Community Analysis in Large Networks: Methods and Applications

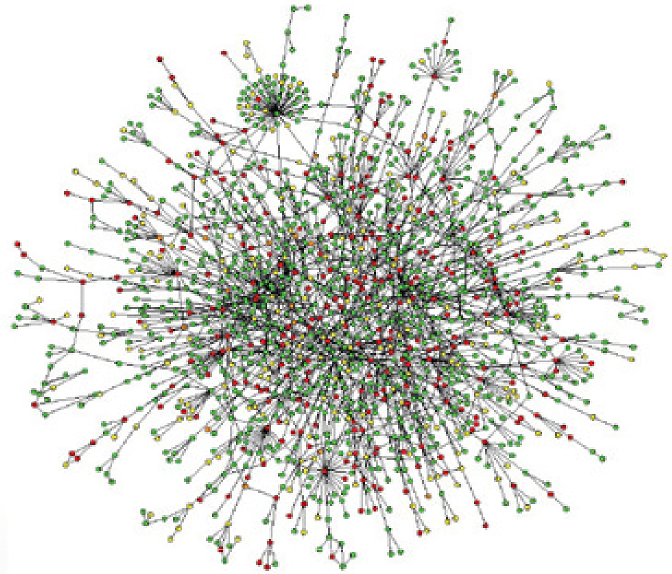
Tanmoy Chakraborty
PhD. Student

Advisors: Dr. Animesh Mukherjee
Prof. Niloy Ganguly

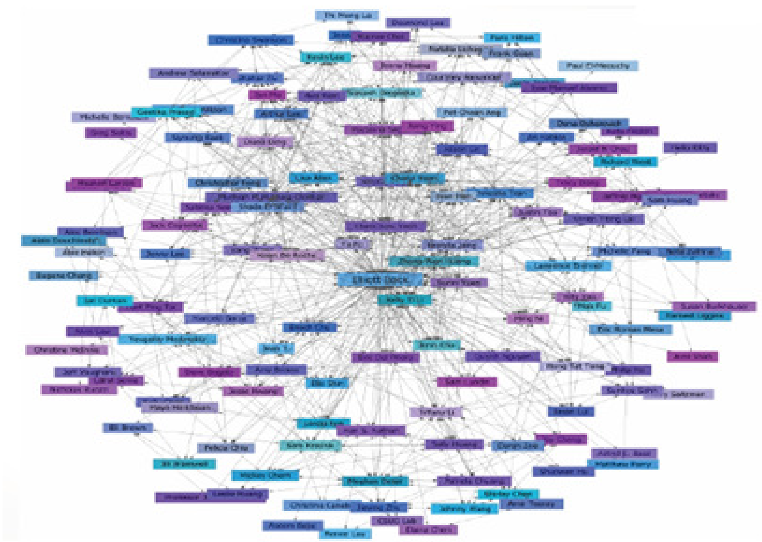
Dept. of Computer Science & Engineering
Indian Institute of Technology, Kharagpur, India

PhD. Defence Seminar, September 22, 2015

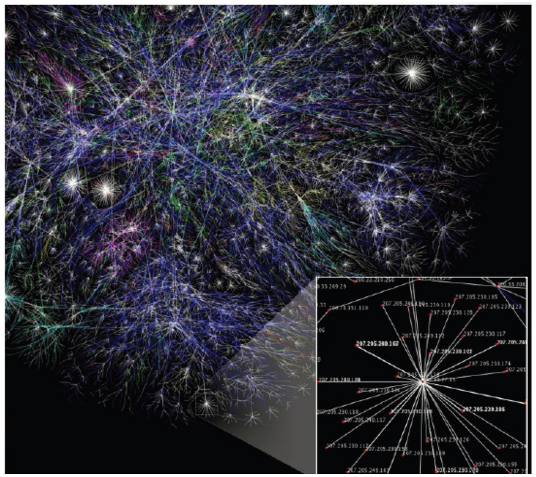
Networks



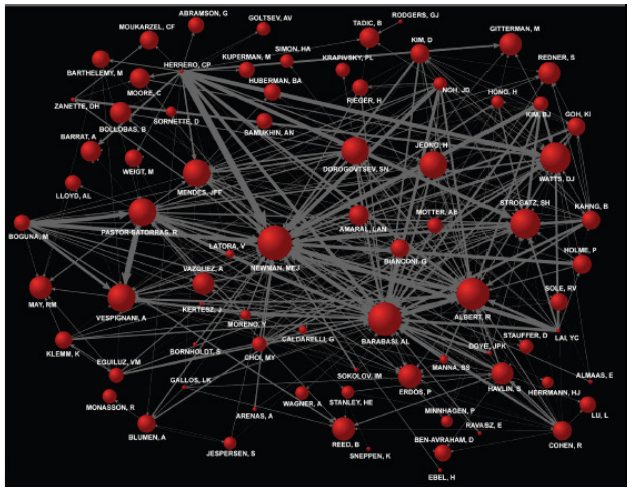
Protein-protein interaction network



Social network



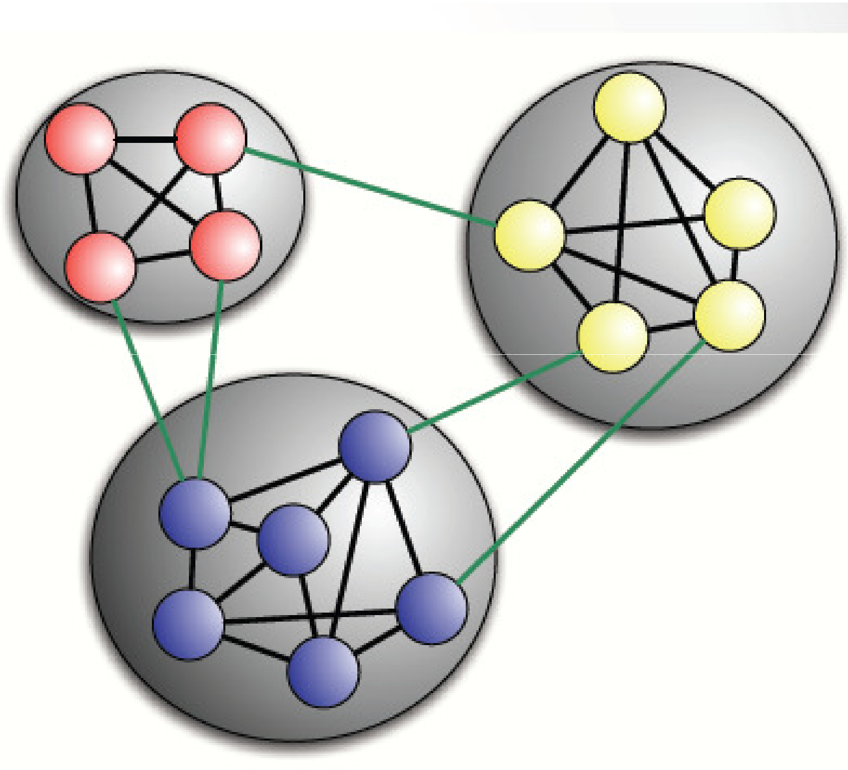
Internet



Citation network

Community Structure

Communities: sets of tightly connected nodes



- People with common interests
- Scholars working on the same field
- Proteins with equal/similar functions
- Papers on the same/related topics
- …

Similar functionality

Questions We Ask

- Why are the algorithms dependent on the **vertex ordering**?
  - Invariant substructure in the networks
  - Characterizing such substructure
- Community: a **local property** or a global property?
  - Heterogeneity of belongingness
  - Quantitative indicators of belongingness
- How do real-world **communities interact**?
  - Evolutionary landscape of evolving communities
  - Modeling real interaction phenomenon
- How do we use community information for **applications**?
  - Analyzing and modeling patterns in networks
  - Designing prediction and recommendation systems

Our Work: Community Analysis

Our research focuses on quantifying "**meaningful communities**" in real networks

S1: Methods:

Design metrics and algorithms

S2: Applications:

Design real systems

Our Work: Overview

	S1: Methods	S2: Applications
G1: Patterns	Nat. Sci. Rep. 13	CACM 15
G2: Analysis/ Modeling	ASONAM 13 SNAM 14 SocialCom 13 COMSNETS 14	JOI 15 CACM 15 SIGKDD 15
G3: Algorithms/ Predictions	SIGKDD 14 IEEE TKDE (submitted) ACM TKDD (submitted)	JCDL 14 ICDE 15

