Automatic Classification and Analysis of Interdisciplinary Fields in Computer Sciences

Tanmoy Chakraborty

Google India PhD Fellow
Indian Institute of Technology, Kharagpur
India

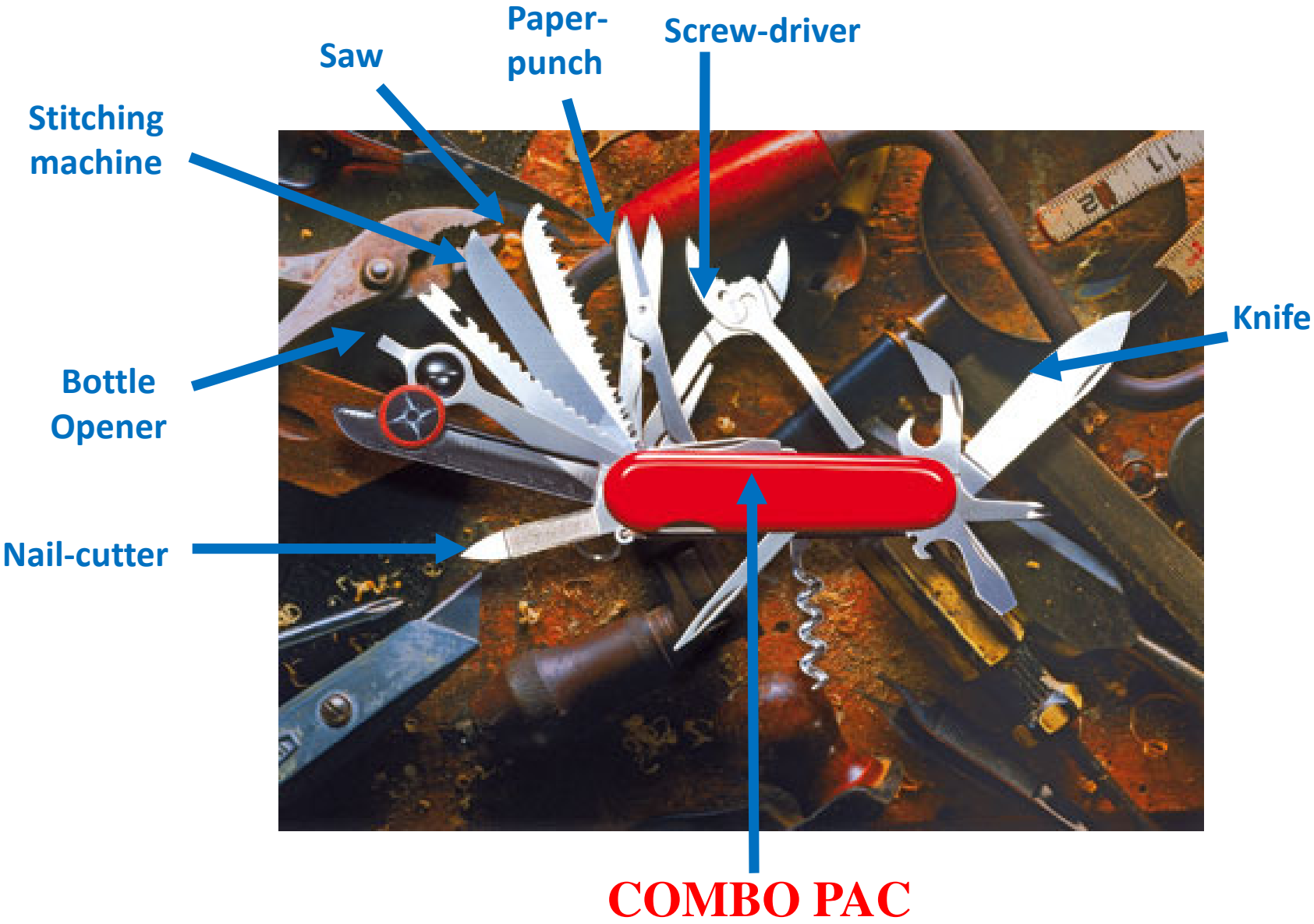
In collaboration with:
Srijan Kumar, M Dastagiri Reddy, Suhansanu Kumar,
Niloy Ganguly, Animesh Mukherjee
IIT-Kgp, India

