Computer Science Fields as Ground-truth Communities: Their Impact, Rise and Fall

Tanmoy Chakraborty
Google India PhD Fellow
Indian Institute of Technology, Kharagpur (IIT-Kgp)
India - 721302

In collaboration with:
Sandipan Sikdar, Niloy Ganguly, Animesh Mukherjee
IIT-Kgp, India

Outline

- Problem definition
- Dataset
- Community scores
- Time-transition of scientific paradigms
- Reasons behind paradigm shift
- Correlation with NSF
- Conclusion
Outline

Problem definition

Dataset
Time transition of scientific paradigms
Reasons behind paradigm shift
Correlation with NSF
Conclusion
Motivation:

Community Detection

- **Communities:**
  
  groups of nodes within which the connection is dense but between which the connection is relatively sparse.

- **Problem in community detection:**
  
  Lack of ground-truth community for evaluating the algorithms
Motivation:
Temporal Interactions among Communities

- **Longitudinal** inter-cluster interactive patterns
- **Dynamics** behind community evolution
- **Temporal authoritative ranking** of communities
Problem Definition

- **Ground-truth Communities**
  - Large citation network of computer science domain
  - Fields $\Rightarrow$ ground-truth communities

- **Temporal analysis:**
  - Temporal Impact of scientific communities
  - Time transition of scientific paradigm
  - Factors behind paradigm shift
  - Predicting forthcoming impactful communities
Outline

Problem definition

Dataset

Time transition of scientific paradigms
Reasons behind paradigm shift
Correlation with NSF
Conclusion
Dataset

- Large **DBLP dump** used in Arnetminer project
  [Tang et al., SIGKDD, 2008]

- **Bibliographic information during 1960-2008**

  - Paper name
  - Author(s)
  - Publication venue
  - Year of publication
  - Abstract
  - References

<table>
<thead>
<tr>
<th># of valid papers</th>
<th>702,973</th>
</tr>
</thead>
<tbody>
<tr>
<td># authors</td>
<td>495,311</td>
</tr>
<tr>
<td>Avg. number of papers/author</td>
<td>3.52</td>
</tr>
<tr>
<td>Avg. number of authors/paper</td>
<td>2.609</td>
</tr>
<tr>
<td># unique venue name</td>
<td>1,705</td>
</tr>
</tbody>
</table>

- **Missing Field information of each paper**
Citation Network

Node (paper)

Link (citation)

Indegree

Outdegree

Distributions
Tagging Dataset

➢ Field Tagging
  o Automated crawling of Microsoft Academic Search
    [http://academic.research.microsoft.com/]

<table>
<thead>
<tr>
<th>AI</th>
<th>Bioinformatics</th>
<th>NLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
<td>Graphics</td>
<td>WWW</td>
</tr>
<tr>
<td>Networking</td>
<td>Comp. Vision</td>
<td>Education</td>
</tr>
<tr>
<td>Database</td>
<td>Data Mining</td>
<td>OS</td>
</tr>
<tr>
<td>Architecture</td>
<td>Security</td>
<td>Simulation</td>
</tr>
<tr>
<td>Software Engg.</td>
<td>IR</td>
<td>HCI</td>
</tr>
<tr>
<td>Machine Learning</td>
<td>Scientific Comp.</td>
<td>Multimedia</td>
</tr>
</tbody>
</table>

24 Fields

➢ Continent Tagging
  o Authors are tagged by one of the three continents
    (North America, Europe, Others)

11.23% papers belong to multiple fields

Publicly available: http://cnerg.org
Outline

Problem definition
Dataset

Time-transition of scientific paradigms

Reasons behind paradigm shift
Correlation with NSF
Conclusion
• Measuring the **impact of each field** (its constituent papers) around a particular year.

