5°EBBC
encontro brasileiro de bibliometria e cientometria
SÃO PAULO 2016
6,7 e 8 de JULHO
Universidade de São Paulo

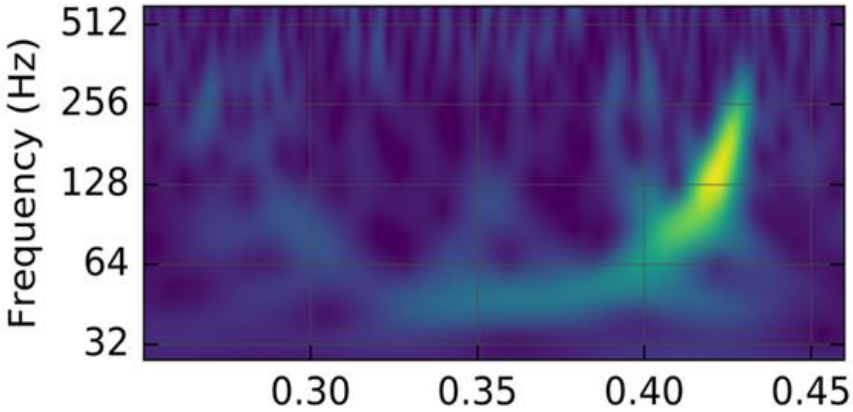
Bibliometrics Research in the Era of Big Data: Challenges and Opportunities

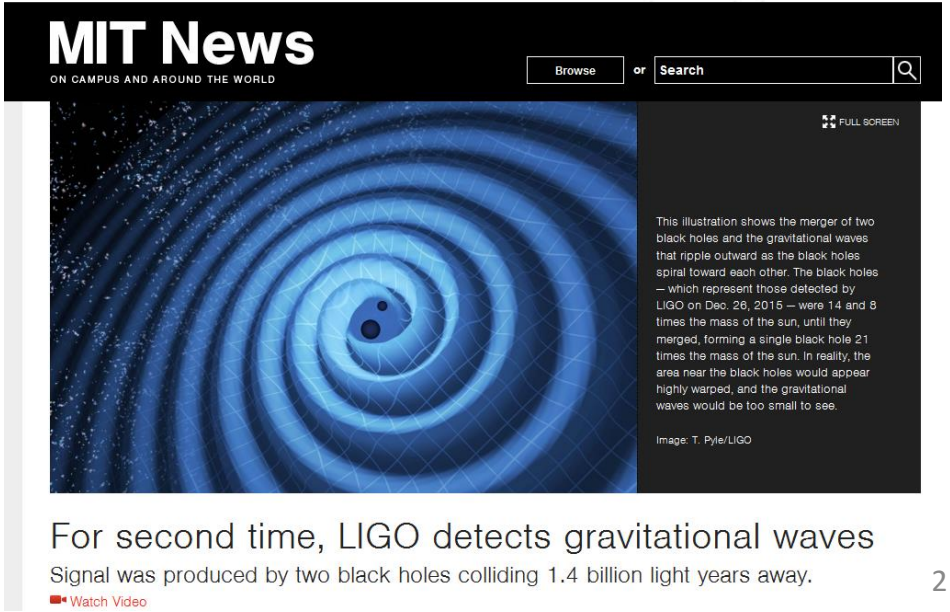
Dietmar Wolfram

University of Wisconsin-Milwaukee

Big Data Research

- LIGO Gravitational Wave Detection Observatory
- 1.7 billion files of data processed, requiring 26PB of storage





