Understanding Persistent Scientific Collaboration

Yi Bu

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

Ying Ding

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

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

Library, Tongji University, Shanghai, China

Xingkun Liang*

Department of Information Management, Peking University, Beijing, China

Dakota S. Murray

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

Corresponding author: Xingkun Liang (lxk@pku.edu.cn).

Understanding Persistent Scientific Collaboration

Abstract: Common sense suggests that persistence is key to success. In academia, successful researchers have been found more likely to be persistent in publishing, but little attention has been given to how persistence in maintaining collaborative relationships affects career success. This paper proposes a new bibliometric understanding of persistence that considers the prominent role of collaboration in contemporary science. Using this perspective, we analyze the relationship between persistent collaboration and publication quality along several dimensions: degree of transdisciplinarity, difference in co-author's scientific age and their scientific impact, and research-team size. Contrary to traditional wisdom, our results show that persistent scientific collaboration does not always result in high-quality papers. We find that the most persistent transdisciplinary collaboration tends to output high-impact publications, and that those co-authors with diverse scientific impact or scientific ages benefit from persistent collaboration more than homogeneous compositions. We also find that researchers persistently working in large groups tend to publish lower-impact papers. These results contradict the colloquial understanding of collaboration in academia and paint a more nuanced picture of how persistent scientific collaboration relates to success, a picture that can provide valuable insights to researchers, funding agencies, policy makers, and mentor-mentee program directors. Moreover, the methodology in this study showcases a feasible approach to measure persistent collaboration.

INTRODUCTION

Popular culture abounds in tales of persistence leading to success, a tale that also echoes through scientific mythos. Madame Curie became the only scientist to win the

Nobel Prizes in both physics and chemistry only after years of tedious work extracting milligrams of radium from pitchblende residue, and Thomas Edison experimented with thousands of materials before discovering that tungsten was the best material for light bulb filament. Persistence has long been thought to be characteristic of success in academia, a characteristic explored by Ioannidis, Boyack, and Klavans (2014) who demonstrated that author's persistent efforts, represented as uninterrupted and continuous presence in publishing, are related to author's high impact and academic career success.

It seems intuitive that persistence would be a *sine qua non* of success in science, although other career factors, such as collaboration, have become increasingly important. Larivière *et al.* (2016), for instance, observed that collaboration in science has been increasing over the past century, and that collaboration is positively correlated with academic quality. Wuchty, Jones, and Uzzi (2007) further suggested that the increasing cost, scale, and complexity of scientific research, along with advances in communication technology have led teamwork to become the norm across many scientific disciplines. Popular culture brims with stories of lone and persistent scientific geniuses: Albert Einstein, Alan Turing, and Madame Curie, just to name a few; but in the contemporary scientific landscape it is teams, not individuals, who drive knowledge production.

Despite the prominence of collaboration in contemporary science, and the widespread cultural emphasis of persistence, collaboration and persistence are largely considered as separate processes. Ioannidis *et al*. (2014) found that those who persistently publish are the most likely to be high-impact authors, but did not explore the role of collaboration. Petersen (2015) is a rare instance of a study of the benefits of long-term and productive collaborative activity, but used a coarse-grained classification and focused on a small number of only the most persistent collaborations. There is a gap in our understanding of the nuances of persistence collaboration, and career success, and a lack of methods and indicators to study such phenomenon.

We propose a new approach to understanding persistence in science, one that properly considers the prominent role of teams. In particular, we explore to what extent persistence in maintaining long-term collaborative relationships impacts the academic success of these collaborations. We develop a generalizable methodology and bibliometric indicators capable of revealing details of collaboration and persistence, and demonstrate their utility by analyzing the nuances in collaboration among a large dataset of computer science researchers. To further understand the nuances of persistence, we also consider a host of other factors that have been consider in previous research on research collaboration, such as the degree of transdisciplinarity between authors (Bu *et al.*, 2017), authors' scientific age (Peacocke, 1993), authors' scientific impact (Amjad *et al.*, 2017), and the size of research teams (Larivière, Gingras, & Sugimoto, 2014).

This article is outlined as follows. We first discuss work related to our study, giving attention to past bibliometric studies of collaboration and persistence. We then detail the data, design, and methods for our analysis. Next, we present our findings concerning the role of persistence in collaboration, and provide our interpretations of these results. Finally, we conclude with a summary of our findings, their limitations, implications, and thoughts for future research.

RELATED WORK

Studies on Persistent Presence

Uninterrupted and continuous presence (UCP) has proven to be an important indicator for measuring persistence in scientific activities. Ioannidis *et al.* (2014) analyzed papers in Scopus that were published between 1996 and 2011 and found that only one percent of authors, labeled UCP authors, persistently published at least one article every year; they concluded that not only do UCP authors receive more citations, but that they also feature high h-indices regardless of their disciplines. They also demonstrated the importance of persistence to the structure, stability, and vulnerability of a scientific career. Wu, Venkatramanan, and Chiu (2016) employed a similar notion of UCP authors to define whom they term as "top active authors", selected by their degree of persistence, and found that these authors who persistently publish in their domains are "representative of overall populations" (p. 1). However, these studies only provide a single perspective of persistence in academia—that of publications compared between UCP and non-UCP authors.

Petersen (2015) conducted a longitudinal study of the benefits of various degrees of collaborative activity towards a scientist's career, especially those benefits resulting from so-called "super ties": long-term relationships where two co-authors have high publication overlap. The author found evidence of a phenomenon termed the "apostle effect", an increase in citations and productivity resulting from extremely strong collaborative ties. But Petersen (2015) analyzed a limited number of scientific careers and used a simple classification system that cannot easily capture the nuanced nature of persistence in collaboration. We expand upon these previous studies by analyzing scientific collaboration, rather than their publications, and by abandoning the coarse classification method of UCP and strong ties in favor of a continuous variable measuring the degree of persistent scientific collaboration.

Transdisciplinary Scientific Collaboration

Among the advantages of transdisciplinary scientific collaboration (TSC) in academia are that they allow researchers to "handle high levels of complexity, tap otherwise isolated sources of local knowledge, foster transformative thinking, and enhance legitimacy" (Xu, Ding, & Malic, 2015, p. 2), to challenge common disciplinary and institutional boundaries (Davoudi & Pendlebury, 2010), and to work as the key pathway to scientific innovation (Gray, 2008; Stokols, 2006). TSC has been helping solve a number of practical problems in various fields such as library and information science (Huang & Chang, 2011), cognitive science (Derry, Schunn, & Gernbacher, 2014), and health science (Lee, McDonald, Anderson, & Tarczy-Hornoch, 2009). Some researchers have noted that TSC suffers from several drawbacks, causing inter-personal friction and requiring extra resource and time investment (Schaltegger *et al.*, 2013), and are often confronted with tremendous practical barriers such as communication among members due to different jargons (Institute of Medicine, 2000); despite these shortcomings, there is little doubt that TSC plays an increasingly crucial role in academic success (Wang, Thijs, & Glanzel, 2015), leading some countries to implement policies encouraging TSC (Woelert & Millar, 2013).