Organization of the Thesis

Methods

Chapter 1: Constant Communities in Networks

Chapter 2: Permanence and Community Structure

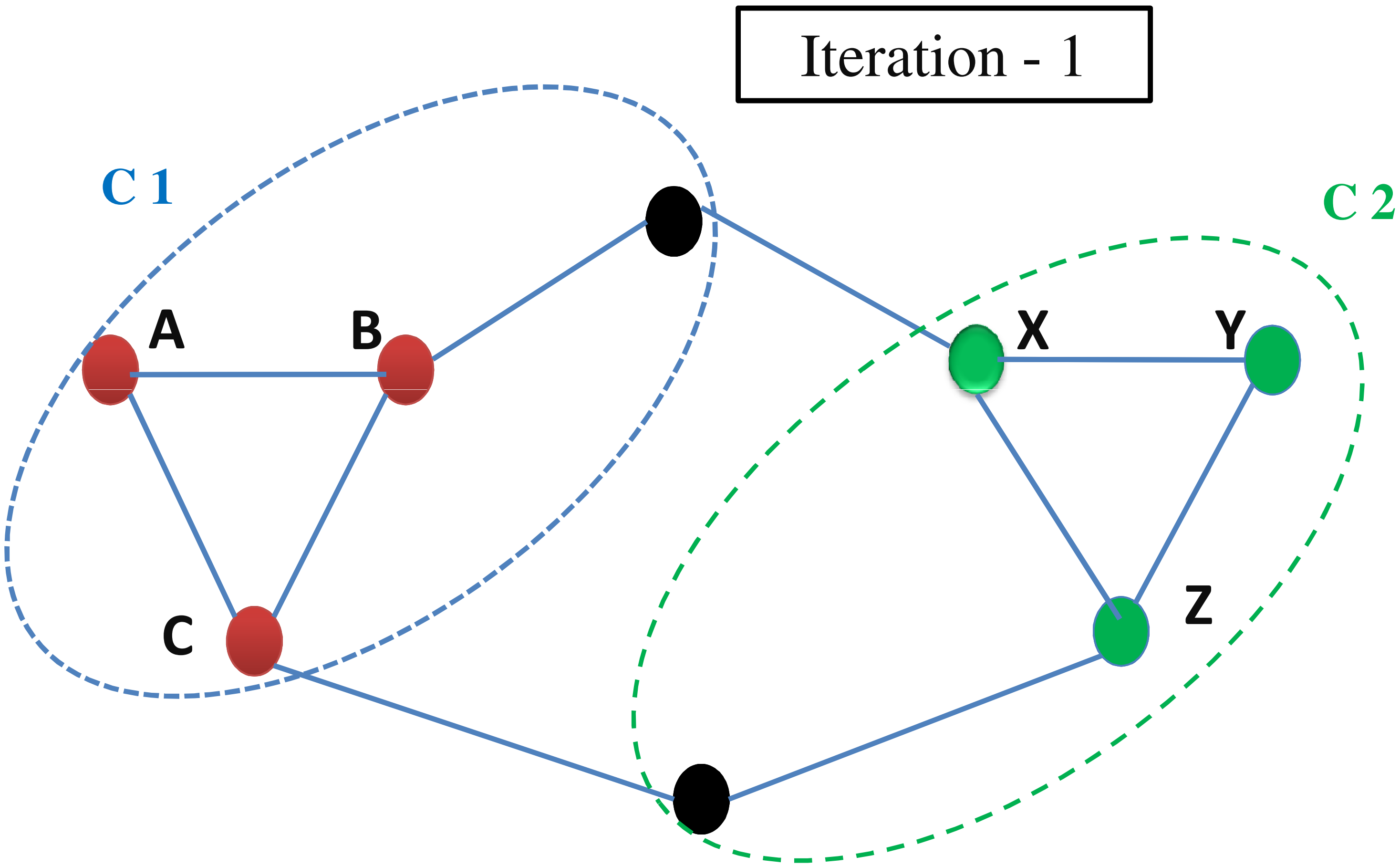
Chapter 3: Analyzing Real-world Communities

