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

The 2013 IEEE/ACM ASONAM, Niagara Falls, Canada, August 25-28, 2013

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

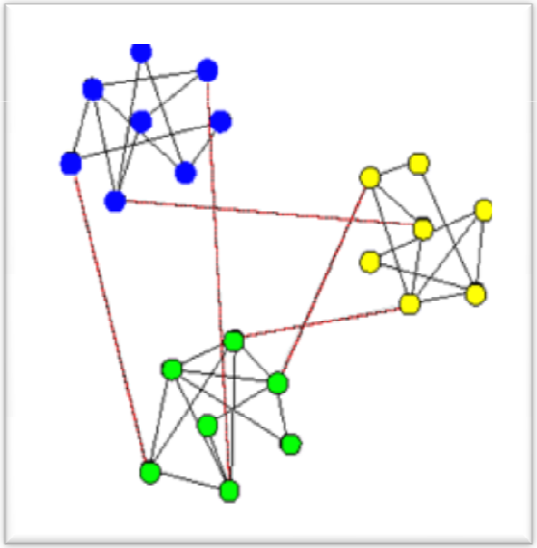
Motivation:
Community Detection

- **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 => 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]

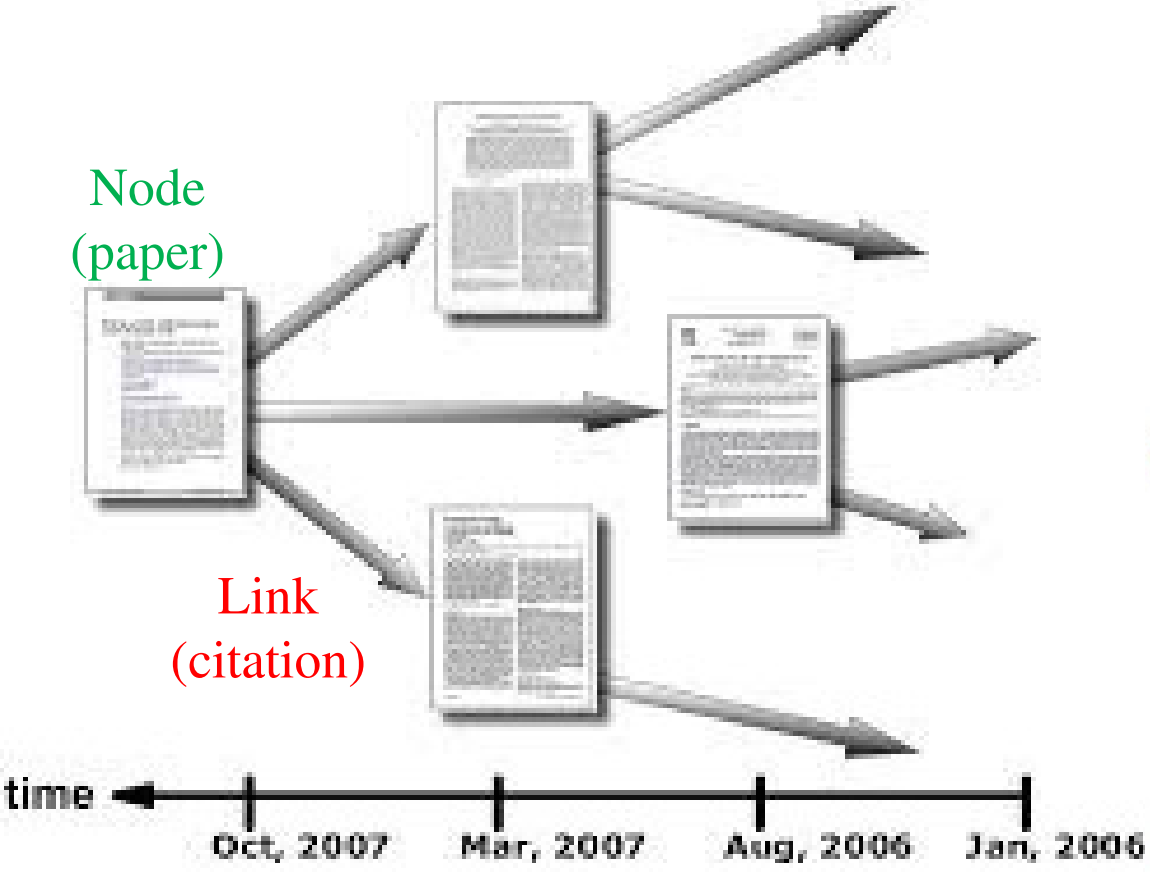
- Bibliographic information during **1960-2008**
  - Paper name
  - Author(s)
  - Publication venue
  - Year of publication
  - Abstract
  - References