Studies exploring the relationships between TSC and success have seldom used a temporal perspective in their analysis; the temporal information, however, could be of importance, as it might affect whether TSC could have higher scientific achievements than non-TSC. This paper fills in this gap by examining how persistence, the temporal perspective, relates to the quality of output resulting from TSC.

Scientific Collaboration and Diverse Scientific Ages/Impacts of Collaborators

When collaborating on scientific publications, labor tends to be distributed based on the academic age of contributors, with younger and less experienced scholars performing the more “technical” tasks, such as performing experiments, while older scholars contribute more to data analysis and preparation of the manuscript (Larivière *et al.*, 2016). Noting that scholars with different levels of experience and expertise make different contributions, some scholars have explored how such diversity of age, impact, and thus contribution might influence the quality of publications throughout life. For example, Amjad *et al.* (2017) found that those collaborating with authoritative authors (AAs) in their first publication have a higher probability to achieve greater impact (measured using the h-index) than those who have never collaborated with AAs, but they have less impact than those who collaborated with AAs only after establishing a stable career.

Other researchers have tried to explore the relationship between collaborator's impact and the quality of their collaboration. For example, Leimu and Koricheva (2005) examined the relationship between citation count of collaborators and the influence of their co-authored articles, but failed to find any significant correlations between them. Similarly, Zhang, Bu, Ding, and Xu (2017) also failed to detect any significant correlation between co-author's citation count and the formation of collaboration in the field of information retrieval. These studies show inconclusive results regarding a relationship between collaboration quality and collaborator's impact difference.

Scientific Collaboration and Research Team Size

A scientific collaboration can be regarded as a research team in which the first and the corresponding author (if they are different) are the leaders while the others are team members (Chinchilla-Rodríguez *et al.*, 2012). Previous studies have focused on the relationships between the impact of scientific collaboration and research team size. For example, Wuchty *et al.* (2007) concluded that publications and patents published by a team tend to receive more citations than those by an individual, and furthermore that "this advantage is increasing over time" (p.1036). Similarly, Guimera, Uzzi, Spiro, and Amaral (2005) used team size as an independent variable in their proposed model for the self-assembly of creative teams and indicated that team size could determine team performance. Larivière *et al.* (2014) expanded on the dataset used by previous studies by including all of the publications from 1900 to 2011 from Science Citation Index, Social Science Citation Index, and Arts and Humanities Citation Index to argue that the more authors an article has, the greater its impact. Moreover, some studies have explored the relationship between collaboration impact and research team composition. On the other hand, Curral *et al.* (2001) argued that large-size teams would have “poorer team processes” (p.199). While large and small teams each have their advantages and disadvantages, Hackman and Vidmar (1970) found that between four to five members is the optimal perceived team size, at least in the realm of business. But teams are complex, with the various dynamics of their formation and operation growing organically around small groups and prominent individuals (Chinchilla-Rodríguez *et al.*, 2012), and so to better understand scientific teams, a more nuanced approach is needed.

METHODOLOGY

Data

The dataset used in this article comes from ArnetMiner (Tang *et al.*, 2008a), which covers 2,092,356 academic articles from the field of computer science published between 1936 and 2014 including 1,207,061 unique authors and 8,024,869 local citation relationships. Author’s names were disambiguated according to Tang, Fong, Wang, and Zhang (2012), in which a unified probabilistic framework is implemented along with both content- and structure-based information and two steps are included, estimating the weights of feature functions and assigning papers to different authors[1]. Collaborations are represented using co-authored papers, of which only papers published between 2001 and 2010 were selected[2], providing a final dataset of 885,562 unique authors, 3,822,638 unique collaboration pairs, 449,875 articles, and 606,843 local citation relationships. The number of citations each article received is calculated

[1] By doing so, the author name disambiguation has a precision rate of 83.01% and a recall rate of 79.54% on the ArnetMiner dataset (Tang *et al.*, 2012).

[2] The ArnetMiner dataset ends in 2014, so we pick 2010 as the ending year of analyses so that the papers published before 2010 could have a period of time window to accumulate their citations (Wang, 2013). Moreover, ten years is a sufficient long-time period for researchers to set up and develop their research. The length of a researcher’s career is usually less than 50 years, and so ten years is significant period in his/her career. These are why we set 2001-2010 as the time periods of our following analyses.

based only on the citation relationships recorded in ArnetMiner, i.e. local citation counts. The count of the yearly number of citations is used as an article's indicator of impact, which minimizes the bias of older papers, which have more time to accumulate citations. The h-index (Hirsch, 2005) is calculated for each author according to his or her publications and citation counts recorded in the dataset.

Methods

The objective of this paper is to analyze the relationships between the impact of collaboration and degrees of persistence in scientific collaboration. Four other variables are used to examine such relationship: degree of transdisciplinarity, difference between collaborators' scientific ages, difference between collaborators' scientific impact, and team size. To measure the degree of transdisciplinarity we use the Author-Conference-Topic (ACT) model (Tang, Jin, & Zhang, 2008b) and cosine similarity. The differences between collaborator's scientific impact and scientific age are calculated as the normalized absolute difference of h-index and normalized absolute difference of the publication year of their first paper, respectively. Size of research team of author pairs is measured by calculating the number of authors in all of the two authors' co-authored publications (including the author pair themselves), divided by the total number of co-authored publications; thus, if a pair of authors appeared as co-authors on three publications which had two authors (only the author pair, and no other team members), four authors, and six authors respectively, then the team size of each collaborator would be four ($= \frac{2+4+6}{3}$). Figure 1 provides a visual overview of the methodology used in this paper.