Chapter 4: Community-based Applications

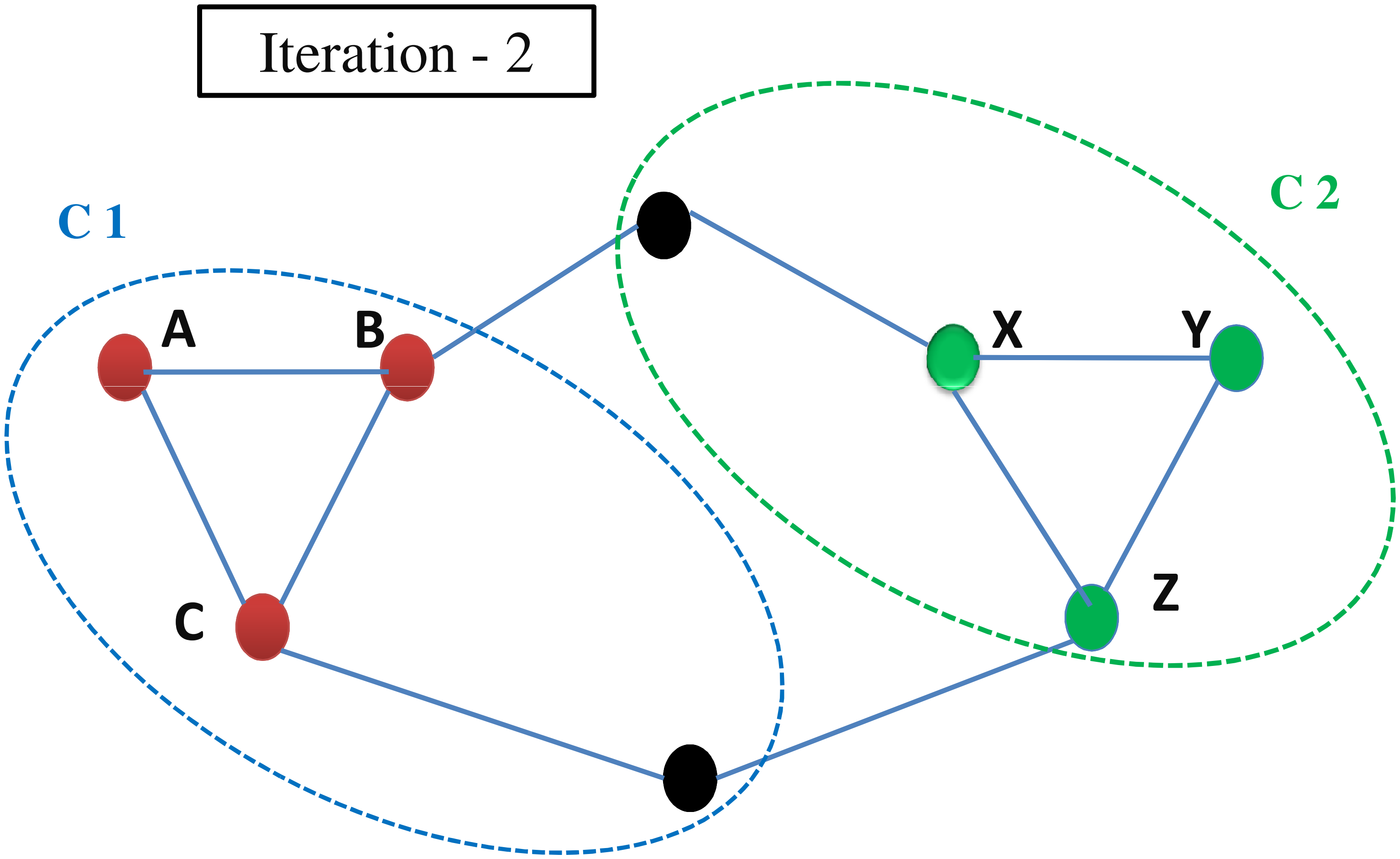
Applications

Chapter 1：
Constant Communities in Networks

Vertex Ordering

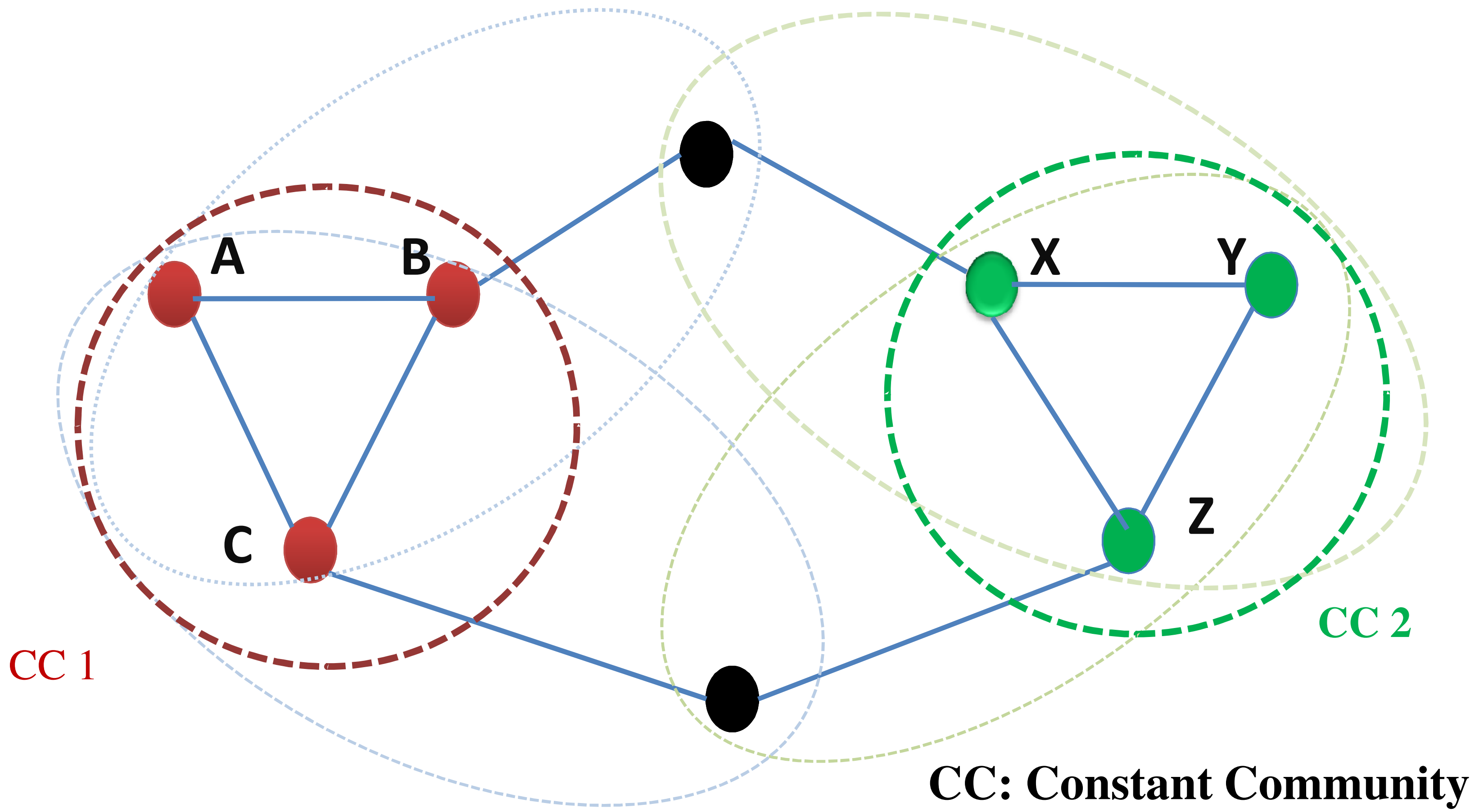


Vertex Ordering



Constant Communities

Combining previous results

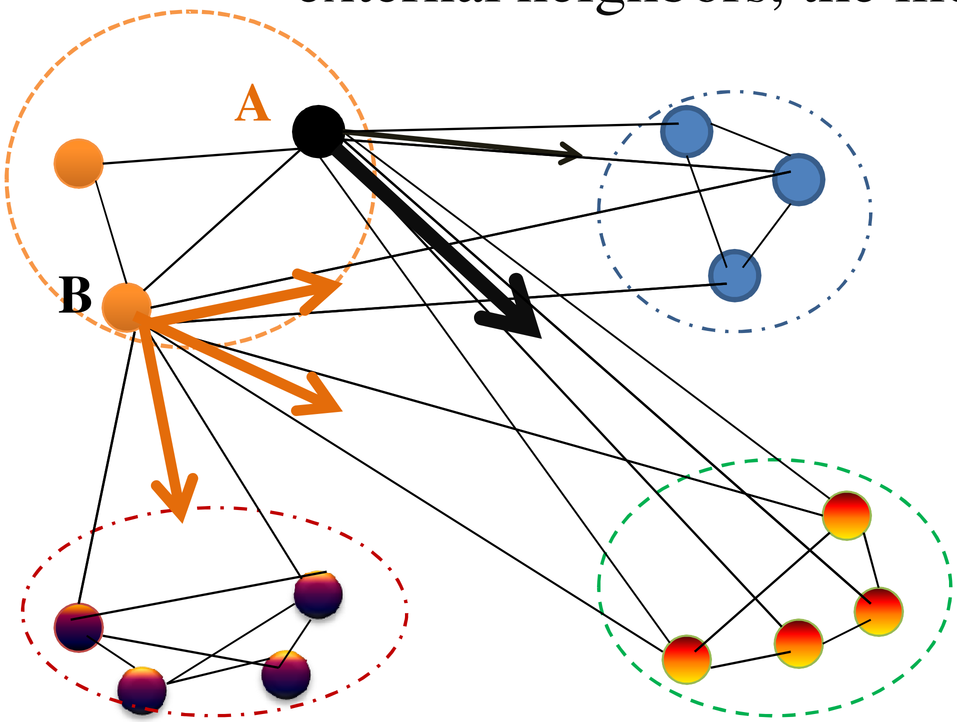


Group of vertices **always remain together** under any vertex ordering

Characterizing Constant Vertices

Two factors:

- **(i) Internal strength:** the more the number of internal neighbors, the more it becomes stable.
- **(ii) Divergence of external pull:** the more distributed the external neighbors, the more it becomes stable.



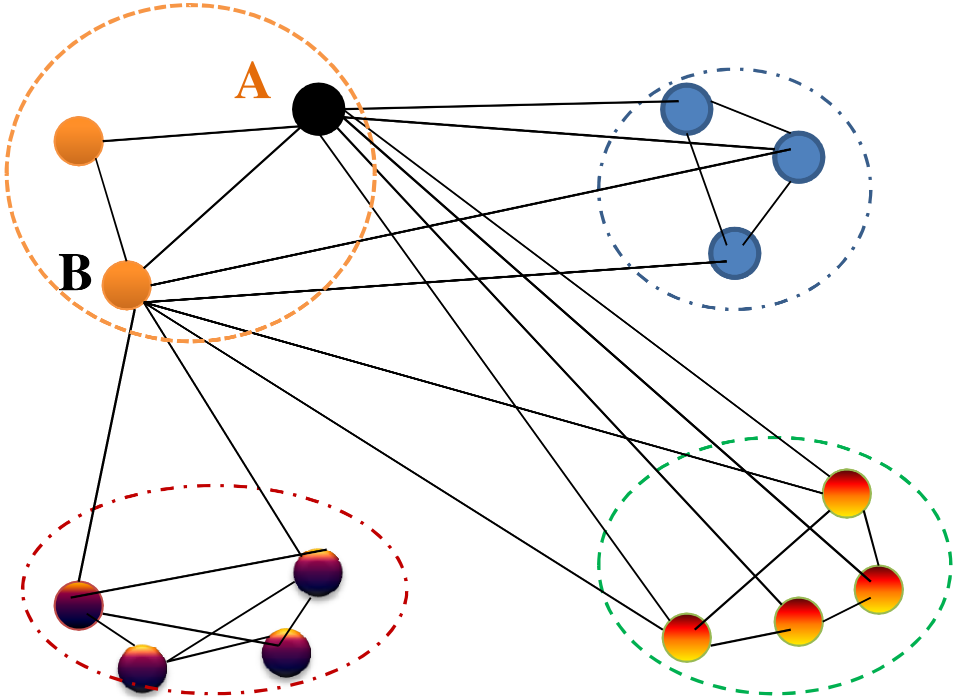
- B is more stable than A

Relative Permanence

$$\Omega(v) = \frac{In(v)}{D(v)} \times \frac{\sum_{i=1}^{k} \frac{1}{ENG_i(v)}}{EN(v)}$$

External neighbor divergence

Internal strength



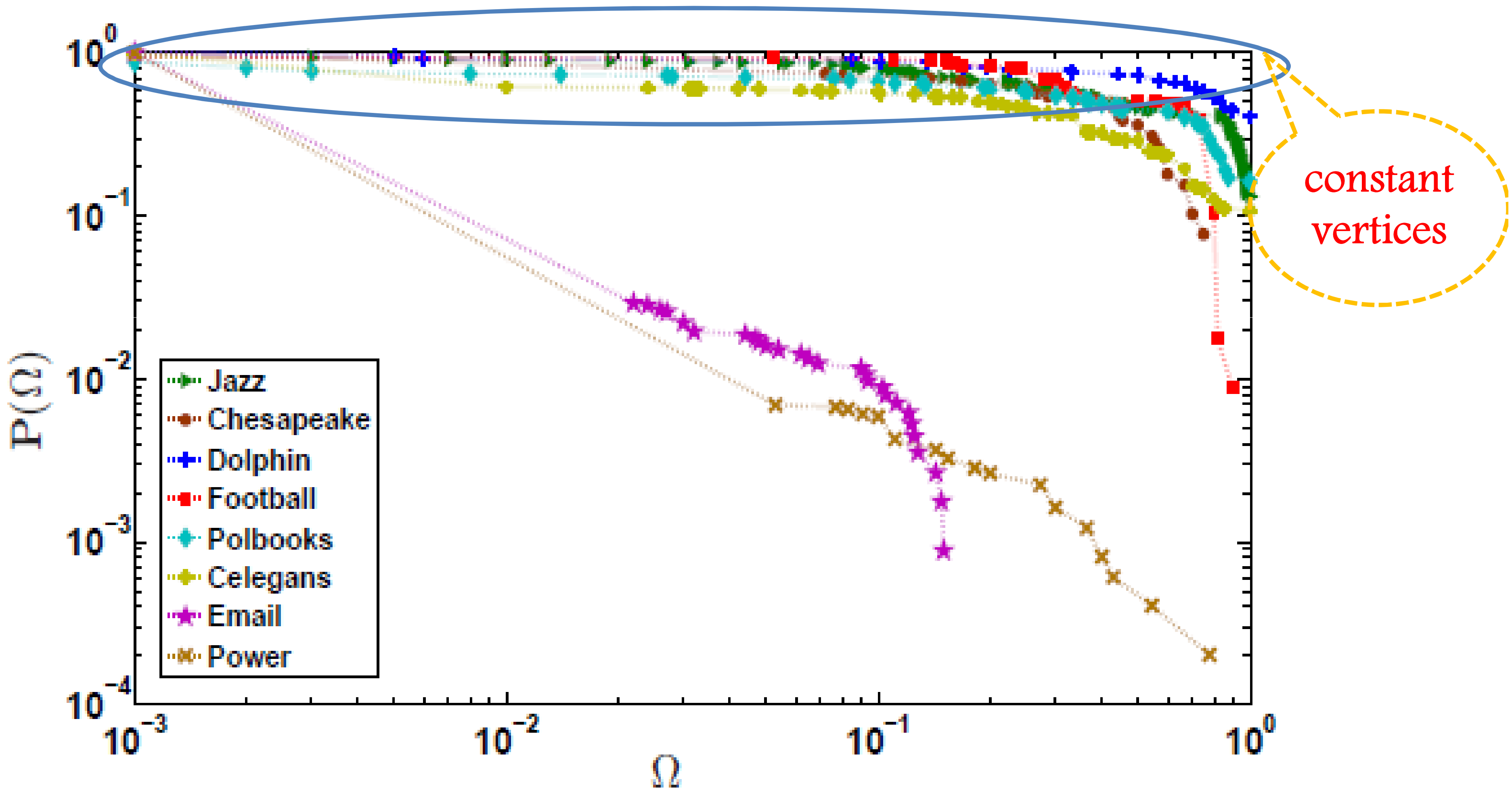
- Ω(v)=Relative permanence
- *In(v)*=# of internal neighbor
- *D(v)*= Degree of *v*
- *K*=#of external neighbor comm
- *EN*(*v*)=#external neighbors of *v*
- $ENG_i(v)$=Connections to *i*th external neighbor community

$$\Omega(A) = \frac{2}{8} \times \frac{\frac{1}{4} + \frac{1}{2}}{6} = \frac{1}{32}$$

