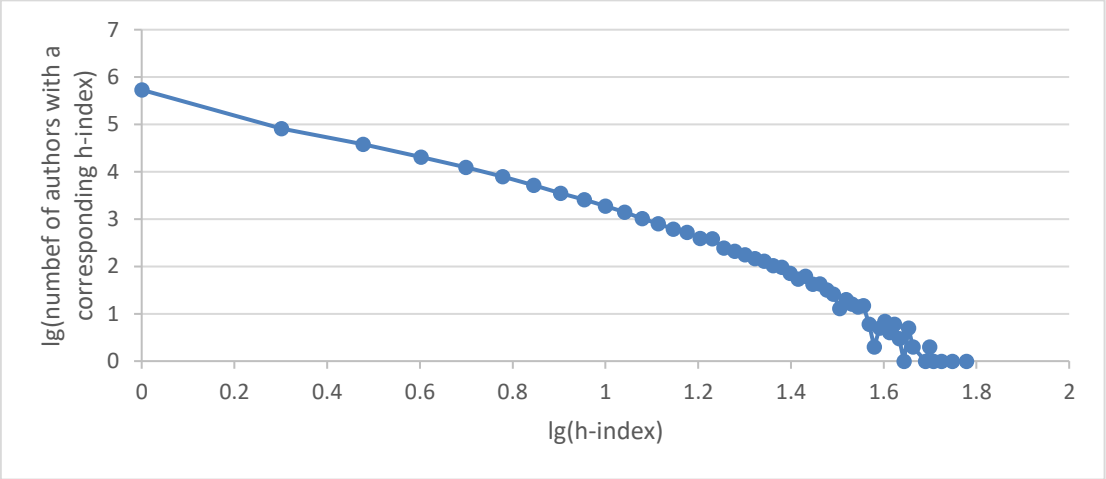

Figure 1. Authors' h-indices and the number of authors with the corresponding h-index in the dataset (log-log scale).

Methods

This paper separately explores three topics, the first being the relation between the impact of researchers and the topic diversity of their collaborators, measured using the Author-Conference-Topic (ACT) model (Tang et al., 2008). Another topic is the relation between the impact of the researchers and the topic diversity of their collaborators, measured using the normalized standard deviation of h-index; note that the topic/impact diversity here is calculated among the scholar's collaborators instead of between the scholar and his/her collaborators. The last topic is the different patterns of scientific collaboration before and after Ph.D. graduation, for which we apply Exponential Factor Analysis (Cattell, 1965) to visualize and analyze various selected

impact indicators.

Topic Diversity Measurement

Figure 2 shows the graphic notation of the ACT model used in this study, in which d represents document, w is word, c is the publication venue of the document, a_d is the set of co-authors, N_d is the number of word tokens in document d, x is the author, z is the topic, A, T, and D are number of unique authors, number of topics, and number of documents, respectively, α, β, and μ are hyper parameters, θ and φ are multinomial distributions over topics and words, respectively, and ψ is a multinomial distribution over publication venues (Tang et al., 2008a). In the output of the ACT model, each author is represented as an m-dimension vector, and each component of the vector represents the distribution possibility of the author on this topic. By employing the ACT model, we can obtain the topic distribution of each author, shown as vectors.

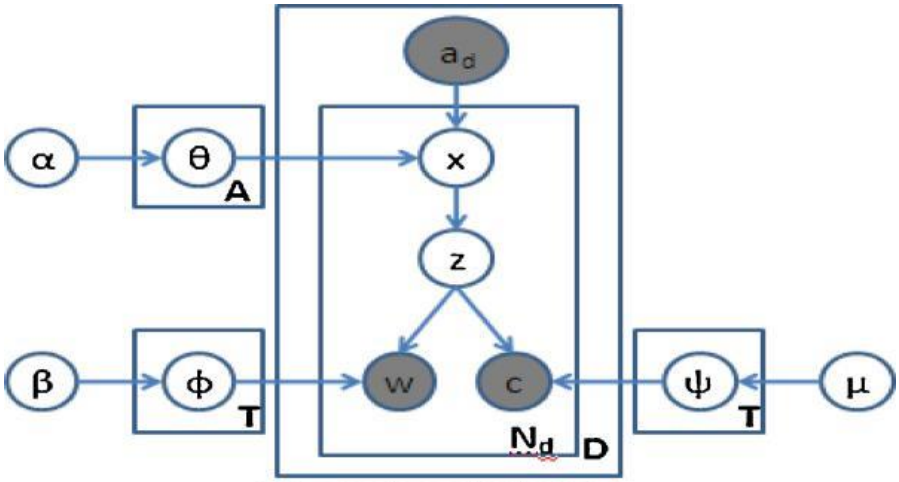

Figure 2. The Author-Conference-Topic (ACT) model (Tang et al., 2008a).

