Hands-on Tutorial

Supported by Microsoft Research
Program overview

• The CADRE project (Val Pentchev)

• Hands on intro to CADRE (Mat Hutchinson)

• Interactive demo with packages and notebooks (Filipi Silva)

• CADRE fellow presentation (Yi Bu)

• Demo for scalability and Reproducibility (Xiaoran Yan)

• Q&A and conclusion
The CADRE project

Val Pentchev
The CADRE team
CADRE Leadership
Partners

University of Iowa Libraries
University of Michigan Libraries
Michigan State University Libraries
University of Minnesota Libraries
Ohio State University Libraries
Penn State University Libraries
Purdue University Libraries
Rutgers University Libraries

Health Partners
Pervasive Technology Institute
Midwest Big Data Hub
South Big Data Hub
West Big Data Hub
Microsoft Research
Web of Science Group

This project was made possible in part by the Institute of Museum and Library Services LG-70-18-0202.
Topic 1

• Content
Topic 2

Content

• Content
Hands on intro to CADRE

Mat Hutchinson
Demo 1

https://github.com/iuni-cadre/ISSI-tutorial
Questions?
Demo 2

https://github.com/iuni-cadre/ISSI-tutorial
Demo 3

https://github.com/iuni-cadre/ISSI-tutorial
Questions?
CADRE Fellows

Xiaoran Yan
CADRE related events

- Apr. 2019
  - 2019 CADRE meeting
  - CADRE Fellowship open
  - 1st Fellows announced
  - ISSI workshop & tutorial

- Sep. 2019

- May. 2020
  - 2020 CADRE meeting
  - BTAA Library Conference 2020
  - 2020 CADRE hack-a-thon
CADRE Fellowship program

• Gain access to the big bibliometric data sets
• Receive data and technical support for your project
• Join the CADRE community on Slack channels, GitHub repositories and other platforms
• Have early access to free cloud computing resources
• Receive travel scholarships
Utilizing Data Citation for Aggregating, Contextualizing, and Engaging with Research Data in STEM Education Research

Researchers: Michael Witt, Loran Carleton Parker, Ann Bessenbacher
Affiliation: Purdue University
MCAP: Mapping Collaborations and Partnerships in SDG Research

Researchers: Jane Payumo, Devin Higgins, Scout Calvert, Guangming He
Affiliation: Michigan State University
The global network of air links and scientific collaboration – a quasi-experimental analysis

Researchers: Katy Börner, Adam Ploszaj, Lisel Record, Bruce Herr II
Affiliation: Indiana University Bloomington and University of Warsaw
Measuring and Modeling the Dynamics of Science Using the CADRE Platform

Researchers: Russell Funk, Michael Park, Thomas Gebhart, Britta Glennon, Julia Lane, Raviv Murciano-Goroff, Matthew Ross, Jina Lee, Erin Leahey

Affiliation: University of Minnesota, University of Pennsylvania, New York University, Boston University, University of Arizona
Comparative analysis of legacy and emerging journals in mathematical biology

Researchers: Marisa Conte, Samuel Hansen, Scott Martin, Santiago Schnell
Affiliation: University of Michigan and University of Michigan Medical School
Systematic over-time study of the similarities and differences in research across mathematics and the sciences

Researcher: Samuel Hansen
Affiliation: University of Michigan
A user story from CADRE fellows
Understanding citation impact of scientific publications through ego-centered citation networks

Researchers: Yi Bu, Chao Min, Ying Ding
Affiliation: Indiana University Bloomington and Nanjing University
Exploring ego-centered citation networks: A technical introduction

Yi Bu¹, Chao Min², and Ying Ding¹

¹: School of Informatics, Computing, and Engineering, Indiana University, U.S.A.
²: School of Information Management, Nanjing University, China
Understanding citation impact of scientific publications

- Citation impact as a type of impact
  - Citation impact among all types of impact
  - Citation impact of scientific publications

- Benefits from understanding citation impact
  - Measuring citation impact offers a useful way of examining the scientific impact of a publication.
  - Measuring citation impact can also assist in understanding knowledge diffusion and the use of information.
Understanding citation impact of scientific publications (cont.)

- Previous ways of understanding citation impact of scientific publications:
  - ✓ Count-based strategies: raw citation count, normalized citation measures...
  - ✓ Network-based strategies: PageRank, EigenFactor...
Understanding citation impact of scientific publications (cont.)

- Local details are missing!
  ✓ “Deep” or “wide” impact?
Understanding citation impact of scientific publications (cont.)

- Local details are missing!
  - How does an article impact other research, and what are the patterns? The direct citations between citing publications (DCCPs) offer a good way to mine how a publication impacts other research.

