Microsoft Research

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

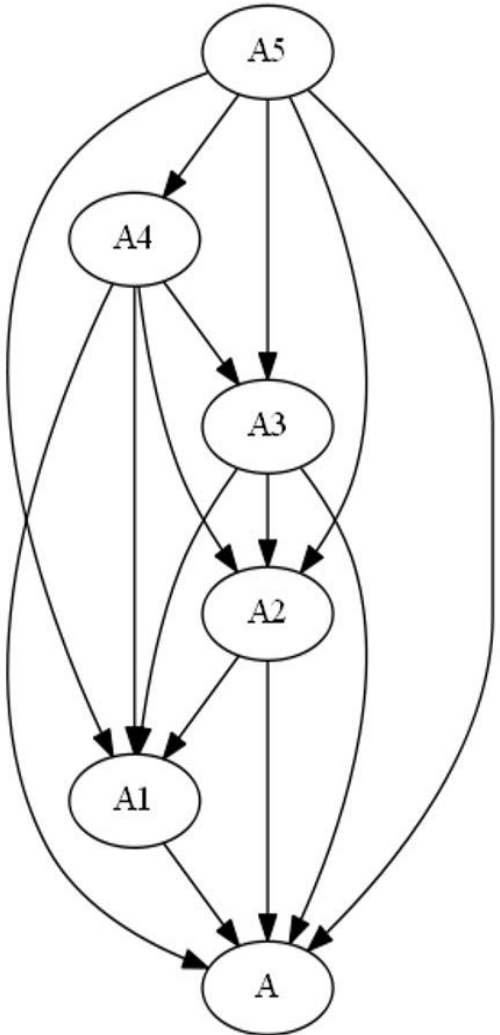
- Scientific impact as a type of impact
- Citation impact as a type of scientific impact
  - ✓ Citation impact among all types of impact
  - ✓ Citation impact of scientific publications
- Benefits from understanding citation impact
  - ✓ Indicator perspective: Measuring citation impact offers a useful 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.

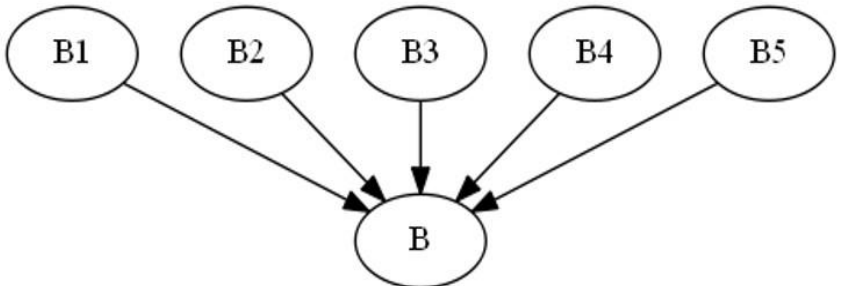
Understanding citation impact of scientific publications (cont.)

- 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?





Understanding citation impact of scientific publications (cont.)

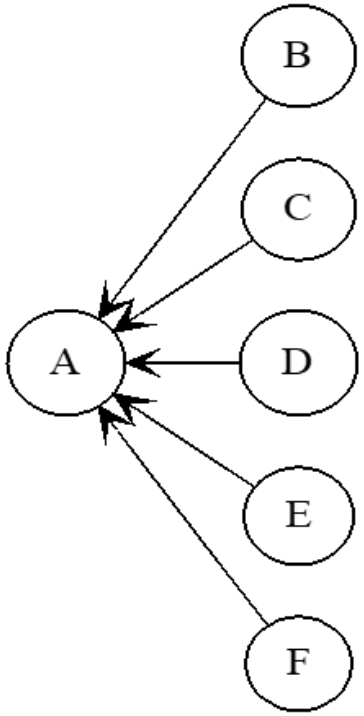
- Local details are missing!
  - ✓ How does an article impact other research, and what are the patterns? The direct citations between citing publications (DCCPs) offer a good way to mine how a publication impacts other research.

		citing publication					
		SSH	BHS	PSE	LES	MCS	subtotal
Cited publication	SSH	11138	224	16	5	37	11420
	BHS	440	1254	2	11	1	1708
	PSE	137	1	19	3	18	178
	LES	57	13	3	11	0	84
	MCS	194	0	17	0	26	237
	subtotal	11966	1492	57	30	82	13627

