Exploring ego-centered citation networks: A technical introduction

Presenter: Yi Bu
School of Informatics, Computing, and Engineering, Indiana University, U.S.A.
Understanding citation impact of scientific publications

• Scientific impact as a type of impact
• Citation impact as a type of scientific impact
  ✓ Citation impact among all types of impact
  ✓ Citation impact of scientific publications
• Benefits from understanding citation impact
  ✓ Indicator perspective: Measuring citation impact offers a useful way of examining the scientific impact of a publication.
  ✓ More general perspective: 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?
• 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>citing 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>11138</td>
<td>224</td>
<td>16</td>
<td>5</td>
<td>37</td>
<td></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></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></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></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></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></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>
<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>
<td>0</td>
</tr>
<tr>
<td>2007</td>
<td>111</td>
<td>0</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>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>753</td>
<td>9</td>
<td>3</td>
<td>0</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>
<td>0</td>
</tr>
<tr>
<td>2011</td>
<td>1310</td>
<td>80</td>
<td>2</td>
<td>1</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>1092</td>
<td>39</td>
<td>3</td>
<td>1</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>1440</td>
<td>187</td>
<td>19</td>
<td>3</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>1110</td>
<td>449</td>
<td>30</td>
<td>2</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>1161</td>
<td>361</td>
<td>12</td>
<td>12</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td>1491</td>
<td>290</td>
<td>44</td>
<td>57</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td>1329</td>
<td>274</td>
<td>63</td>
<td>5</td>
<td>67</td>
<td></td>
</tr>
</tbody>
</table>

Published year and discipline distributions of citing publications of $h$-index article’s DCCPs
Ego-centered citation networks as a tool to understand citation impact
Preliminary research questions

• Do DCCPs occur frequently?
• How does DCCPs differ 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 MAG tables

- SQL extraction as a .txt file:

```python
import psycopg2
conn = psycopg2.connect(database = 'core_data', user = 'buyi', password = )
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 str(row[0]) in paper_id_set and str(row[1]) in paper_id_set:
        lines.append(f'{row[0]}======{row[1]}').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()
```
Difficulty 1: How to extract DCCPs?

Direct citations to A

Direct citations between citing publications (from the perspective of A)

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>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 QUITE 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