A Preliminary Analysis of Geography of Collaboration in Data Papers by S&T Capacity Index

Chen, Pei-Ying Indiana University Bloomington, USA | peiychen@iu.edu
Li, Kai Renmin University of China, People's Republic of China | kai.li@ruc.edu.cn
Jiao, Chenyue University of Illinois at Urbana-Champaign, USA | cjiao4@illinois.edu

ABSTRACT
Geography is one of the defining factors in scientific collaboration. Despite the voluminous evidence for how geographical proximity shapes the formation of collaboration in research articles, it has been rarely examined in the emerging genre of data papers, one that describes research data and has enjoyed growing attention in the data-driven paradigm of research. This poster presents preliminary findings from our project that aims to evaluate the geographical dynamics behind the production of data papers. We analyze how researchers from different countries collaborate with one another using 6,821 data papers published in *Scientific Data* and *Data in Brief* between 2014 and 2020. We found that data papers rely heavily upon domestic collaboration and the collaboration pattern largely mirrors that of research articles, although some distinctiveness was also observed. We discuss future work in conclusion, with the ultimate goal of opening a more meaningful conversation about the relationship between the data-driven paradigm and knowledge production.

KEYWORDS
data papers; geography; scientific collaboration

INTRODUCTION
The famous quote from Louis Pasteur that "science knows no country" illustrates the universalism of the scientific system. However, it is also obvious that "science takes place" (Olechnicka, Ploszaj & Celińska-Janowicz, 2018; p. 4). While empirical studies have shown that geographical proximity is positively related to the formation of collaboration networks (Hoekman, Frenken & van Oort, 2009; Pan, Kaski & Fortunato, 2012), cultural/linguistic affinity as well as historical and socio-economic factors are also important determinants in international collaboration (Zitt, Bassecoulard & Okubo, 2000).

As an emerging scientific genre that describes research data objects, data papers introduce a novel mode of knowledge production and presentation under the data-driven paradigm of research (Li & Jiao, 2022). However, we are yet to know whether there are distinct geographical dynamics in the production of data papers (as compared to research articles) in terms of collaboration pattern and correspondence between authors' physical locations and their subjects of study. In this poster, we report preliminary findings on the geography of collaboration among data paper authors, as the first step towards a more thorough understanding of the impact of the data-driven paradigm on knowledge production.

DATA AND METHODS
In this study, we collected 504 data papers published in *Scientific Data* and 6,332 in *Data in Brief* from Scopus on November 15, 2020. The two journals were selected as the two leading exclusively data journals (Kim, 2020; Walters, 2020). We extracted country names from authors' affiliations, using the *countrycode* R package (Arel-Bundock, Enevoldsen & Yetman, 2018) to examine the collaboration pattern at the country level. After excluding papers with no identifiable country-level information on authors' affiliations, the final sample consisted of 6,821 data papers for the present analysis.

We distinguished between single-author, domestic collaboration, as well as leading versus supporting author(s) in international collaboration based on the number of authors, number of countries, and order of countries in authors' affiliation information per paper. To facilitate discussion, we adopted the S&T Capacity Index developed by Wagner, Brahmakulam, Jackson, Wong and Yoda (2001) that classifies 150 countries/territories into four groups: scientifically advanced countries (SAC), scientifically proficient countries (SPC), scientifically developing countries (SDC), and scientifically lagging countries (SLC). Our sample comprises 146 countries/territories, including 22 SAC, 23 SPC, 20 SDC, and 61 SLC.

PRELIMINARY FINDINGS
Over two thirds of the data papers (67.7%) in our sample result from domestic collaboration (i.e., co-authors in the same country), 28.6% international collaboration, and 3.7% single-author papers. The most productive country is the United States (n=1,330), which authored papers more than twice as much as China (n=612), the second most productive country. Figure 1A shows the top 20 most productive countries and their collaboration profiles. While just over half of them are SAC, it is notable the remaining half is composed of not only SPC (China, India, Spain,

85th Annual Meeting of the Association for Information Science & Technology | Oct. 29 – Nov. 1, 2022 | Pittsburgh, PA.

Brazil, and South Africa) but also SDC (Iran and Colombia) and SLC (Nigeria and Malaysia). Moreover, there seems to be a trade-off between % of domestic collaboration and % supporting authorship in international collaboration, as the top five countries in the latter are also the bottom five in the former. If we consider only the number of leading and supporting authorship in international collaboration, it appears that SAC and SPC are more likely to be in supporting roles, while countries with more leading than supporting authorship tend to be SDC and SLC (Figure 1B). As both figures show, however, there are considerable within-group variations, indicating there are complex factors that weigh in on collaboration dynamics.

A further examination of the collaboration pattern by S&T capacity in leading and supporting authorship roles reveals that SAC are the most popular collaborating partners for countries across all four groupings, especially among SPCs—nearly 70% of the SPC-led papers have co-authors in SAC. While researchers tend to collaborate with their counterparts in countries with higher S&T capacities, a significant share (26%) of collaboration between SLCs is also present (Figure 1C). In fact, SAC and SLC are both groups with relatively high proportions of within-group collaboration. Although numerous studies have noted the frequent collaboration between SACs (Gazni, Sugimoto & Didegah, 2012; Wagner et al., 2001), the collaboration between SLCs is less well documented, let alone in data papers. Given that the majority of SLCs are in the Global South, which is shown to have comparative advantages in disciplines related to natural resources and infectious diseases (Miao et al., 2022), it would be interesting to investigate how the observed collaboration patterns in data papers vary by discipline.