<table>
<thead>
<tr>
<th>Cited publication</th>
<th>SSH</th>
<th>BHS</th>
<th>PSE</th>
<th>LES</th>
<th>MCS</th>
<th>subtotal</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSH</td>
<td>1138</td>
<td>224</td>
<td>16</td>
<td>5</td>
<td>37</td>
<td>11420</td>
</tr>
<tr>
<td>BHS</td>
<td>440</td>
<td>1254</td>
<td>2</td>
<td>11</td>
<td>1</td>
<td>1708</td>
</tr>
<tr>
<td>PSE</td>
<td>137</td>
<td>1</td>
<td>19</td>
<td>3</td>
<td>18</td>
<td>178</td>
</tr>
<tr>
<td>LES</td>
<td>57</td>
<td>13</td>
<td>3</td>
<td>11</td>
<td>0</td>
<td>84</td>
</tr>
<tr>
<td>MCS</td>
<td>194</td>
<td>0</td>
<td>17</td>
<td>0</td>
<td>26</td>
<td>237</td>
</tr>
<tr>
<td>subtotal</td>
<td>11966</td>
<td>1492</td>
<td>57</td>
<td>30</td>
<td>82</td>
<td>13627</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>year</th>
<th>SSH</th>
<th>BHS</th>
<th>PSE</th>
<th>LES</th>
<th>MCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2007</td>
<td>111</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2008</td>
<td>455</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2009</td>
<td>753</td>
<td>9</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2010</td>
<td>1155</td>
<td>19</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2011</td>
<td>1310</td>
<td>80</td>
<td>2</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>2012</td>
<td>1092</td>
<td>39</td>
<td>3</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>2013</td>
<td>1440</td>
<td>187</td>
<td>19</td>
<td>3</td>
<td>41</td>
</tr>
<tr>
<td>2014</td>
<td>1110</td>
<td>449</td>
<td>30</td>
<td>2</td>
<td>31</td>
</tr>
<tr>
<td>2015</td>
<td>1161</td>
<td>361</td>
<td>12</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>2016</td>
<td>1491</td>
<td>290</td>
<td>44</td>
<td>57</td>
<td>60</td>
</tr>
<tr>
<td>2017</td>
<td>1329</td>
<td>274</td>
<td>63</td>
<td>5</td>
<td>67</td>
</tr>
</tbody>
</table>
Understanding citation impact of scientific publications (cont.)
Ego-centered citation networks as a tool to understand citation impact
Preliminary research questions

• Do DCCPs occur frequently?
• How does DCCPs different in papers with different citation impacts and in different years?
Preliminary results: The universality of DCCPs
Preliminary results (cont.)
Technical details: Extracting citing relationships from the raw WoS tables

• SQL extraction as a .txt file:

```python
import psycopg2
conn = psycopg2.connect(database = 'core_data', user = 'buyi', password = 'xxxxxxxxxxx')
cur = conn.cursor()
cur.execute("SELECT paper_id, paper_reference_id FROM mag_core.paper_references;")
outFile = open("mag_citing.txt", "w")
lines = ['citing id====cited id']
for row in cur:
    if row[0] in paper_id_set and str(row[1]) in paper_id_set:
        lines.append('({}==={})'.format(str(row[0]), str(row[1])))
    if len(lines) % 100000 == 0:
        outFile.write(\n'.join(lines) + 'n')
        lines = []
outFile.write(\n'.join(lines) + 'n')
cur.close
```

• .txt file to a Python dictionary:

✓ if paper in paper_citing.keys()
Difficulty 1: How to extract DCCPs?

Sample output:

<table>
<thead>
<tr>
<th>Id of A-type paper (focal)</th>
<th>Id of B-type paper</th>
<th>Id of C-type paper</th>
</tr>
</thead>
</table>

Direct citations to A

Direct citations between citing publications (from the perspective of A)
Difficulty 1: How to extract DCCPs? (cont.)

• This task is computationally expensive:
  ✔ In MAG, we have ~0.1 billion papers. The below Python script will perhaps take forever...

```python
indirect_citation = defaultdict(list)
for paper in paper_year.keys(): # for papers that have pub_year information
    for citing_paper_1 in paper_citing[paper]:
        for citing_paper_2 in paper_citing[paper]:
            if citing_paper_1 in paper_citing[citing_paper_2]:
                temp = []
                temp.append(citing_paper_1)
                temp.append(citing_paper_2)
                indirect_citation[paper].append(temp)
```
Difficulty 2: Self-citations in ego-centered citation networks?

- If two papers (A and B) share at least one co-author and B cites A, such citation is called a self-citation (first-order self-citation).

- How about these circumstances, when B cites A?
  - A and B don’t share co-authors, but A and C do, and B and C do. [second-order self-citations]
  - A and B don’t share co-authors, but A and C do, B and D do, and C and D do. [third-order self-citations]
  - This indicates how researchers’ social distance impacts on their self-citation patterns.

- How to technically achieve these?
Difficulty 2: Self-citations in ego-centered citation networks?

• Completing this task is also computationally expensive:
  ✓ Deriving n-order self-citations need to know **the shortest paths and their lengths** in the co-authorship and citation networks
  ✓ Such networks are quite huge (hundreds of millions of nodes in the citation network, and millions of nodes in the co-authorship network)
Questions?

Presenter: Yi Bu, Indiana University

Email: buyi@iu.edu
Website: https://buyi08.wixsite.com/yi-bu
Scalability & Reproducibility

Xiaoran Yan
Difficulty 1: How to extract DCCPs?
Difficulty 1: How to extract DCCPs? (cont.)