2013 ASE/IEEE SocialCom, Washington D.C., USA, September 8-14, 2013

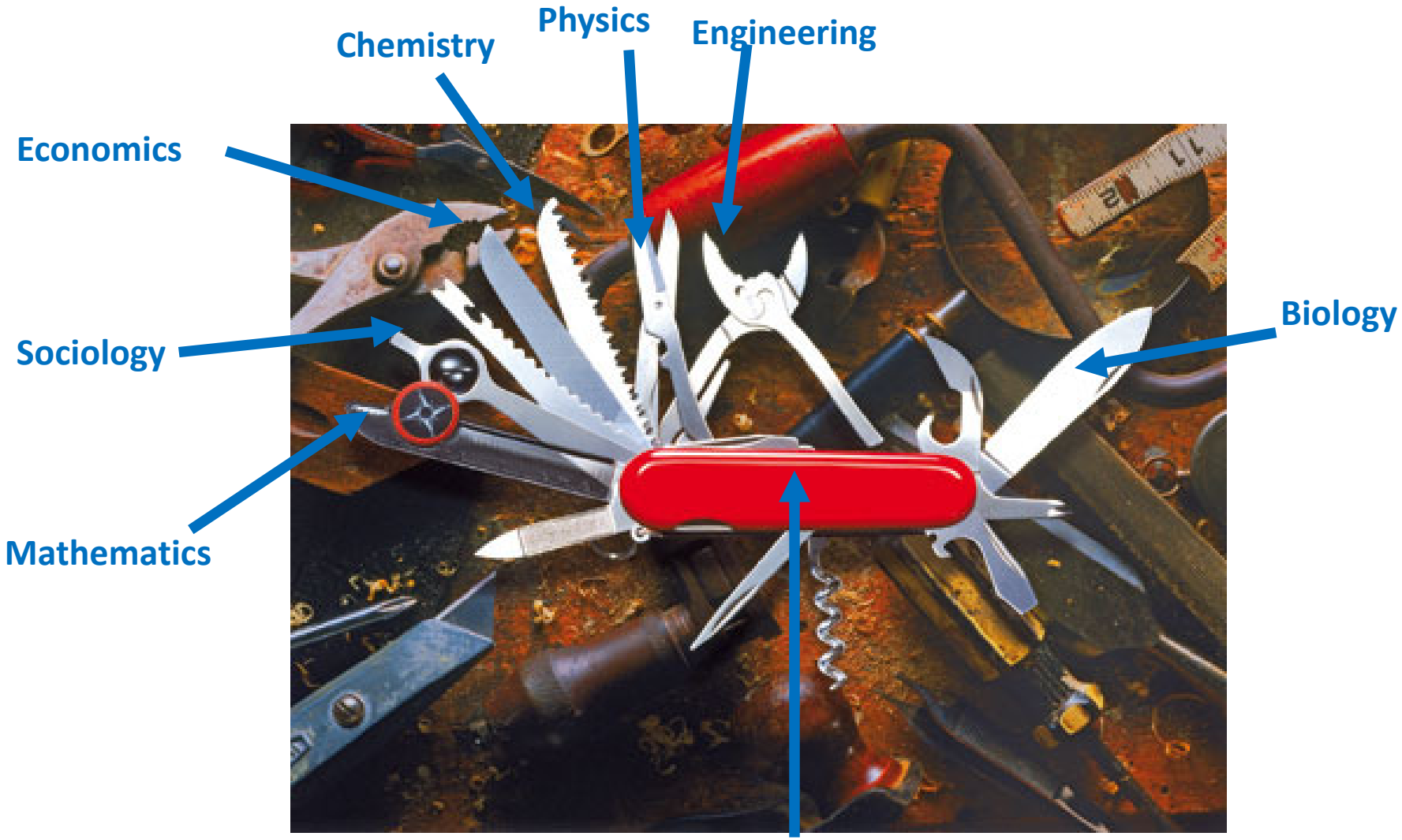
Outline

- Problem Definition
- Dataset
- Indicators of Interdisciplinarity
- Unsupervised Classification Model
- Evolution Landscape of Interdisciplinarity
- Core-periphery Analysis
- Conclusion

How to seek a good toolkit ?