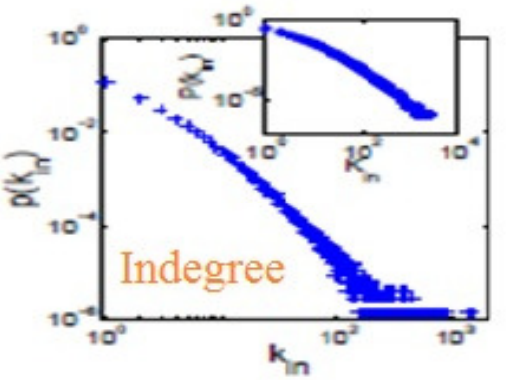
# of valid papers	702,973
# authors	495,311
Avg. number of papers/author	3.52
Avg. number of authors/paper	2.609
# unique venue name	1,705

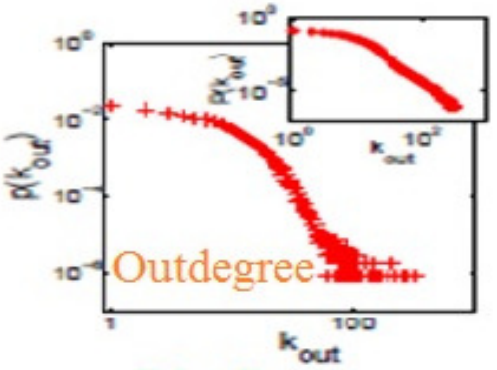
- Missing

  Field information of each paper

Citation Network







Distributions

Tagging Dataset

➢ Field Tagging

- Automated crawling of Microsoft Academic Search

[http://academic.research.microsoft.com/]

24 Fields

AI	Bioinformatics	NLP
Algorithm	Graphics	WWW
Networking	Comp. Vision	Education
Database	Data Mining	OS
Dist Comp.	Prog. Lang.	Embedded Sys.
Architecture	Security	Simulation
Software Engg.	IR	HCI
Machine Learning	Scientific Comp.	Multimedia

11.23% papers belong to multiple fields

➢ Continent Tagging

- Authors are tagged by one of the **three continents (North America, Europe, Others)**