We then apply cosine similarity to measure the research topic similarities between the authors' collaborators. Specifically, the cosine between topic distributions of the collaborators is calculated. Suppose that author A has n collaborators, a_1, a_2, …,

a_n. The number of topics we select is m (equal to the number of components in the vector). The research topic vector of A's collaborators, $a_i(i = 1,2, \dots, n)$, is represented as $(t_{i,1}, t_{i,2}, \dots, t_{i,m})$. The topic diversity of A's collaborators, $TD(A)$, is calculated as:

$$TD(A) = \begin{cases} \frac{2}{n(n-1)} \sum_{i \neq j} cos[(t_{i,1}, t_{i,2}, \dots, t_{i,m}), (t_{j,1}, t_{j,2}, \dots, t_{j,m})], & n > 1 \\ 0, & n = 0,1 \end{cases} \quad (1)$$

where $cos[(t_1, t_2, \dots, t_m), (t_{i,1}, t_{i,2}, \dots, t_{i,m})]$ refers to the cosine similarity between these two vectors:

$$cos[\big((t_{i,1}, t_{i,2}, \dots, t_{i,m}\big), (t_{j,1}, t_{j,2}, \dots, t_{j,m})] = \frac{\sum_{k=1}^{m} t_{i,k} t_{j,k}}{\sqrt{\sum_{k=1}^{m} {t_{i,k}}^2} \cdot \sqrt{\sum_{k=1}^{m} {t_{j,k}}^2}} \quad (2)$$

Impact Diversity Measurement

We use the normalized standard deviation (NSD) to indicate the degree of impact diversity among the collaborators an author works with, where the h-index is applied to indicate the impact of the collaborators. According to Hirsch (2005), if a researcher with n publications has h publications that have received $\geq h$ times of citations and has other $(n - h)$ publications that have received $\leq h$ times of citations, his/her h-index is defined as h. Mathematically, assume that author A has n collaborators, $a_1, a_2, \dots, a_n$, and the h-index of collaborator a_i is h_i $(i = 1,2, \dots, n)$; author A has collaborated with a_i for x_i times. The impact diversity of A's collaborators, $ID(A)$, which is de facto the h-index NSD of A's collaborators, should be calculated as:

$$ID(A) = \frac{\sum_{i=1}^{n} \left(h_i x_i - avg(A)\right)^2}{\sum_{i=1}^{n} x_i} \quad (3)$$

where $avg(A) = \frac{\sum_{i=1}^{n} h_i x_i}{\sum_{i=1}^{n} x_i}$. Essentially NSD (i.e. $ID(A)$) here is different from the

standard deviation since the number of collaborations instead of the number of distinct collaborators is calculated.

Scientific Collaboration before and after Ph.D. Graduation

We select several collaboration-related indicators of high-impact authors to identify the different roles these indicators play during different stages of scientific careers. Specifically, Table 1 shows the indicators that can be divided into three types: general indicators (Type I), including the number of publications, the number of citations, and h-index; indicators showing scientific careers *before* Ph.D. graduation (Type II); and indicators showing scientific careers *after* Ph.D. graduation (Type III). Type II and III indicators include: simple indicators, such as the number of publications/citations before/after Ph.D. graduation; collaboration with authoritative authors (AAs, see details in the end of this paragraph) before/after Ph.D. graduation, such as the number of collaborations with AAs and the number of unique non-AA collaborators; collaboration with non-authoritative authors (non-AAs) before/after Ph.D. graduation, such as the number of collaborations with non-AAs and the number of unique non-AA collaborators; the number of single-authored publications before/after Ph.D. graduation, and; the impact of different types of collaborations before/after Ph.D. graduation, using the average number of citations as measurements. Note that in Table 1, we select "authoritative authors" (AAs) as those whose h-indices are 40 or more, and the number of AAs in this dataset is 35. This criterion of selecting AAs was adopted by Amjad et al. (2017) with the same dataset.

Table 1. Scientific-collaboration-related indicators explored in this article.

No.	Abbr.	Meaning
General indicators (Type I)		
1	paper_count	Total number of papers one author has published during his/her whole scientific career

2	citation_number	Total citation counts of one author's publications during his/her whole scientific career
3	h_index	The value of one author's h-index
Indicators showing scientific careers before Ph.D. graduation (Type II)		
4	PaperCountB	Number of papers one author has published before Ph.D. graduation
5	CitationB	Citation counts of one author's papers published before Ph.D. graduation
6	CoauAATimesB	Number of collaborations with AAs before Ph.D. graduation
7	CoauUniAACountB	Number of unique AA collaborators before Ph.D. graduation
8	IndipenCountB	Number of single-authored papers one author has published before Ph.D. graduation
9	CoauNoAATimesB	Number of collaborations with non-AAs before Ph.D. graduation
10	CoauUniAuCountB	Number of unique non-AA collaborators before Ph.D. graduation
11	AvgCitWithAAB	Average citation counts of one author's papers collaborated with AAs before Ph.D. graduation
12	AvgCitIndipenB	Average citation counts of one author's single-authored papers published before Ph.D. graduation
13	AvgCitWithNoAAB	Average citation counts one author's papers collaborated with non-AAs before Ph.D. graduation
Indicators showing scientific careers after Ph.D. graduation (Type III)		
14	PaperCountA	Number of papers one author has published after Ph.D. graduation
15	CitationA	Citation counts of one author's papers published after Ph.D. graduation
16	CoauAATimesA	Number of collaborations with AAs after Ph.D. graduation
17	CoauUniAACountA	Number of unique AA collaborators after Ph.D. graduation
18	IndipenCountA	Number of single-authored papers one author has published after Ph.D. graduation
19	CoauNoAATimesA	Number of collaborations with non-AAs after Ph.D. graduation
20	CoauUniAuCountA	Number of unique non-AA collaborators after Ph.D. graduation
21	AvgCitWithAAA	Average citation counts of one author's papers collaborated with AAs after Ph.D. graduation
22	AvgCitIndipenA	Average citation counts of one author's single-authored papers published after Ph.D. graduation
23	AvgCitWithNoAAA	Average citation counts of one author's papers collaborated with non-AAs after Ph.D. graduation

Note: All of the values of the indicators shown in Table 1 are calculated at the end of 2014.