Presentation Organization

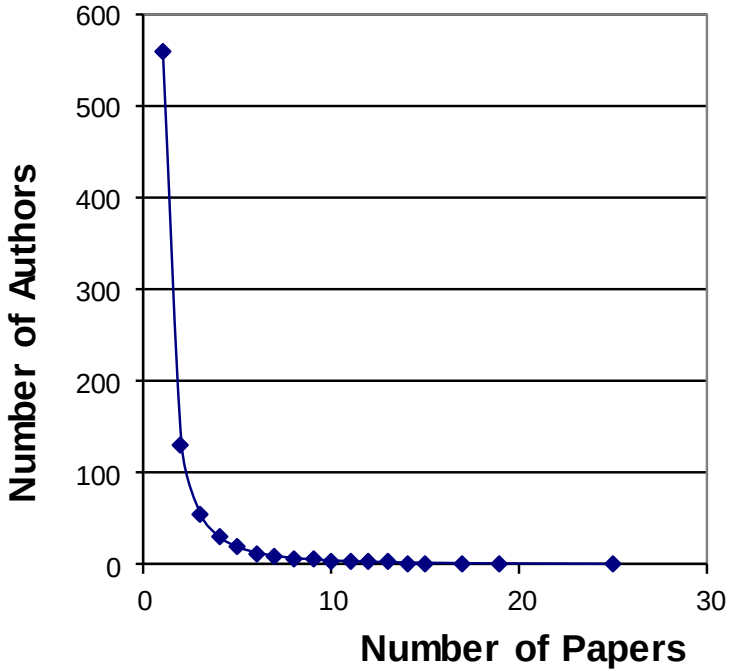
- Overview
- Current Issues
- Big data aspects of network-based & text-based bibliometric analysis
- Opportunities & ongoing challenges

Early Bibliometric Datasets

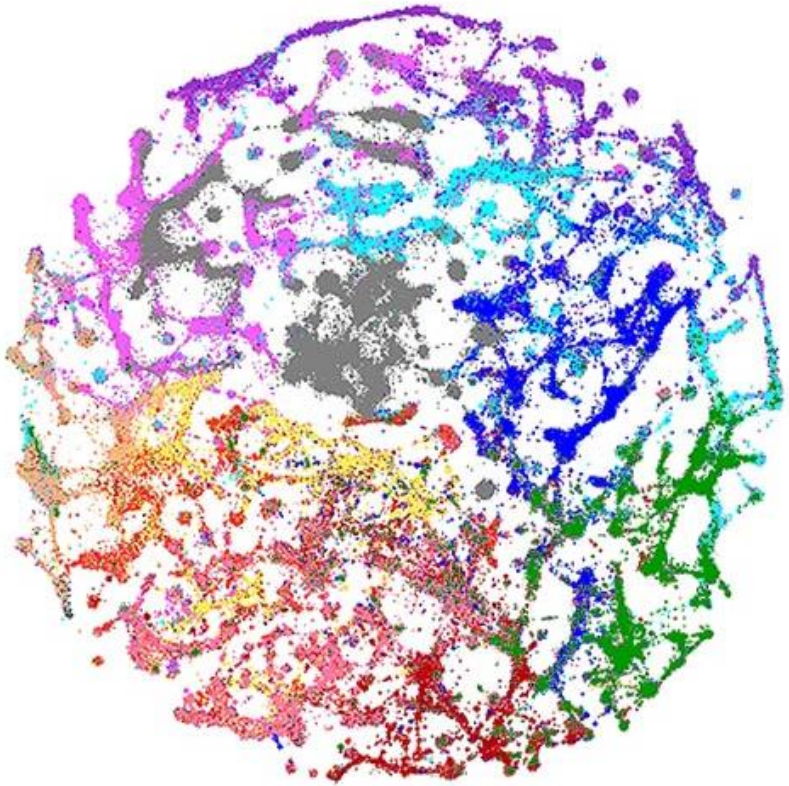
- **Lotka** (1926) – used 6891 & 1325 authors for his datasets
- **Bradford** (1934) – used 326 journals with 1332 references
- **Zipf** (1949) – generally used < 10K words types

> "Kaeding's total bulk of nearly 11 million running words so far overshoots a sample of optimum size that it is of little practical use to us." Section 3.IV

Visualization of Early Bibliometric Data



The relationship between authors and publications



Big Data in Bibliometrics Today

The relationship between 22M documents covering all areas of scholarship

http://www.mapofscience.com/

What has Changed?

- ✓ Much more available data
- ✓ Better data processing methods
- ✓ New analytical tools for numeric & textual data

What is “Big” in Bibliometrics?