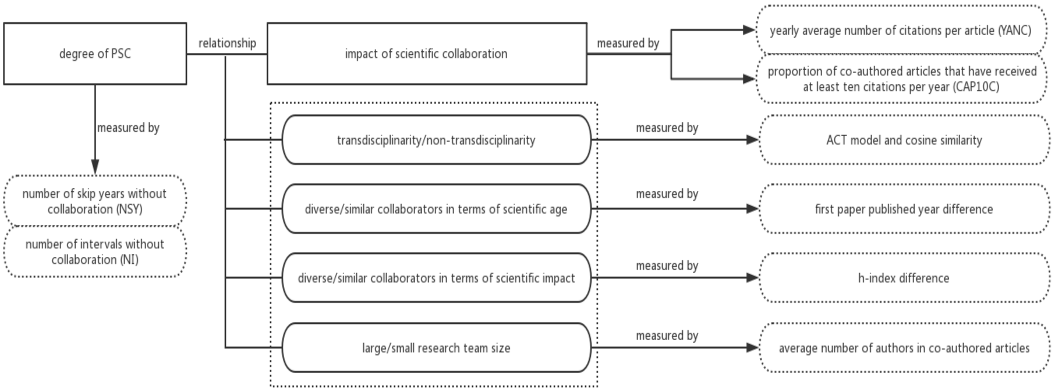
degree of PSC
relationship
impact of scientific collaboration
measured by
yearly average number of citations per article (YANC)
proportion of co-authored articles that have received at least ten citations per year (CAP10C)
measured by
number of skip years without collaboration (NSY)
number of intervals without collaboration (NI)
transdisciplinarity/non-transdisciplinarity
measured by
ACT model and cosine similarity
diverse/similar collaborators in terms of scientific age
measured by
first paper published year difference
diverse/similar collaborators in terms of scientific impact
measured by
h-index difference
large/small research team size
measured by
average number of authors in co-authored articles

Figure 1. Overview of methods.

Measuring the Impact of Collaboration

We calculate the yearly average number of citations per article received (YANC), which indicates the impact of co-authored articles. We also calculate the proportion of co-authored articles that have received at least ten citations per year (CAP10C), which is equal to the number of co-authored articles that have received at least ten citations per year divided by the number of co-authored articles two collaborators have written.

Measuring the Degree of Transdisciplinarity

The Author-Conference-Topic (ACT) model (Tang *et al.*, 2008b) is employed to measure an author’s research topic whereby each author is represented by a distribution of *topics* and each topic is represented by a distribution of *words*. Word distributions are modeled using the titles and abstracts of an author’s publications. A fixed number of latent topics are learned from the titles and abstracts of all publications in the dataset; we found 50 topics to work well, each representing sub-fields of computer science. A vector containing 50 components is calculated for every author, each of which contain *topic distributions*, or the probabilities of terms appearing in that author’s abstracts and titles being “generated” by the corresponding topic (Tang *at al*. 2008b). Degree of transdisciplinarity is operationalized as topic similarity, which is measured by calculating the cosine similarity between author's vectors. The more similar their topics, the less transdisciplinary their collaboration; and *vice versa*.

Measuring the Degree of Persistent Scientific Collaboration (PSC)

Table 1 shows three examples of scientific collaboration between 2001 and 2010, where the number in each cell represents the number of publications the author pair published during the corresponding year. For instance, authors A_1 and A_2 collaborated on four papers in 2002 but did not collaborate in 2003-2005. When measuring the degree of persistent scientific collaboration (PSC), a natural approach is to employ *the number of skip years without collaboration* (NSY), which refers the number of years they have zero co-published articles within a given time period. Similar to Ioannidis *et al.* (2014), the smaller NSY two authors have, the more persistent their collaboration. For example, there are five years that authors A_1 and A_2 did not collaborate between 2001 and 2010, so their NSY is five. Similarly, NSY between authors A_3 and A_4 as well as A_5 and A_6 is five and six, respectively. Using NSY as a measure of persistence, we see that collaboration between A_5 and A_6 is the least persistent among these three pairs.

Table 1. An example of calculation on the degree of PSC.

	2001	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**
$(\boldsymbol{A_1}, \boldsymbol{A_2})$	3	4	0	0	0	1	2	0	0	1
$(\boldsymbol{A_3}, \boldsymbol{A_4})$	2	0	2	0	1	0	4	0	0	1
$(\boldsymbol{A_5}, \boldsymbol{A_6})$	5	0	0	0	0	4	0	0	3	1

Although NSY between (A_1, A_2) and (A_3, A_4) are identical, the nature of their PSC is different; the collaboration between authors A_1 and A_2 could be seen as less persistent, because the number of years without collaboration is less disperse, and there are more consecutive years with no collaboration. During these consecutive years, it is likely that their collaboration has been interrupted and they might have stopped working with each other, while the single years without collaboration between authors A_3 and A_4 might indicate that they are still collaborating, but that their projects require greater time investment. The more instances of consecutive years that two authors do not collaborate, the higher the probability that collaboration is interrupted. Therefore, besides NSY, we employ another measure—*the number of intervals without collaboration* (NI), defined as the number of contiguous time periods that two authors have no joint publications. For Table 1, the NI of an author pair would be calculated as the number of intervals of consecutive zero(s) in their row. Given identical values for NSY, the greater an author pair's NI, the greater the degree of their PSC. For example, (A_1, A_2) has an NI of two because there are two intervals with no collaboration (2003-2005 and 2008-2009); (A_3, A_4) has an NI of four (2002, 2004, 2006, and 2008-2009), so while each author pair has the same NSY, (A_3, A_4) has the greater NI, and thus the greater degree of PSC.

Given these assumptions and analyses, the mathematic definition of the degree of PSC is as follows. Assume that for N years (annotated as Year $y_1, y_2, \ldots, y_N$), collaborations are counted as potential PSC records. In these N years, authors i and j have co-authored $p_{i,j}$ articles ($p_{i,j} \geq 0$). Specifically, they have collaborated $p_{i,j,q}$ times in the year of y_q $(q = 1,2, \ldots, N)$. We can represent their numbers of collaborations in each year among the N-year time using a vector $\overrightarrow{P_{i,j}}$:

$$\overrightarrow{P_{i,j}} = (p_{i,j,1}, p_{i,j,2}, \ldots, p_{i,j,N}) \tag{1}$$

where $\sum_{q=1}^{N} p_{i,j,q} = p_{i,j}$. Essentially, during their N years' collaborations between authors i and j, we define s_{ij} as NSY, which is equal to the number of zeros among all of the components in $\overrightarrow{P_{i,j}}$.

On the other hand, in $\overrightarrow{P_{i,j}}$, we define $\overrightarrow{P_{i,j}}$'s consecutive sub-vector $\overrightarrow{S_{i,j,m}}$ that contains

u components ($m \leq N, 1 \leq u \leq N, 1 \leq x_1 < x_2 < \cdots < x_u \leq N$), $p_{i,j,x_1}, p_{i,j,x_2}, \ldots, p_{i,j,x_u}$, as a vector catering to the following criteria:

$$\begin{cases} p_{i,j,x_1} = p_{i,j,x_2} = \cdots = p_{i,j,x_u} = 0 \\ p_{i,j,x_1-1} \neq 0 \; (x_1 \neq 1) \; OR \; x_1 = 1 \\ p_{i,j,x_u+1} \neq 0 \; (x_u \neq N) \; OR \; x_u = N \end{cases} \tag{2}$$

The count of sub-vectors, $\overrightarrow{S_{\imath,\jmath,m}}$, that caters to these criteria is defined as $v_{i,j}$ (i.e. $max(m) = v_{i,j}$) which essentially represents the number of intervals without collaboration (NI) between i and j within the given N years.

