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."

-- NCBI. (2017). Growth of GenBank and WGS, http://www.ncbi.nlm.nih.gov/genbank/statistics.

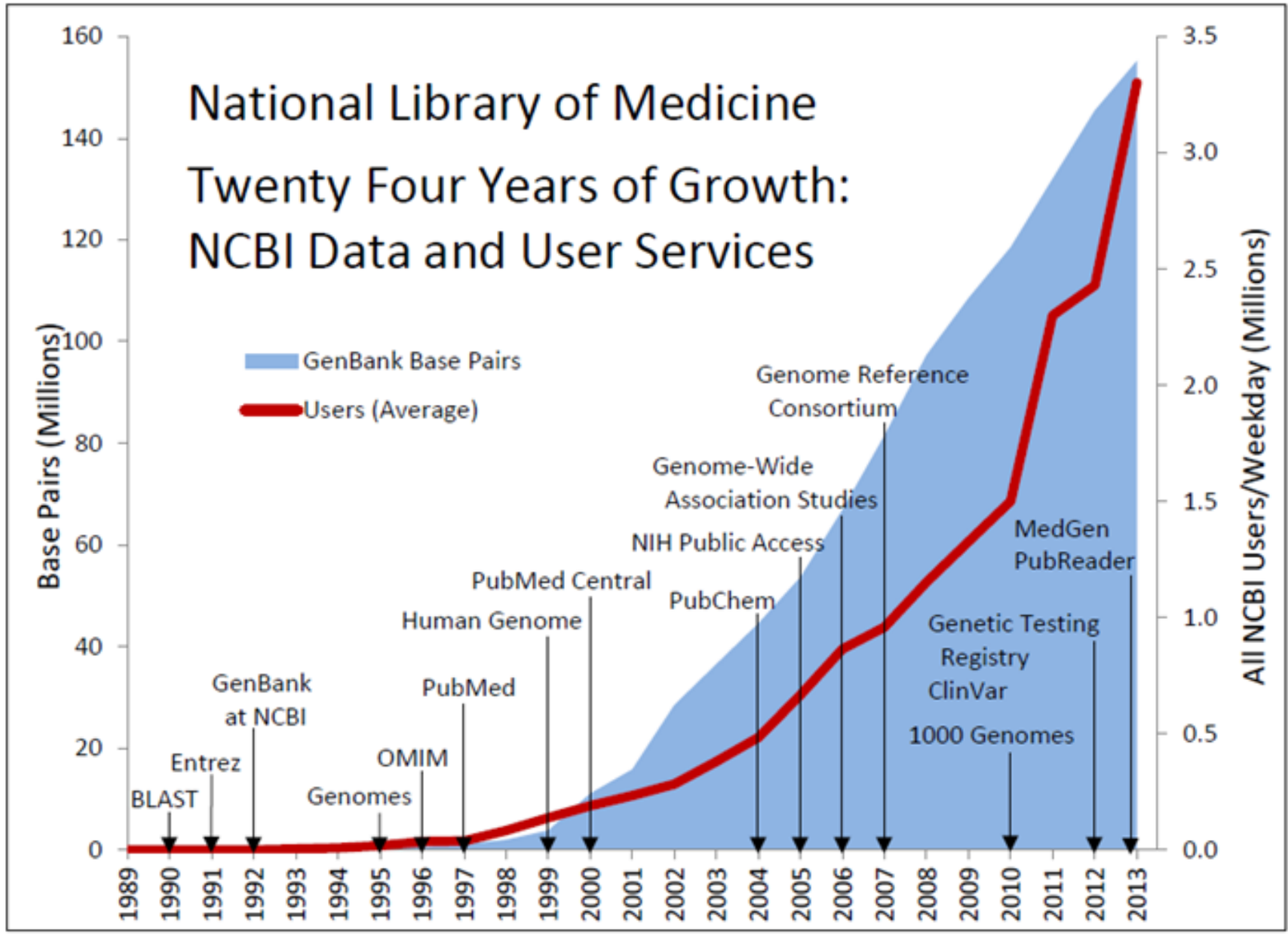


Image credit: https://www.nlm.nih.gov/about/2015CJ.html

GenBank's big metadata as a source for quantitative studies of team science

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
```

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

```
SOURCE      Bacillus subtilis subsp. subtilis str. 168
  ORGANISM  Bacillus subtilis subsp. subtilis str. 168
            Bacteria; Firmicutes; Bacilli; Bacillales; Bacillaceae; Bacillus.