Interdisciplinarity



In The Lines of Great Thinkers

"*We are not students of some subject matter, but students of problems. And problems may cut right across the borders of any subject matter or discipline.*"

– Karl Popper

"*Interdisciplinary research is the only way to do research in current times.*"

– Fritjof Capra

Outline

Problem Definition

Dataset
Indicators of Interdisciplinarity
Unsupervised Classification Model
Evolution dynamics of Interdisciplinarity
Core-periphery Analysis
Conclusion

Problem Definition

- Proper **quantitative indicators** of Interdisciplinarity
- **Unsupervised classification** of core and interdisciplinary fields
- **Evolution** dynamics of interdisciplinarity
- **Core-periphery analysis** of citation network

Outline

Problem Definition

Dataset

Indicators of Interdisciplinarity
Unsupervised classification model
Evolution dynamics of Interdisciplinarity
Core-periphery Analysis
Conclusion

Dataset

- Large **DBLP dump** used by Chakraborty et al. *(ASONAM, 13)*

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

- Bibliographic information during **1960-2008**

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

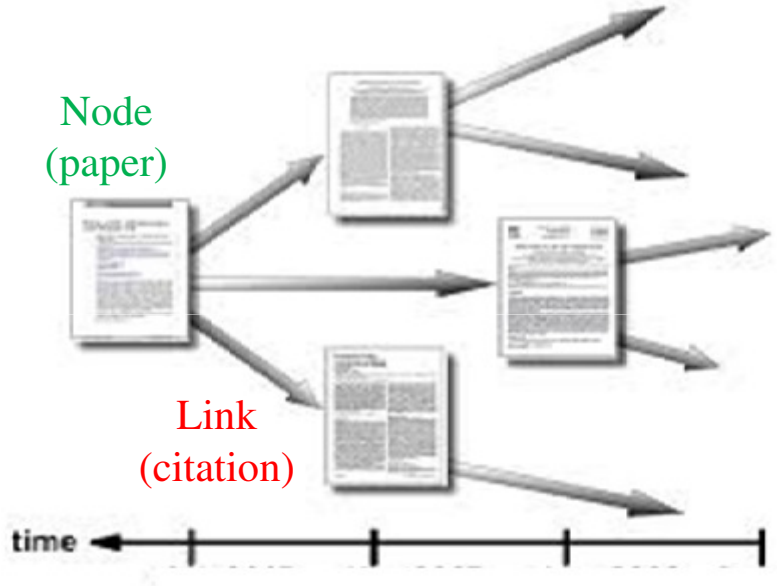
# of valid papers	702,973
# authors	495,311
# unique venue name	1,705

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

Citation network

- **Aggregated Network: 1960-2005**



- **Time-stamp wise Networks:**

  5 years sliding window (60-64, 61-65, 62-66, ..., 2001-2005)

Outline

Problem definition
Dataset

Indicators of Interdisciplinarity

Unsupervised classification model
Evolution dynamics of interdisciplinarity
Core-periphery organization
Conclusion

Indicators of Interdisciplinarity

- Reference Diversity Index (RDI)
- Citation Diversity Index (CDI)
- Membership Diversity Index (MDI)
- Attraction Index

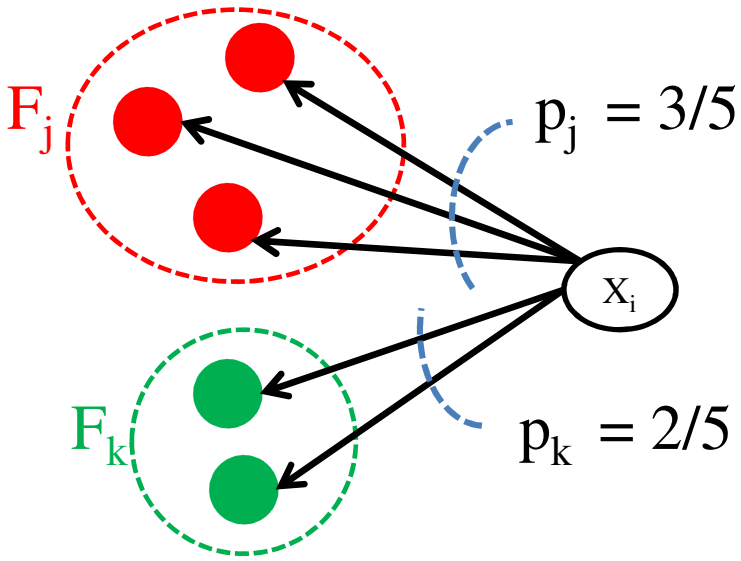
- Most of the indices are **Entropy** based measures
- **More Entropy => More diversity**