The degree of PSC between i and j, $D_{i,j}$, indicating how persistent their collaboration is, is defined as:

$$D_{i,j} = N - s_{ij} + \lambda v_{i,j} \tag{3}$$

where λ $(0 < \lambda < 1)$ is a parameter to fit the model. $D_{i,j} \in (1, N]$, if we remove those collaboration pairs who have no collaboration record in the given N years. **Note that an author pair can only have ONE value of degree of PSC**. For example, in Table 1, if we set $\lambda = 0.5$, we can calculate the degree of PSC for each author pair as six $(= 10 - 5 + 0.5 \times 2)$, seven $(= 10 - 5 + 0.5 \times 4)$, and five $(= 10 - 6 + 0.5 \times 2)$, respectively. Table 2 shows the distribution of author pairs in terms of degree of PSC.

Table 2. Distributions of author pairs in degree of PSC.

Degree of PSC	1.0-2.0	2.5-4.0	4.5-6.0	6.5-8.0	8.5-10.0
Proportion of author pairs	0.47	0.25	0.18	0.09	0.01

Measuring the Collaborator Diversity in terms of Impact and Scientific Age

To measure the difference between collaborator's impacts, we calculate the absolute difference between each author's h-index. The absolute difference is then normalized by the value of the maximum absolute difference among all author pairs' h-indices. To measure the scientific age difference of collaboration, we use the number of years between each author's first publications, and normalize by the maximum absolute difference of scientific ages among all author pairs' scientific ages.

Suppose that ξ author pairs having collaborated with each other within the N consecutive years, ap_1 containing authors $ap_{1,1}$ and $ap_{1,2}$, ap_2 containing authors $ap_{2,1}$ and $ap_{2,2}$, …, ap_ξ containing authors $ap_{\xi,1}$ and $ap_{\xi,2}$, are selected. For $ap_{k,1}$ and $ap_{k,2}$ $(k = 1,2,\dots,\xi)$, we annotate their h-indices as $h_{k,1}$ and $h_{k,2}$, respectively. The absolute difference between their h-indices, ad_k, should be calculated as:

$$ad_k = \left|h_{k,1} - h_{k,2}\right| \tag{4}$$

The normalized absolute difference of h-indices, nad_k, is derived as:

$$nad_k = \frac{ad_k}{max(ad_1,ad_2,\dots,ad_\xi)} \tag{5}$$

where $max(ad_1, ad_2, \dots, ad_\xi)$ refers to the maximum value among ad_1, ad_2, …, and ad_k.

Meanwhile, we annotate $ap_{k,1}$ and $ap_{k,2}$ who published their first articles in year $y_{k,1}$ and $y_{k,2}$, respectively, and the absolute difference between their scientific ages, $ad_k{}'$, is calculated as:

$$ad_k{}' = \left|y_{k,1} - y_{k,2}\right| \tag{6}$$

Similarly, we can calculate the normalized absolute difference between their scientific ages, $nad_k{}'$, as:

$$nad_k{}' = \frac{ad_k{}'}{max(ad_1{}',ad_2{}',\dots,ad_\xi{}')} \tag{7}$$

Measuring Team Size

Suppose authors $ap_{k,1}$ and $ap_{k,2}$ have completed t_k co-authored articles within N consecutive years. These co-authored articles, $w_1, w_2, \ldots, w_{t_k}$, have $co_1, co_2, \ldots, co_{t_k}$ authors, respectively ($co_1, co_2, \ldots, co_k \geq 2$ because "collaboration" requires at least two researchers). The average team size of the collaboration between $ap_{k,1}$ and $ap_{k,2}$, ATS_k, is calculated as:

$$ATS_k = \frac{1}{t_k} \sum_{r=1}^{t_k} co_r \tag{8}$$

Essentially ATS_k is equal to the average number of authors in the co-authored articles published by the given author pairs. These teams are *not* necessarily a *constant* set of researchers, meaning that the identities of authors appearing alongside the author pair are irrelevant, only the number of co-authors is important. We use this mathematical definition to measure research-team size of collaborating authors.

Correlation Analysis

To explore the potential relationships among these variables, we employ Pearson's r to implement two-side correlation analysis. For all co-author pairs, we represent their degrees of PSC as $\overrightarrow{DoP} = (DoP_1/N, DoP_2/N, \ldots, DoP_\sigma/N)$ where σ is the total number of co-author pairs and the components serve as each of their degrees of PSC normalized by the number of years considered in the experiment. Similarly, we can build the vectors representing the degree of transdisciplinarity, impact and scientific age difference, team size, and YANC of all co-authored pairs as:

$$\overrightarrow{DoT} = (DoT_1, DoT_2, \ldots, DoT_\sigma),$$

$$\overrightarrow{ID} = (nad_1, nad_2, \ldots nad_\sigma),$$

$$\overrightarrow{SAD} = (nad_1{}', nad_2{}', \ldots nad_\sigma{}'),$$

$$\overrightarrow{TS} = (ATS_1/max(ATS), ATS_2/max(ATS), \dots ATS_\sigma/max(ATS)),\text{ and}$$

$$\overrightarrow{PCC} = (PCC_1, PCC_2, \dots PCC_\sigma),$$

respectively. Each component of these vectors is the corresponding value of certain variable, some of which need to be normalized before further processing. To examine the potential correlation between the degree of PSC and YANC under different scenarios (degree of transdisciplinarity, impact and scientific age difference, and team size), we then use Pearson's correlation coefficient to calculate the correlation between the $\overrightarrow{PCC}$ and several vectors, including $\overrightarrow{DoT} + \overrightarrow{DoP}$, $\overrightarrow{ID} + \overrightarrow{DoP}$, $\overrightarrow{SAD}$+$\overrightarrow{DoP}$, and $\overrightarrow{TS} + \overrightarrow{DoP}$.

RESULTS AND DISCUSSION

Overview

Figure 2 shows the results of the YANC and CAP10C among different degrees of PSC groups. We can see that groups with a generally higher degree of PSC have more YANC and CAP10C but the middle-high (degree of PSC between "6.5-8.0") group has the highest number of YANC and CAP10C.