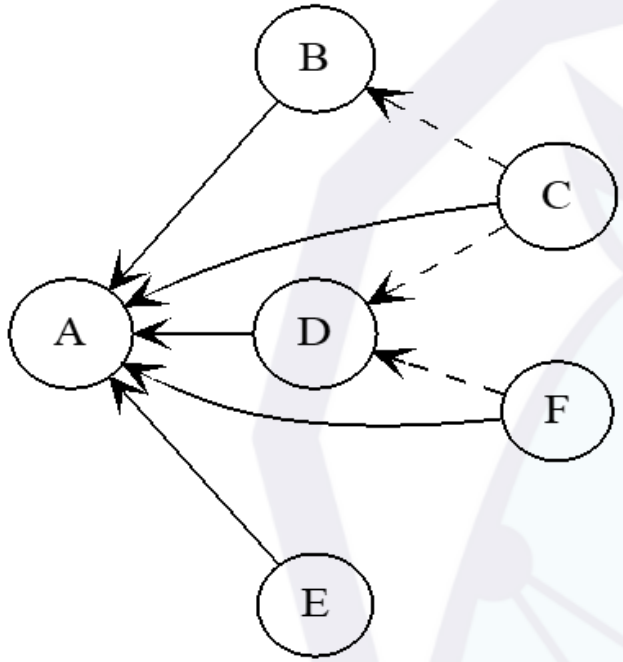
year	SSH	BHS	PSE	LES	MCS
2006	13	0	0	0	0
2007	111	0	0	0	0
2008	455	0	2	2	4
2009	753	9	3	0	0
2010	1155	19	0	1	0
2011	1310	80	2	1	12
2012	1092	39	3	1	9
2013	1440	187	19	3	41
2014	1110	449	30	2	31
2015	1161	361	12	12	13
2016	1491	290	44	57	60
2017	1329	274	63	5	67

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



(a)

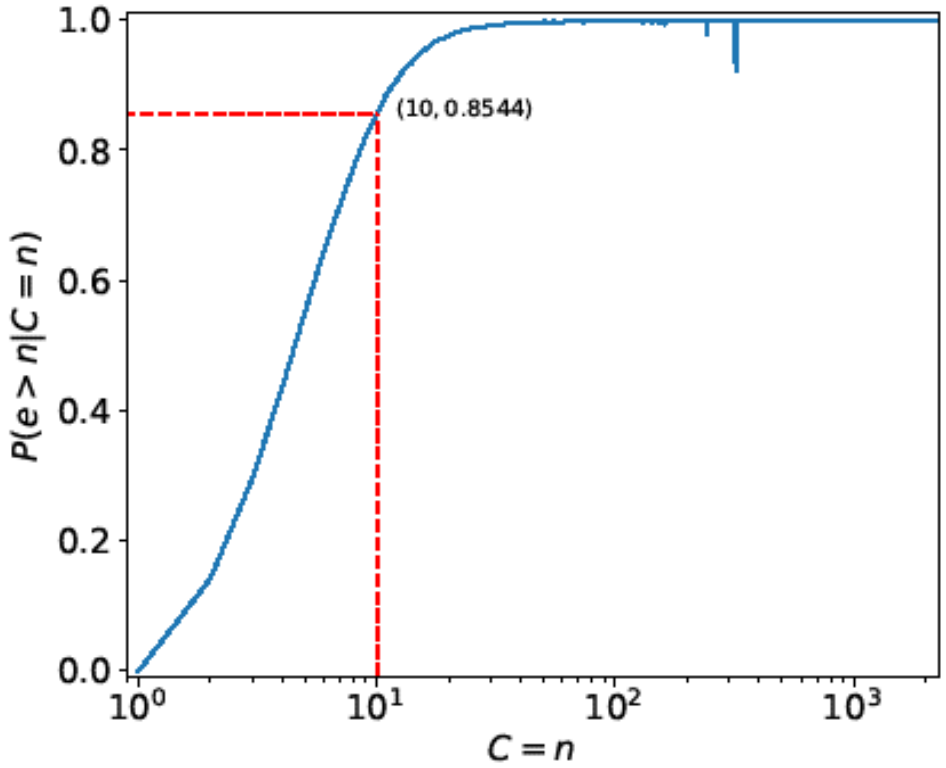


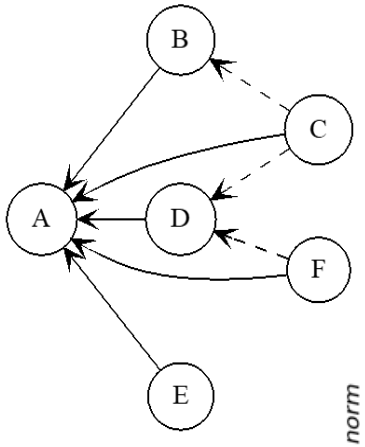
(b)