- Bibliographic datasets are usually more bounded than data used in other scientific disciplines
  - Recorded discourses (articles, books) are still manageable
- For other areas, very large sets are available
  - Full text corpora
  - Web links
  - Social media data

Data Issues

1. Data accessibility
2. Size & dimensionality of datasets
3. Data analysis & summarization approaches for
    - Entities of interest (authors, papers, journals) based on network or text analysis
    - Revealing overt & hidden relationships

1) Data Accessibility

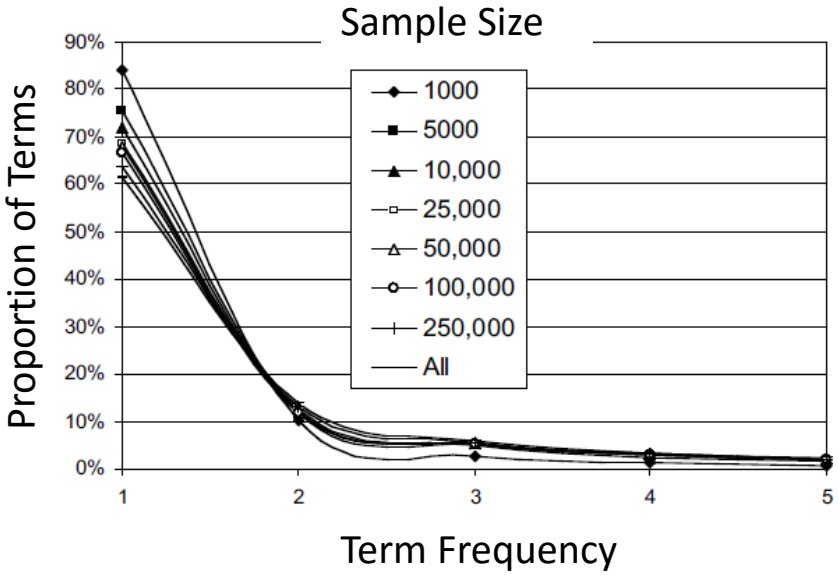
- Database providers may limit access to full data
  - If available, cost may be a limiting factor
- Privacy issues for personal & transaction log data
  - Few providers are now willing to share data
- Storage availability
  - Storage is cheap these days, so is data transfer

2a) Size: Is More Data Always Better?

- Complete data is a good idea … but is it good data?
  - Addresses the population and not just a sample
- Raw data may require extensive cleaning and standardization if from different sources
  - More data $\Rightarrow$ More initial processing
- "… brute force computation with big data may lead to false discoveries and spurious correlations …" Prathap (2014)

Example: The Impact of Dataset Size

- Dataset size may affect conclusions drawn for some types of bibliometric datasets



Ajiferuke, I., Wolfram, D., & Famoye, F. (2006). Sample size and informetric model goodness-of-fit outcomes: A search engine log case study. *Journal of Information Science*, 32(3), 212-222.

2b) Dimensionality: High Data Dimensionality Issues

- Large datasets & complex relationships lead to high dimensional data representation
  (e.g., vector space model)
- High dimensionality $\Rightarrow$ Larger computational overhead
- Need ways to reduce dimensionality without losing essence of relationships

