ISSI 2019

SAPIENZA
UNIVERSITÀ DI ROMA

17th INTERNATIONAL CONFERENCE ON SCIENTOMETRICS & INFORMETRICS

ISSI2019

with a Special STI Indicators Conference Track

2-5 September 2019
Sapienza University of Rome, Italy

PROCEEDINGS

VOLUME II

EdizioniEfesto

PROCEEDINGS OF THE 17TH CONFERENCE OF THE INTERNATIONAL SOCIETY FOR SCIENTOMETRICS AND INFORMETRICS

© Edizioni Efesto - ISBN 978-88-3381-118-5 - August 2019
Printed in Italy

Editors: *Giuseppe Catalano, Cinzia Daraio, Martina Gregori, Henk F. Moed and Giancarlo Ruocco*

Graphic cover design: *Francesco Manzo* | *graframan.com*
Cover photo: ©*Fayee - stock.adobe.com*

Investigating scientific collaboration through the sequence of authors in the publication bylines and the diversity of collaborators

Yi Bu[1], Chenwei Zhang[2], Yong Huang[3], Cassidy R. Sugimoto[4], and Zaida Chinchilla-Rodríguez[5]

[1] *buyi@iu.edu*
Center for Complex Networks and Systems Research, School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, U.S.A.

[2] *zhang334@indiana.edu*
School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, U.S.A.
Qiqihar Institute of Engineering, Qiqihar, Heilongjiang, China

[3] *yonghuang1991@whu.edu.cn*
School of Information Management, Wuhan University, Wuhan, Hubei, China

[4] *sugimoto@indiana.edu*
School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, U.S.A.

[5] *zaida.chinchilla@csic.es*
Instituto de Políticas y Bienes Públicos (IPP), Consejo Superior de Investigaciones Científicas (CSIC), Madrid, Spain

Abstract

In scientometrics, it is critical to investigate the patterns of scientific collaboration and how these patterns result in different impacts. In this research, we investigate the relationship between the sequence of authors in the publication bylines and the diversity of their collaborators. The diversity of collaborators is quantified with two dimensions, namely topic and impact diversities. Using the ArnetMiner dataset containing ACM-indexed publications in computer science, we find that the following two patterns tend to lead higher-impact scientific publications: (1) greater topic diversity of collaborators plus more tendency to work as leading authors (including first and/or corresponding authors); and (2) less topic diversity of collaborators plus less tendency to work as leading authors. Meanwhile, from the perspective of impact diversity, the results of our empirical study show that authors who work as more leading authors and collaborate with less impact diversity researchers have tendencies to receive more citations than those with collaborators with greater impact diversity. We also detect different patterns of authors' sequence and diversity of their collaborators before and after their Ph.D. graduation.

Background and Research Objectives

Scientific collaboration is prevalent in various disciplines (Wu *et al.*, 2019). Scientometricians have made great efforts to understanding scientific collaborations from different perspectives, such as scale-free networks (Newman, 2001), homophily and transitivity (Zhang *et al.*, 2018), dependency vs. autonomy (Chinchilla *et al.,* 2018a), geographical proximity (Katz, 1994), science of team science (Stokols *et al.*, 2008), temporal aspects (Bu *et al.*, 2018c), and labour of contribution (Lu *et al.*, 2018).

In bibliometrics and scientometrics, co-authorships are often applied as an important measurement for scientific collaboration. The sequence of co-authors identifies details on "who is accountable for the integrity of the reported study and who deserves what amount of credit for the work" (p. 359), as well as their contributions (He *et al.*, 2012). The patterns of authors' sequence reveal practical implications for scientists, funding providers, and research evaluators; thus, it is crucial to paint a more nuanced picture on the sequence of authors, their collaborators, and the impacts of their co-authored publications (Chinchilla *et al.*, 2019). Another branch of study in scientific collaboration focuses on the diversity of collaborators,

a.k.a., members in a research team. For instance, Bu *et al.* (2018b) studied the relationship between an author's impact and his/her collaborators' diversities, namely research topic and impact diversities. They found that high-impact authors tend to have more diverse collaborators in these two dimensions. Likewise, Zhang *et al.* (2019) concluded that the diversity on team members' productivity and scientific ages will increase the team performance. Similarly, a temporal-based analysis demonstrated that co-authors with diverse scientific impact or scientific ages benefit from persistent collaboration more than homogeneous compositions, a.k.a., less diversity (Bu *et al.*, 2018a).

