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.

### Summary

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

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
### Method and Approach

#### Publication and Author Data:

<table>
<thead>
<tr>
<th>Publication Title</th>
<th>Author Name</th>
<th>Date</th>
<th>Faculty Appointments</th>
<th>Keywords</th>
<th>Department</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Python Processing:

Model authors as vectors over vocabulary collected from Titles and Keywords

<table>
<thead>
<tr>
<th>Vocabulary: [&quot;dogs&quot;,&quot;we&quot;,&quot;hot&quot;,&quot;cats&quot;,&quot;like&quot;]</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Smith, Sarah: [0, 1, 0, 1, 1]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sarah's Publication Title: &quot;we like cats&quot;</td>
</tr>
<tr>
<td>Baker, Bill: [1, 1, 1, 0, 1]</td>
</tr>
<tr>
<td>Bill's Publication Title: &quot;we like hot dogs&quot;</td>
</tr>
</tbody>
</table>

#### 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)

### 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

- | Year | Authors |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td></td>
</tr>
<tr>
<td>201</td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td></td>
</tr>
</tbody>
</table>
Understanding Duke Research Based on Large-Scale Faculty Publication Records

**Final Products**

- **Similarity Network**
  - Length between authors is proportional to similarity

- **Co-Author Network**
  - Node-size depends on number of publications with Scholar of Interest

- **Co-Investigator Network**
  - Node size depends on total grant count

![Sample Network Datatable](sample-network-datable.png)

**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