REFERENCE   1
  AUTHORS   Kunst,F., Ogasawara,N., Moszer,I., Albertini,A.M., Alloni,G.,
            Azevedo,V., Bertero,M.G., Bessieres,P., Bolotin,A., Borchert,S.,
            Borriss,R., Boursier,L., Brans,A., Braun,M., Brignell,S.C.,
            Bron,S., Brouillet,S., Bruschi,C.V., Caldwell,B., Capuano,V.,
            Carter,N.M., Choi,S.K., Codani,J.J., Connerton,I.F., Cummings,N.J.,
            Daniel,R.A., Denizot,F., Devine,K.M., Dusterhoft,A., Ehrlich,S.D.,
            Emmerson,P.T., Entian,K.D., Errington,J., Fabret,C., Ferrari,E.,
            Foulger,D., Fritz,C., Fujita,M., Fujita,Y., Fuma,S., Galizzi,A.,
            Galleron,N., Ghim,S.Y., Glaser,P., Goffeau,A., Golightly,E.J.,
            Grandi,G., Guiseppi,G., Guy,B.J., Haga,K., Haiech,J., Harwood,C.R.,
            Henaut,A., Hilbert,H., Holsappel,S., Hosono,S., Hullo,M.F.,
```

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

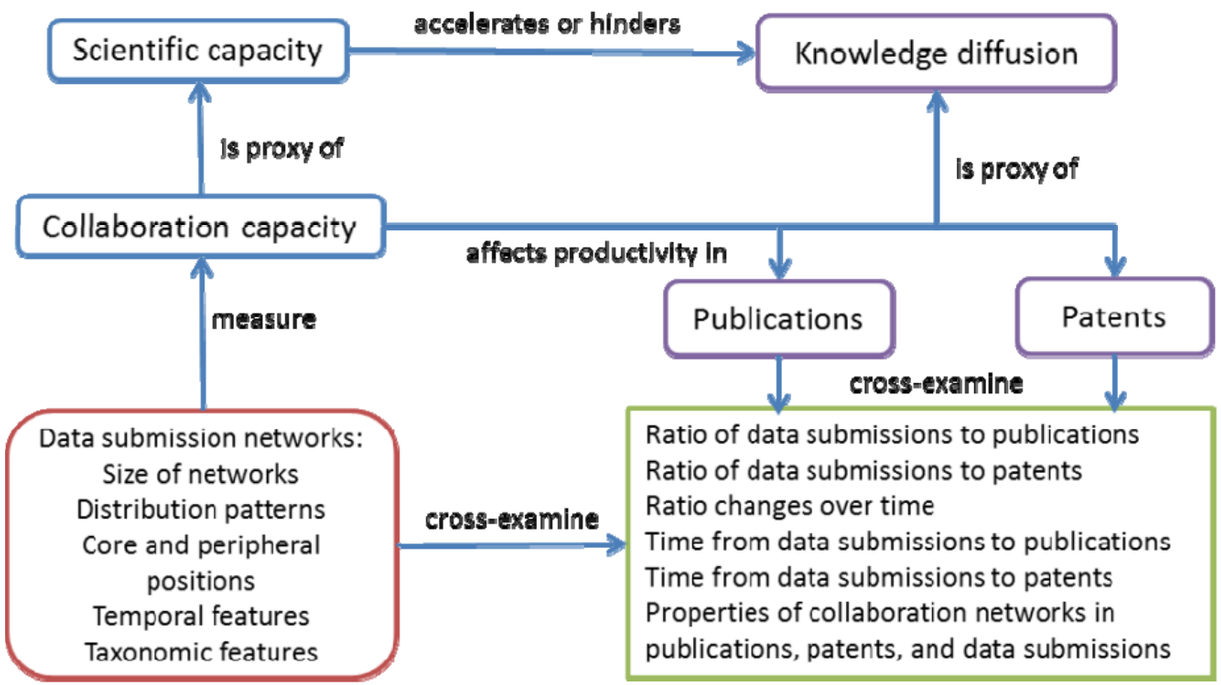
F. Kunst, N. Ogasawara [...] A. Danchin

Nature **390**, 249–256 (20 November 1997)
doi:10.1038/36786
Download Citation

Received: 16 July 1997
Accepted: 29 September 1997
Published: 20 November 1997

```
            Viari,A., Wambutt,R., Wedler,E., Wedler,H., Weitzenegger,T.,
            Winters,P., Wipat,A., Yamamoto,H., Yamane,K., Yasumoto,K., Yata,K.,
            Yoshida,K., Yoshikawa,H.F., Zumstein,E., Yoshikawa,H. and