We consider these indicators as essentially twofold. On the one hand, we divide the publications of researchers based on whether they are collaborative articles and whether the coauthors include AAs, forming three types—coauthored with AAs, coauthored with non-AAs, and working independently—to separately examine the different features of collaborations before and after researchers' Ph.D. graduation. On the other hand, in the two types working collaboratively, we examine whether there are any differences between two similar but distinct indicators—the number of collaborations and the number of unique collaborators.

The Exploratory Factor Analysis (EFA) method is employed here (Cattell, 1965), as it is an important method of mining a few latent factors representing the observed variables. It is different from Principal Component Analysis (PCA), a frequently used method in factor analysis, which seeks some orthogonal components to replace the observed variables. Generally, we can simply show the difference between PCA (Formula 4) and EFA (Formula 5) as follows:

$$Principle_i = \sum_{j=1}^{n} p_j \cdot OV_j \ (i \in \{1,2,\dots,num\}, j \in \{1,2,\dots,n\}) \qquad (4)$$

$$OV_j = \sum_{k=1}^{m} q_k \cdot factor_k \ (k \in \{1,2,\dots,m\}, j \in \{1,2,\dots,n\}) \qquad (5)$$

Here, $Principle_i$ represents the *i*th component which resulted from PCA, OV_j is the *j*th observed variables in dataset, $factor_k$ is the *k*th factor which resulted from EFA, p_j is the value of the *j*th parameters, q_k is the value of the *k*th factor, m is the number of factors, n is the number of observed variables, and num is the number of components obtained by PCA. From Formulas 4 and 5, we can see that EFA represents the variables as the linear combination of common factors while PCA uses

principal components as the linear combination of the variables, showing a different direction of representation. Theoretically, the difference of the principle in the two methods is that PCA utilizes prior information while EFA does not. From both mathematical and theoretical perspectives, this difference gives EFA a better capability to interpret the factors and their internal relationships without prior information.

RESULTS AND DISCUSSION

Collaborators' Topic Diversity

The results of our study on topic diversity of collaborators are shown in Figure 3, where the degree of darkness is proportional to the number of authors who exhibit certain collaborators' topic similarity and h-indices. It is clear that collaborators of high-impact authors have fewer topic similarities, in that their collaborators are more diverse in research areas. This indicates that their collaborators can have relatively different research focuses, where the higher impact an author has, the more research topic diversity his/her collaborators have. In fact, the "difference" in terms of research topic, to some extent, leads to trans-topic or even transdisciplinary studies. Transdisciplinary study yields specific benefits, such as handling high levels of complexity and enhancing innovation (Xu et al., 2015a). Basically, researchers in different fields can broaden their domain horizons and lead transdisciplinary studies that benefit researchers by encouraging novel ideas and fresh perspectives. When traditional Library and Information Studies (LIS) researchers, for example, work with those in computer science, they can learn advanced algorithms and gradually use them in their own works. Bringing the Author-Conference-Topic model (Tang et al., 2008a) into scientific collaboration research (Ding, 2011) is a good case for this. Also, researchers in different research areas may provide better explanations, broader backgrounds, clearer logic, and deeper discussions to a project, which will benefit

resulting articles. Essentially, an author with more collaborators has a higher level of sociability, which to a large extent helps them achieve success. Our findings suggest that high-impact authors tend to pursue diverse-topic collaborations, which confirms the conclusions of Schaltegger et al. (2013) and Xu, Ding, Song, and Chambers (2015b).