Similarly, the viability and productivity of diversity is impacted by the support it receives from institutional leaders and research funding agencies (Stokols *et al.,* 2019). Many universities, governments, and funding agencies encourage and require cross-disciplinary applicant teams to submit collaboration plans as part of their research proposals (Wang & Shapira 2015; Zhang *et al.*, 2018). However, there are risks especially related with publishing and the allocation of credit in the peer review and academic reward system, institutional barriers, and funding requirement (Bromhan *et al.,* 2016).

Concerning the academic reward system, there exists a lack of credit given to interdisciplinary research in the context of promotion and tenure and limits to career advancement and publishing (Roy *et al.* 2013). Among obstacles are negative perceptions of interdisciplinary research by traditional disciplinary specialists and consequently, troubles publishing because research does not adhere to or fit neatly within traditional disciplinary frameworks (Rafols *et al.,* 2012), and in general, problems related to the peer review system (Wagner *et al.*, 2019). Evidences also suggests that it takes longer for scholars doing interdisciplinary research to establish themselves in their careers (Rhoten & Parker, 2004), and that scholars can be less productive, possible due to cognitive and collaborative challenges associated with such research, which is counterproductive especially in early career stages (Leahey *et al.,* 2017). Institutional review processes may be deeply rooted in disciplinary approaches to evaluation and only senior researchers, who face less-rigid performance evaluations, are better equipped for the complexity associated to with leading and publishing interdisciplinary research projects and publications. Goring *et al.* (2014) coincide in how the current reward structure in academia and other institutions may be misaligned with the current practice of interdisciplinary collaborative science, especially for early career researchers. They advocate for developing strategies behind team building and the requirements for understanding philosophical underpinning to promote interdisciplinary collaborative success. In this research-in-progress paper, we investigate the relationship between the sequence of authors in the publication bylines and the diversity of their collaborators. The diversity of collaborators is measured in two aspects, research topic and impact diversities.

Methodology

Similar to our previous work (Bu *et al.*, 2018a, 2018b), in this paper, we employ the ArnetMiner dataset (Tang *et al.*, 2018b) containing ~2M ACM-indexed computer science publications, as well as ~1.2M authors of these publications and ~8M citation relations between these publications. The authors' names were disambiguated according to the algorithm proposed by Tang *et al.* (2012). Some descriptive statistics can be found in some previous work (e.g., Amjad *et al.*, 2017; Bu *et al.*, 2018b). We follow Bu *et al.* (2018b, 2018c) to focus on articles published between 2001 and 2010, which results in ~450K publications, ~885K distinct authors, nearly 4M different collaboration pairs, and ~606K local citation relations. Note that the ignorance of global citation (e.g., citations from publications outside the current dataset) relations is one of the limitations of the current study, partly because transdisciplinary citations will be missing. We follow Bu *et al.* (2018b) to quantify two

dimensions of diversity for an author's collaborators, namely research topic and impact diversities. In terms of research topic diversity of an author's collaborators, we run Author-Conference-Topic model (Tang *et al.*, 2008a), an extended Latent Dirichlet allocation (LDA) model, on our dataset and calculate the cosine similarity between the topic vectors of an authors' collaborators. As for impact diversity of an author's collaborators, 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 (Hirsch, 2005) is applied to indicate the impact of the collaborators. We are also interested in the sequence of an author in his/her publication's byline. In computer science, last authors tend to be corresponding authors of a certain publication. Hence, first authors and last authors of the computer science publications are regarded as leading authors (Chinchilla-Rodríguez *et al.*, 2019) in the current paper.

Preliminary Results

Figure 1 contains two sub-figures. In the left sub-figure, the horizontal axis represents the percentage of an author's working as leading authors (i.e., first or last authors), while the vertical axis indicates the cosine similarity of collaborators' research topic—the greater two collaborators' research topic cosine similarity is, the less diversity they are. The color is proportional to the average number of citations received by the corresponding publications. In the left sub-figure, one can find that the top left and the bottom right corners of the heat map feature the most darkness, which demonstrates two patterns that tend to lead higher-impact scientific publications: (1) A greater topic diversity of collaborators plus more tendency to work as leading authors (including first and/or corresponding authors); and (2) a less topic diversity of collaborators plus less tendency to work as leading authors.