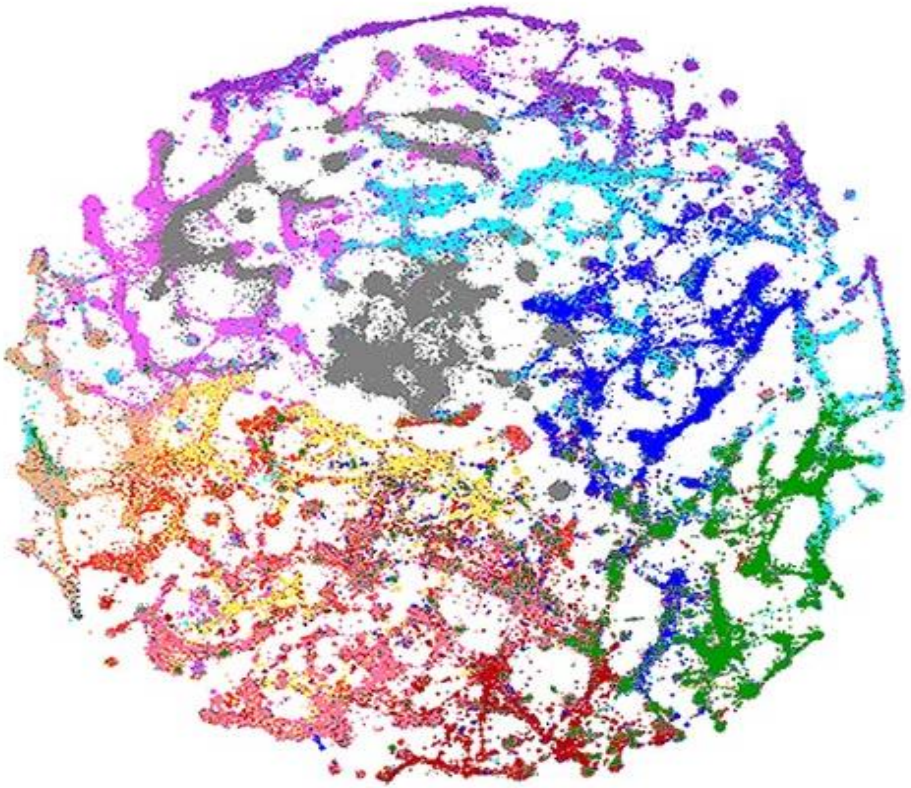
3) Processing Large Datasets

- Can apply dimensionality reduction based on statistical methods
  - Visualization (2 & 3 dimensions)
  - Factor analysis
  - Clustering
- Use mining and pattern detection methods to reveal hidden relationships
- Different methods may be needed for link-based data (e.g., citations, collaboration) and textual data