Reference Diversity index (RDI)

RDI of a paper X_i = $RDI(X_i) = -\sum_j p_j \log p_j$

p_j = proportion of references of X_i citing the papers of field F_j

More RDI, more interdisciplinarity



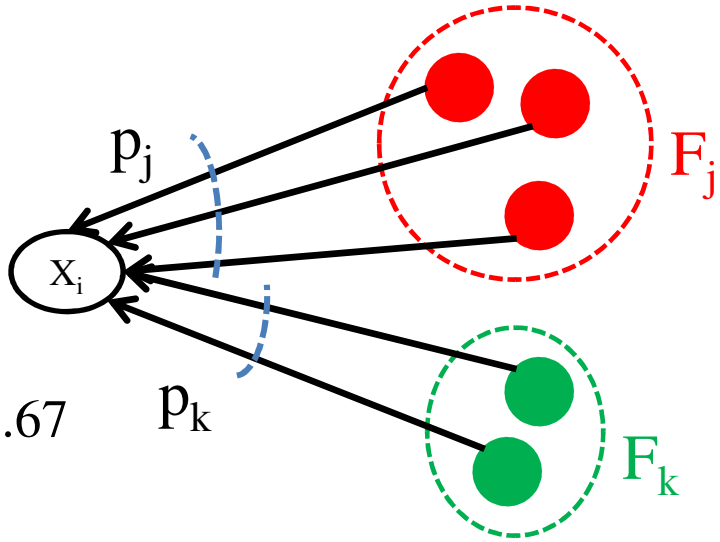
$RDI(X_i)$ = - 3/5 log (3/5) – 2/5 log (2/5)
= 0.67

Citation Diversity Index (CDI)