            Danchin,A.
  TITLE     The complete genome sequence of the gram-positive bacterium
            Bacillus subtilis
  JOURNAL   Nature 390 (6657), 249-256 (1997)
   PUBMED   9384377
REFERENCE   2  (bases 1 to 14210)
  AUTHORS   Glaser,P.
  TITLE     Direct Submission
  JOURNAL   Submitted (25-JUN-1997) Philippe Glaser, Regulation de l'Expression
            Genetique, Institut Pasteur, 28 Rue du Dr Roux, Paris, 75724,
            France
```

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

Methods

- Source: metadata describing molecular sequences in GenBank

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

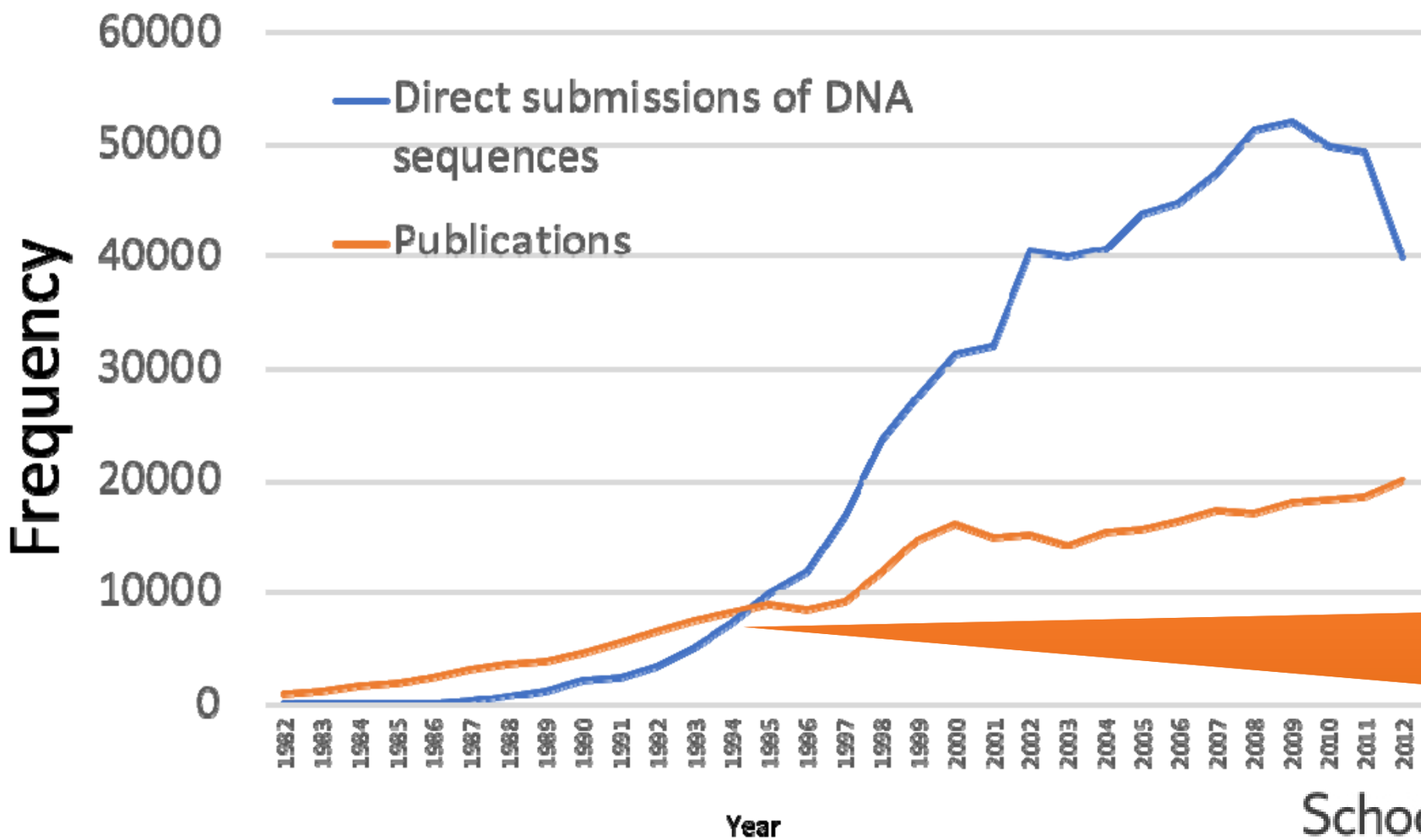
Purpose: using descriptive stats and visualization techniques to look for patterns, structures, and problems

- 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

Data submissions vs. publications

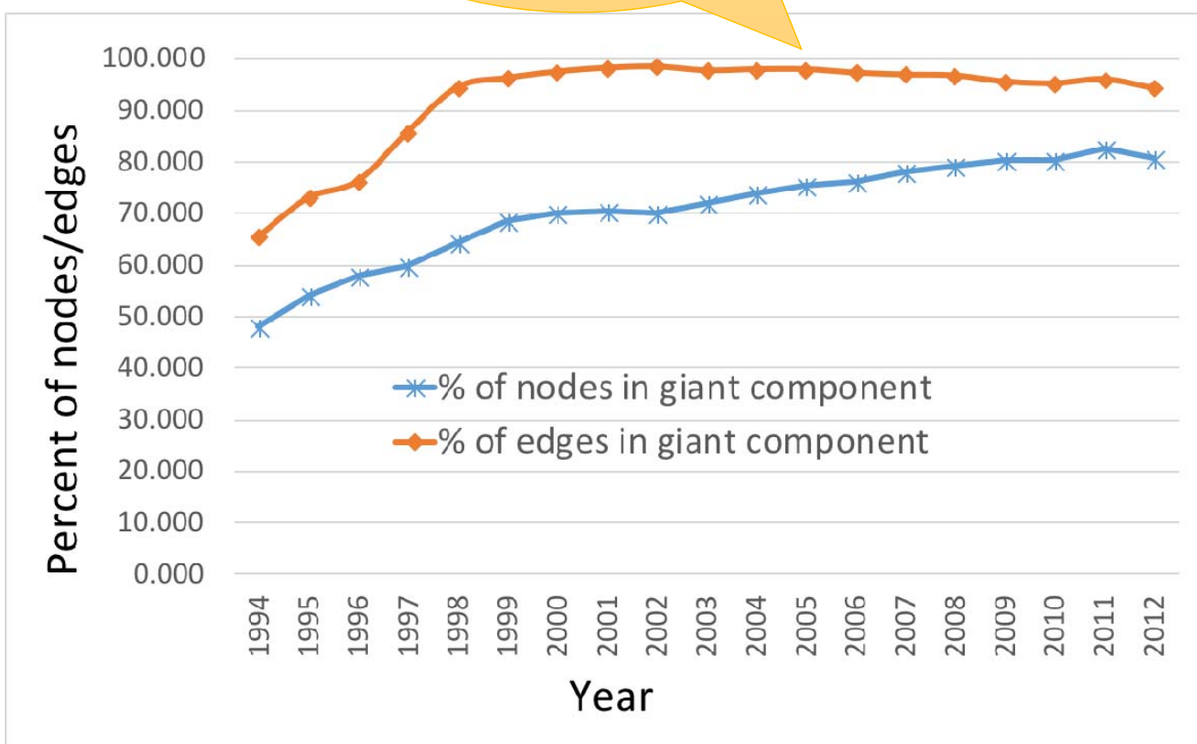