Data Analysis Challenges

- Many techniques are available
- Which is best, most reliable and most valid?
  - Different methods provide different outcomes
  - Some will provide a “goodness of fit” or loss function to indicate reliability of the outcome

Map of Human Knowledge



What does this tell us?

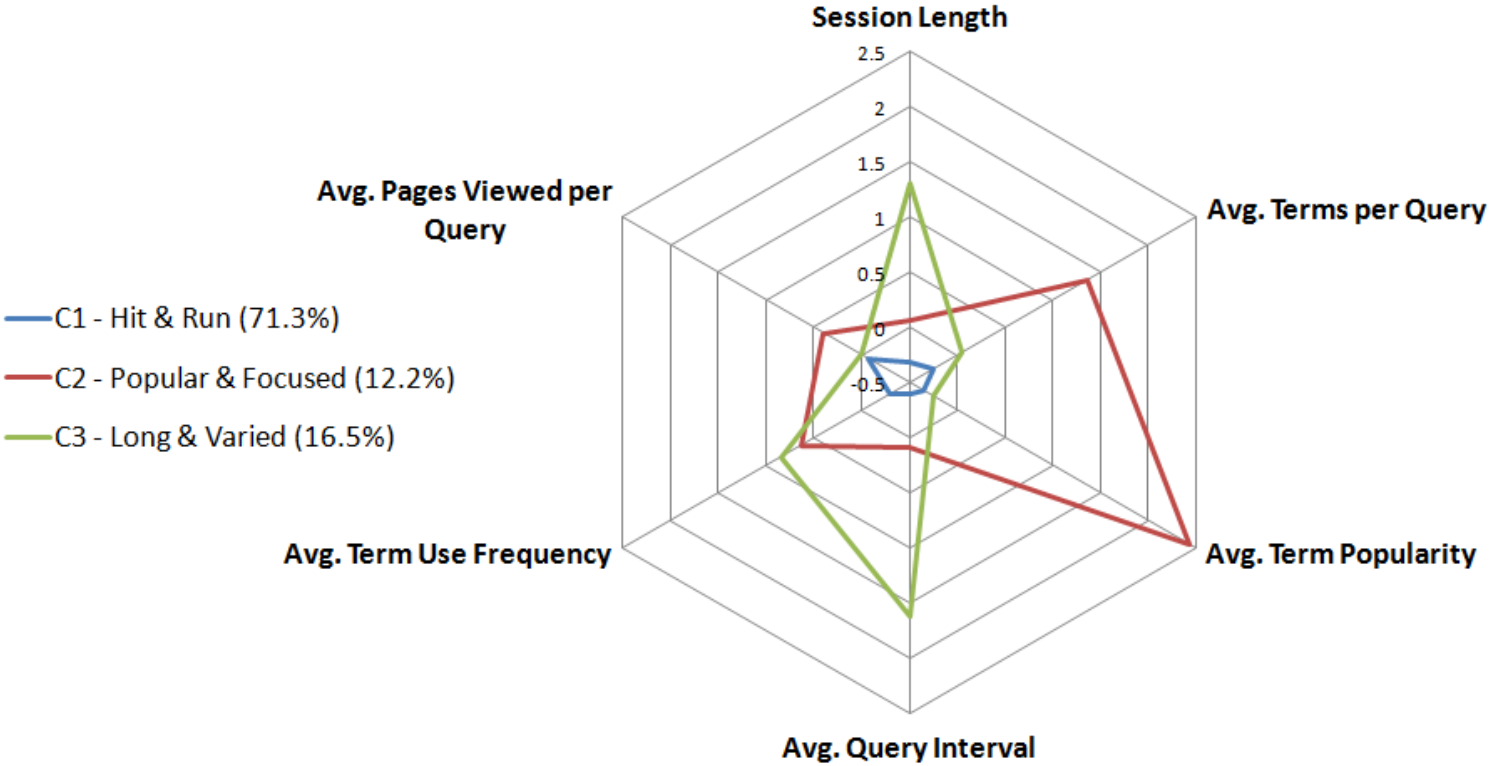
Does it prove anything?