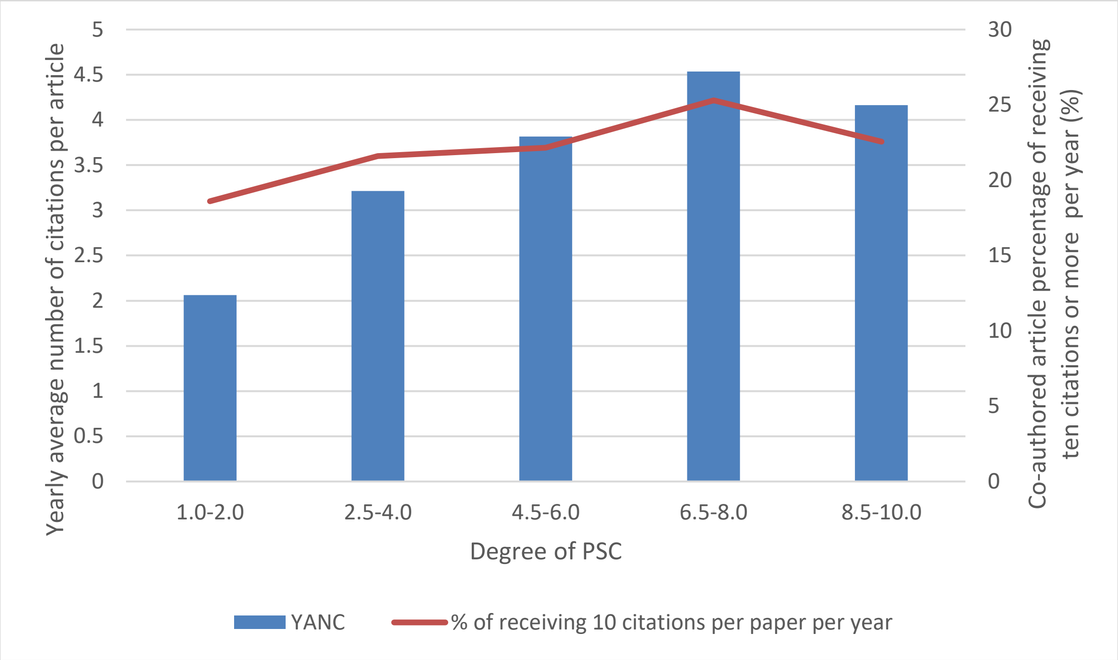

Figure 2. The yearly average number of citations per article (YANC) and the percentage of receiving ten citations or more per paper per year (CAP10C) for different degrees of PSC groups ($\lambda = 0.5$, the same below).

It is not necessarily confirming that the degree of PSC increases with YANC and CAP10C (which would indicate the influence of their co-authored researchers), but rather that groups with moderately high, but not extreme, degree of PSC could have access to more opportunities to increase their YANC and CAP10C. These results emphasize the importance of maintaining continuous collaboration in academia—in fact, persistent collaboration with fewer interruptions can establish strong trust between collaborators and lead to long-term success and sustainability. While collaborators with a low degree of PSC have to spend time together to become acquainted with one another's research and personality, collaborators with middle-high degree of PSC have worked together persistently, allowing them to optimize their research process. Their high level of familiarity in research helps them to better utilize and share the resources between each other, e.g., existing code, datasets, algorithms, software and tools, wonderful notes or ideas, and people (such as colleagues, students, and friends) who are specialized in different domain areas. Access to a wider pool of resources can likely clear some barriers to research and might make research easier for collaborators to tackle research questions and produce high-quality and innovative research. Moreover, in medical research and other "cumulative sciences" where cumulative production of information is mandatory (Ioannidis *et al.*, 2014), PSC is expected in order to accrue more research resources and achieve success in their careers.

Contrary to traditional wisdom, Figure 2 shows that the author pairs with the highest degree of PSC do not show better research performance than the middle-high degree groups, indicating that too much focus on specific collaborators might narrow the perspectives of scholars, or lead scholars to become complacent in their topics and ideas (Pope, 2016). This finding could also result from an effect similar to that described by Uzzi (2006), wherein too much embeddedness in the same relationships can limit collaboration efficiency. While collaboration can result in the mutual exchange of knowledge and skills between involved researchers, there may be diminishing returns to working too persistently with the same collaborator. Thus, highly persistent collaboration may stifle their potential by clinging too closely to a small number of relationships, rather than expanding their network and gaining access to new knowledge from other researchers.

Table 3 shows the results of correlation analysis, in which we can see the transdisciplinarity and impact diversity plus PSC has more significant correlation with YANC than other variables. The details of relationship among these variables will be shown in the following sections.

Table 3. Correlation analysis results.

Variables	*r*	*p*
(PSC + transdisciplinarity) and YANC	0.35	***
(PSC + impact diversity) and YANC	0.27	***

(PSC + scientific age diversity) and YANC	0.28	**
(PSC + team size) and YANC	-0.19	**

*Note: ***:<0.001; **:<0.01.*

Transdisciplinary Scientific Collaboration and the Degree of PSC

Figure 3 shows the results of our analysis of TSC, where the horizontal axis represents the degree of PSC of author pairs, the vertical axis maps topic similarly (where similar topics would be close to one, while dissimilar, “transdisciplinary” topics would be closer to zero) between authors in a pair, and the intensity of the color is proportional to the YANC of the co-authored publications of authors pairs with the corresponding characteristics. The YANC appears highest for collaborators who have moderate persistence but similar topics, but that this advantage quickly diminishes for the most persistent non-TSC. That is to say, collaborators with similar research interests do not need to continuously collaborate with each other, but maintaining a medium degree of collaboration appears crucial for higher impact. While PSC might allow collaborators to become familiar and accumulate academic resources, too much focus on specific collaborators might limit a researcher’s potential and reduce persistent collaboration benefits. But a researcher having many ephemeral collaborations might lead the authors to invest too much time in getting familiar with collaborators, thus benefitting little from the collaboration. Although previous studies (Gray, 2008; Xu *et al.*, 2015) have noted that distinct collaborators are important to research success because they provide broader perspectives and expertise necessary to tackle complex problems, these studies failed to capture the drawbacks and nuances that the temporal element, persistence, provides.

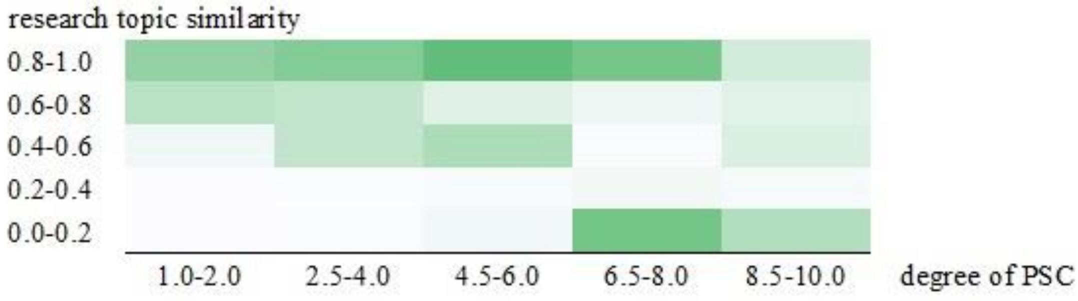

Figure 3. The collaborators' research topic similarities and their degrees of PSC (the more topic similarity two authors have, the less transdisciplinary they are; the more intense the color of a cell, the higher the YANC of the co-authored articles written by the authors with the corresponding topic similarity and degree of PSC).