The right heat map of Figure 1 reveals the relationship between the percentage of working as leading authors (first and last authors of publications) of an author and his/her collaborators' impact diversities. The only difference between this heat map and its left one is the vertical axis—the current sub-figure shows the NSD of collaborators' *h*-indices while the left cosine similarity of their research topic. We find that authors who work as more leading authors and collaborate with less impact diversity researchers have tendencies to receive more citations than those with collaborators with greater impact diversity.

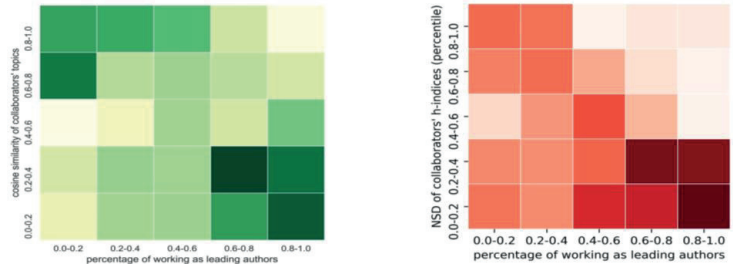

Figure 1. The relationship between the percentage of working as leading authors (first and last authors of publications) of an author and his/her collaborators' research topic (left) and impact (right) diversities. The darkness of cells shows the average number of citations received by the corresponding publications. Note that NSD does not range from zero to one, thus we represent its percentile instead in the vertical axis of the right sub-figure.

Following Amjad *et al.* (2017) and Bu *et al.* (2018d), we also investigate the difference of distributions of authors' sequence before and after their Ph.D. graduation, an important milestone in their scientific career. Due to the limitation of our dataset, we only employ a small sub-set in the dataset (~1K authors), in which we can find the authors' Ph.D. graduation year online. Table 1 shows the basic descriptions, where one can see that before Ph.D.

graduation, more than 70% of an author's publications are first-authored, but the number decreases to ~20% after he/she receives the doctoral degree. Reversely, the percentage of their last-authored publications increases from 8.2% to 39.5% after an author's graduation. The finding makes sense. Before Ph.D. graduation, students tend to work under their supervisor—researchers who tend to lead a study and work as the corresponding authors (more often than not, last authors)—and students themselves tend to write manuscripts, conduct empirical studies, and implement ideas, as pointed out by DeCastro *et al.* (2013) as well as Pachalen and Bhattacharya (2015). Yet, after Ph.D. graduation, authors might have their own students/postdocs, at which stage they might start to lead a certain study and work as corresponding authors (Gingras *et al.*, 2008; Rowlands & Nicholas, 2006). We also quantify research topic and impact diversity of authors' collaborators before and after they receive their doctoral degrees, as shown in the right part of Table 1, where the research topic diversity is equivalent to one minus the cosine similarity of collaborators' research topics and the impact diversity equals to the NSD of collaborators' *h*-indices. Specifically, one can find that the values for both of the two dimensions of diversities increase after an author got his/her Ph.D., though not quite obvious. The dual increasing found in the right part of Table 1 echoes our previous findings in Bu *et al.* (2018b).

Table 1. The distribution of first-, last-, and other-authored publications' percentage before and after authors' Ph.D. graduation, as well as their diversities.

	First-authored publication	*Last-authored publications*	*Middle-authored publications*	*Research topic diversity*	*Impact diversity*
Before graduation	73.8%	8.2%	18.0%	0.48	2.45
After graduation	21.4%	39.5%	39.1%	0.60	2.86

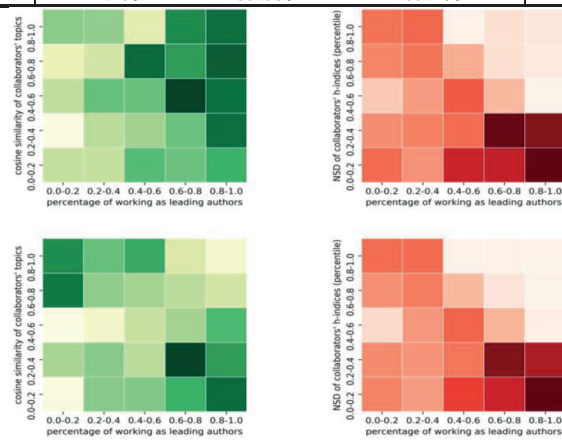

Figure 2. The relationship between the percentage of working as leading authors (first and last authors of publications) of an author and his/her collaborators' research topic (left) and impact (right) diversities. The above row indicates those before Ph.D. graduation, whilst the bottom row shows those after Ph.D. graduation.