Exploration of Large Datasets

- Software and statistical tools permit summarization of large datasets

- Can rely on numeric or textual features of data
  - Bibliographic data (surrogates, full text)
  - Information system content & usage
  - Social media

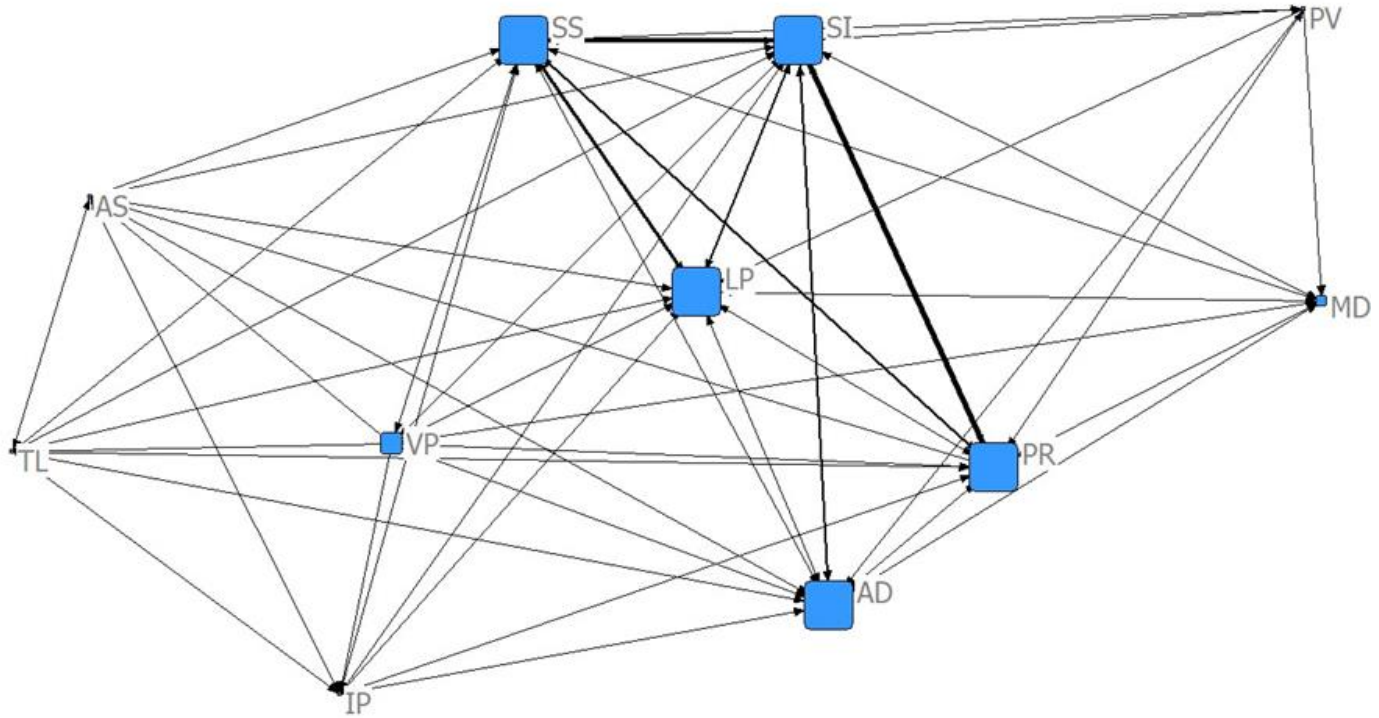
Clustering of Search Session Patterns



Based on 163,526 search sessions

Wolfram, D., Wang, P., & Zhang, J. (2009). Identifying Web search session patterns using cluster analysis: A comparison of three search environments. *Journal of the American Society for Information Science and Technology*, 60(5), 896-910.

Network Analytic Approaches: Search Action Relationships



Han, H., & Wolfram, D. (2016). An exploration of search session patterns in an image-based digital library. *Journal of Information Science*, 42(4), 477-491.

Using Network Analysis Software

- Software can visualize and analyze network-based relationships

- Many examples can handle large datasets
  - Pajek, Gephi (general network analysis)
  - CiteSpace, Sci^2, VOSviewer (metrics-focused)

Identifying Hidden Relationships & Patterns

- Data mining (Thelwall, 2001)
- Text mining (Song & Chambers, 2014)
- Clustering/classification (Glänzel & Schubert, 2003)
- Community detection (Bohlin et al., 2014)

Search Session Pattern Mining

Pattern sequence					# of sessions
SI	SI				9831
SS	SI				8252
SS	SS				8212
SI	SI	SI			7495
SI	SS				7219
SS	SI	SS			6059
SS	SS	SS			5871
SI	SI	SI	SI		5843
SS	SS	SI			5779
SI	SS	SI			5777
PR	PR				5357
LP	SS				5270
PR	SI				5181
SI	PR				5116
SI	SI	SS			4920
SS	SI	SS	SI		4821
SI	SI	SI	SI	SI	4768
LP	SI				4729
LP	LP				4655

Han, H., & Wolfram, D. (2016). An exploration of search session patterns in an image-based digital library. *Journal of Information Science*, 42(4), 477-491.

Network-based vs. Text-based Metrics

- Citations & collaborations form the foundation of traditional comparative analysis
- Downside: No link $\Rightarrow$ No relationship
- Language can expand relationship possibilities
  - Term co-occurrence
  - Topic modeling
  - Identify hidden patterns with text mining

Dealing with Large Text-based Corpora

- Language-based methods have greatly benefitted bibliometrics research
  - Natural Language Processing (NLP)
  - Text mining
  - Topic modeling
- Methods scale reasonably well
- Need large enough corpora to provide reliable outcomes