• This task is computationally expensive:
  ✓ In MAG, we have ~0.1 billion papers. The below Python script will perhaps take forever...

```python
indirect_citation = defaultdict(list)
for paper in paper_year.keys():  # for papers that have pub_year information
    for citing_paper_1 in paper_citing[paper]:
        for citing_paper_2 in paper_citing[paper]:
            if citing_paper_1 in paper_citing[citing_paper_2]:
                temp = []
                temp.append(citing_paper_1)
                temp.append(citing_paper_2)
                indirect_citation[paper].append(temp)
```
CADRE’s solution

• An easy to use graphical interface of a query builder with preview functionality

• A unified engine with optimized combinations of solutions based on relational/graph/document databases

• For users who want intuitive and quick access of data, no programing skills required

• In development: APIs for power users
CADRE’s solution

- Access over 220 million scientific publications
- Effortlessly query data and analyze results
- Reproduce research & leverage tools
CADRE’s solution

- Databases
- GUI-query
- Notebooks
- RAC
Questions?

Presenter: Xiaoran Yan, Indiana University

Email: yan30@iu.edu
CADRE’s solution

Access over 220 million scientific publications

Effortlessly query data and analyze results

Reproduce research & leverage tools
The reproducibility “Crisis”

Spectrum of Reproducibility

Computational -> Statistical -> Empirical

Current solutions

Matlab code of minimum absolute spectral similarity

6 commits 1 branch

Branch: master  New pull request

every.x: Update README.md
LFR.mat  Add files via upload
README.md  Update README.md
absSpecSim.m  Add files via upload
main.m  Add files via upload
sparsify.py  Added a Python implementation

Publications

https://arxiv.org/abs/1806.07479
Source code: https://github.com/IU-AMBITION/MASS

https://www.nature.com/articles/s41598-018-31202-1

Matlab function for the Minimum absolute spectral similarity (MASS)
Big data pipelines in the industry
CADRE’s solution

- Databases
- GUI-query
- Notebooks
- RAC
Empowered by the open-source ecosystem

The old way: Applications on host

- App
- App
- Libraries
- Kernel

Heavyweight, non-portable
Relies on OS package manager

The new way: Deploy containers

- App
- Libraries
- App
- Libraries
- App
- Libraries
- Kernel

Small and fast, portable
Uses OS-level virtualization
Reproducible notebooks on Kubernetes

Turn a Git repo into a collection of interactive notebooks

Have a repository full of Jupyter notebooks? With Binder, open those notebooks in an executable environment, making your code immediately reproducible by anyone, anywhere.
Demo 5

https://github.com/iuni-cadre/ISSI-tutorial
### The CADRE ecosystem

| 3\(^{rd}\) party | • Plugins and extensions  
| | • Computing resources  
| | • Other data sets  |
| RAC | • Package marketplace  
| | • Derivatives data  
| | • Pipeline builder  |
| CADRE core | • Centralized databases  
| | • Data API  
| | • Coding environment  |
Reproducible notebooks on Kubernetes

Turn a Git repo into a collection of interactive notebooks

Have a repository full of Jupyter notebooks? With Binder, open those notebooks in an executable environment, making your code immediately reproducible by anyone, anywhere.
SHARED BIGDATA-GATEWAY FOR RESEARCH LIBRARIES (SBD-GATEWAY)

FEDERATED SINGLE SIGN ON
- Custom
- SAML
- Shibboleth
- Globus

WEB QUERY INTERFACE
- Query GUI

RESEARCH ASSET COMMONS
- Saved and Shared Results, Tools and Datasets
  - Visualizations
  - Annotations
  - Data Derivatives
  - Metadata

APPLICATION PROGRAMMING INTERFACE (API)
- Text Analytics
  - Analytics Tool
  - Analytics Tool
  - Analytics Tool
- Database Analytics
  - Analytics Tool
  - Analytics Tool
  - Analytics Tool

CLOUD COMPUTE GATEWAY
- Machine Learning
  - Analytics Tool
  - Analytics Tool
  - New Analytics Tool
- Data Processing and Querying
  - elasticsearch
  - Spark
  - Reaktame
  - SPARQL
  - presto
  - Cypher

CLOUD STORAGE
- Indexed Storage
  - Relational Databases
  - Graph Databases
- Raw Format Data Lake Storage
  - Web of Science
  - MS Academic Graph

AUTHORIZATION
- GRANULAR DATASET PERMISSIONS

LOCAL AND INSTITUTIONAL STORAGE
- Databases

CLOUD AND LOCAL COMPUTE RESOURCES
- Ceph
- sigma
- tableau.js
- kibana
- Gephi
- R
- php
- jetstream
- Commercial Cloud Services

Notebooks
- Databases
CADRE related events

- Apr. 2019
  - 2019 CADRE meeting
  - CADRE Fellowship open
- Sep. 2019
  - 1st Fellows announced
  - ISSI workshop & tutorial
- May. 2020
  - 2020 CADRE meeting
  - BTAA Library Conference 2020
  - 2020 CADRE hackathon
Contact Us

https://cadre.iu.edu

cadre@iu.edu

@CADRE_Project