We duplicate our experiments shown in Figure 1 on authors that we know their Ph.D. graduation years, and separately consider different patterns in terms of percentage of leading authors and their diversities (research topic and impact diversities) before and after their Ph.D. graduation. The above row in Figure 2 indicates those before Ph.D. graduation, whilst the bottom row shows those after Ph.D. graduation. As shown in the top left sub-figure, one can observe that Ph.D. student who have tendency to work as leading authors with lower research topic diversity of collaborators tend to have higher-impact work, which is different from the

pattern revealed in the left sub-figure in Figure 1. Yet, the left bottom sub-figure of Figure 2 looks quite similar to the left sub-figure in Figure 1, partly because of the dominant number of publications authored by post-doctoral researchers among our dataset. As for the impact diversity of collaborators, we surprisingly find that the right two sub-figures in Figure 2 look similar to the right sub-figure in Figure 1, indicating a uniform pattern between authors' sequence and their collaborators' impact diversity, regardless of before or after the authors' Ph.D. graduation.

Conclusion Remarks and Future Work

In this research-in-progress paper, we investigate the relationship between the sequence of authors in the publication bylines and the diversity of their collaborators. The diversity of collaborators is quantified with two dimensions, namely topic and impact diversities. There are many potential implications and applications regarding the finding and the approach of this study. As pointed out by Klein and Falk-Krzensinski (2017), for instance, the Computing Research Association has been grounding generic recommendations in the information science, computing, and engineering fields in a project named *Promotion and Tenure of Interdisciplinary Faculty*. In the project, they not only highlight the interdisciplinarity in job interviews but also emphasize their proposed collaboration-based center/institute to "seek advice on how to balance participation on large team projects with work that establishes a strong individual reputation" (p. 1056).

Using the ArnetMiner dataset containing ACM-indexed publications in computer science, we find that the following two patterns tend to lead higher-impact scientific publications: (1) greater topic diversity of collaborators plus more tendency to work as leading authors (including first and/or corresponding authors); and (2) less topic diversity of collaborators plus less tendency to work as leading authors. Meanwhile, from the perspective of impact diversity, the results of our empirical study show that authors who work as more leading authors and collaborate with less impact diversity researchers have tendencies to receive more citations than those with collaborators with greater impact diversity. We also detect different patterns of authors' first- and last-authored publications before and after their Ph.D. graduation.

There are some future works following this paper. Firstly, we are going to distinguish first and corresponding authors more in detail, as well as other impact related indicators to ensure the robustness of the findings. Secondly, we will follow Bu *et al.* (2018d) to set up several milestones (e.g., Ph.D. graduation, 5 years after Ph.D. graduation) and will consider authors' scientific collaborations with "giants." Thirdly, we will conduct similar empirical studies in various disciplines and implement more comparisons among disciplines with computer science. Moreover, many other issues should be considered in our following-up study, such as collaborators' contribution (Lu *et al.*, 2018) and joint effect of mobility and scientific collaboration (Chinchilla-Rodríguez *et al.*, 2018b). Furthermore, the empirical results shown in Table 1 might be biased. The small sub-dataset used is derived from their "Group A" in Amjad *et al.* (2017), those who have ever collaborated with "giants" at least once in their career. The potential effects of collaborating with "giants" on their author sequence patterns might exist, and this issue will also be researched and discussed in our future work by increasing the sample size for those authors that their Ph.Ds. are known, as well as adding an extra category to distinguish between postdocs and senior researchers are crucial for the reliability of the results.

References

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.