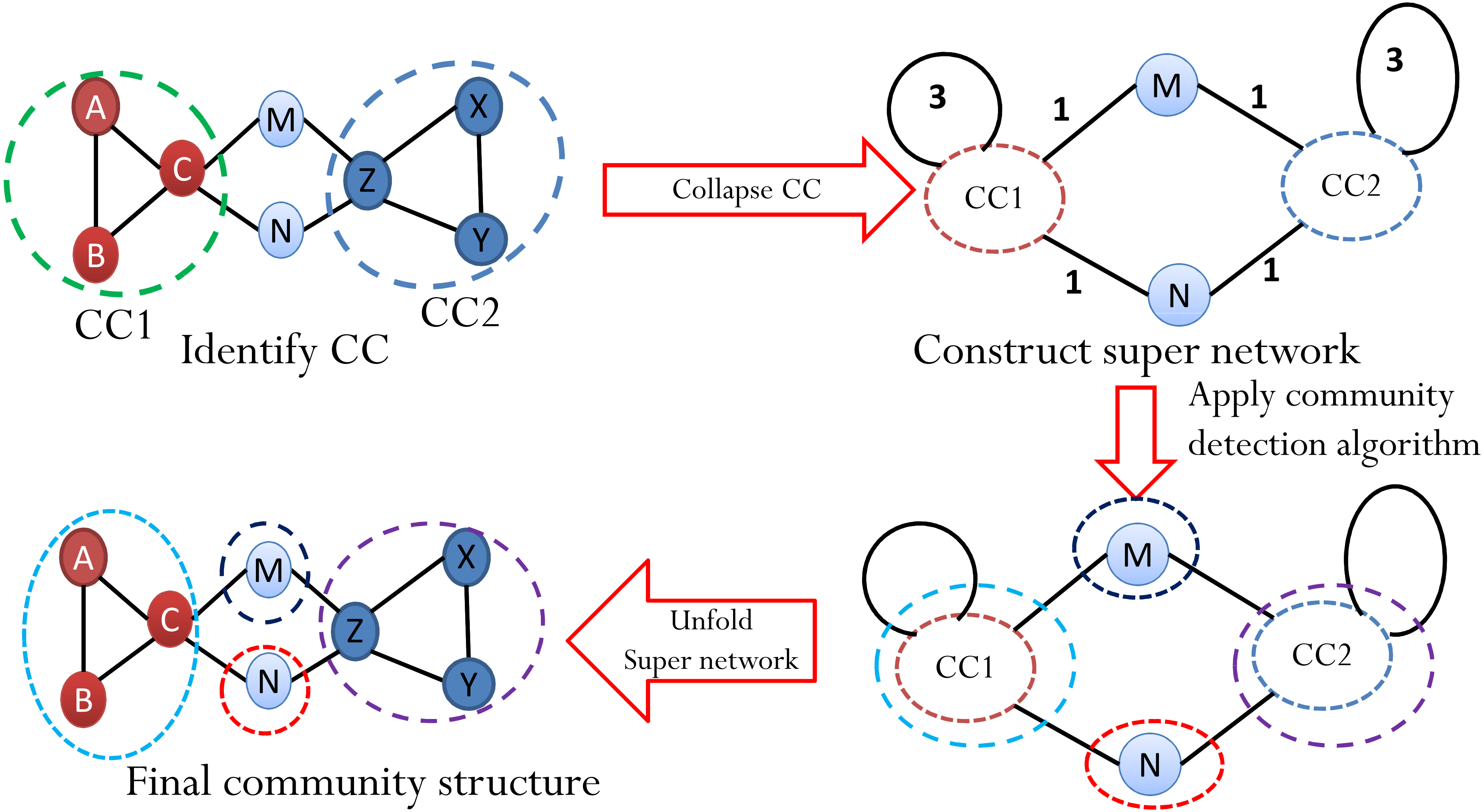
$$\Omega(B) = \frac{2}{8} \times \frac{\frac{1}{2} + \frac{1}{2} + \frac{1}{2}}{6} = \frac{1}{16}$$

Discussion

Distribution of Relative Permanence



Improving Community Detection Algorithms



Modularity (Q) Improvement on Real Networks

Networks	Louvain			CNM		
	Improvement of Q (%)	Variance of Q (- CC)	Variance of Q (+ CC)	Improvement of Q (%)	Variance of Q (- CC)	Variance of Q (+ CC)
Polbook	3.34	1.74e-5	1.2e-32	1.20	2.25e-5	0
Dolphin	1.30	1.76e-5	0	1.90	0.9e-10	0
Football	2.45	2.01e-5	0	3.05	7.25e-8	6.4e-10
Email	4.80	6.89e-5	0.9e-12	5.80	1.7e-8	1.36e-12

Chapter 2:

Permanence and Community Structure

Modularity

$$Q = \frac{1}{m} \sum_{c=1}^{n_c} \left(\underbrace{\boxed{l_c}}_{\text{Actual edges}} - \underbrace{\boxed{\frac{d_c^2}{4m}}}_{\text{Expected edges}} \right)$$

M. E. J. Newman, M. Girvan, PRE , 2004

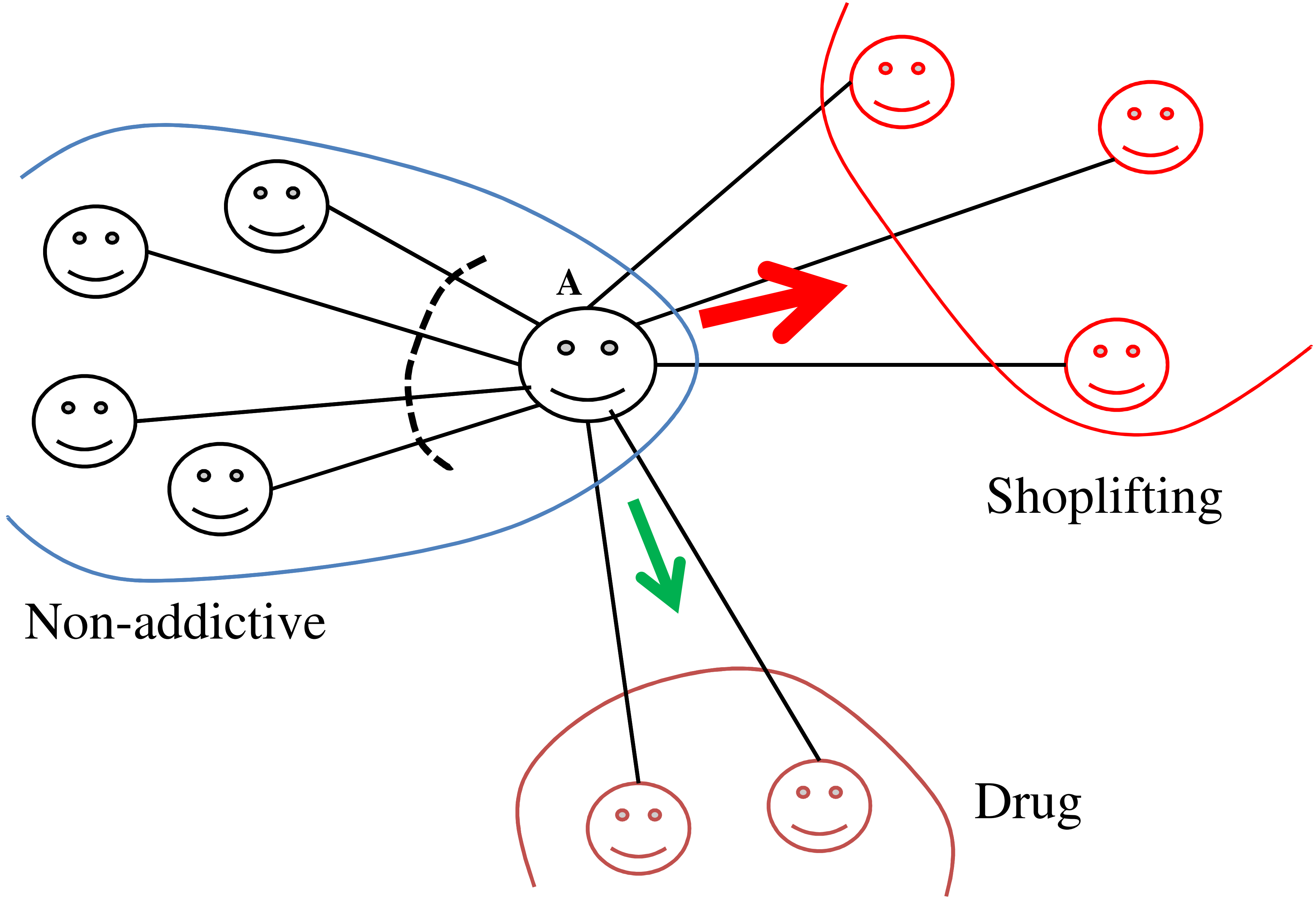
M. E. J. Newman, PRE, 2004

m = # edges
n_c = # communities
l_c = # internal edges in community c
d_c = sum of degrees of all nodes in c