- CDI of a paper X_i at time t_i =

$$CDI_{t_i}(X_i) = -\sum_j p_j \log p_j$$

p_j = proportion of citations received by X_i from the papers of field F_j



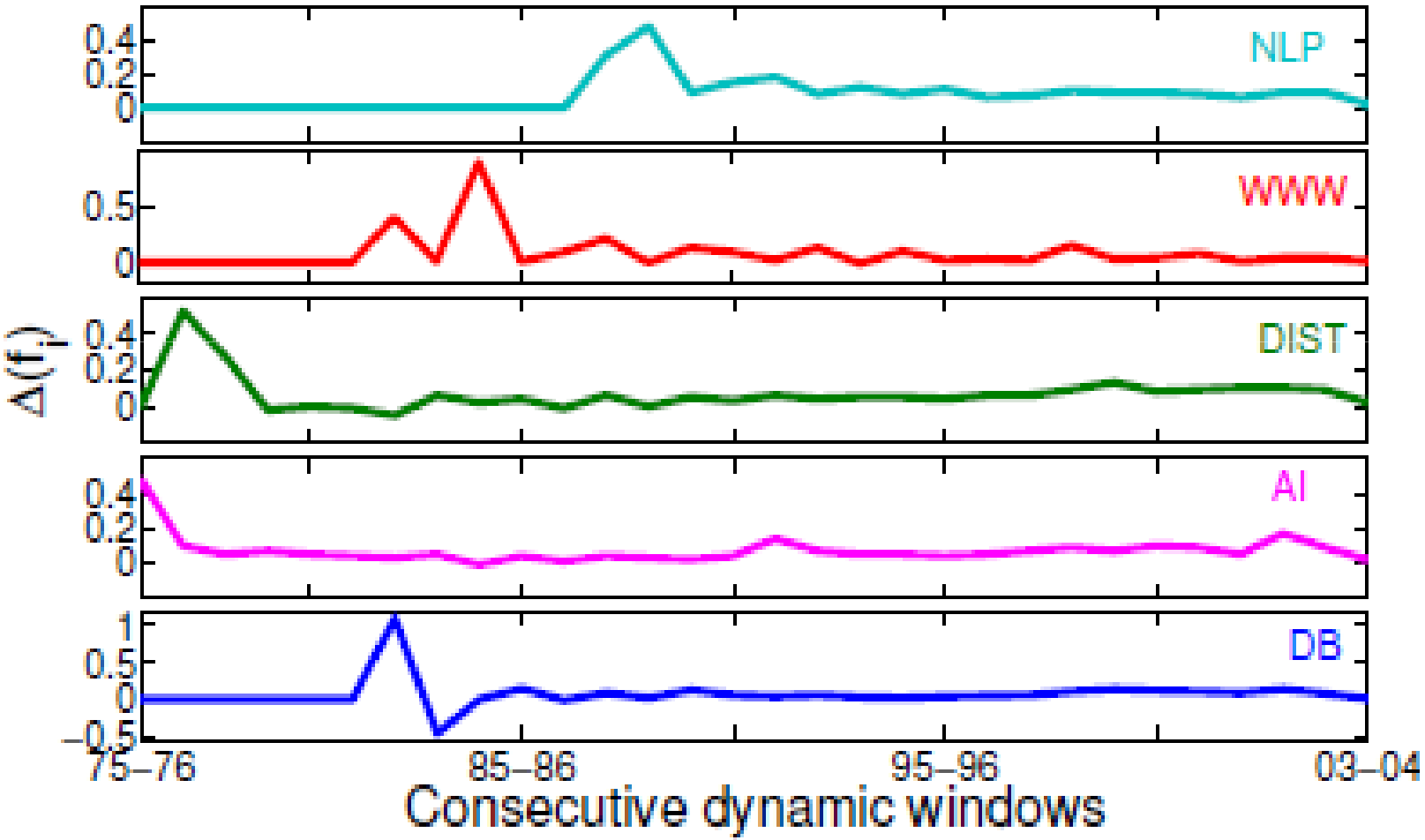
CDI(X_i) = 0.67

➢Drift:

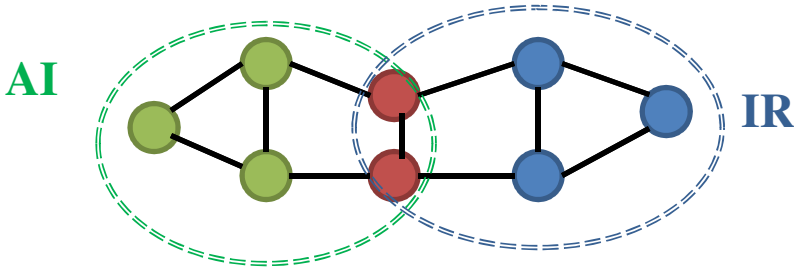
Drift of CDI between two successive time windows =

$$\Delta_{t_i}(f_i) = CDI_{t_{i+1}}(f_i) - CDI_{t_i}(f_i)$$

Spikes in CDI



Membership Diversity Index (MDI)



1. Identify overlapping communities
   [Xie et al., ICDM, 2011]

2. Tag the communities by the fields (major field in a group)

3. For each field f_i in the dataset,
   3.1 Observe the belongingness of all papers in different field-tagged communities
   3.2 Measure MDI

$$MDI(f_i) = -\sum_j p_j \log p_j$$

where, p_j is the fraction of overlapped papers of field f_i belonging to the communities tagged as f_j

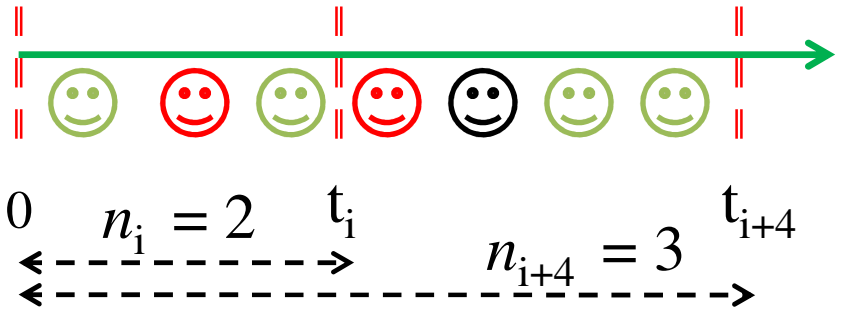
More MDI, more interdisciplinarity

External Evidence: Attraction index

$$\chi_f = \frac{n_{i+4} - n_i}{c_i}$$

- n_i: # unique authors up to the year t_i (in field f)
- n_{i+4}: # unique authors up to the year t_{i+4} (in field f)
- c_i: # publications in f in the time window (t_{i+4} - t_i)

