Understanding Duke Research Based on Large-Scale Faculty Publication Records

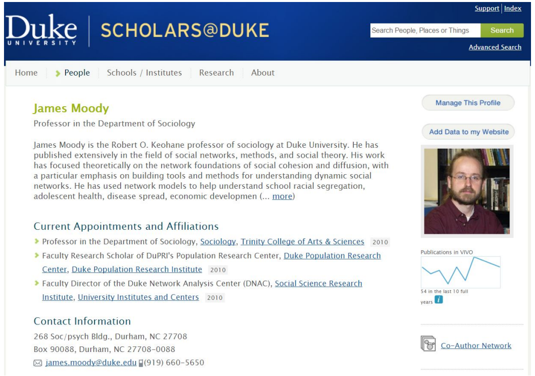
Summary

Collaboration is essential at all stages of the scientific process at Duke. However, at such a large, diverse university, finding collaborators and analyzing past collaborations can be a cumbersome process. This project seeks to ease these challenges. We expand upon current visualizations in the Scholars@Duke database and provide new data visualizations for greater insight into collaboration.

What Exists on Scholars@Duke

Data at Scholars@Duke

- Scholars@Duke collects information on Duke scholars, including data on publications, grants, and faculty
- The data allows scholars to be grouped into several network graphs
- Includes over 230,000 publications, 8,000 scholars, and 22,000 grants
- Scholars@Duke: Large database of Publications and Artistic Works, Grants, and Researchers



James Moody

Professor in the Department of Sociology

James Moody is the Robert O. Keohane professor of sociology at Duke University. He has published extensively in the field of social networks, methods, and social theory. His work has focused theoretically on the network foundations of social cohesion and diffusion, with a particular emphasis on building tools and methods for understanding dynamic social networks. He has used network models to help understand school racial segregation, adolescent health, disease spread, economic developmen (... more)

Current Appointments and Affiliations

- Professor in the Department of Sociology, Sociology, Trinity College of Arts & Sciences 2010
- Faculty Research Scholar of DuPRI's Population Research Center, Duke Population Research Center, Duke Population Research Institute 2010
- Faculty Director of the Duke Network Analysis Center (DNAC), Social Science Research Institute, University Institutes and Centers 2010

54 in the last 10 full years

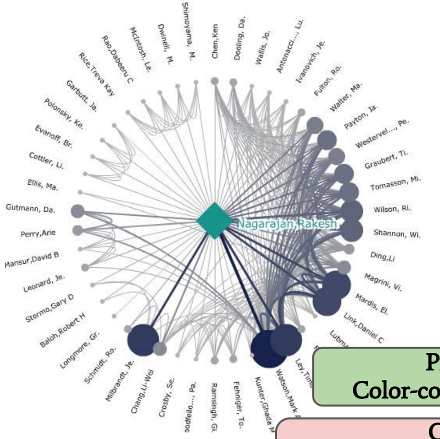
Contact Information

268 Soc/psych Bldg., Durham, NC 27708
Box 90088, Durham, NC 27708-0088
✉ james.moody@duke.edu (919) 660-5650

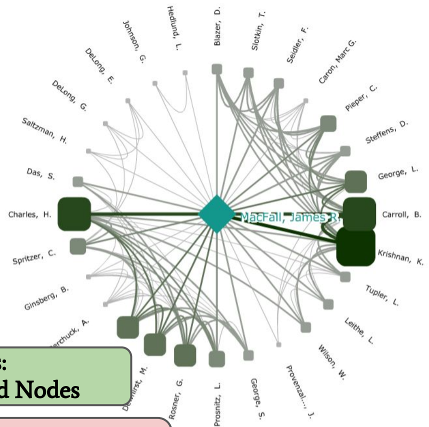
Scholars@Duke currently uses the Open Source VIVO application to visualize Scholars' data in two ways:

1. Co-Author networks to display collaborations on publications
2. Co-Investigator networks to show grant collaborations

Current Co-Author Network



Current Co-Investigator Network



Pros:
Color-coded Nodes

Cons:
Missing explanation of connections
Lacks spatial information

Understanding Duke Research Based on Large-Scale Faculty Publication Records

Method and Approach

Publication and Author Data:

Publication Title
Author Name Date
Faculty Appointments
Keywords Department

Python Processing:
Model authors as vectors over vocabulary collected from Titles and Keywords

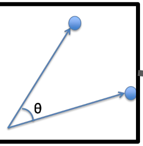
Vocabulary: ["dogs","we","hot", "cats","like"]

Smith, Sarah: [0, 1, 0, 1, 1]
Sarah's Publication Title: "we like cats"

Baker, Bill: [1, 1, 1, 0, 1]
Bill's Publication Title: "we like hot dogs"

Comparing Author Vectors:
Cosine Similarity used to approximate the angle between author vectors

$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

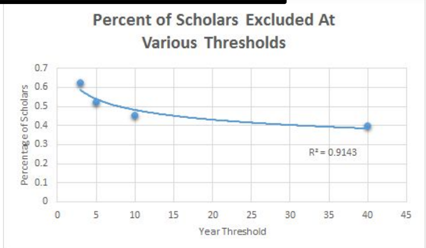


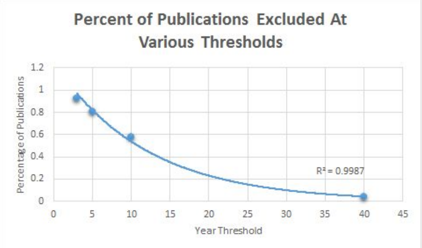
Visualizing Similarity and Co-Authorship:
Compare similarities between all authors and find co-authorship information