Global Measure

- Total internal connections
- Total external connections

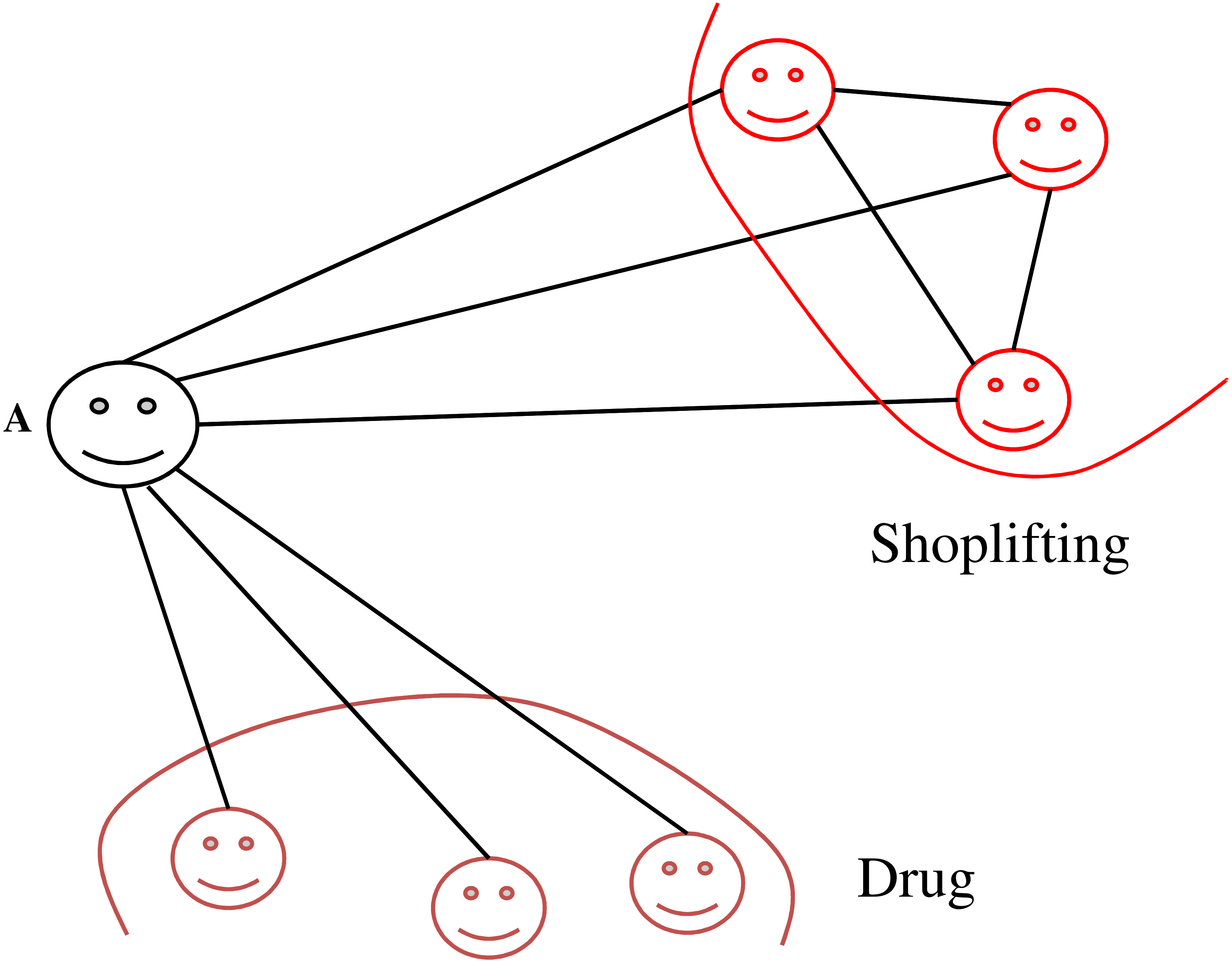
Our Perspective of a Community


A
Shoplifting
Non-addictive
Drug

Heuristic I

Total Internal connections > **maximum external connections to any one of the external communities**

Modularity, Conductance, Cut-ratio consider total external connections


A
Shoplifting
Drug

Heuristic II

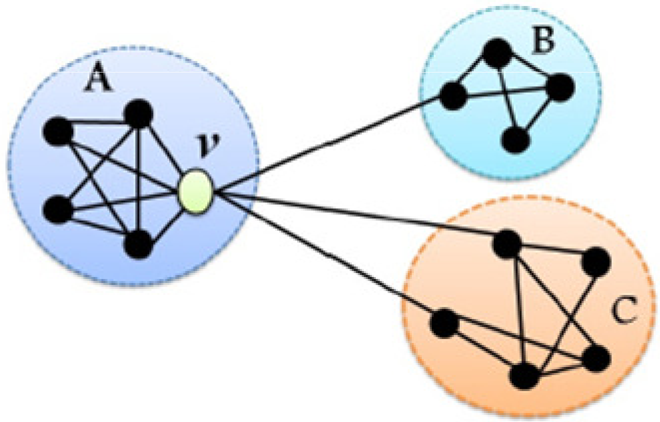
Internal neighbors should be highly connected => high clustering coefficient among internal neighbors

Modularity, conductance and cut-ratio do not consider clustering coefficient

Permanence

$$Perm(v) = \left[\frac{I(v)}{E_{\max}(v)} \times \frac{1}{D(v)}\right] - (1 - C_{in}(v))$$

$$\begin{cases} I(v) = \text{internal deg of } v \\ D(\text{v}) = \text{degree of } v \\ E_{max}(v) = \text{Max connection to an external neighbor} \\ C_{in}(v) = \text{clustering coefficient of internal neighbors} \end{cases}$$

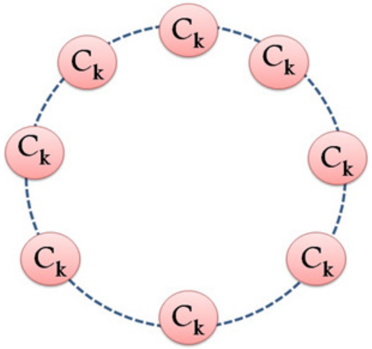


$Perm(v)$=0.12

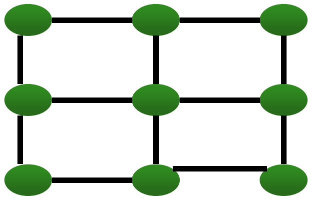
$I(v)=4$, $D(v)=7$, $E_{max}(v)=2$
$C_{in}(v)=5/6$

Discussion

Permanence



Permanence ~ 1



Permanence = 0

Wrong vertex-to-community assignment

Permanence ~ -1

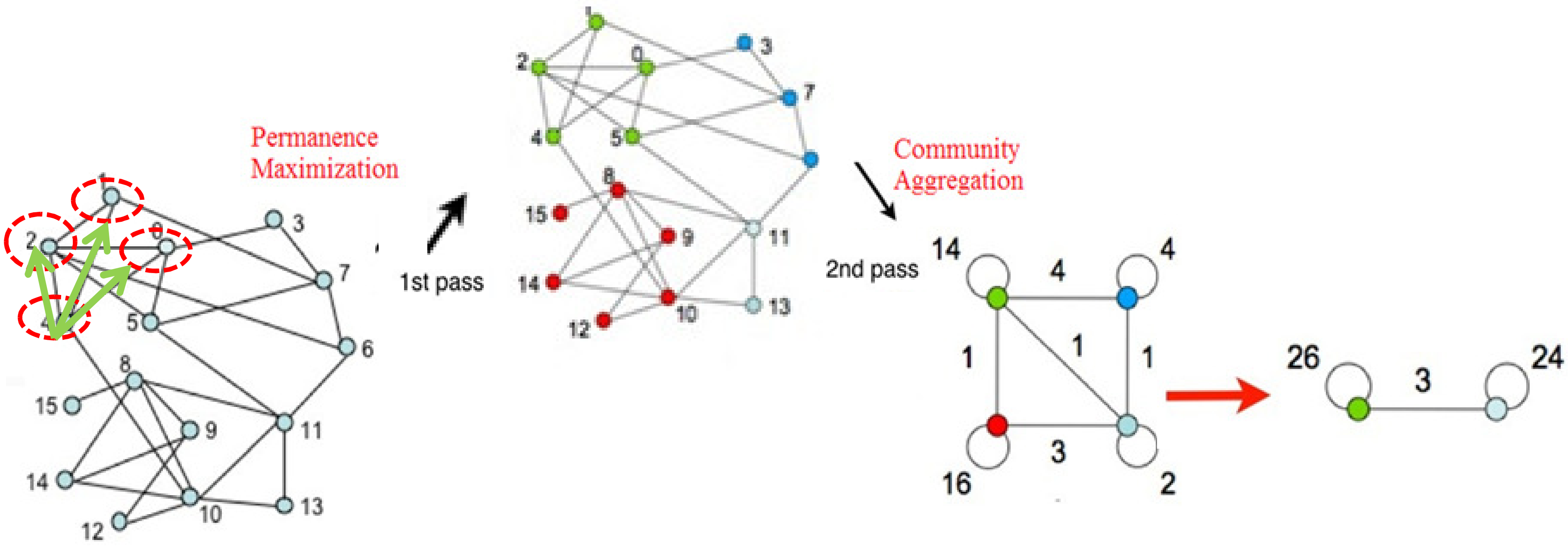
MaxPerm：
Non-overlapping Community Detection Algorithm

Major Limitations

- **Limitations of optimization algorithms**
  - Resolution limit *(Fortunato & Barthelemy, PNAS, 07)*
  - Degeneracy of solutions *(Good et al., PRE, 10)*
  - Asymptotic growth *(Good et al., PRE, 10)*

MaxPerm: Community Detection Based on Maximizing Permanence

- Follow similar strategy used in Louvain algorithm (a greedy modularity maximization) *(Blondel et al., J. Stat. Mech, 07)*
- We only consider those communities having size >=3



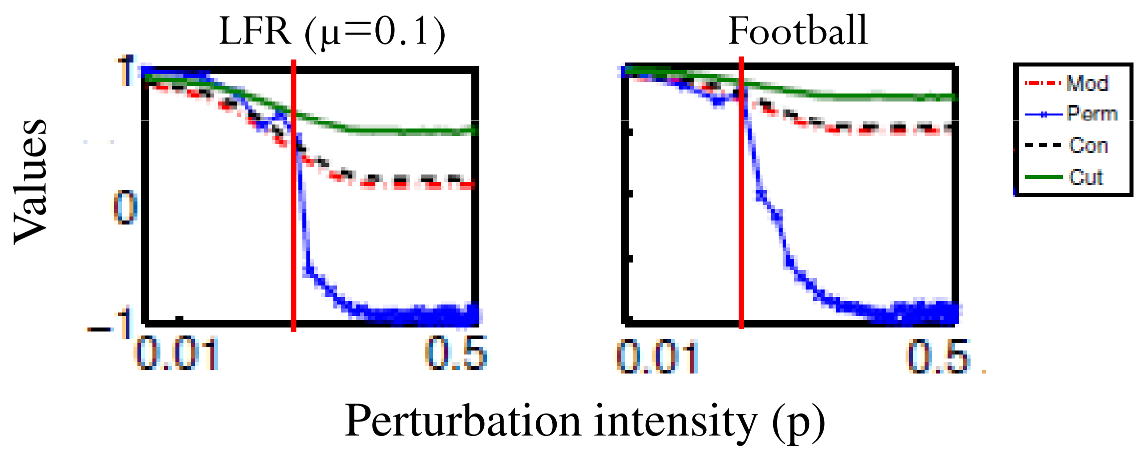
Experimental Results

Algo	LFR (μ=0.1)	LFR (μ=0.3)	Football	Railway	Coauthorship
Louvain	0.02	0.00	0.02	0.14	0.00
FastGrdy	0.00	$\boxed{0.87}$	0.01	$\boxed{0.37}$	0.14
CNM	0.14	$\boxed{0.40}$	0.30	0.00	0.05
WalkTrap	0.00	0.00	0.02	0.02	0.01
Infomod	0.06	0.08	0.19	0.04	0.00
Infomap	0.00	0.00	0.02	0.02	0.03

Table: Improvement of our algorithm w.r.t the other algorithms (averaged over all validation measures)

More about Permanence

- **Permanence** is **not very sensitive** to **minor perturbation**, but **very sensitive** after a **certain threshold**



- **Permanence finds small-size communities**
  - Identify singleton (act as junction in Railway n/w) and small communities (subfields in Coauthorship n/w)

Theoretical Issues

Resolution limit

If a vertex is **very tightly connected** to a community and **very loosely connected** to another community, highest permanence is obtained when it joins the community to which it is more connected.