More χ , more interdisciplinarity



$$\chi_f = 1/4$$

Outline

Problem definition
Dataset
Indicators of Interdisciplinarity

Unsupervised Classification Model

Evolution dynamics of interdisciplinarity
Core-periphery organization
Conclusion

Unsupervised Classification Model

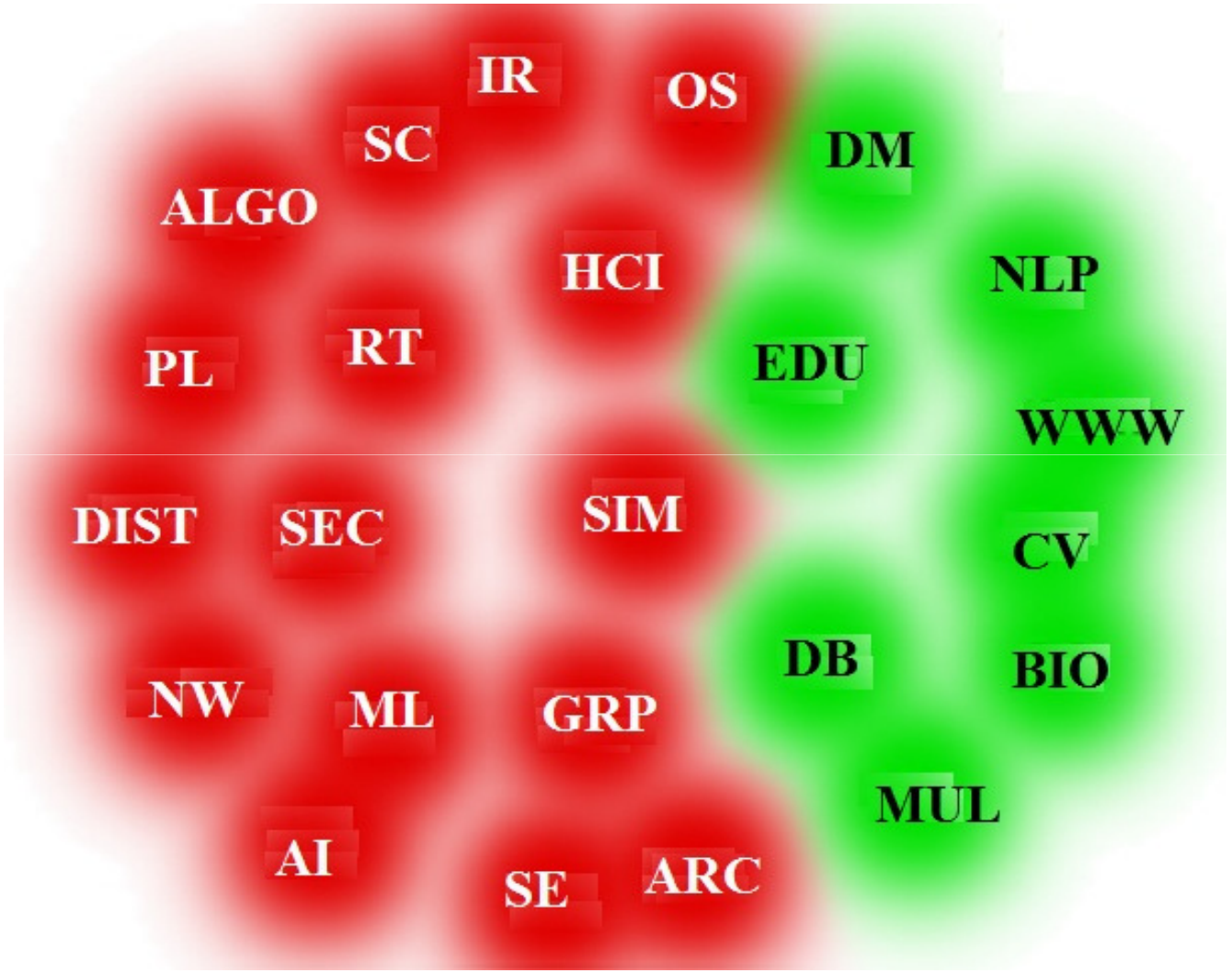
- **A field** is represented by **a vector of size 4** indicating four features

- **Adjacency matrix A** of size 24×24

  A(i,j)= Cosine similarity of field *i* and *j*

- Clustering algorithm proposed by Waltman et al. *(J. Informetrics, 2010)*

Result of the Classification



Outline

Problem definition
Dataset
Indicators of Interdisciplinarity
Unsupervised classification model