Publicly available: http://cnerg.org

Outline

Problem definition

Dataset

Time-transition of scientific paradigms

Reasons behind paradigm shift

Correlation with NSF

Conclusion

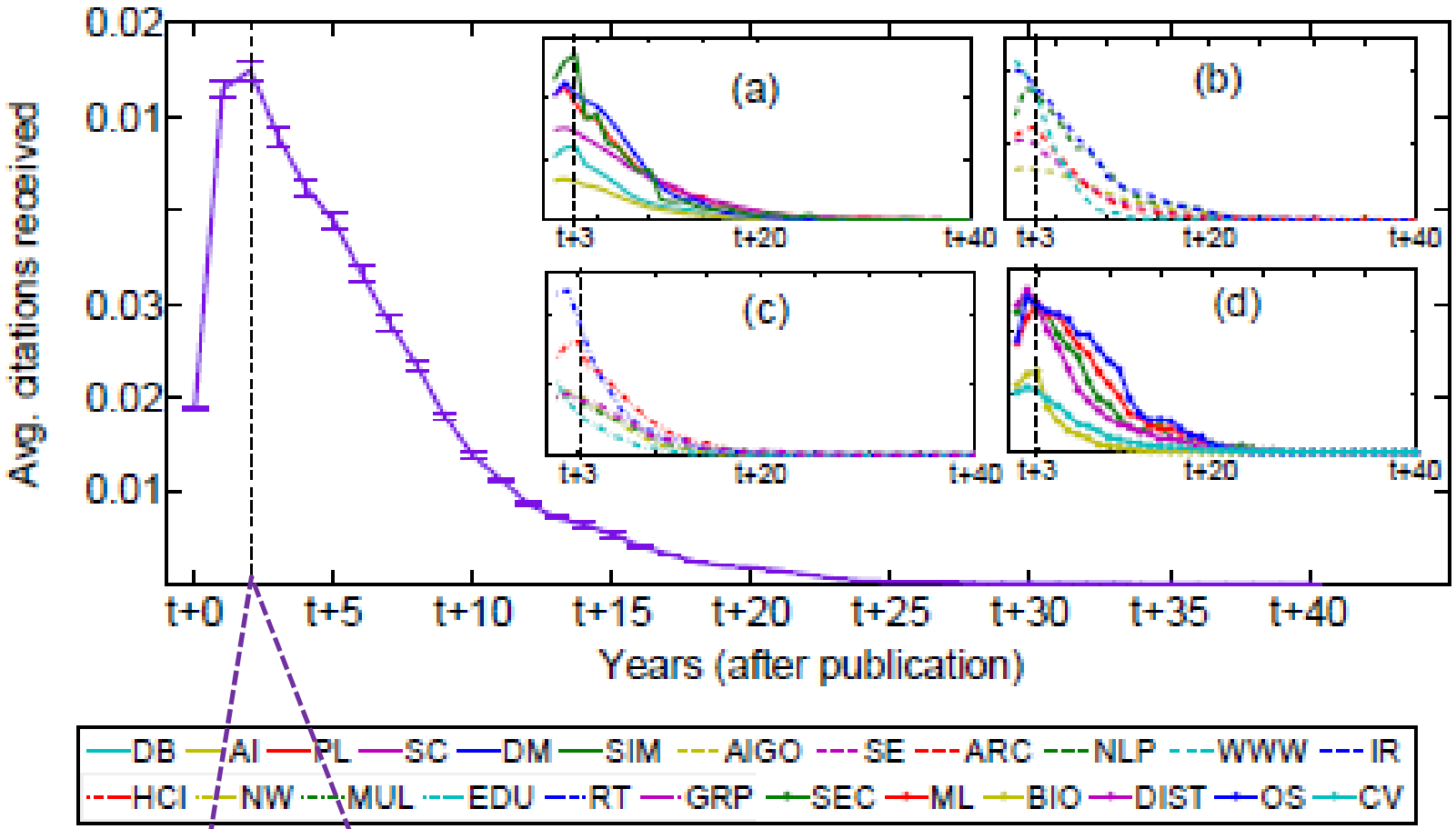
- Measuring the **impact of each field** (its constituent papers) around a particular year.