Degeneracy of solution

if a vertex is **sufficiently loosely connected** to its neighbouring communities and has equal number of connections to each community, then in most cases it will remain as **singleton**, rather than arbitrarily joining any of its neighbour groups.

Asymptotic growth of value

All the parameters of parameters are **independen**t of the **symmetric growth** of network size and the number of communities.

Analytical proofs: http://cnerg.org/permanence

Metric for Overlapping Communities

Overlapping Permanence (OPerm)

$$P^c_{ov}(v) = \frac{I^c(v)}{E_{max}(v)} \times \frac{1}{D(v)} - (1 - c^c_{in}(v)) \cdot \frac{I^c(v)}{I(v)}$$

Generalized metric

D(v)=degree of v

$E_{max}(v)$=Max connection to an external community

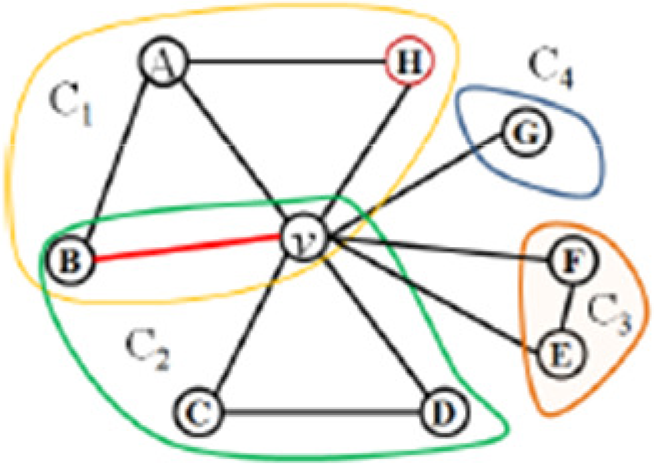
$C^c in(v)$=clustering coeff. of internal neighbors of v in c

$I(v)$ = # of internal neighbors of v

$$I^c(v) = \sum_{e \in \Gamma_v^e} 1/x_e$$

Γ_v^e = internal edges of v in community

χ_e = # of communities edge e shares



D(v)=8, I(v)=5

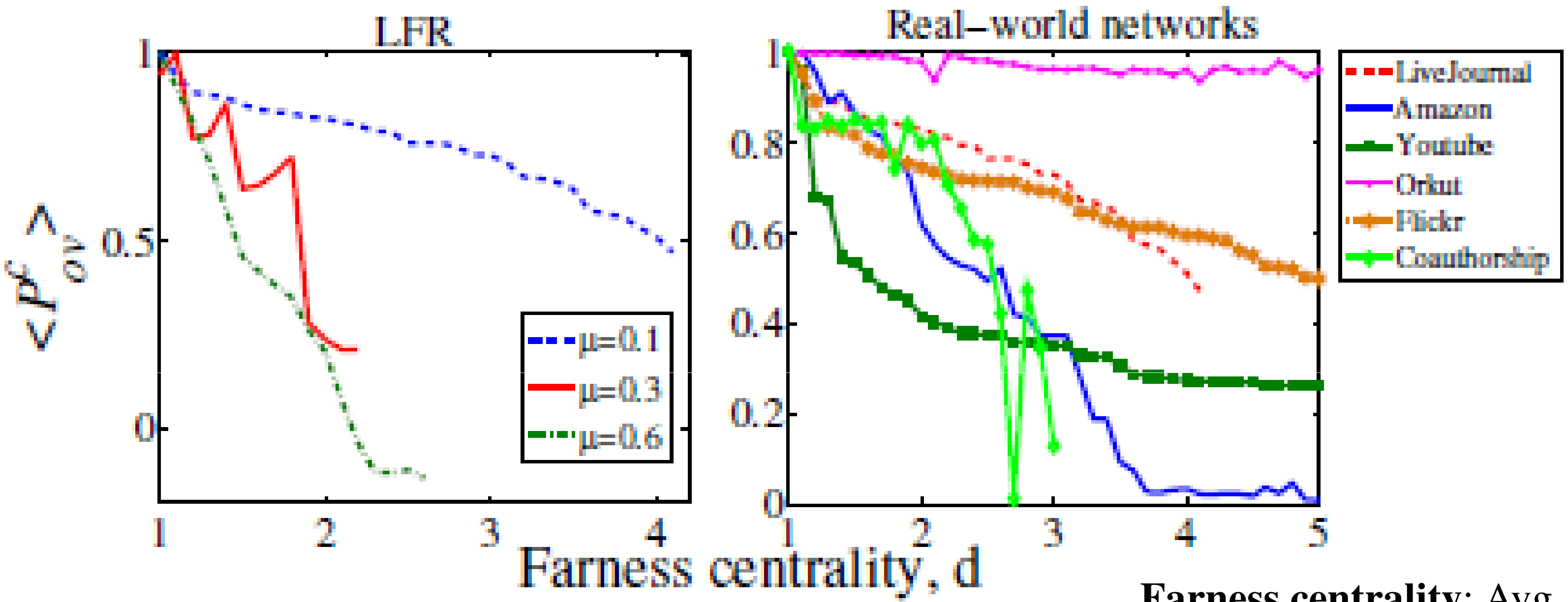
$$P^{c1}_{ov}(v) = \frac{1+1+1/2}{2\times 8} - (1-2/3)\times\frac{1+1+1/2}{5} = -0.01$$

$P^{c2}_{ov}(v) = -0.18$ $\quad P_{ov}(v) = P^{c1}_{ov}(v) + P^{c2}_{ov}(v) = -0.19$

$$P_{ov} = \frac{1}{|V|}\sum_{v\in V} P_{ov}(v)$$

Inference from OPerm Values

Core-periphery Structure within Communities

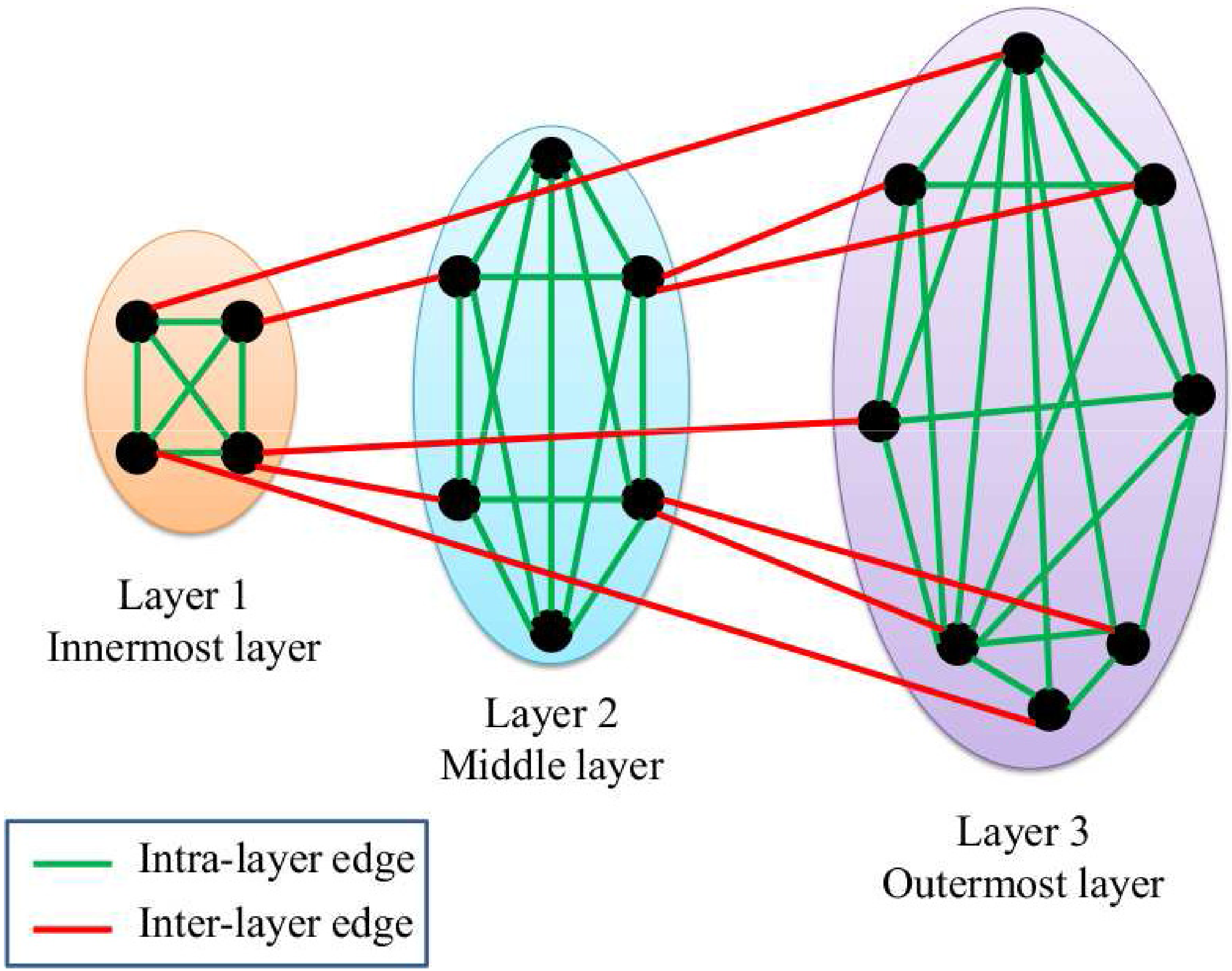


Farness centrality: Avg. shortest path of each vertex within a community.