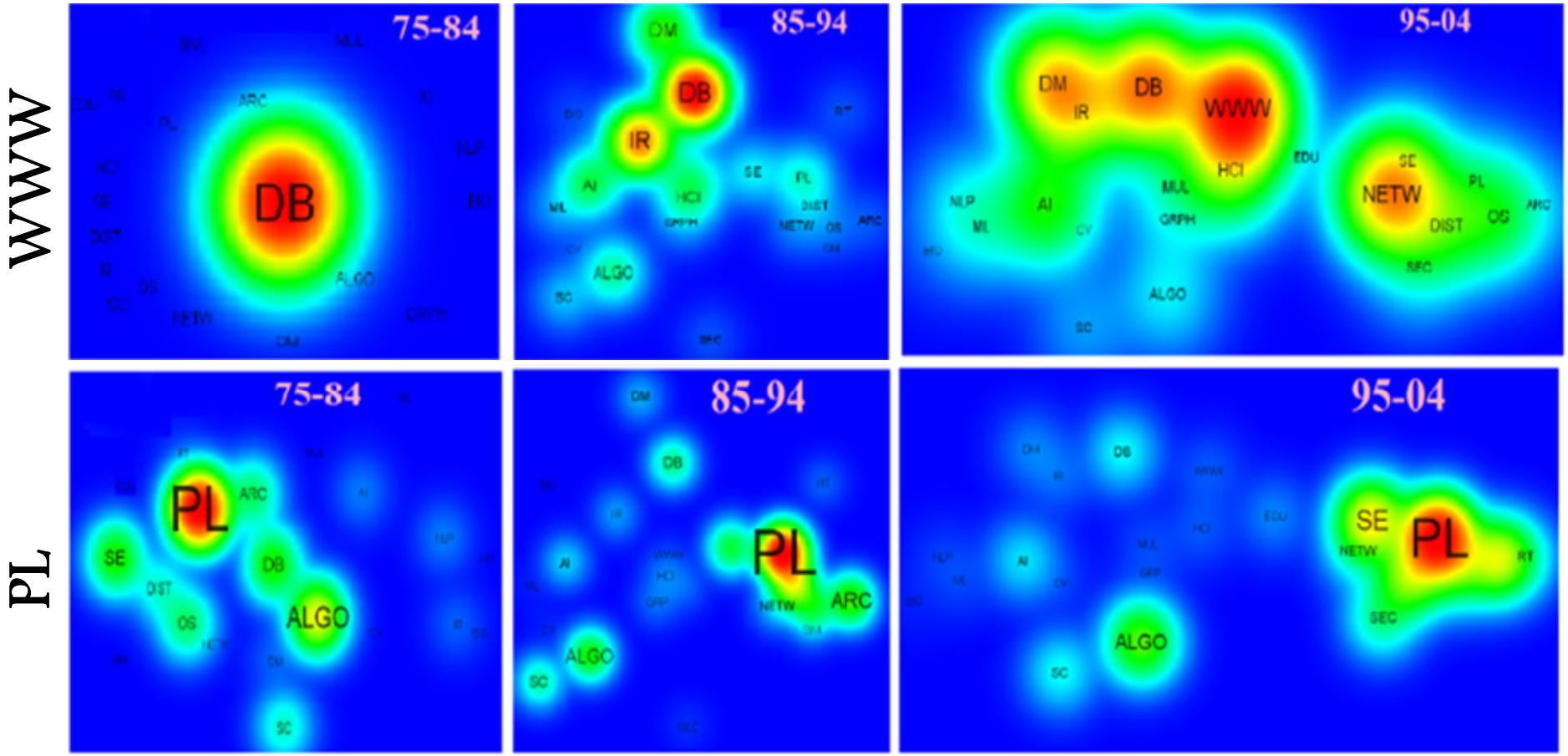
Evolution Landscape of Interdisciplinarity

Core-periphery organization
Conclusion

Evolution dynamics

- Construct a field-field citation network
- 24 nodes in each time-stamp
- Draw directed and weighted edges based on citations
- Observe the citation distribution across the fields

Evolutionary Landscape



- Fields are grouped based on the **connection proximity**
- The **size of the font** indicates the **relative importance** (# of incoming citations) of a field