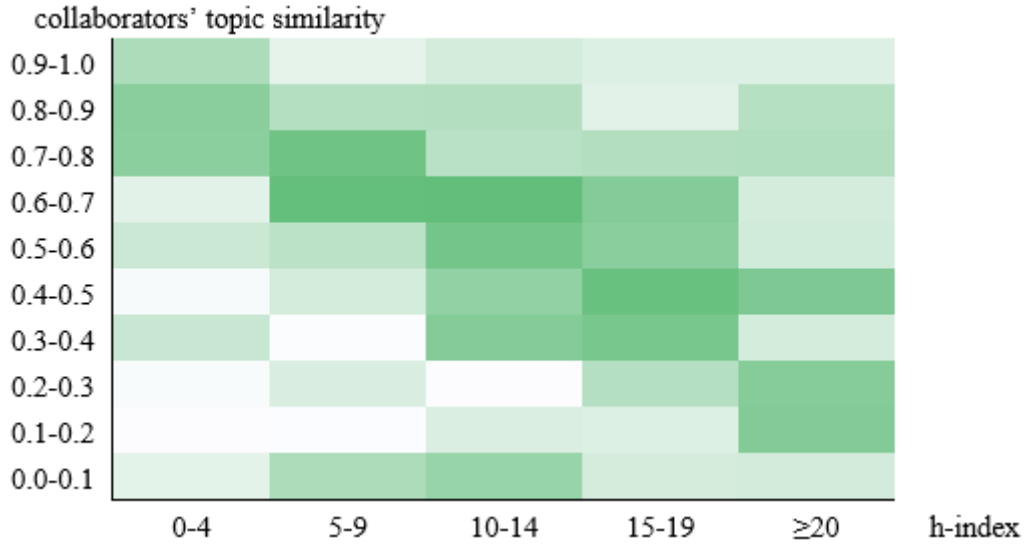

Figure 3. The authors' h-indices and their collaborators' topic similarity/diversity (the larger the topic similarities of authors' collaborators have, the less diverse their collaborators are in terms of research topic; the darker an area is in this figure, the more researchers whose collaborators have corresponding h-index and topic similarity).

Collaborators' Impact Diversity

The results of analyzing the impact diversity of collaborators are shown in Table 2, in which the larger NSD indicates that the author's collaborators are more diverse in terms of impact (h-index). Obviously, high-impact authors show higher NSD, which indicates that the influence of the higher-impact authors' collaborators is more diverse. This finding is different from conventional wisdom, where researchers are thought to collaborate with high-impact authors because they can provide profound knowledge, experience, and resources throughout the collaborations (Adegbola, 2013). But high-impact authors are often senior researchers in a given discipline while

low-impact authors may be junior scholars or students. When collaborating with high-impact authors, a researcher may not lead the whole research process but may instead focus on certain details of the work; yet in the cases that researchers' h-indices are higher than their collaborators', they might lead the whole research efforts and take the responsibility of revising the manuscript and guiding collaborators (Kumar and Ratnavelu, 2016). The fact that they handle macro-level issues (e.g. leading the research) during collaborations with lower-impact collaborators shows a more significant role they play than when collaborating with high-impact collaborators.

Table 2. The authors' h-indices and their collaborators' impact diversity (the larger NSD refers to collaborators who show more diversity in terms of impact).

h-index	0-4	5-9	10-14	15-19	⩾20
NSD of collaborators	1.43	3.72	6.94	8.80	10.57

Our interpretation of this data is that researchers who have collaborators with more diverse h-indices might have sufficiently exercised their research ability. For example, authors only working on empirical studies may not read sufficient numbers of papers on the topic, which is like the phenomenon of labor division in socioeconomics. As a result, their research would be limited and their theoretical background could be weak despite the fact that they have good operational skills. On the other hand, it is more likely that those who only focus on literature reviews or paper writing tend to show inadequate ability to model or enact relevant practices. Both of these cases could be discussed under the Cannikin Law (Buckets Effect), which suggests that a wooden bucket's capacity is determined by its shortest plank, or "the chain is only as strong as the weakest link." This implies that averages have little meaning, in that you cannot necessarily offset failure by corresponding success. True scientific success requires comprehensiveness and few weaknesses. Different types of tasks exercise the scholar's ability and skills in distinct aspects and ways (Krapf, 2015), which can lead

to high possibilities for their future success. It is therefore much better to “grow up” through research processes by undertaking different tasks that make the researchers “stronger” and more mature than to collaborate with specific impact researchers (e.g. high- or medium-impact researchers). In other words, high-impact authors collaborate with impact-diverse researchers, through which they gain rich experiences and benefit from different skills.

Scientific Collaboration before and after Ph.D. graduation

Figure 4 shows the results of EFA in a two-dimensional graph, in which each node represents an indicator and the coordinate values of a node are equal to its EFA loading values, which are normalized between 0 and 1. We can see that most Type II indicators (ends with “B”) feature a larger value on the vertical axis (VA) but a smaller value on the horizontal axis (HA) (on the left or the upper-left part of the figure). Type III indicators (ends with “A”) have a larger value on the HA but a smaller value on the VA (on the right or the lower-right part of the figure). The position difference between the indicators before and after Ph.D. graduation shows the distinguished scientific collaboration patterns among different stages of a career (Hamermesh, 2015). As for Type I indicators, both of the values on the two axes are large (on the upper-right of the figure). Hence we interpret HA as the “after-Ph.D.-graduation dimension” and VA as the “before-Ph.D.-graduation dimension.”

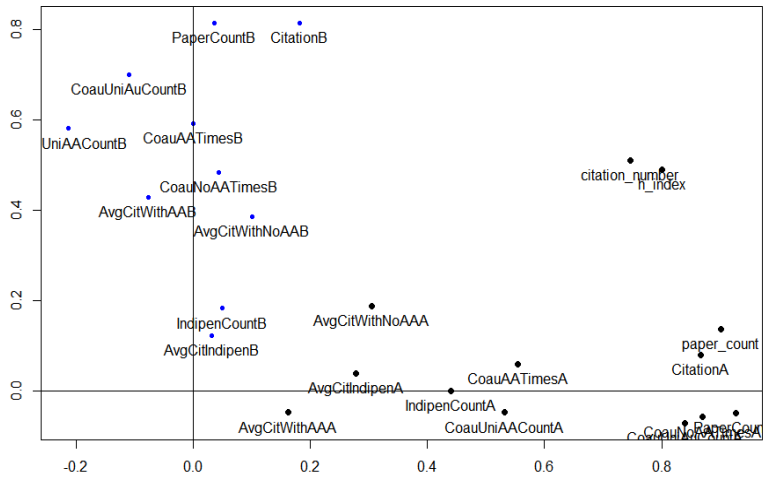

Figure 4. EFA results visualized in a two-dimensional graph.