The most transdisciplinary collaboration, appearing at the bottom of Figure 3, has an overall lower YANC than non-transdisciplinary collaboration. Furthermore, TSC only becomes effective when allowed sufficient persistence, indicating that higher-impact publications tend to come from collaborators from diverse research areas maintaining a high degree of PSC.

Meanwhile, the TSC with the least degree of persistence has some of the lowest YANC, and thus the weakest research performance compared to other types of collaboration. Although transdisciplinary collaboration has potentials to produce high-quality research (Gary, 2008; Stokols, 2006), seldom research has explored their faults; the benefits of TSC may only manifest given sufficient time and persistence. This temporal characteristic may be related to the fact that transdisciplinary collaboration is more time-consuming (Schaltegger *et al.*, 2013) and are faced with more barriers such as differences in attitude, jargons, publishing and professional organizations, career trainings, and leadership (Institute of Medicine, 2000). Even worse, although transdisciplinary collaboration is often encouraged at a policy level, they are not sufficiently supported under current funding structures, structures that also don't consider the importance of persistence (Bromham *et al.*, 2016; Woelert &

Millar, 2013). As Domik and Fischer (2011) noted, the short-term research performance of transdisciplinary scientific collaboration is limited, which, along with our findings, highlights the importance of continuing to adequately support persistent transdisciplinary collaboration.

Because we add the temporal dimension from a scientometric perspective, these findings also to some extent supplement the theory of structural holes (Burt, 1995) and weak ties (Granovetter, 1973). Both theories emphasize the potential benefits of less homogeneous neighbors in networks that TSC consists of. Our findings imply that to reveal the benefit of the structure holes or weak ties, some persistence, might be necessary, at the cost of time and effort to maintain persistence.

Collaborator Diversity in terms of Scientific Age and Impact, and the Degree of PSC

Figures 4 and 5 show the results of the analysis of persistence and author’s difference in scientific age and impact, where the horizontal axis represents the degree of PSC, the vertical axis maps the difference in scientific age (Figure 4), or the difference in scientific impact (Figure 5). Values of the vertical axis that are close to zero indicate similar ages, while values close to one indicate larger age differences. The intensity of the color in each cell is proportional to the YANC value of author pairs corresponding to the given characteristics. In each figure, those author pairs that have large differences in either age or impact, but also along with a high degree of persistence, have the best research performance. In these cases, the two collaborators might have advisor-advisee or senior-junior relationships, in which the senior researchers (or advisors) are likely to enhance their junior’s performance by contributing knowledge, theories, skills, and research experiences (Adegbola, 2013). Persistent collaboration (high degree of PSC) between the seniors and juniors might help produce more high-impact publications, a finding that echoes Muschallik and Pull (2016) who found that mentees involved in formal mentoring programs were more productive compared to those who were not. A potential implication of this finding is that universities should provide support for advisor-advisee and senior-junior (e.g. full professor and assistant professor) relationships and encourage their persistent collaboration. Along with financial aid, universities should also consider offering human-resources supports, such as supplying more opportunities to attract external, experienced, and high-impact researchers to collaborate with advisees and junior researchers. Chinchilla-Rodríguez *et al.* (2012) revealed the incredibly heterogeneous dynamics that affect the formation, composition, and production of scientific groups; our findings do not consider all of these factors and dynamics, but they offer an important step to understanding the nuances of collaboration and teams in science, nuances that have not been previously explored at this large scale.

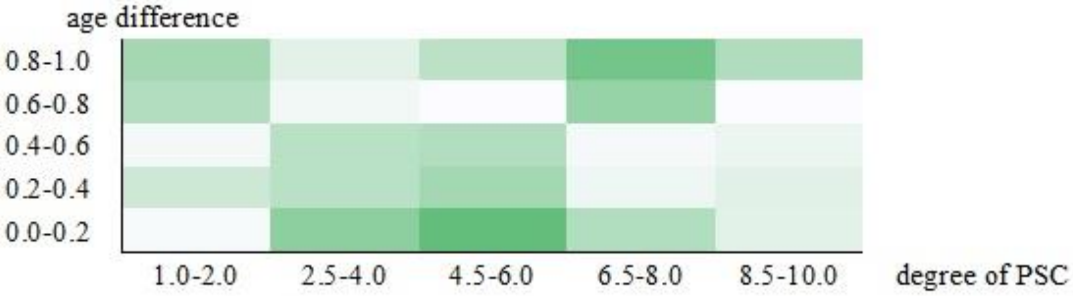

Figure 4. The collaborators' scientific age differences and their degrees of PSC (the more intense the color of a cell, the higher the YANC of the co-authored articles written by the authors with the corresponding scientific age difference and degree of PSC).

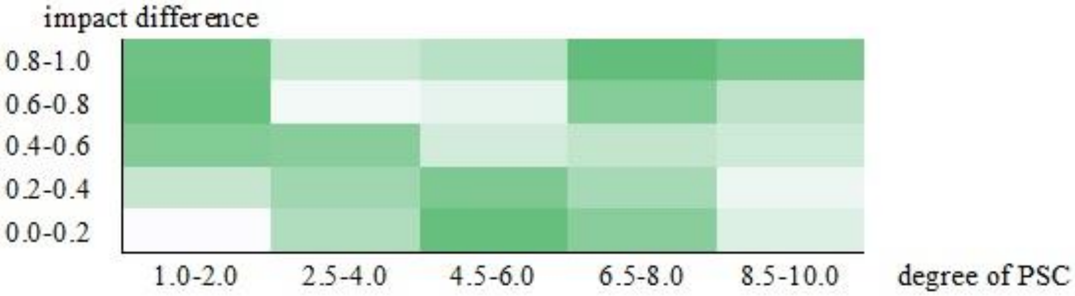

Figure 5. The collaborators' scientific impact differences and their degrees of PSC (more intense the color of a cell, the higher the YANC of the co-authored articles written by the authors with the corresponding scientific impact difference and degree of PSC).

Figures 4 and 5 also indicate that for collaborators with dissimilar scientific ages, but especially those with dissimilar scientific impacts (senior and junior co-authors), group with a low degree of PSC also tends to have good research performance. We interpret the occasional collaboration with high YANC as associated with the “halo effect”, otherwise known as preferential attachment (Barabási *et al.*, 2002), wherein junior researchers that collaborate with an authoritative author (“giant”) of the discipline will attract more citations than they otherwise would have.

