Collaboration Capacity: Measuring the Impact of Cyberinfrastructure-Enabled Collaboration Networks

Jian Qin, Jeff Helmsley, Sarah Bratt
School of Information Studies
Syracuse University

SciTS 2018, Galveston, Texas, May 21-23, 2018
Background: Data-intensive biological research

“From 1982 to the present, the number of bases in GenBank has doubled approximately every 18 months.”


Image credit: https://www.nlm.nih.gov/about/2015CJ.html
GenBank’s big metadata as a source for quantitative studies of team science
Collaboration across countries, labs, and fields

- Big problems, big data (and big metadata), and big teams
- Relations between data production and paper publication
- Large scale studies of collaboration networks to find patterns, structures, and empirical evidence for in-depth exploration

The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*

F. Kunst, N. Ogasawara, A. Danchin


doi:10.1038/36786

Download Citation

Received: 16 July 1997
Accepted: 29 September 1997
Published: 20 November 1997
The collaboration capacity framework

**Collaboration capacity:** the ability of an individual researcher or a team of researchers to collaborate throughout the data production and publication lifecycle and sustain a network of collaborators over time.

**Assumptions:**
- Collaboration capacity is a proxy for studying scientific capacity
- Data, publication, and patent together can be used as a proxy for studying knowledge diffusion
- Collaboration capacity significantly affects the level of research productivity and extent of knowledge diffusion

**Assumptions:**
- Collaboration capacity is a proxy for studying scientific capacity
- Data, publication, and patent together can be used as a proxy for studying knowledge diffusion
- Collaboration capacity significantly affects the level of research productivity and extent of knowledge diffusion
Methods

• Source: metadata describing molecular sequences in GenBank

• Exploratory data analysis (EDA)
• Social Network Analysis (SNA)

Purpose: using descriptive stats and visualization techniques to look for patterns, structures, and problems

• Based on our framework, datasets generated include:
  • Size of collaboration networks for data submission
  • Extent of knowledge diffusion
  • Rate of knowledge diffusion
Findings

• Connectedness of collaboration networks
• Ratio of data submissions to publications
As early as 1994, the number of data submissions surpassed that of publications.
Connectedness vs. distributedness

Authors remain well connected over time

While more clusters of smaller communities emerged
Ratio of submissions to publications

- x axis: # of authors who submitted sequence data
- y axis: # of authors who published a paper associated with the data submissions
- After 1998, more authors were involved in data production than those in paper publications
- Significant increment in productivity:
  - Before 1998, majority had a range between 20 publications and 50 data submissions
  - Since 2008, a sizable # of authors had a high productivity in the range of 50~100 publications and 100~300 data submissions
A sharp increase in the average ratio of submission to publication: signaling a turning point for microbiology to become data-intensive science?
Discussion

• Big problems require as well as generate big data that need the orchestration of big teams
• Collaboration capacity framework currently contains quantitative metrics and can incorporate team science metrics in future research
• More data mining can be done to the GenBank metadata records:
  • What are the common team formation by sector, institution, and research field?
  • What team sizes sustained a high productivity?
  • How did the team sizes and connectedness correlate with productivity and knowledge diffusion?
  • What other measures can be applied to study collaboration capacity?
Conclusion

- Metadata in science data repositories (GenBank, Protein Bank, etc.) offers a great source for quantitative study of team science
  - New dimensions to look at team work in basic science research
  - At very large scales
  - Can be combined with local team data for productivity and impact assessment

- Implications of big metadata analytics on team science research will need to be explored further
Acknowledgements

The authors thank the support of NSF SciSIP grant #1561348.

Chensen Wang and Suchitra Deekshitula provided technical assistance for data processing and computational analysis.