To better analyze the different roles these indicators play, we establish the relationships between the indicators and axes using the radar map, noted in Figure 5. Certain indicators loading heavier on the HA or VA shows that this indicator plays a more important role in this dimension. We can see that the h-index and total citation number of an author have heavier loads on both HA and VA, suggesting that these two indicators are effective in evaluating career trajectory regardless of being before or after Ph.D. graduation; thus high performance on these indicators is associated with greater scientific success. Total number of publications only shows a heavy load on HA instead of both. Also, the position that h-index occupies in Figure 4 is between that of total number of citations and publications, which verifies the essence of h-index, by combining scholars' productivity and impact levels as measured by the number of publications and citations their work has received. Figure 5 also confirms the findings shown in Figure 4, that all of the Type II indicators exhibit much heavier loads on VA (red line in Figure 5, representing "before Ph.D. graduation"), while those of Type III have much heavier loads on HA (blue line in Figure 5, representing "after Ph.D. graduation").

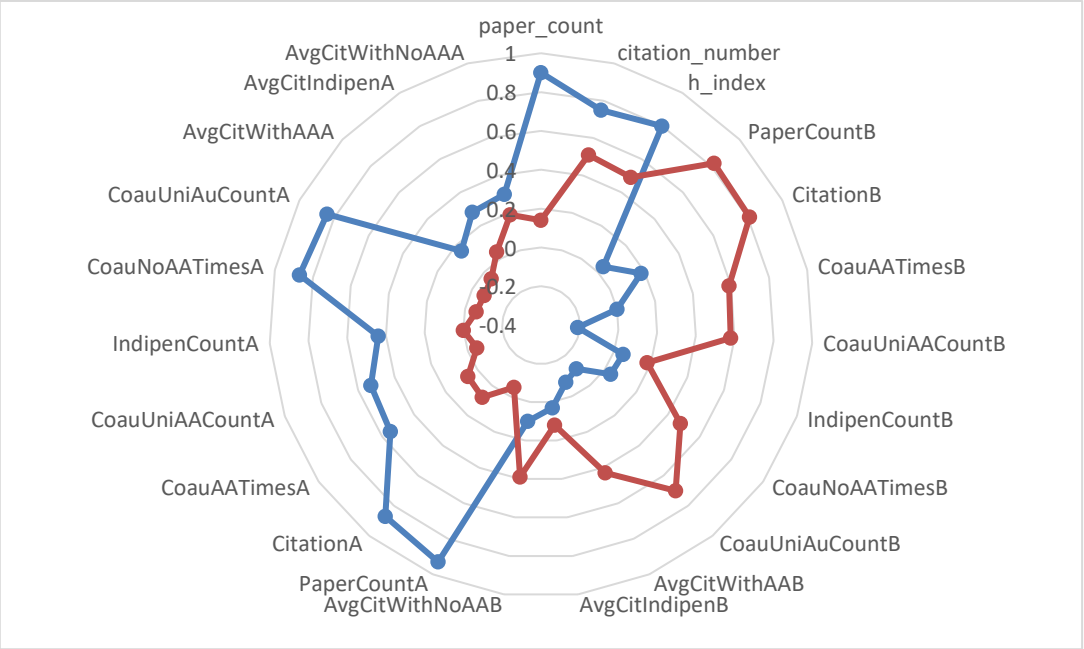

Figure 5. Relationships between the indicators and their loads on HA and VA (Blue: HA, after Ph.D. graduation; red: VA, before Ph.D. graduation).

Among all of the Type II indicators, the number of publications and the number of citations before Ph.D. graduation load mostly on HA, while two "independence-related" indicators, that of the number of single-authored publications ("IndipenCountB") and their average citations ("AvgcitIndipenB") load to a much lesser degree. This implies that it is *not* wise for Ph.D. candidates to publish articles *on their own*, which is probably because their research is immature and needs more guidance of senior researchers or mentors at that time. Otherwise, the impact of such sole-authored articles of junior scholars might not be high. This result is similar to Muschallik and Pull (2016)'s study, in which they found that mentees in formal mentoring programs are more productive than comparable researchers not involved with such programs. Moreover, Figure 5 shows that three "AA-related" and three "non-AA-related" Type II indicators also have approximate loading values on VA. Collaboration with other researchers may thus be a form of preparation for success before Ph.D. graduation, regardless of whether they are AAs or not. But there is no doubt that collaborating with AAs before Ph.D. graduation still plays an important

role in a researcher's career.