• **Local citation density** is important

But

What should be the time window?
Average Inward Citations

Peaks within 3 years from publication, then declines
Authority of a Field

**Inwardness** of a field $f_i$ at time $t$

$In(f_i^t) = \sum_{j \neq i} w_{j \rightarrow i}^t$

where,

$w_{j \rightarrow i}^t = \frac{C_{j \rightarrow i}^t}{P_i^t}$

$C_{j \rightarrow i}^t$ = number of citations received by the papers of field $f_i$ from field $f_j$

$P_i^t$ = number of papers in field $f_i$

$1 \leq t \leq 3$ (current year + next 3 years)

We only consider cross-field citations
Scientific Paradigm Shift: Time Transition Diagram

- Rise in inwardness & decline near transition throughout
- Second ranked field emerges as the leader in the next window.
Outline

Problem definition
Dataset
Time transition of Scientific paradigms

Reasons behind paradigm shift

Correlation with NSF
Conclusion
Probable Reasons

1. Collaboration
2. High impact papers
3. Support from Backup fields
Reason 1: Collaborations

- Rank top fields based on:
  - Collaborative papers (papers with multiple authors)
  - Multi-continental papers
  - Diversity of a papers (average number of fields in which authors have worked)

Rank of the top fields increases after 1981
Reason 2: High Impact papers

Frac. of top and second rank fields among the 10% high impact papers

- 82% cases $\Rightarrow$ fraction of top ranked field’s papers declines and second ranked field rises at the transition point.
Reason 3: Citations from Backup Fields

- **Backup fields**: fields that provide citations to other fields

- In 75% cases, citation patterns from the top backup fields decline at the transition period \(\rightarrow\) citations get distributed among the fields.
Outline

Problem definition
Dataset
Community scores
Scientific paradigm shift through cross-citation interactions
Reasons behind paradigm shift

Correlation with NSF

Conclusion
National Science Foundation (NSF)

- US government agency that supports fundamental research and education

- The NSF receives about 40,000 research proposals each year, and funds about 10,000 of them.

- NSF has its own submission/acceptance history in each year and these proposals can be categorized into fields.
Funding Statistics of NSF

<table>
<thead>
<tr>
<th>Yrs</th>
<th>Inwardness results</th>
<th>Proposal submitted</th>
<th>Proposal awarded</th>
</tr>
</thead>
<tbody>
<tr>
<td>03</td>
<td>AI/IR/NW</td>
<td>NW/AI/HCI</td>
<td>NW/ALGO/SE</td>
</tr>
<tr>
<td>04</td>
<td>AI/IR/NW</td>
<td>AI/HCI/RT</td>
<td>RT/ARC/DIST</td>
</tr>
<tr>
<td>05</td>
<td>AI/IR/NW</td>
<td>AI/ML/HCI</td>
<td>GRP/SE/ALGO</td>
</tr>
<tr>
<td>06</td>
<td>IR/ML/AI</td>
<td>ML/ALGO/SEC</td>
<td>ALGO/SEC/ML</td>
</tr>
<tr>
<td>07</td>
<td>ML/AI/ALGO</td>
<td>ALGO/ML/HCL</td>
<td>ALGO/HCI/SEC</td>
</tr>
<tr>
<td>08</td>
<td>ML/AI/ALGO</td>
<td>ML/ALGO/SE</td>
<td>ALGO/ML/SE</td>
</tr>
</tbody>
</table>

During 2003-2008, top three fields based on
(i) Our prediction
(ii) proposal submission statistics
(iii) award statistics
Correlations with NSF Funding

- \( \text{Correlation}(\zeta) = \frac{s}{n} \);
  
  \( s = \) similarity pair (at least one out of top three)
  
  \( n = \) number of years = 46
Outline

Problem definition
Dataset
Community scores
Scientific paradigm shift through cross-citation interactions
Reasons behind paradigm shift
Correlation with NSF

Conclusion
Insights

- Computer Science Fields $\Rightarrow$ ground-truth communities
- Temporal community interactions $\Rightarrow$ scientific paradigm shift.
- Citation information $\Rightarrow$ Dynamics of community evolution
- Predicted results **perfectly correlates** with the proposal submission statistics, and **partially correlates** with funds awarded.
Acknowledgements

- **Financial Support:** Google India Pvt. Ltd.
- **Travel support:** Dept. of Science & Technology, Govt. of India
- **Providing NSF dataset:** Mr. Ansumana Cooper, NSF, US
- **Technical support:** All the members of CNeRG, IIT-Kgp
Thank You

http://cse.iitkgp.ac.in/~tanmoyc/
http://cnerg.org/