Bu, Y., Ding, Y., Liang, X., & Murray, D.S. (2018a). Understanding persistent scientific collaboration. *Journal of the Association for Information Science and Technology, 69*(3), 438-448.
Bu, Y., Ding, Y., Xu, J., Liang, X., Gao, G., & Zhao, Y. (2018b). Understanding success through the diversity of collaborators and the milestone of career. *Journal of the Association for Information Science and Technology, 69*(1), 87-97.
Bu, Y., Murray, D. S., Ding, Y., Huang, Y., & Zhao, Y. (2018c). Measuring the stability of scientific collaboration. *Scientometrics, 114*(2), 463-479.
Bu, Y., Murray, D. S., Xu, J., Ding, Y., Ai, P., Shen, J., & Yang, F. (2018). Analyzing scientific collaboration with "giants" based on the milestones of career. *Proceedings of the Association for Information Science and Technology, 5*5(1), 29-38.
Chinchilla-Rodríguez, Z., Miguel, S., Perianes-Rodríguez, A., Sugimoto, C.R. (2018a). Dependencies and autonomy in research performance: examining nanoscience and nanotechnology in emerging countries. *Scientometrics, 115*(3), 1485–1504.
Chinchilla-Rodríguez, Z., Bu, Y., Robinson-García, N., Costas, R., & Sugimoto, C.R. (2018b). Travel bans and scientific mobility: Utility of asymmetry and affinity indexes to inform science policy. *Scientometrics, 116*(1), 569-590.
Chinchilla-Rodríguez, Z., Sugimoto, C.R., & Larivière, V. (2019). Follow the leader: On the relationship between leadership and scholarly impact in international collaborations. *PLoS One*.
DeCastro, R., Sambuco, D., Ubel, P.A., Stewart, A., & Jaqsi, R. (2013). Mentor networks in academic medicine: Moving beyond a dyadic conception of mentoring for junior faculty researchers. *Journal of the Association of American Medical Colleges, 88*(4), 488-496.
Gingras, Y., Larivière, V., Macaluso, B., & Robitaille, J.-P. (2008). The effects of aging on researchers' publication and citation patterns. *PLoS One, 3*(12), e0004048.
Goring, S.J., Weathers, K.C., Dodds, W.K., Soranno, P.A., Sweet, L.C., Cheruvelil, K.S., ... & Utz, R.M. (2014). Improving the culture of interdisciplinary collaboration in ecology by expanding measures of success. *Frontiers in Ecology and the Environment, 12*(1), 39-47.
He, B., Ding, Y., & Yan, E. (2012). Mining patterns of author orders in scientific publications. *Journal of Informetrics, 6*(3), 359-367
Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academic Science of United States of America, 102*(46), 16569-16572.
Katz, J.S. (1994). Geographical proximity and scientific collaboration. *Scientometrics, 31*(1), 31-43.
Klein, J.T., & Falk-Krzensinski, H.J. (2017). Interdisciplinary and collaborative work: Framing promotion and tenure practices and policies. *Research Policy, 46*(6), 1055-1061.
Leahey, E., Beckman, C.M., & Stanko, T.L. (2017). Prominent but less productive: The impact of interdisciplinarity on scientists' research. *Administrative Science Quarterly*, 62(1), 105-139
Lu, C., Ding, Y., Zhang, Y., Bu, Y., & Zhang, C. (2018). Types of scientific collaborators: A perspective of author contribution network. In *Proceedings of iConference 2018*, March 25-28, 2018, Sheffield, U.K.
Newman, M.E.J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America, 98*(2), 404-409.
Packalen, M., & Bhattacharya, J. (2015). Age and the trying out of new ideas. National Bureau of Economic Research. DOI: 10.3386/w20920.
Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., & Stirling, A. (2012). How journal rankings can suppress interdisciplinary research: A comparison between innovation studies and business & management. *Research Policy*, *41*(7), 1262-1282
Rowlands, I., & Nicholas, D. (2006). The changing scholarly communication landscape: An international survey of senior researchers. *Learned Publishing, 19*(1), 31-55.
Roy, E.D, Morzillo, A.T, Seijo, F, Reddy, S.M.W, Rhemtulla, J.M, Milder, J.C, …, & Martin S.L. (2013). The elusive pursuit of interdisciplinarity at the Human-Environment Interface, *BioScience,* 63: 745-753.
Stokols, D., Hall, K.L., Taylor, B.K., & Moser, R.P. (2008). The science of team science: Overview of the field and introduction to the supplement. *American Journal of Preventive Medicine, 35*(2), S77-S89.
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.
Wagner, C. S., Whetsell, T. A., & Mukherjee, S. (2019). International research collaboration: Novelty, conventionality, and atypicality in knowledge recombination. *Research Policy, 48*(5), 1260-1270.
Wu, L., Wang, D., & Evans, J.A. (2019). Large teams develop and small teams disrupt science and technology. *Nature, 566*(7744), 378.
Zhang, C., Bu, Y., & Ding, Y. (2019). Does diversity of team members affect scientific success of a team? A preliminary study. In *Proceedings of iConference 2019*, March 31-April 3, 2019, Washington D.C., U.S.A.
Zhang, C., Bu, Y., Ding, Y., & Xu, J. (2018). Understanding scientific collaboration: Homophily, transitivity, and preferential attachment. *Journal of the Association for Information Science and Technology, 69*(1), 72-86.