In terms of Type III indicators, the number of publications and citations after Ph.D. graduation also load heavily on HA, which is similar to the corresponding "before Ph.D. graduation" indicators. Different from Type II indicators in which AA-related and non-AA-related indicators are both important, AA-related indicators affect scientific careers less than the others, regardless of the number of collaborations ("CoauAATimes"), the number of unique collaborators ("CoauUniAACountA"), or the average number of citations ("AvgCitWithAAA"). This indicates that for researchers who have earned their doctoral degrees or have been established as experienced scholars, collaborating with authoritative authors may not play as powerful a role as it did prior to their Ph.D. graduation. Our interpretation of this data is that there is potential for the "halo effects" of AAs to outshine junior scholars and thus can fetter their further development to some extent. Meanwhile, indicators showing the number of collaborations with AAs (or non-AAs) load about ten percent more than that showing the number of unique collaborators with AAs (or non-AAs), suggesting that it is necessary for a number of papers to be coauthored with certain fixed AAs (or non-AAs). Based on the premise that for an approximate number of collaborations with AAs (or non-AAs), it is not good to pursue more AA (or non-AA) collaborators probability because "short-term collaborations" have limited promotions on scholars' career success. Moreover, the number of single-authored publications has a greater influence on a scientific career after Ph.D. graduation compared with that in Type II, before graduation, indicating that high-impact authors tend to collaborate less after their Ph.D. graduation compared to before their graduation. This finding is similar to that of Kumar and Ratnavelu (2016), in which older researchers are found to have significantly fewer coauthored articles than younger researchers. Yet neither of the two indicators showing the number of single-authored papers before or after Ph.D. graduation loads much in EFA results. This shows that leading research and

accomplishing it alone can only to a small extent benefit a junior researcher's career, as scientific collaboration can more easily lead to ultimate success (Bozeman and Boardman, 2014).

CONCLUSIONS

This paper analyzes scientific careers from the perspective of collaborators' diversity and the Ph.D. career milestone. To identify the collaborators' diversity, we calculate the research topic and impact diversity of the collaborators based on a computer science dataset. The h-index is employed to measure the impact of the researchers. The results of our empirical study show that high-impact authors have a tendency to collaborate with researchers focusing on different research topics. Contrary to conventional wisdom, we find that high-impact authors not only have abundant experience working with other high-impact authors but also collaborate with authors who have a medium h-index, meaning that high-impact authors' collaborators have more impact diversity, in terms of h-index, than others. Moreover, from the analyses on scientific-collaboration indicators related to Ph.D. graduation, we conclude that collaborating with authoritative authors plays an important role before researchers' Ph.D. graduation, while working with non-authoritative authors is more important after their Ph.D. graduation. This finding highlights the drawback of traditional wisdom of a researchers' need to "stand on the shoulder of giants" for success, and the importance of working with other researchers at different stages of their scientific careers.

This paper discovers a pattern of high-impact researchers in terms of their scientific collaborations, that is, trying to collaborate with more diverse scholars in terms of research topic as well as impact. Furthermore, one of the implications drawn from the results reflecting collaborations among different stages of careers is that advisors of Ph.D. students could provide them more opportunities to collaborate with other

experienced researchers, and not limit them to their advisor. A common phenomena in current Ph.D. program, especially in computer science field, is that the advisor prefers to see their doctoral students spending most of their time in the laboratory and only focusing on the group task. Actually the results from this article reveal its drawback, aka. will not benefit the young doctoral students at all.

One of the limitations of this research is that we only measure the diversity between the researchers' collaborators, but not the diversity between the researchers and their collaborators. In the future, detailed exploration of the research topic diversity and impact diversity between the researchers and their collaborators could be implemented. We will further explore the relationships between researchers' impact and their collaborators' diversity in various areas, such as affiliation, background, or geographical diversity. We also intend to further divide researchers' scientific careers into several milestones, allowing us to explore with more depth the characteristics of successful scientific career development.

ACKNOWLEDGEMENTS

The International Joint Research Project is funded by the National Natural Science Foundation of China (#71420107026). The full title of this project is "Research on Knowledge Organization and Service Innovation in the Big Data Environments". The authors are very grateful for the constructive comments and helpful suggestions from three anonymous reviewers.

REFERENCES

Adegbola, M. (2010). Nurses collaborating with cross disciplinary networks: Starting to integrate genomics into practice. *Journal of National Black Nurses Association, 21*(1), 46-49.

Adegbola, M. (2013). Scholarly tailgating defined: A diverse, giant network. *The ABNF Journal: Official Journal of the Association of Black Nursing Faculty in Higher Education, Inc, 24*(1), 17-20.

Amjad, T., Ding, Y., Xu, J., Zhang, C., Daud, A., Tang, J., & Song, M. (2017). Standing on the shoulders of giants. *Journal of Informetrics, 11*(1), 307-323.

Baker, V. L., & Pifer, M. J. (2011). The role of relationships in the transition from doctoral student to independent scholar. *Studies in Continuing Education, 33*(1), 5-17.

de Beaver, D., & Rosen, R. (1979). Studies in scientific collaboration: Part II. Scientific co-authorship, research productivity and visibility in the French scientific elite, 1799-1830. *Scientometrics, 1*(2), 133-149.