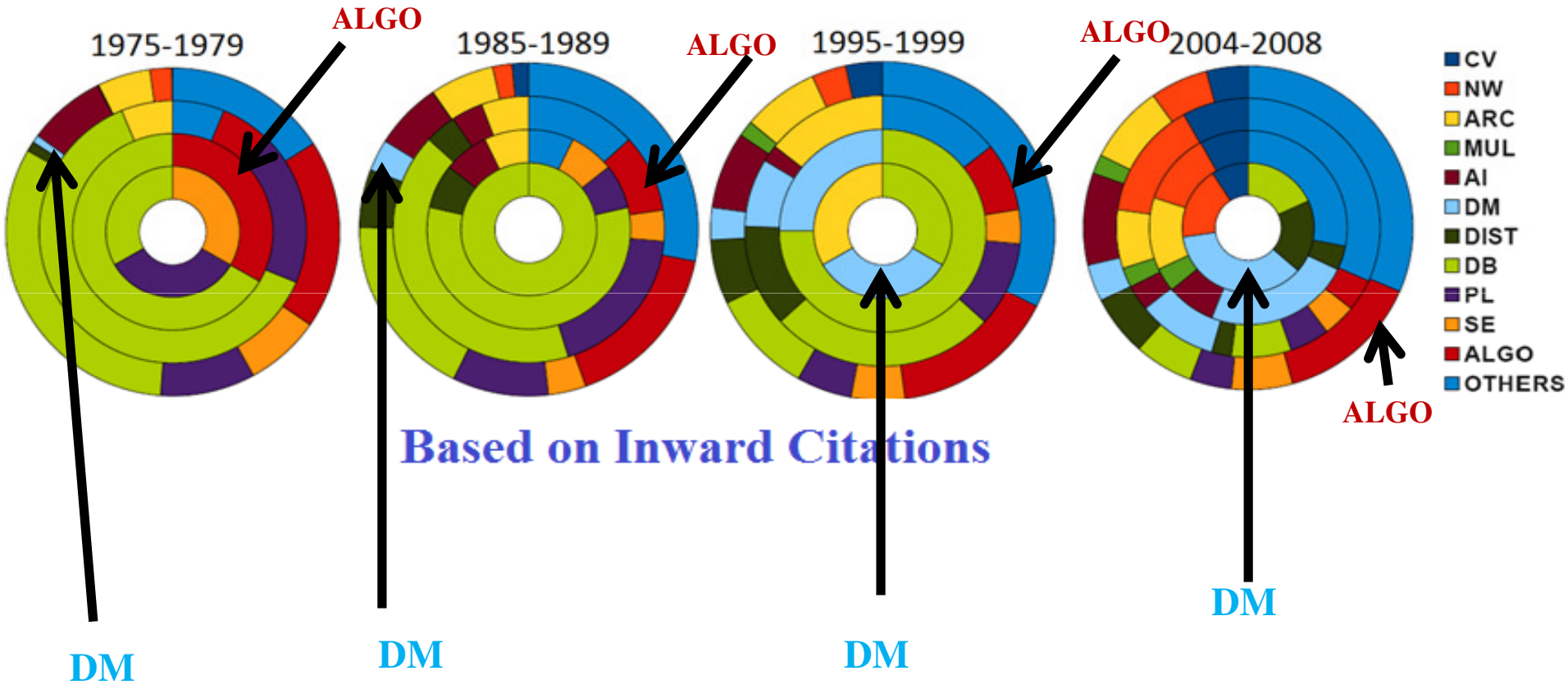
Outline

Problem definition
Dataset
Indicators of Interdisciplinarity
Unsupervised classification model
Evolution dynamics of interdisciplinarity

Core-periphery Analysis

Conclusion

Core-periphery Analysis



Outline

Problem definition
Dataset
Indicators of Interdisciplinarity
Unsupervised classification model
Evolution dynamics of interdisciplinarity
Core-periphery organization

Conclusion

Conclusions

- Quantitative indications of interdisciplinary
- Automated scheme to identify interdisciplinary fields
- Evolutionary landscape depicts the cross-hybridization among fields
- K-core analysis shows the steady movements of interdisciplinary field at the core

➢ Future Directions:

- Identify interdisciplinary papers
- Predicting the possible fields to be intermingled next.

Acknowledgements

- Financial Support: Google India Pvt. Ltd.

- Technical support:

Complex Network Research Group (CNeRG), IIT-Kgp

http://cnerg.org/

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