- **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

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

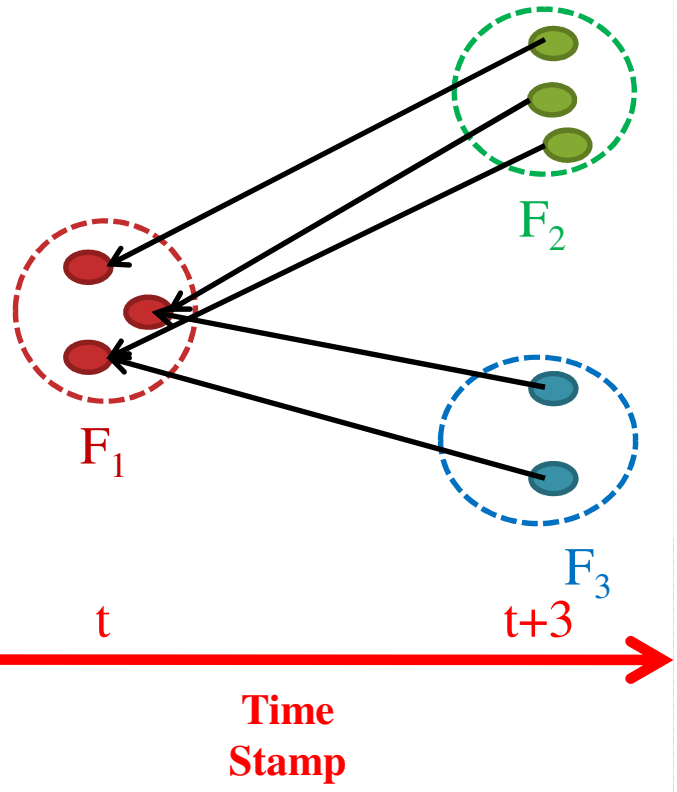
where,

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

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

P_i^t = # of papers in field f_i

$1 <= t <= 3$ (current year + next 3 years)