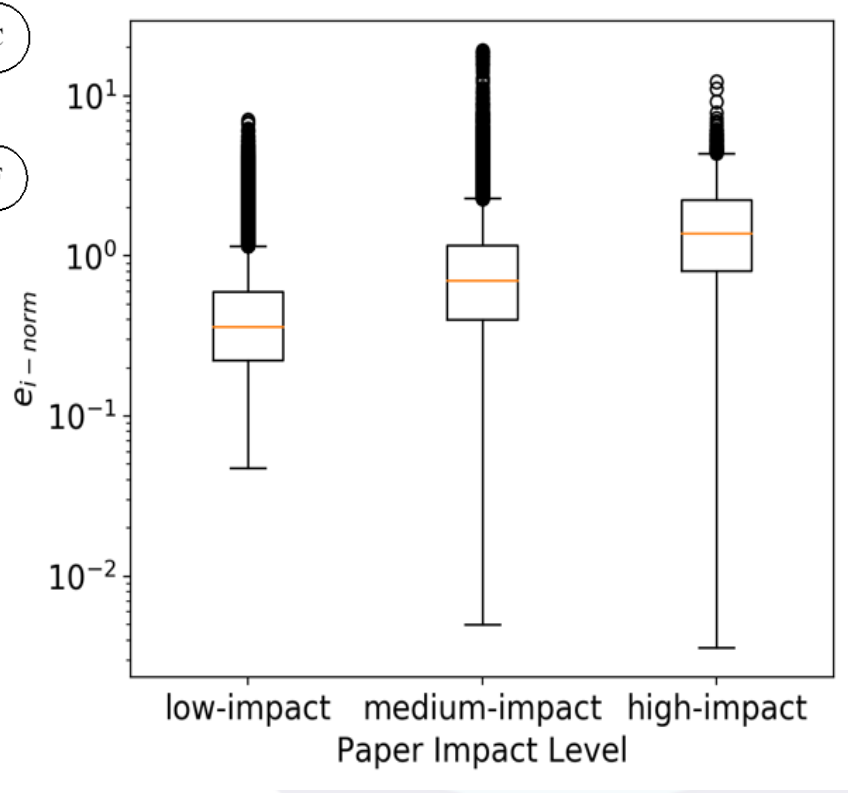
Preliminary research questions

- Do DCCPs occur frequently?
- How does DCCPs differ in papers with different citation impacts and in different years?

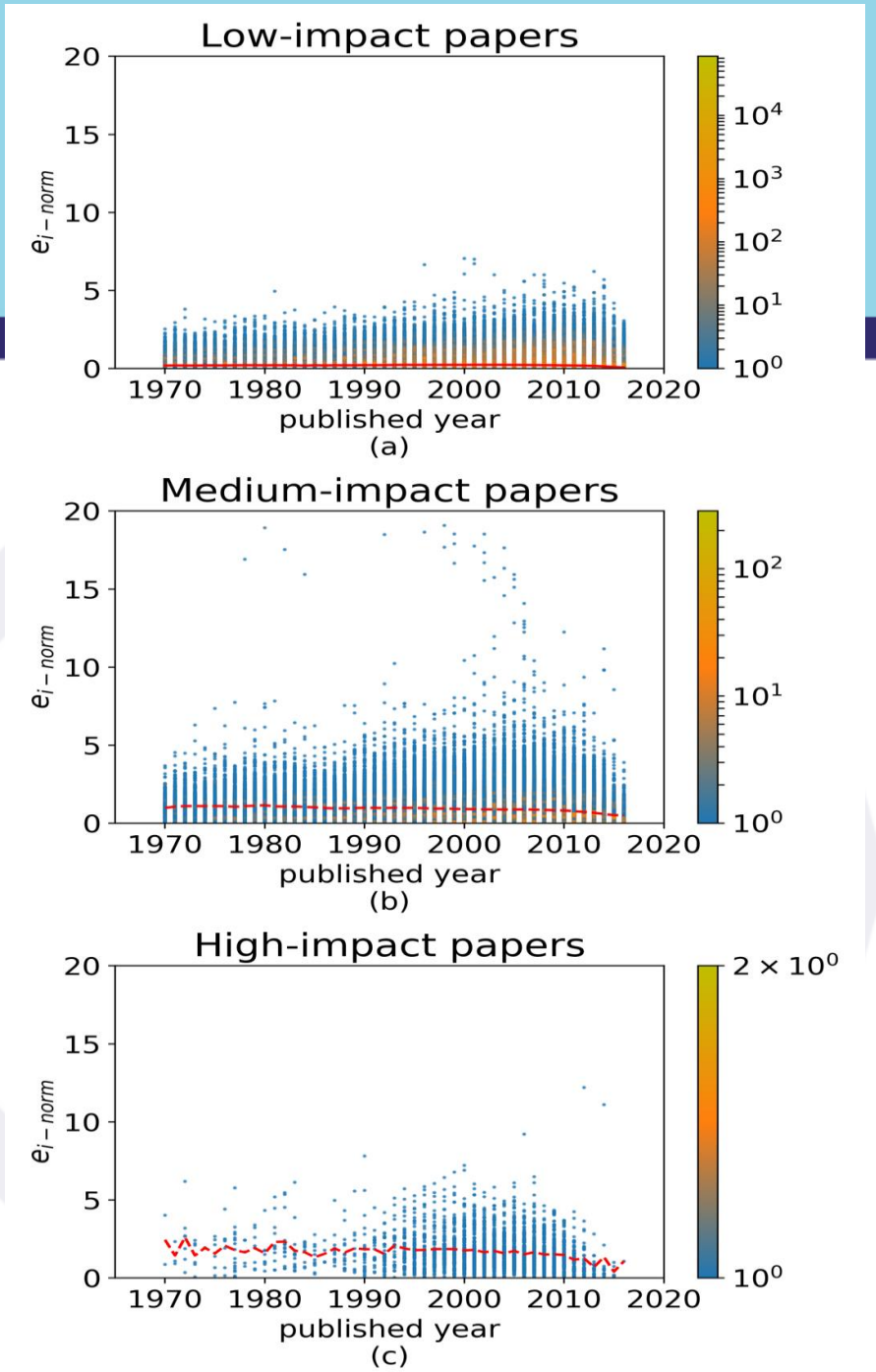
Preliminary results: The universality of DCCPs