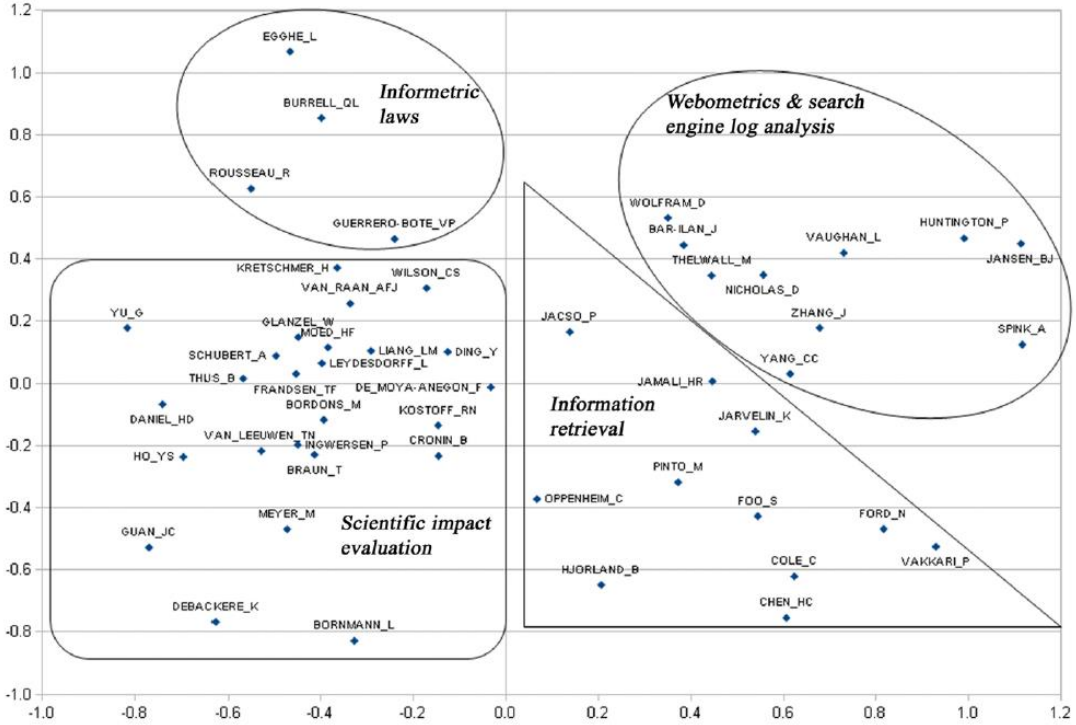
1) Co-word Analysis

- Longstanding use in metrics research
  (e.g., Braam & Moed, 1991; Ding, Chowdhury & Foo, 1997)
- Simple to use
- Independence assumption limitations
- Information retrieval matching methods can be used

2) Topic Modeling

- Applications of topic modeling
  - Tang et al. (2008) – applied Latent Dirichlet Allocation to academic search
  - Lu & Wolfram (2012) – compared author research similarity using topic modeling, co-authorship & co-citation
  - Ding & Song (2014) – measuring scholarly impact

Author-Topic Modeling for Author Research Relatedness



An A-T model produced more coherent groupings of prolific authors in information science than co-citation analysis

Lu, K., & Wolfram, D. (2012). Measuring author research relatedness: A comparison of word-based, topic-based and author co-citation approaches. *Journal of the American Society for Information Science and Technology,* 63(10), 1973-1986.

3) Text Mining

- Can be combined with bibliometric methods
  - Citation mining for user research profiling (Kostoff et al., 2001)
  - Clustering of scientific fields (Janssens, 2007)
  - Knowledge structure of bioinformatics (Song & Kim, 2013)

- Text mining techniques are integrated into some bibliometric mapping software, including
  - VOSviewer (http://www.vosviewer.com/)
  - CiteSpace (http://cluster.cis.drexel.edu/~cchen/citespace/)

Ongoing Issues

- Access to high quality data
- Processing overhead – techniques can demand high performance computing resources
- Multiple ways to address big data summarization with different outcomes
- Better methods needed for assessing outcomes

Future Directions

- Complexities of bibliometric datasets lend themselves to Information Retrieval techniques
  - Resulting “big data” require data and text processing or mining techniques to identify overt & hidden patterns
- Topic modeling and other text-based methods show great promise for providing complementary approaches to citation & co-authorship data
  - Computational overhead to train models is still high

For More Information

- Ding, Y., Rousseau, R., & Wolfram, D. (Eds.). (2014). *Measuring scholarly impact: Methods and practice*. Berlin: Springer.
- Rousseau, R. (2012). A view on" big data" and its relation to Informetrics. *Chinese Journal of Library and Information Science*, 5(3), 12-26.

Thank you
Obrigado

Email: dwolfram@uwm.edu
LinkedIn: www.linkedin.com/in/DietmarWolfram/
Academia.edu: uwm.academia.edu/DietmarWolfram/