Bird, C., Barr, E. T., Nash, A., Devanbu, P. T., Filkov, V., & Su, Z. (2009). Structure and dynamics of research collaboration in computer science. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pp. 826-837, April 29-May 2, 2009, Sparks, NV., U.S.A.

Bozeman, B., & Boardman, C. (2014). *Research collaboration and team science: A state-of-the-art review and agenda*. New York: Springer.

Bredereck, R., Chen, J., Faliszewski, P., Guo, J., Niedermeier, R., & Woeginger, G. (2014). Parameterized algorithmics for computational social choice: Nine research challenges. *Tsinghua Science and Technology, 19*(4), 358–373.

Bridle, H., Vrieling, A., Cardillo, M., Araya, Y., & Hinojosa, L. (2013). Preparing for an interdisciplinary future: A perspective from early-career researchers. *Futures*, 53, 22-32.

Cattell, R. B. (1965). Factor analysis: An introduction to essentials (II): The role of factor analysis in research. *Biometrics*, 21, 405-435.

Chinchilla-Rodriguez, Z., Ferligoj, A., Miguel, S., Kronegger, L., & de Moya-Anegon, F. (2012). Blockmodeling of co-authorship networks in Library and Information Science in Argentina: A case study. *Scientometrics*, *93*(3), 699-717.

Costas, R, Nane, T, & Larivière, V. (2015). Is the year of first publication a good proxy of scholar's academic age? In *Proceedings of the 15th International Conference on Scientometrics and Informetrics*, pp. 988-998, June 29-July 3, 2015, Istanbul, Turkey.

Dannefer, D., Uhlenberg, P., Foner, A., & Abeles, R. P. (2005). On the shoulders of a giant: The legacy of Matilda White Riley for gerontology. *The Journals of Gerontology: Social Sciences, 60*(6), S296-S304.

Davoudi, S. & Pendlebury, J., (2010). Evolution of planning as an academic discipline, *Town Planning Review, 81*(6), 613-644.

Derry, S., Schunn, C., & Gernsbacher M. (2014). *Interdisciplinary collaboration: An emerging cognitive science*. Hove: Psychology Press.

Ding, Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informetrics, 5*(1), 187-203.

Ding, Y., & Stirling, K. (2016). Data-driven discovery: A new era of exploiting the literature and data. *Journal of Data and Information Science*, *1*(4), 1-9.

Ebadi, A., & Schiffauerova, A. (2015). How to receive more funding for your research? Get connected to the right people! *PLoS ONE, 10*(7), e0133061.

Freeman, R. B., & Huang, W. (2014). Collaborating with people like me: Ethnic co-authorship within the U.S. (No. w19905). Cambridge, MA: *National Bureau of Economic Research*.

Hamermesh, D. S. (2015). Age, cohort and co-authorship. Cambridge, MA: *National Bureau of Economic Research*.

Hart, R. L. (2000). Co-authorship in the academic library literature: A survey of attitudes and behaviors. *Journal of Academic Librarianship, 26*(5), 339-345.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceeding of the National Academic Science of United States of America, 102*(46), 16569-16572.

Holgate, S. A. (2012). How to collaborate. *Science (Careers).* Retrieved at http://www.sciencemag.org/careers/2012/07/how-collaborate.

Huang, J., Zhuang, Z., Li, J., & Giles, C. L. (2008). Collaboration over time: characterizing and modeling network evolution. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pp. 107-116, February 11-12, 2008, Palo Alto, CA., U.S.A.

Huang, M., & Chang, Y. (2011). A study of interdisciplinarity in information science: Using direct citation and co-authorship analysis. *Journal of Information Science*, 37(4), 369-378.

Institute of Medicine. (2000). *Committee on building bridges in the brain, behavioral, and clinical sciences. Bridging disciplines in the brain, behavioral, and clinical sciences.* Pellmar, T. C., & Eisenberg, L, Eds. Washington, DC: National Academies Press.

Kessel, F., & Rosenfield, P. (2008). Toward transdisciplinary research: Historical and contemporary perspectives. *American Journal of Preventive Medicine*, *35*(2 suppl), S225-S234.

Kindermann, R., & Snell, J. L. (1980). *Markov Random Fields and their applications*. Providence, RI: *American Mathematical Society*.

Kram, K. E., & Isabella, L. A. (1985). Mentoring alternatives: The role of peer relationships in career development. *Academy of Management Journal, 28*(1),

110-132.

Krapf, M. (2015). Age and complementarity in scientific collaboration. *Empirical Economics*, *49*(2), 751-781.

Kraut, R., Egido, C., & Galegher, J. (1988). Patterns of contact and communication in scientific research collaboration. In *Proceedings of the 1988 ACM Conference on Computer-Supported Cooperative Work*, pp. 1-12, September 26-28, 1988, Portland, OR., U.S.A.

Kumar S., & Ratnavelu, K. (2016). Perceptions of scholars in the field of Economics on co-authorship associations: Evidence from an international survey. *PLoS ONE, 11*(6), e0157633.

Lee, E., McDonald, D., Anderson, N., & Tarczy-Hornoch, P. (2009). Incorporating collaboratory concepts into informatics in support of translational interdisciplinary biomedical research. *International Journal of Medical Informatics*, *8*(1), 10-21.