As early as 1994, the number of data submissions surpassed that of publications

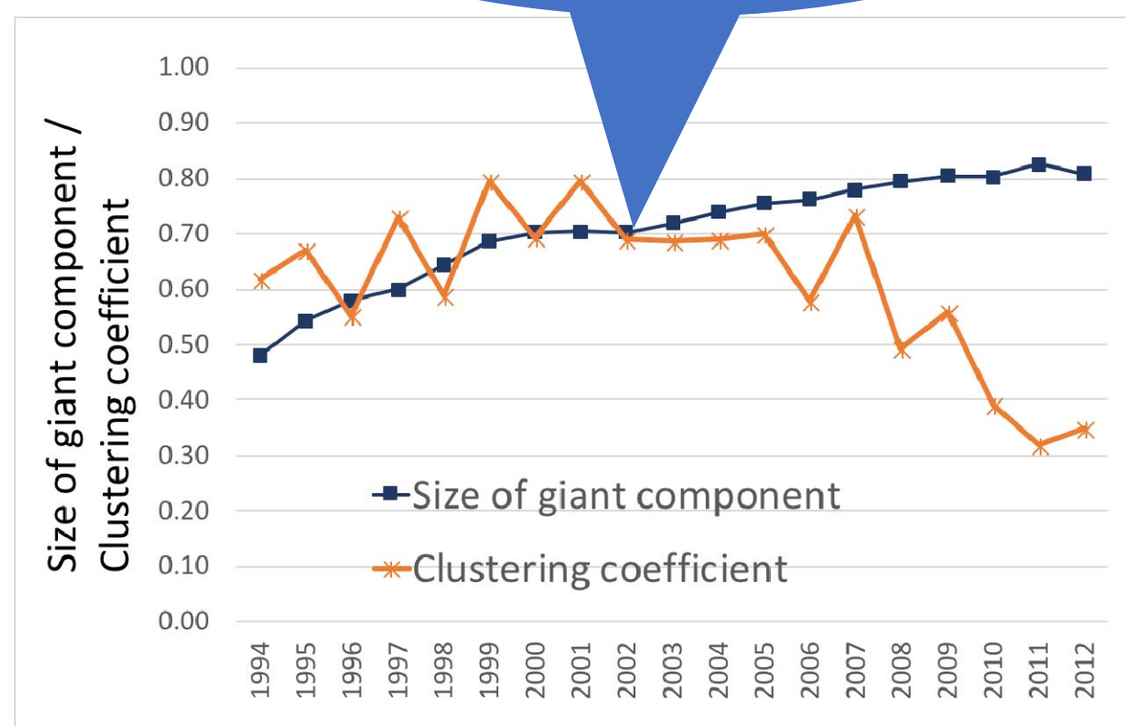
School of Information Studies
Syracuse University

Connectedness vs. distributedness

Authors remain well connected over time

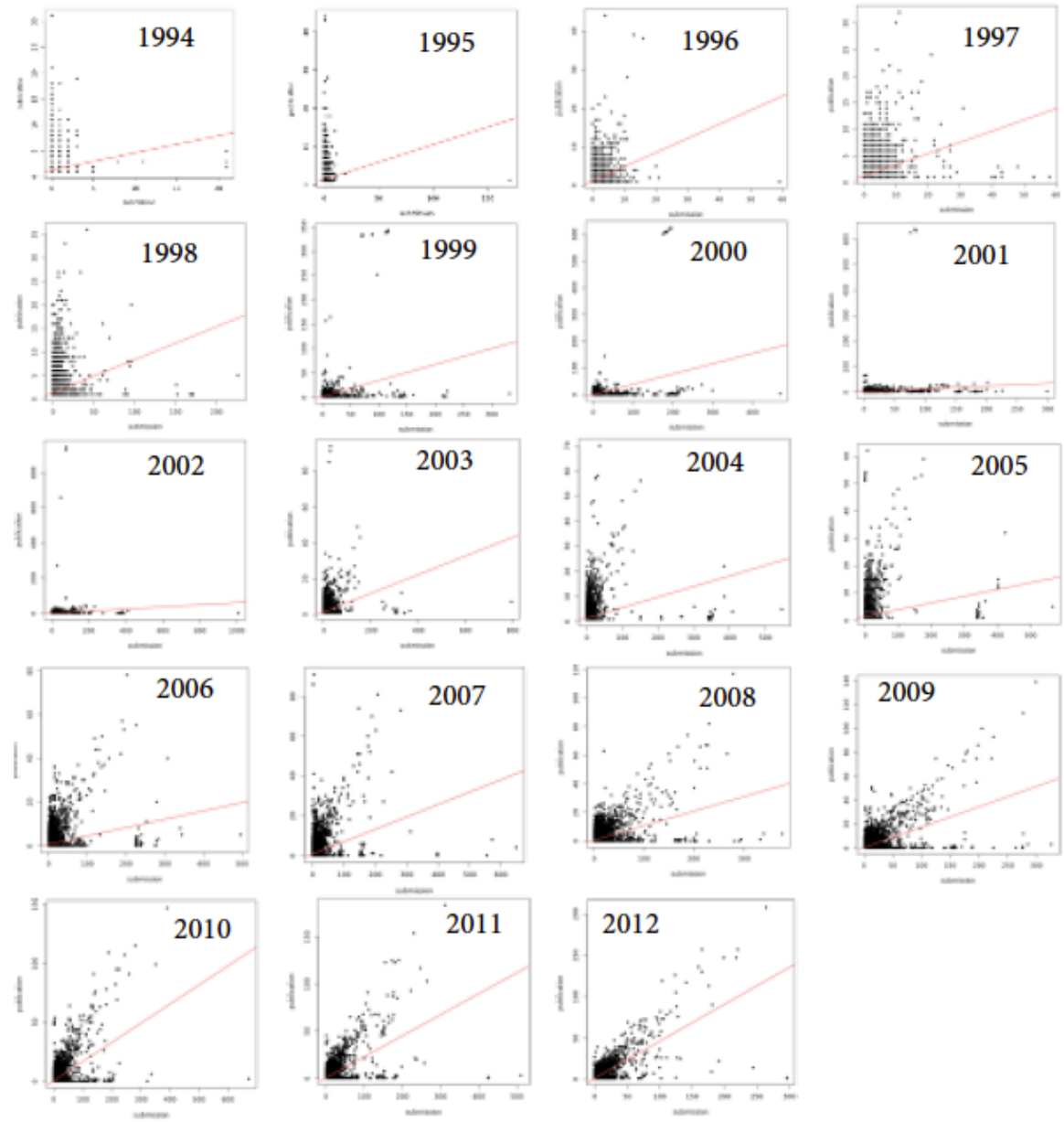


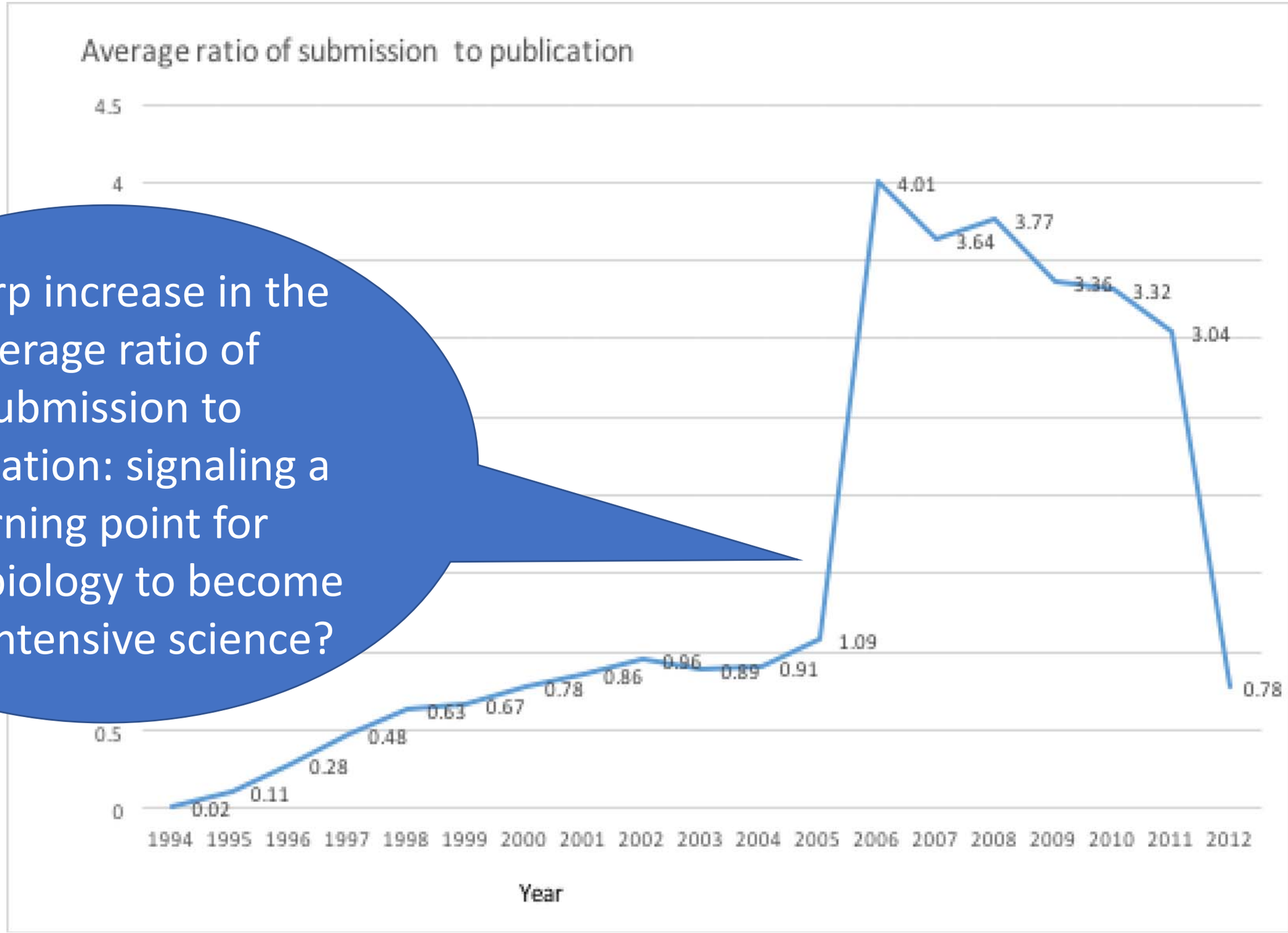
While more clusters of smaller communities emerged



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.