High Performers Emerged from Data-to-Knowledge Pathways

Jian Qin[1], Jeff Hemsley[1], Sarah Bratt[2], Alex Smith[1]
[1] School of Information Studies, Syracuse University, Syracuse, NY 13244, USA
[2] School of Information, University of Arizona, Tucson, AZ 85721, USA

Keywords: Collaboration capacity, Metadata analytics, Research performance

Extended Abstract

Collaboration capacity rerfers to “the ability of a researcher, or group of researchers, to garner collaborators and maintain a productive relationship among the collaborators” (Qin, Hemsley, Bratt, 2018 & 2022). It is a framework of metrics developed for assessing the effectiveness of collaboration enablers in facilitating successful collaborations, fostering the growth of Scientific & Technical (S&T) human capital, and more importantly, accelerating innovations and new discoveries. The magnitude and increment/decrement of collaboration capacity over time can be an indicator of the impact and effectiveness of its enablers: science policy, cyberinfrastructure, and S&T human capital. While this framework (Figure 1) is theoretically reasoned and still under development, we face the challenge in how to operationalize the assessment of effectiveness and impact of collaboration capacity enablers. If we consider the measures in the data production and datta-to-knowledge boxes in Figure 1 as independent variables, what would be a suitable measure for the dependent variable Collaboration Capacity (CC)?

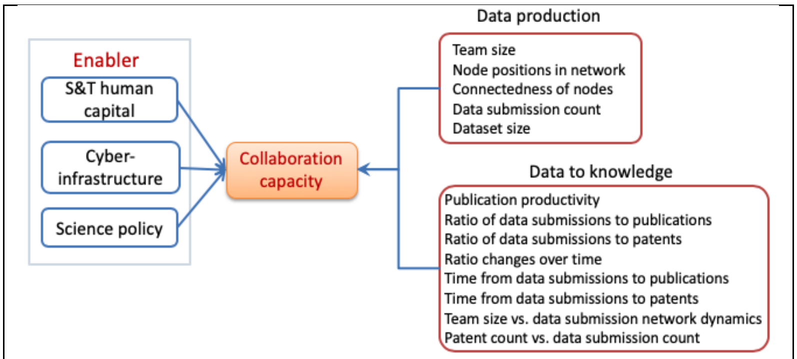
Enabler
S&T human capital
Cyber-infrastructure
Science policy
Collaboration capacity
Data production
Team size
Node positions in network
Connectedness of nodes
Data submission count
Dataset size
Data to knowledge
Publication productivity
Ratio of data submissions to publications
Ratio of data submissions to patents
Ratio changes over time
Time from data submissions to publications
Time from data submissions to patents
Team size vs. data submission network dynamics
Patent count vs. data submission count

Figure 1. The collaboration capacity framework version 2.0 (Suorce: Qin et al., 2023)

One solution we used is the number of new collaborators a researcher can garner within a certain time period (Qin, Hemsley, & Bratt, 2022). However, there could be exceptions, for example, a researcher may have a stable collaboration circle over time, i.e., little or no change in the number of new collaborators, but remained productive, or a researcher has a relatively small team size over time but made disruptive discoveries (Wu, Wang, & Evans, 2019). A high collaboration capacity can be corrrelated with performance, but is not sufficient to conclude that the greater the collaboration capacity, the better the performance. The goal of measuring collaboration capacity is to find an optimal point at which the enablers (science policy, cyberinfrastructure, and S&T human capital) can help maximize researcher’s collaboration capacity effectively. Obviously, by looking at the new collaborators added or reduced during a itme intetrval alone would not be enough.

This presentation shares an experiment result that examines whether the meaure of new collaborators gained over titme is an appropriate measure for CC by selecting the high performers to analyze their pathways from engagement in data production to publications. The argument is that the more the researchers engaged in data production in their early career, the more likely they are the active users of cyberinfrastructure (for data submissions and using data discovery and sharing services) and participants of their mentors’ projects. It would be no surprise if these researchers stayed in academia or career researcher paths and emerged as high performers at some point of their career. It is from this reasoning that we assume high performers tend to have an optimal collaboration capacity. To express this assumption in quantitative measures, we formulate two hypotheses:

H1: Authors who were active in data submissions in their early career are more likely to generate a high number of publications (high performers) later in their career.

H2: High performers measured by number of publications tend to have an optimal collaboration capacity.

To test these two hypotheses, we will first locate the range of the optimal collaboration capacity by examining multiple factors, including network statistical properties and qualitative tracking for selected high performers' career mobility and publication impact.

We will use the metadata collected from GenBank for both sequence submissions and associated publications for testing the two hypotheses. Methods used to collect and process the GenBank metadata has been reported in Qin, Hemsley, and Bratt (2022). The data spans from 1984-2021, which provides sufficient time frames for tracking authors through their career and identify change patterns in both data production and publication domains. Authors in this dataset are grouped into four performance groups based on the number of publications: consistently high, low-to-high, high-to-low, and consistently low. For the hypothesis testing, we will focus on the consistently-high and low-to-high performance groups.

In addition, we plan to triangulate funding (we have NIH funding data matched to GenBank authors in our dataset), team sizes, career pathways (data-creators-turned-researchers) for selected authors to conduct in-depth analysis. Work for tracking authors with different levels of performance is well underway.

Studying high performers emerged from data-to-knowledge pathways offers a different perspective on how effectvely the science policy, cyberinfrastructure, and S&T human capital help high performers to reach where they are. It adds a new understanding of the mechanisms of collaboration enablers and their tangible relations for the generation of high performers.

References

Hemsley, J., Qin, J., & Bratt, S. (2020). Data to knowledge in action: A longitudinal analysis of GenBank metadata. In: *Proc. Assoc. Info. Sci. Tech.,* https://doi-org.libezproxy2.syr.edu/10.1002/pra2.253

Qin, J., Bratt, S., Hemsley, J., & Smith, A.O. (2023). Metadata analytics: A methodological discussion. In: International Society of Scientometrics and Informetrics (ISSI) 2023 Conference, Bloomington, IN, July 3-5, 2003.

Qin, J., Hemsley, J., & Bratt, S. (2022). The structural shift and collaboration capacity in GenBank networks: A longitudinal study. Quantitative Science Study, 1-20. DOI: https://doi.org/10.1162/qss_a_00181; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9012484/

Qin, J., Hemsley, J., & Bratt, S. (2018). Collaboration capacity: Measuring the impact of cyberinfrastructure-enabled collaboration networks. Science of Team Science (SCITS) 2018 Conference, Galveston, Texas, May 21-24, 2018.

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, 566 (Feb. 21): 378-382.