Li and Baker : 5 Co-Authorships
Smith and Alvarez : 2 Co-Authorships
~~~~
Li and Alvarez : .280 Similarity
Baker and Smith : .872 Similarity

## Challenges

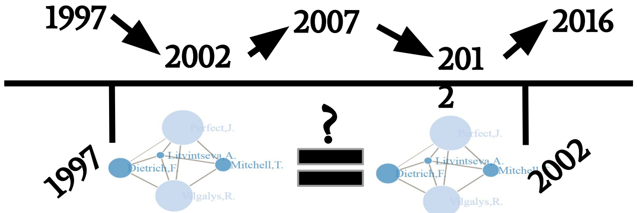
1. **Efficiency**:
   a. Replacing Author's sparse word vectors with more efficient dictionaries
   b. Large number of publications and authors
2. **Balancing Capabilities with Client and User Interests**:
   a. Providing multiple visualizations
   b. Using non-article works
   c. Title vs. Abstract based similarity
   d. Year restriction on publications (see graphs)


Percent of Scholars Excluded At Various Thresholds


Percent of Publications Excluded At Various Thresholds

## Testing the Algorithm

- Devised a means to check if similar authors later appeared as co-authors, beginning with 20 years ago and checking every 5 years to the present date
- Repeated for 6 people up to present date with no affirmative matches

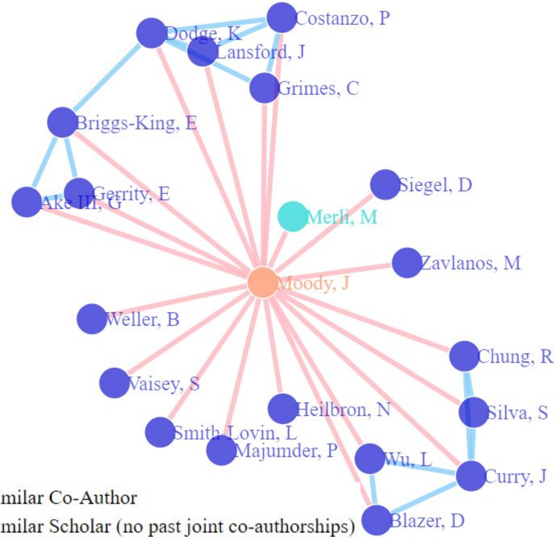

~~~~

Understanding Duke Research Based on Large-Scale Faculty Publication Records

Final Products

Similarity Network

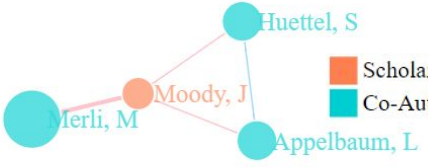
Length between authors is proportional to similarity



■ Similar Co-Author
■ Similar Scholar (no past joint co-authorships)
■ Scholar of Interest

Co-Author Network

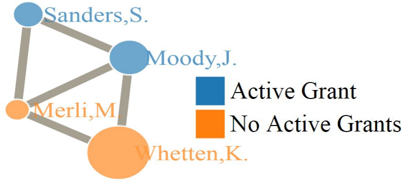
Node-size depends on number of publications with Scholar of Interest



■ Scholar of Interest
■ Co-Author

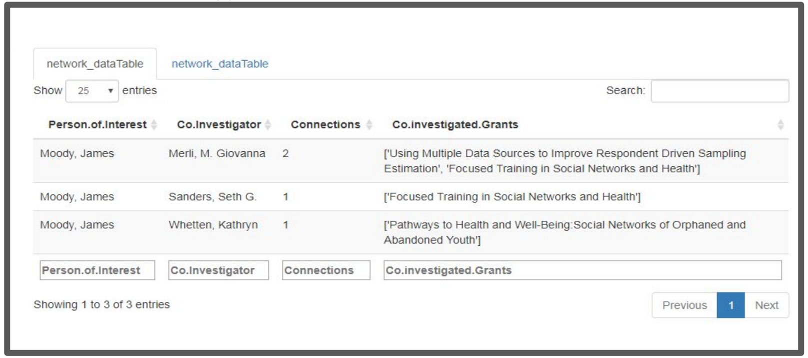
Co-Investigator Network

Node size depends on total grant count



■ Active Grant
■ No Active Grants

Sample Network Datatable



Person.of.Interest	Co.Investigator	Connections	Co.investigated.Grants
Moody, James	Merli, M. Giovanna	2	['Using Multiple Data Sources to Improve Respondent Driven Sampling Estimation', 'Focused Training in Social Networks and Health']
Moody, James	Sanders, Seth G.	1	['Focused Training in Social Networks and Health']
Moody, James	Whetten, Kathryn	1	['Pathways to Health and Well-Being:Social Networks of Orphaned and Abandoned Youth']
Person.of.Interest	Co.Investigator	Connections	Co.investigated.Grants

Showing 1 to 3 of 3 entries

Future Uses and Extensions

- Investigate other topic modeling and similarity analysis methods
- Meta-analyses of scholarship from department to department
- Including scholarly collaborations beyond the Duke Network
- Visualize grant amounts, analytical data (e.g. team interdisciplinarity).
- Re-purpose authors' vocabulary for alternative uses such as:
  - formation of new committees (PhD, special topics, grant teams)
  - Identification of relevant scholars for peer-review and policy analysis

Questions?
John Benhart, john.benhart@duke.edu
Esko Brummel, sab126@duke.edu