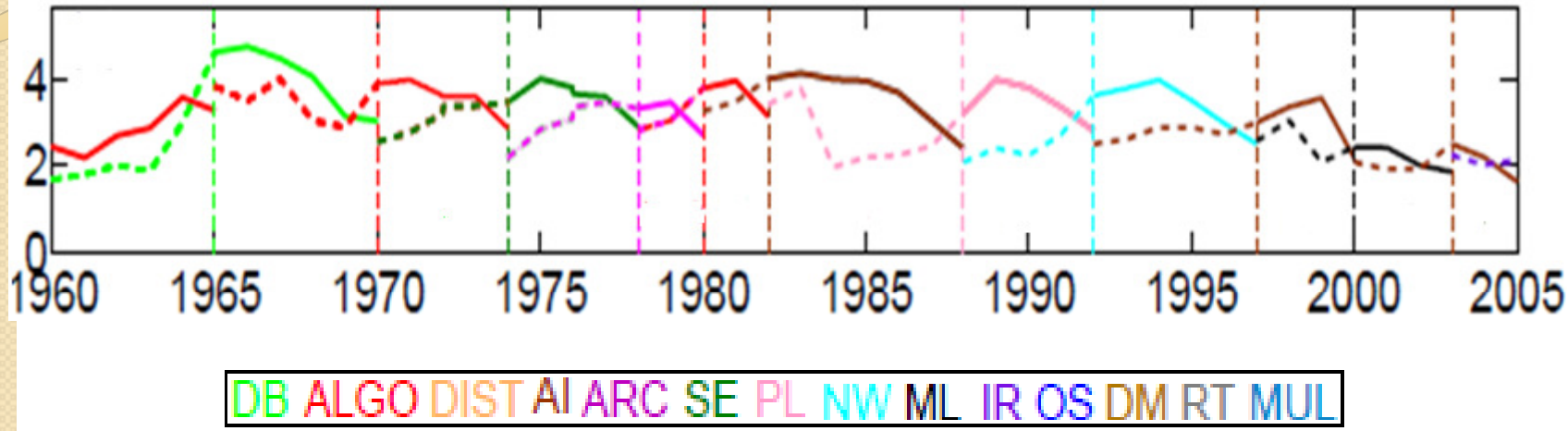


In(F1t)=5/3

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

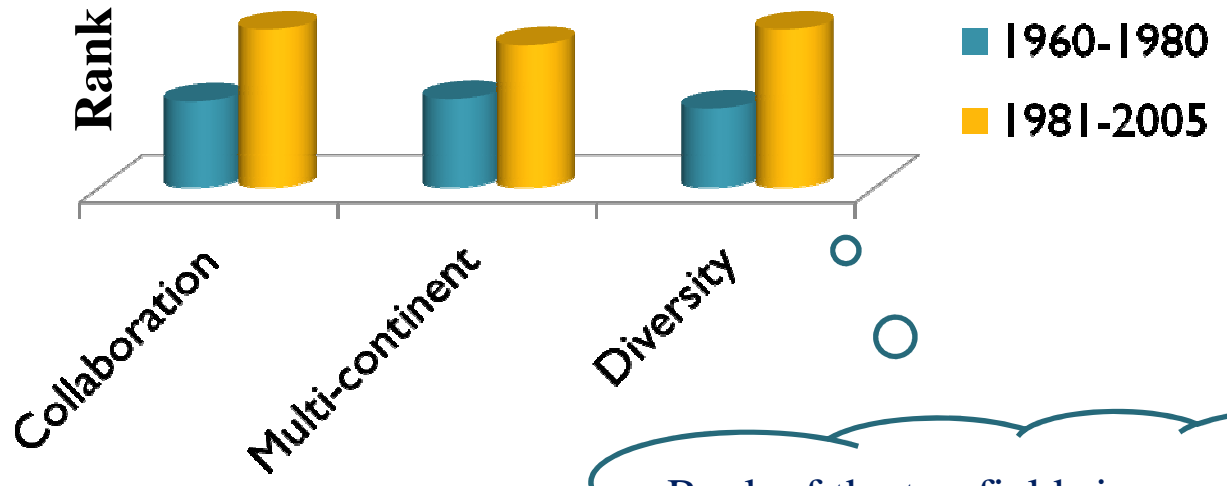
Conclusion

Probable Reasons

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

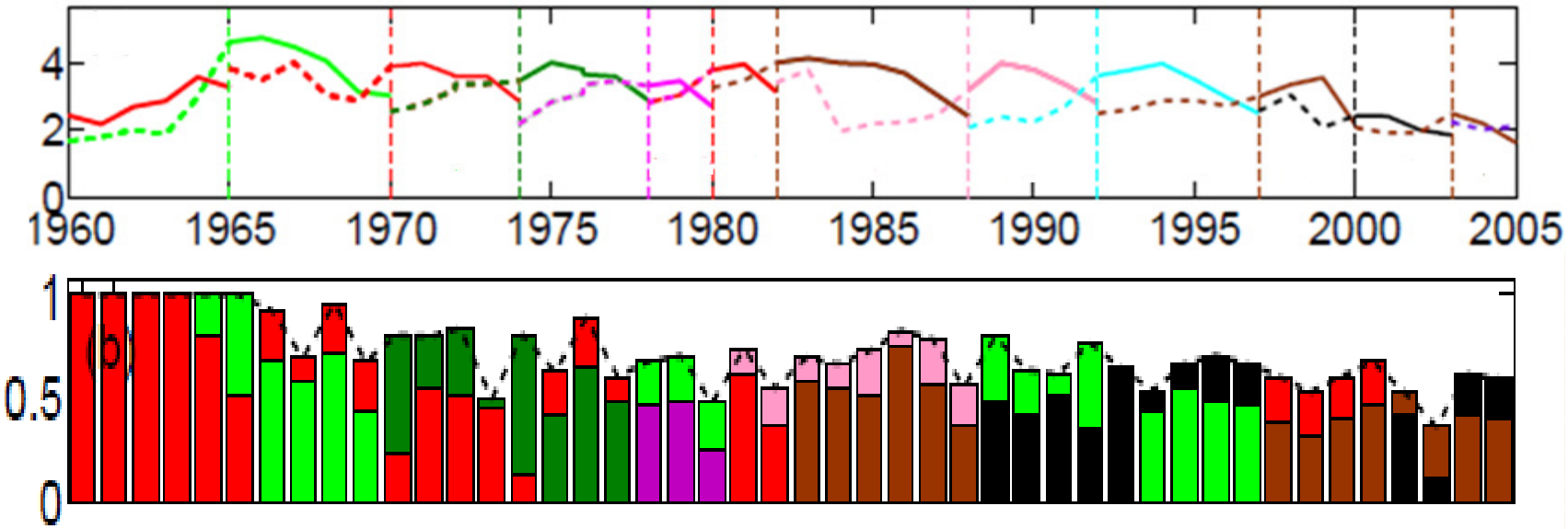
Reason 1: Collaborations

- 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

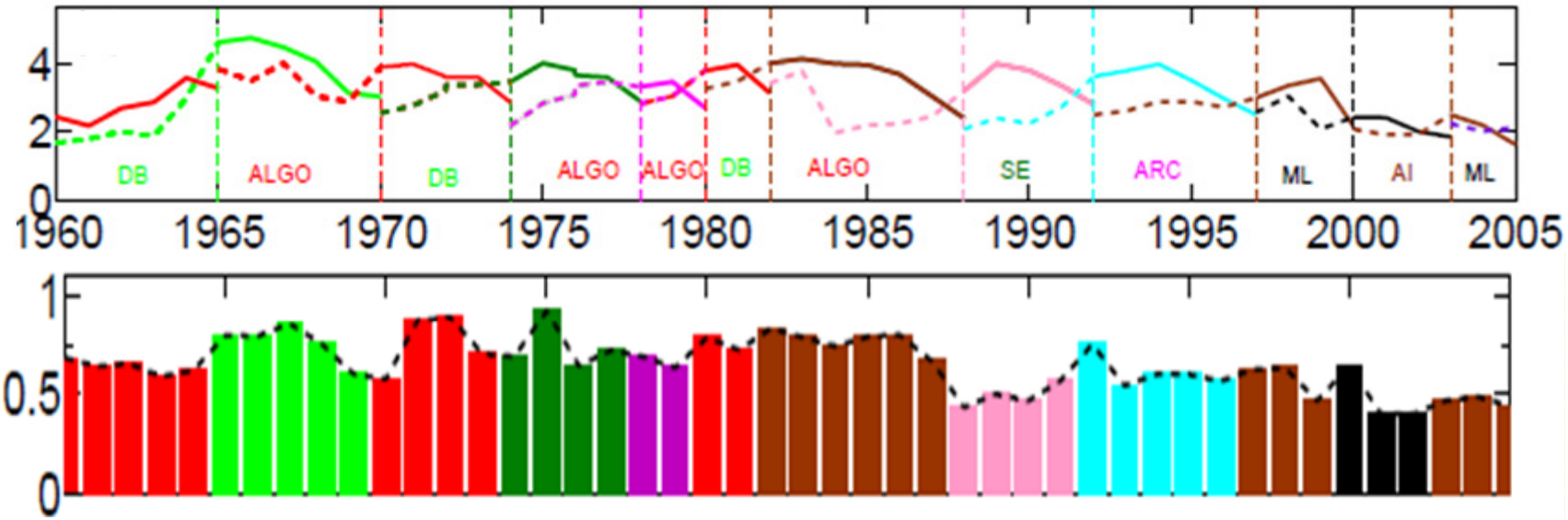
Reason 2: High Impact papers



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

- **82%** cases → 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 → **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**

www.nsf.gov

- 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

Yrs	Inwardness results	NSF	
		Proposal submitted	Proposal awarded
03	AI/IR/NW	NW/AI/HCI	NW/ALGO/SE
04	AI/IR/NW	AI/HCI/RT	RT/ARC/DIST
05	AI/IR/NW	AI/ML/HCI	GRP/SE/ALGO
06	IR/ML/AI	ML/ALGO/SEC	ALGO/SEC/ML
07	ML/AI/ALGO	ALGO/ML/HCL	ALGO/HCI/SEC
08	ML/AI/ALGO	ML/ALGO/SE	ALGO/ML/SE