Preliminary results (cont.)



Technical details: Extracting citing relationships from the raw MAG tables

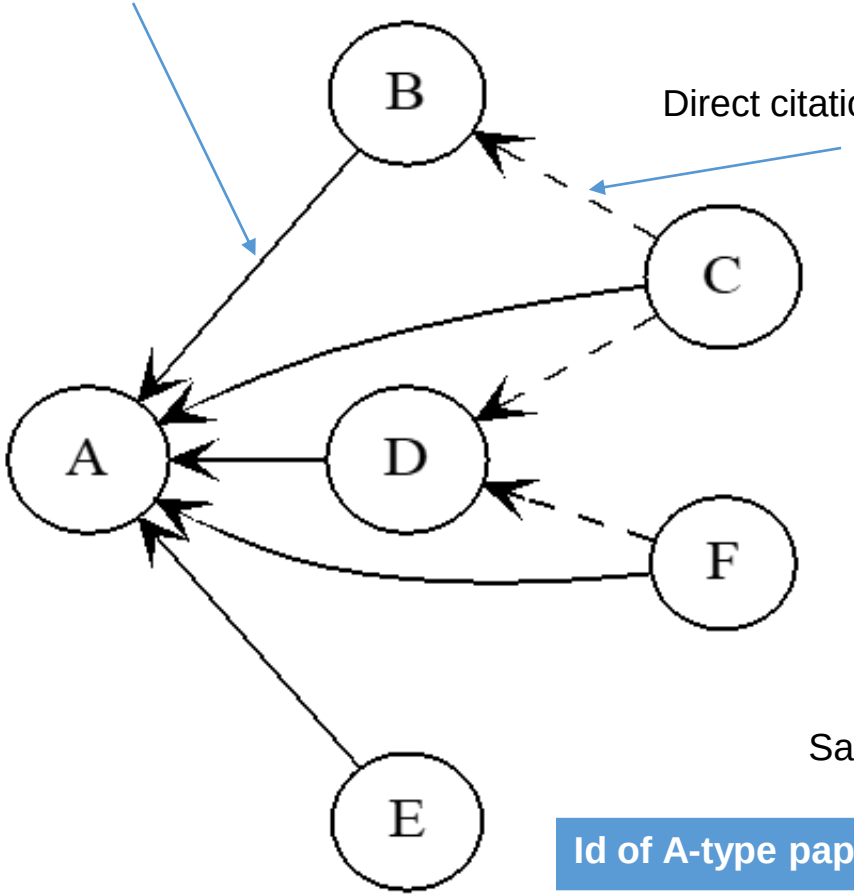
- SQL extraction as a .txt file:

```
import psycopg2
conn = psycopg2.connect(database = 'core_data', user = 'buyi', password = [redacted])
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('{:}======{:}'.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()
  - ✓ from collections import defaultdict
  - ✓ **paper_citing = defaultdict(list)**

Difficulty 1: How to extract DCCPs?

Direct citations to A



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

Sample output:

Id of A-type paper (focal)	Id of B-type paper	Id of C-type paper

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…

```
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