Assortativity

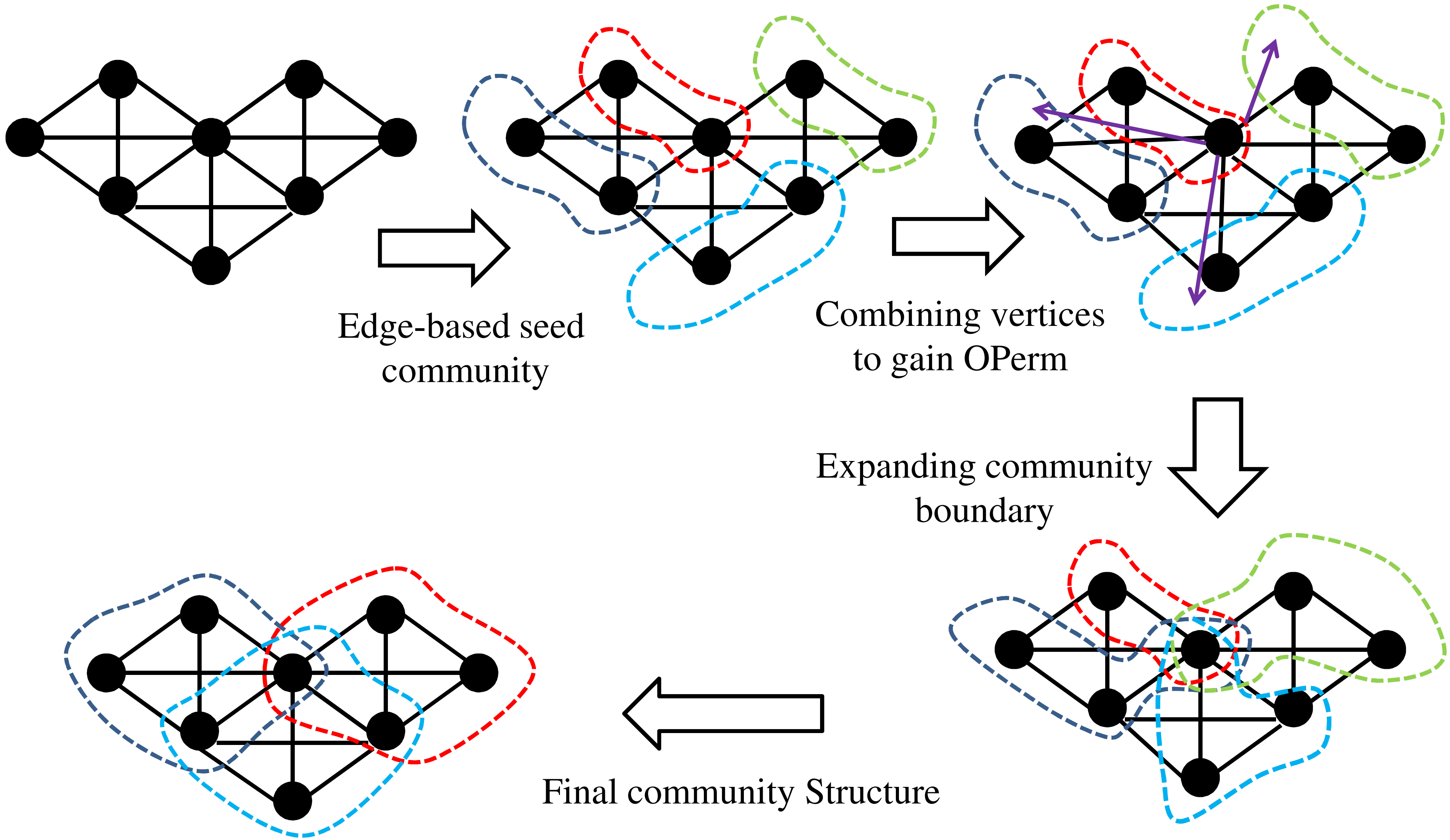
Assortativity	LFR (0.1)	LiveJournal
Degree-based	-0.045	0.037
OPerm-based	0.645	0.465

Layers within a Community



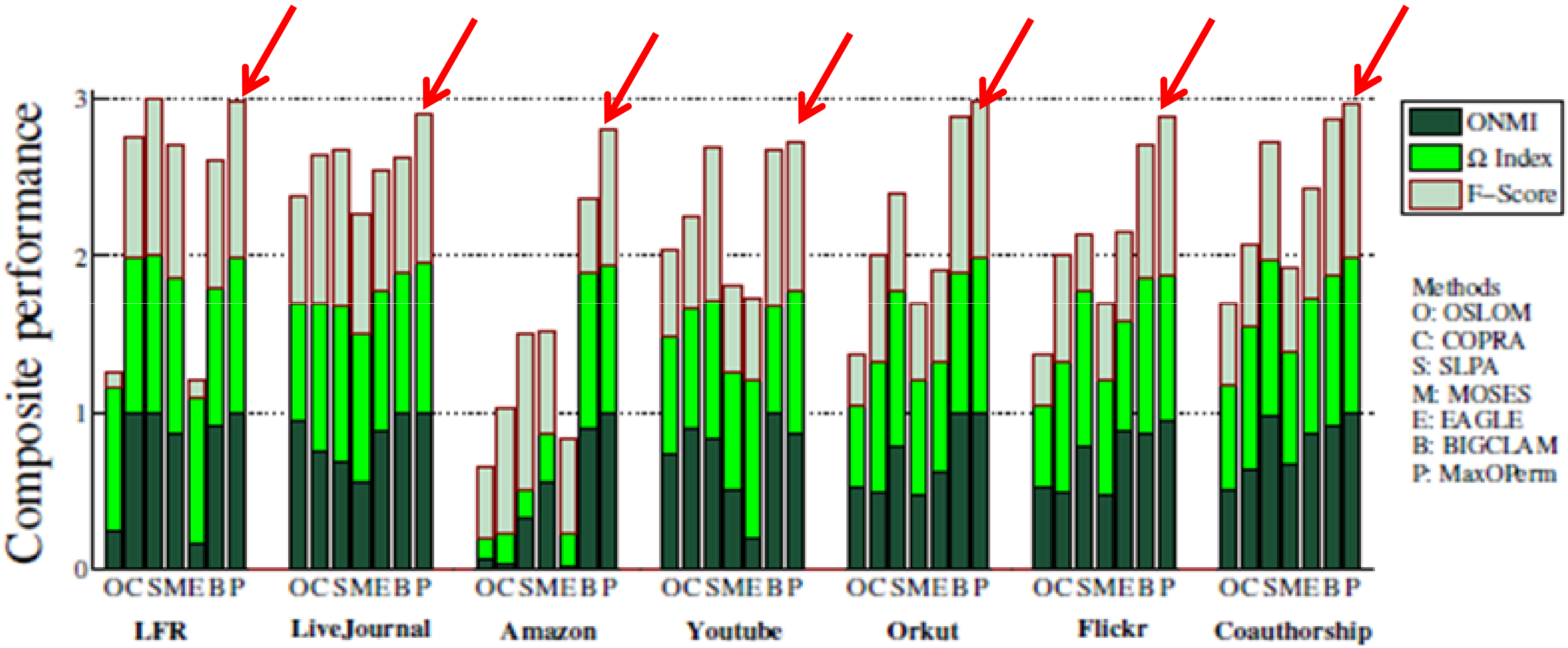
MaxOPerm：
Overlapping Community Detection Algorithm

MaxOPerm: Framework



Experiment Results

(Evaluation with Ground-truth Communities)



Chapter 3:

Analyzing Real-world Communities

Publication Dataset

- Crawled entire Microsoft Academic Search
- Papers in Computer Science domain
- Basic preprocessing

Basic Statistics of papers from 1960-2010	Values
Number of valid entries	3,473,171
Number of authors	1,186,412
Number of unique venues	6,143
Avg. number of papers per author	5.18
Avg. number of authors per paper	2.49

Publication Dataset

Available Metadata
Title
Unique ID
Named entity disambiguated authors' name
Year of publication
Named entity disambiguated publication venue
Related research field(s)
References
Keywords
Abstract