During 2003-2008 , **top three fields** based on

(i) Our prediction

(ii) proposal submission statistics

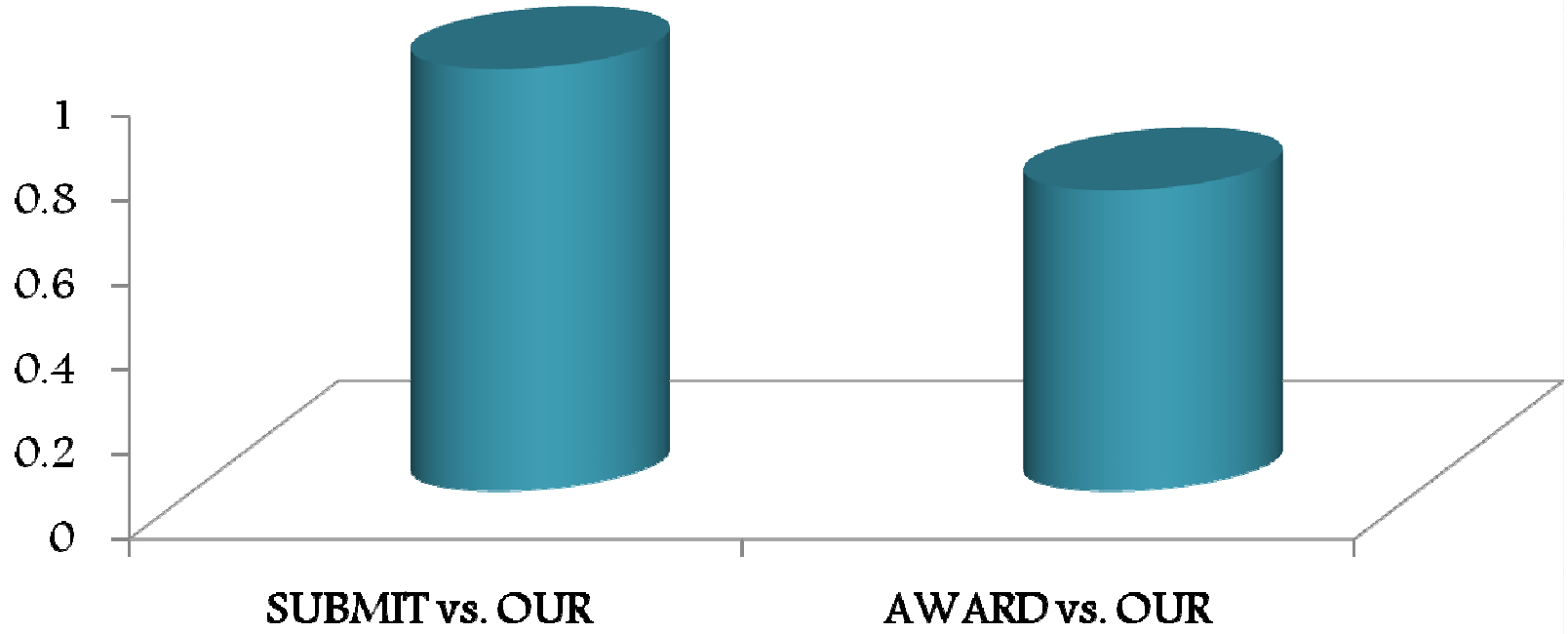
(iii) award statistics

Correlations with NSF Funding

- ***Correlation*(ζ) = s / n ;**

  s = similarity pair (at least one out of top three)

  n = no of year = 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 => ground-truth communities
- Temporal community interactions => scientific paradigm shift.
- Citation information => 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**

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