Moreover, both collaborators that have either similar scientific ages or similar impacts, likely colleague-colleague relationships, as well as a medium degree of PSC have good research performance. These results indicate that colleagues require at least some persistence to reach their potential, but that too much persistence may lead to negative effects, likely resulting from relationship complacency and the narrowing of research perspectives resulting from the focus on specific collaborators.

Our results confirm past studies of scientific age and collaboration, finding that the quality of collaboration varies with the differences in scientific age between collaborators, likely related to differences between mentor-mentee, senior-junior, and colleague-colleague relationships (e.g., Amjad *et al.*, 2017). But in addition to supporting past findings, our inclusion of the temporal perspective presents a more complex and nuanced image of how collaborator diversity relates to persistence and publication quality.

Research Team Size and the Degree of PSC

Figure 6 shows the results of our analysis of research team size and PSC, where the horizontal access represents the degree of PSC, the vertical axis maps intervals of team size, and each cell contains the YANC of author pairs corresponding to the given characteristics. Author pairs with the largest average sizes of teams have great research performance for short, low-persistence collaboration, but as the degree of PSC increases, performance of large teams quickly suffers. Curral *et al.* (2001) also remarks that large teams pressured by a “high requirement to innovate” may manifest “poorer team processes”, decreasing their performance (2001, p. 187). Similarly, Hsiehchen, Espinoza, and Hsieh (2015) found that an increase in team size above a certain threshold often negatively impacts the group’s performance, possibly due to disappearing opportunities for effective interaction between individuals, or some members being pushed to ancillary or otherwise isolated roles. These past findings along with our own are evidence that long-term collaboration within large teams is difficult, and thus may be less likely to produce high-quality publications.

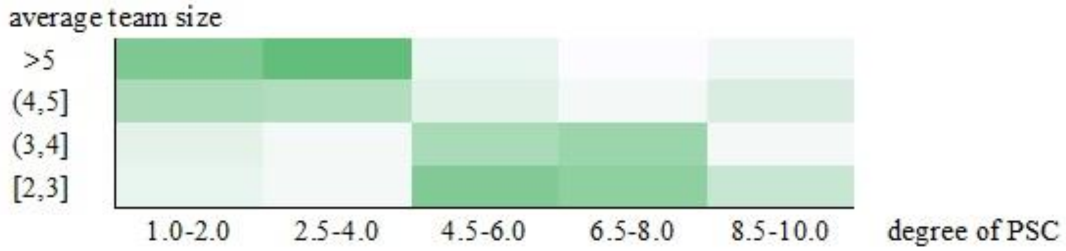

Figure 6. The research team size and collaborators’ degrees of PSC (the more intense the color of a cell, the higher the YANC of the co-authored articles written by the authors with the average team size and degree of PSC).

Those author pairs represented in Figure 6 that have smaller average team sizes appear to be more likely to output high-quality publications given a medium degree of PSC. We interpret this finding to mean that a small research team allows members to become acquainted with one another only after some years of persistent collaboration. However, if two authors often work in a small team and maintain a high degree of PSC, the narrow research perspective might limit their capacity for creativity and innovation, hindering their ability to produce quality and novel publications.

Hackman and Vidmar (1970) provided an ideal team size of between four and five members; our results add a caveat to this assessment, at least for computer science researchers—four to five members in a team may be ideal in the beginning, but not necessarily forever. We find that persistence affects different sizes of teams in distinct ways, and again support the notion that the temporal component allows for a better understanding of collaboration, an understanding that may benefit project instructors and research team leaders who seek to maximize high-quality output and research performance.

CONCLUSIONS

This paper proposes a novel bibliometric perspective to analyze persistent scientific collaboration. Using this perspective, we analyze the relationships between the co-authored articles' impact and the degree of persistence of collaboration (PSC) along four dimensions: degree of transdisciplinarity, difference in scientific age and impact, and research team size. Both traditional wisdom and past research (Ioannidis *et al.* 2014) indicate that persistence is closely related to success, but when we adopt the collaborative perspective, our paper suggests that such claims fail to capture the complexities of persistence and collaboration. We find that collaborators with a middle-high degree of PSC have a tendency to receive more citations, that transdisciplinary collaboration is found to maintain a high degree of PSC so as to publish high-impact articles, and that non-transdisciplinary collaboration requires only a medium degree of PSC. As for those collaborators with larger difference in scientific age or impact (measured by h-index), both higher and lower degree of PSC can lead to good research performance, but likely for different reasons; for collaborators with smaller difference in scientific age or impact, a medium degree of PSC is better. From the perspective of research team size, we find that collaborators having co-authored high-impact papers in large teams tend to maintain a small degree of PSC while those in small teams tend to feature a medium-high degree of PSC. Contrary to the conventional view of persistence, our findings reveal the phenomenon to be far more nuanced than previously imagined, and hint at the complexity and sociality of scientific collaboration that might be dynamically and simultaneously affected by an unknown number of internal and external factors.

This study has several implications for both scientific policy makers and researchers. Transdisciplinary collaboration needs more persistence to produce high-impact outputs than non-transdisciplinary collaboration, and policy should be crafted that considers this relationship—specifically, supports for transdisciplinary collaboration should emphasize persistence and be sustained over longer periods of time. Our findings also demonstrate that collaborators whom are diverse in terms of scientific age and impact tend to write high-quality papers if they maintain a high degree of PSC; as such, academic departments should design mentor-mentee programs that encourage persistent collaboration between participants, and provide resources and opportunities that allow junior assistant professors (or junior researchers) to communicate and collaborate persistently with senior researchers. This paper also highlights different collaboration strategies for working in large or small groups, strategies which department deans, project directors, and research leaders may find useful to optimize performance. Specifically, scholars should be encouraged to collaborate in small teams, but should avoid collaborating persistently as members of large teams. Collaboration, especially transdisciplinary collaboration, is often encouraged by funders and organizational leaders, and the results of this paper might allow them to craft policy which reflects the roles of persistence, age, impact, and team size.

On the other hand, from a methodological perspective, the approach provided in this article could be adopted and duplicated to measure the PSC as well as other related topics. Moreover, this method could also be developed and improved, for example, by adding more series-related variables such as the *yearly rate of change of collaboration count* when calculating the degree of PSC, as the collaboration number for each year is essentially regarded as a series mathematically.