Lee, S., & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science, 35*(5), 673-702.

Leimu, R., & Koricheva, J. (2005). Does scientific collaboration increase the impact of ecological articles? *BioScience, 55*(5), 438-443.

Loannidis, J. P. A., Boyack, K. W., & Klavans, R. (2014). Estimates of the continuously publishing core in the scientific workforce. *PLoS ONE*, *9*(7), e0101698.

Luukkonen, T., Persson, O., & Sivertsen G. (1992). Understanding patterns of international scientific collaboration. *Science Technology Human Values*, *17*(1), 101-126.

Mccaughrean, M., Zinnecker, H., Andersen, M., Meeus, G., & Lodieu, N. (2002).

Standing on the shoulder of a giant: ISAAC, Antu, and star formation. *The Messenger*, 109, 28-36.

Milojević, S. (2010). Modes of collaboration in modern science: Beyond power laws and preferential attachment. *Journal of the American Society for Information Science and Technology*, *61*(7), 1410-1423.

Muschallik, J., & Pull, K. (2016). Mentoring in higher education: Does it enhance mentees' research productivity? *Education Economics*, *24*(2), 210-223.

Newman, M. E. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5200-5205.

Packalen, M., & Bhattacharya, J. (2015). Age and the trying out of new ideas. Cambridge, MA: *National Bureau of Economic Research*.

Pao, M. L. (1982). Collaboration in computational musicology. *Journal of the American Society for Information Science*, *33*(1), 38-43.

Pohl, C. (2005). Transdisciplinary collaboration in environmental research. *Futures*, *37*(10), 1159-1178.

Pohl, C. (2007). From science to policy through transdisciplinary research. *Environmental Science and Policy*, *11*(1), 46-53.

Pravdić, N., & Oluić-Vuković, V. (1986). Dual approach to multiple authorship in the study of collaboration/scientific output relationship. *Scientometrics*, *10*(5-6), 259-280.

Quatman, C., & Chelladurai, P. (2008). Social network theory and analysis: A complementary lens for inquiry. *Journal of Sport Management*, *22*(3), 338-360.

Schaltegger, S., Beckmann, M., & Hansen, E. G. (2013). Transdisciplinarity in corporate sustainability: Mapping the field. *Business Strategy and the*

Environment, *22*(4), 219-229.

Sie, R. L., Drachsler, H., Bitter-Rijpkema, M., & Sloep, P. (2012). To whom and why should I connect? Co-author recommendation based on powerful and similar peers. *International Journal of Technology Enhanced Learning*, *4*(1-2), 121-137.

de Solla Price, D. J., & de Beaver, D. (1966). Collaboration in an invisible college. *American Psychologist*, *21*(11), 1011-1018.

Stokols, D., Harvey, R., Gress, J., Fuqua, J., & Phillips, K. (2005). In vivo studies of transdisciplinary scientific collaboration: Lessons learned and implications for active living research. *American Journal of Preventive Medicine*, *28*(2), 202-213.

Tang, J., Fong, A. C. M., Wang, B., & Zhang, J. (2012). A unified probabilistic framework for name disambiguation in digital library. *IEEE Transaction on Knowledge and Data Engineering*, *24*(6), 975-987.

Tang, J., Jin, R., & Zhang, J. (2008a). A topic modeling approach and its integration into the random walk framework for academic search. In *Proceeding of the Eighth IEEE International Conference on Data Mining*, pp. 1055-1060, December 15-19, 2008, Pisa, Italy.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008b). ArnetMiner: Extraction and mining of academic social networks. In *Proceedings of the fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.990-998, August 24-27, 2008, Las Vegas, NV., U.S.A.

Thurman, P. W., & Birkinshaw, J. (2006). Scientific collaboration results in higher citation rates of published articles. *Pharmacotherapy*, *26*(6), 759-767.

Wang, C., Han, J., Jia, Y., Tang, J., Zhang, D., Yu, Y., & Guo, J. (2010). Mining advisor-advisee relationships from research publication networks. In *Proceedings of the Sixteenth ACM SIGKDD International Conference on Knowledge*

Discovery and Data Mining (SIGKDD'2010), pp. 203-212, July 25-28, 2010, Washington D.C., U.S.A.

Wickson, F., Carew, A. L., & Russell, A. W. (2006). Transdisciplinary research: Characteristics, quandaries and quality. *Futures, 38*(9), 1046-1059.

Xu, J., Ding, Y., & Marlic, V. (2015a). Author credit for transdisciplinary collaboration. *PLoS ONE, 10*(9), e0137968.

Xu, J., Ding, Y., Song, M., & Chambers, T. (2015b). Author credit-assignment schemas: A comparison and analysis. *Journal of the Association for Information Science and Technology, 67*(8), 1973-1989.

Zhang, C., Bu, Y., & Ding Y. (2016). Understanding scientific collaboration from the perspective of collaborators and their network structures. In *iConference 2016 Partnership with Society*, March 20-23, 2016, Philadelphia, PA., U.S.A.

Zuckerman, H. (1967). Nobel laureates in science: Patterns of productivity, collaboration, and authorship. *American Sociological Review, 32*(3), 391-403.