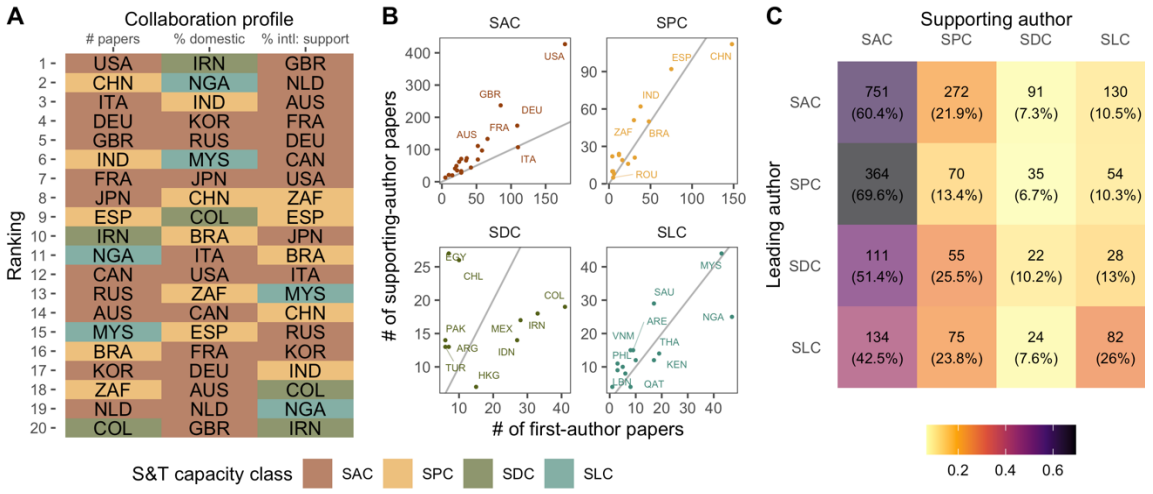

Figure 1. (A) Collaboration profile among top 20 most productive countries. (B) First- vs. supporting-author papers among countries with at least 19 papers (median). (C) Collaboration pattern by S&T capacity. *Note*: (1) Qatar is manually assigned to SLC. (2) The gray lines represent diagonals with intercept=0 and slope=1.

CONCLUSION

In this poster, we present preliminary findings from our project that aim to investigate the geographical dynamics in the production of data papers. Early evidence suggests that the overall collaboration pattern resembles that of the general scientific research, especially the predominance of domestic collaboration and the high concentration of collaboration not only between SACs but also between other less developed countries and SACs. However, we also found a relatively high proportion of within-group collaboration among SLCs, which warrants further investigation into possible cross-discipline variation.

Based on these results, there are three possible directions to be taken in future research. First, we will examine the geographic proximity between authors' affiliations and their subject of study to better understand whether such proximity plays a role in the division of labor in collaboration. Second, we will take discipline into consideration to illustrate the more granular co-authorship patterns in data papers. Third, we will compare the collaboration patterns of data papers with those of research articles more systematically to get a better idea of the distinct characteristics of data papers as a new academic genre.

ACKNOWLEDGMENTS

We thank three anonymous reviewers who wrote and provided helpful comments on previous versions of this document.

REFERENCES

Arel-Bundock, V., Enevoldsen, N. & Yetman, C. (2018). countrycode: An R package to convert country names and country codes. *Journal of Open Source Software, 3*(28), 848. doi:10.21105/joss.00848

Gazni, A., Sugimoto, C. R. & Didegah, F. (2012). Mapping world scientific collaboration: Authors, institutions, and countries. *Journal of the American Society for Information Science and Technology, 63*(2), 323-335. doi:10.1002/asi.21688

Hoekman, J., Frenken, K. & van Oort, F. (2009). The geography of collaborative knowledge production in Europe. *The Annals of Regional Science, 43*(3), 721-738. doi:10.1007/s00168-008-0252-9

Kim, J. (2020). An analysis of data paper templates and guidelines: Types of contextual information described by data journals. *Science Editing, 7*(1), 16-23. doi:10.6087/kcse.185

Li, K. & Jiao, C. (2022). The data paper as a sociolinguistic epistemic object: A content analysis on the rhetorical moves used in data paper abstracts. *Journal of the Association for Information Science and Technology, 73*(6), 834-846. doi:10.1002/asi.24585

Miao, L., Murray, D., Jung, W.-S., Larivière, V., Sugimoto, C. R. & Ahn, Y.-Y. (2022). The latent structure of global scientific development. *Nature Human Behaviour*. doi:10.1038/s41562-022-01367-x

Olechnicka, A., Ploszaj, A. & Celińska-Janowicz, D. (2018). *The geography of scientific collaboration*. Routledge.

Pan, R. K., Kaski, K. & Fortunato, S. (2012). World citation and collaboration networks: Uncovering the role of geography in science. *Scientific Reports, 2*(1). doi:10.1038/srep00902

Wagner, C. S., Brahmakulam, I. T., Jackson, B. A., Wong, A. & Yoda, T. (2001). *Science & technology collaboration: Building capacity in developing countries?*. RAND Corporation. Retrieved from https://www.rand.org/pubs/monograph_reports/MR1357z0.html

Walters, W. H. (2020). Data journals: Incentivizing data access and documentation within the scholarly communication system. *Insights: the UKSG journal, 33*. doi:10.1629/uksg.510

Zitt, M., Bassecoulard, E. & Okubo, Y. (2000). Shadows of the past in international cooperation: Collaboration profiles of the top five producers of science. *Scientometrics, 47*(3), 627-657. doi:10.1023/a:1005632319799