Among the limitations of this study is that it only separately examines the relationships between the degree of PSC and several factors, but fails to offer a combined analysis of all factors. Other limitations relate to the nature of the data; the findings of this paper are to some extent dependent on the coverage and quality of the data source. One such limitation is that the citation count used in this paper is actually the *local* citation count, which might bias current results by excluding citations from outside fields that might be using the methods and techniques developed by computer scientists. Following this, the publications analyzed in this study are limited to computer science; future studies could apply these techniques to other disciplines, examining the role that disciplinary culture plays in persistence, and also examine collaboration that occurs between two culturally or methodologically distinct disciplines, such as computer science and sociology. Our methodology is limited in that it only operates on pairs of author, and does not consider larger groupings; our algorithm makes no differentiation between one author who always collaborates with the same three co-authors on every paper, and another author whose every paper has a different set of three co-authors.

Future research related to PSC may aim to improve upon our methodology, more specifically, capturing subtle and important relationships. Or else researchers might work to identity advisor-advisee and colleague-colleague collaboration and explore the patterns between type of relationship, persistence, and quality. Moreover, future researchers can more closely explore how PSC affects other aspects of researcher’s career, such as altering research topics and bolstering a junior scholar’s impact. Scientific collaboration is complex, and the addition of the temporal component allows researchers to explore how the subtle factors lead to the success of a collaboration.

ACKNOWLEDGEMENTS

We thank Chao Lu and Sanchari Das for their helpful suggestions and our four anonymous reviewers for their insightful comments.

REFERENCES

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.

Barabási, A.L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, *311*(3), 590-614.

Bromham, L., Dinnage, R., & Hua, X. (2016). Interdisciplinary research has consistently lower funding success. *Nature, 534*(7609), 684-687.

Bu, Y., Ding, Y., Xu, J., Liang, X., Gao, G., & Zhao, Y. (2017). Understanding success through the diversity of collaborators and the milestone of career. *Journal of the Association for Information Science and Technology*, DOI: 10.1002/asi.23911.

Burt, R. (1995). *Structural holes: The Social structure of competition*. Cambridge, Massachusetts: Harvard University Press.

Chinchilla-Rodríguez, Z., Ferligoj, A., Miguel, S., Kronegger, L., & Moya-Anegón, F.

(2012). Blockmodeling of co-authorship networks in library and information science in Argentina: a case study. *Scientometrics, 93*(3), 699-717.

Curral, L.A., Forrester, R.H., Dawson, J.F., & West, M.A. (2001). It's what you do and the way that you do it: Team task, team size, and innovation-related group processes. *European Journal of Work and Organizational Psychology, 10*(2), 187-204.

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., & 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.

Domik, G., & Fischer, G. (2011). Transdisciplinary collaboration and lifelong learning: Fostering and supporting new learning opportunities. In C.S. Calude, G. Rozenberg, & A. Salamaa (Ed.), *Rainbow of Computer Science* (pp. 129-143). Boulder, CO: Springer.

Granovetter, M.S. (1973). The strength of weak ties. *American Journal of Sociology, 78*(6), 1360-1380.

Gray, B. (2008). Enhancing transdisciplinary research through co-authored leadership. *American Journal of Preventive Medicine, 35*(2), S124-S132.

Guimera, R., Uzzi, B., Spiro, J., & Amaral, L.A.N. (2005). Team assembly mechanism determine collaboration network structure and team performance. *Science, 308*(5722), 697-702.

Hackman, J.R., & Vidmar, N. (1970). Effects of size and task type on group performance and member reactions. *Sociometry*, *33*(1), 37-54.

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

Hsiehchen, D., Espinoza, M., & Hsieh, A. (2015). Multinational teams and diseconomies of scale in collaborative research. *Science Advances*, *1*(8), e1500211.

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, D.C.: National Academies Press.

Ioannidis, J.P.A., Boyack, K.W., & Klavans, R. (2014). Estimates of the continuously publishing core in the scientific workforce. *PLoS ONE, 9*(7), e101698.

Larivière, V., Desrochers, N., Macaluso, B., Mongeon, P., Paul-Hus, A., & Sugimoto, C.R. (2016). Contributorship and division of labor in knowledge production. *Social Studies of Science*, *46*(3), 417-435.

Larivière, V., Gingras, Y., & Sugimoto, C.R. (2014). Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology, 66*(7), 1323-1332.

Lee, E.S., McDonald, D.W., Anderson, N., & Tarczy-Hornoch, P. (2009).

Incorporating collaboratory concepts into informatics in support of translational interdisciplinary biomedical research. *International Journal of Medical Informatics, 78*(1), 10-21.

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

Miguel, S.E., Chinchilla-Rodríguez, Z., González, C., & Moya Anegón, F. (2012). Analysis and visualization of the dynamics of research groups in terms of projects and co-authored publications. A case study of library and information science in Argentina. *Information Research, 17*(3), 524-546.

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

Peacocke, A. R. (1993). *Theology for a scientific age: being and becoming--natural, divine, and human*. Grove City, OH: Fortress Press.

Petersen, A.M. (2015). Quantifying the impact of weak, strong, and super ties in scientific careers. *Proceedings of the National Academy of Sciences of United of America*, *112*(34), E4671-E4680.

Pope, A. (2016). How much collaboration is too much? Retrieved from http://www.cmswire.com/digital-workplace/how-much-collaboration-is-too-much/

Schaltegger, S., Beckmann, M., & Hansen, E.G. (2013). Transdisciplinarity in corporate sustainability: Mapping the field. *Business Strategy and the Environment, 22*(4), 219-229.

Stokols, D. (2006). Towards a science of transdisciplinary action research. *Community Psychology, 38*(1), 63-77.

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., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008a). 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.

Tang, J., Jin, R., & Zhang, J. (2008b). 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.

Uzzi, B. (1996). The sources and consequences of embeddedness for the economic performance of organizations: The network effect. *American Sociological Review*, *61*(4), 674-698.

Wang J. (2013). Citation time window choice for research impact evaluation. *Scientometrics, 94*(3), 851-872.

Wang, J., Thijs, B., & Glanzel, W. (2015). Interdisciplinarity and impact: Distinct effects of variety, balance, and disparity. *PLoS ONE, 10*(5), e0127298.

Woelert, P., & Millar, V. (2013). The “paradox of interdisciplinarity” in Australian research governance. *Higher Education, 66*(6), 755-767.

Wu, Y., Venkatramanan, S., & Chiu, D.M. (2016). Research collaboration and topic trends in Computer Science based on top active authors. *PeerJ Computer Science, 2*, e41.

Wuchty, S., Jones, B.F., Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science, 316*(5827), 1036-1039.

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

Zhang, C., Bu, Y., Ding, Y., & Xu, J. (2017). Understanding scientific collaboration: Homophily, transitivity, and preferential attachment. *Journal of the Association for Information Science and Technology*, DOI:10.1002/asi.23916.