Available @ http://cnerg.org

Ground-truth Communities

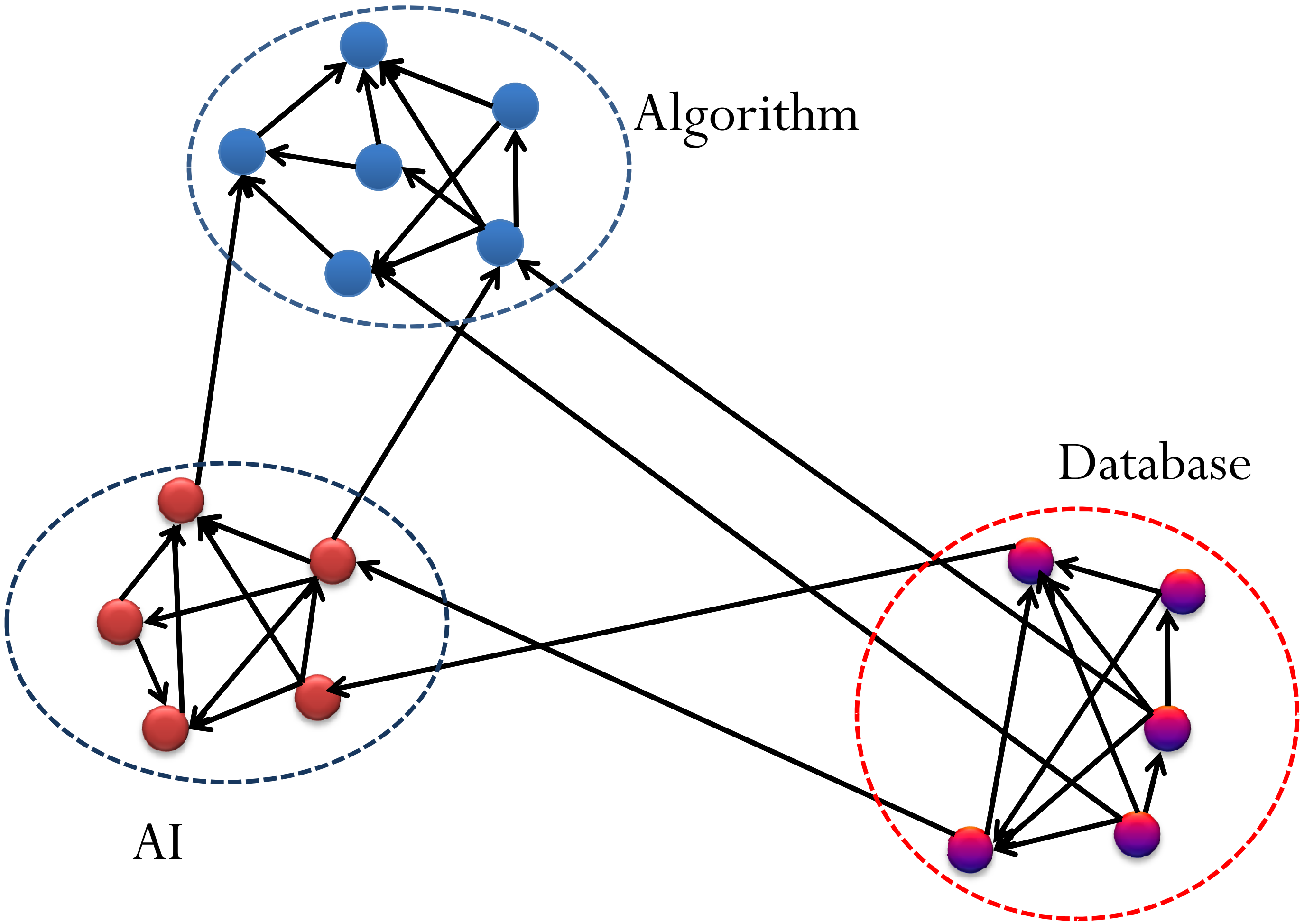


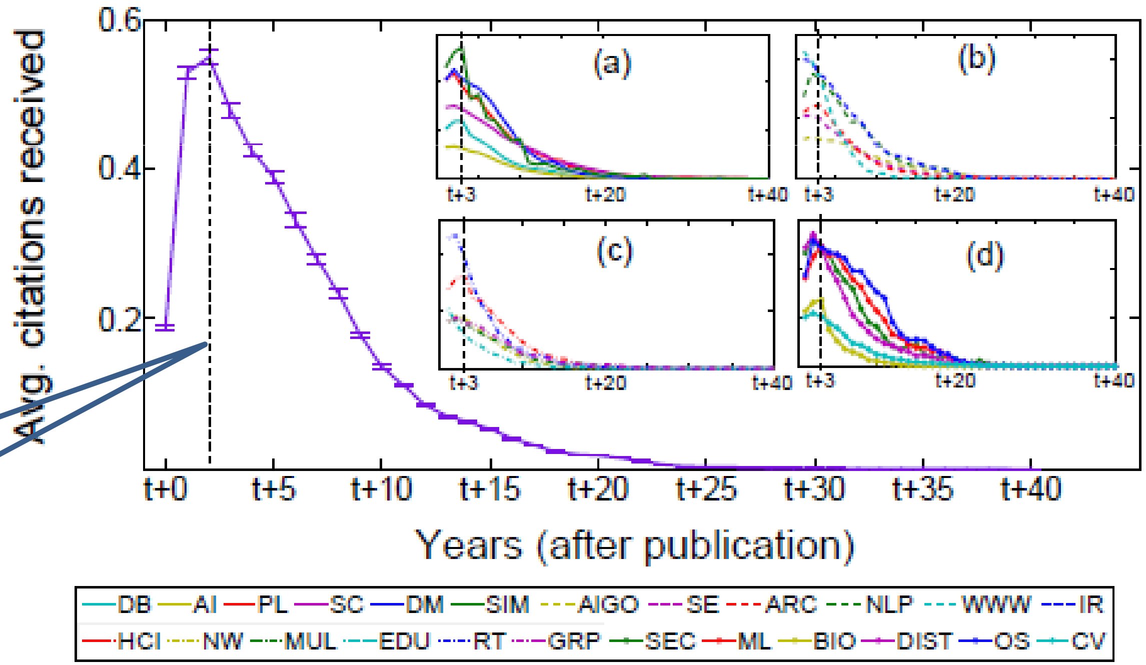
Fig.: Citation network with ground-truth communities

"Impact" of a Field (Community)

- Measuring the **impact of each field** (its constituent papers) around a particular year.
- **Local citation density** is important

(Guns & Rousseau, J. info, 09)
(Jin et al. Chin. Sci. Bull., 07)

Peaks within 3 years from publication, then declines



Average Inward Citations

"Impact" of a Field (Community)

Inwardness of a field f_i at time t

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

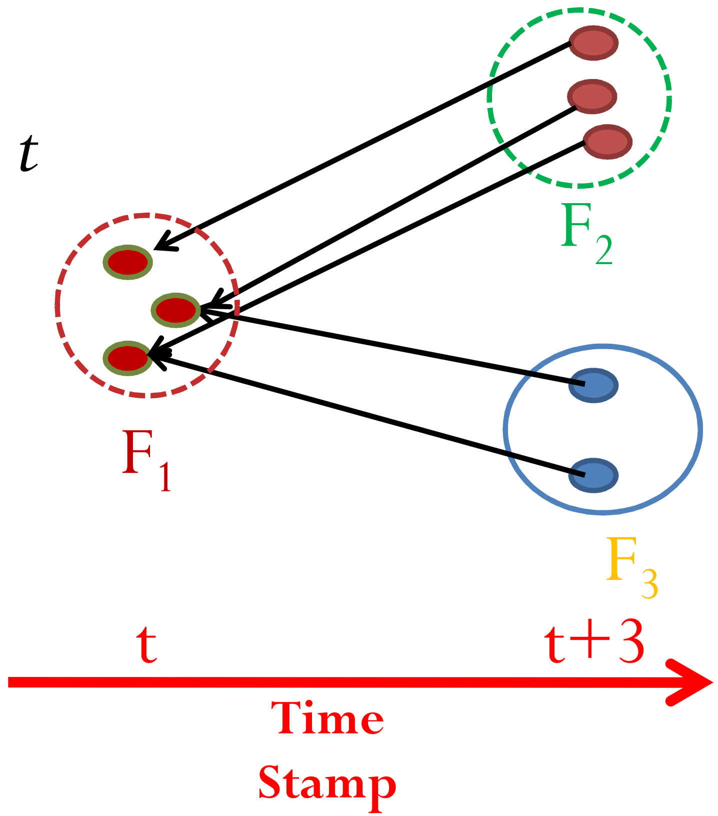
where,

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

$C_{j \to i}^t$ = # 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)

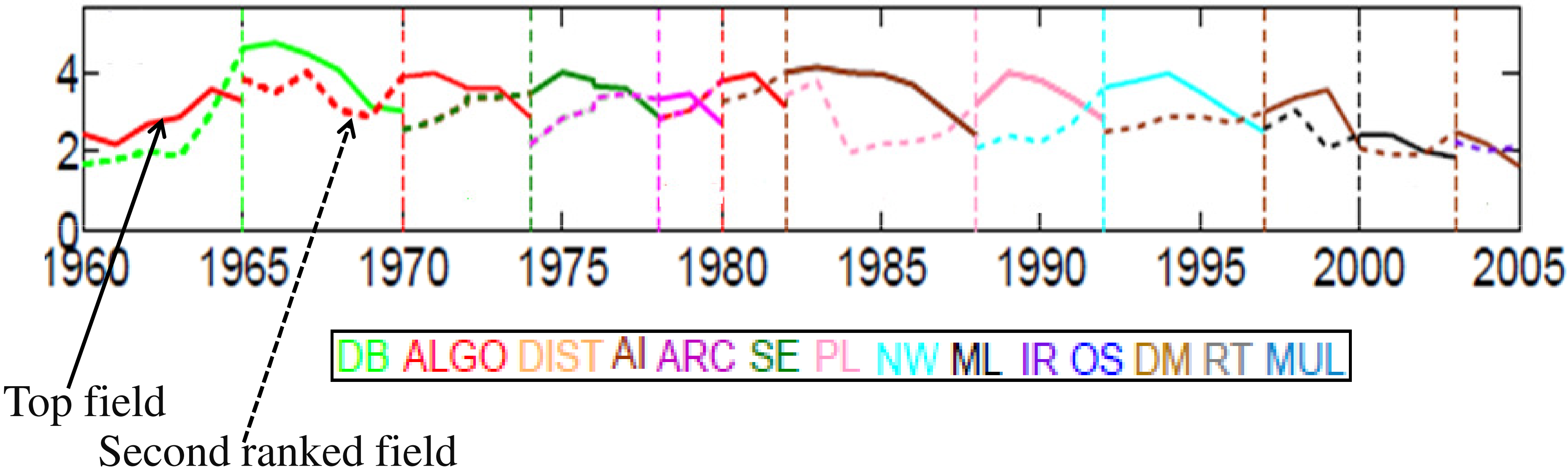


In(F₁ᵗ)=5/3

N.B.: 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

Cause Analysis

- Impact of highly-cited papers
- Impact of collaboration
- Impact of top back-up fields
- Effect of seminal papers

External Evaluation of Inwardness:

Our results have high correlation with the project submission statistics of NSF

Effect of Interdisciplinary Research

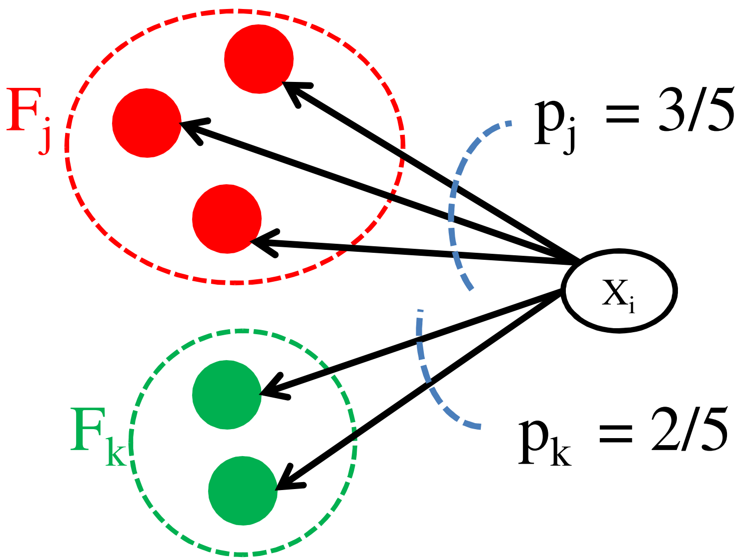
How to measure the degree of interdisciplinarity of a field?

Reference Diversity index (RDI)

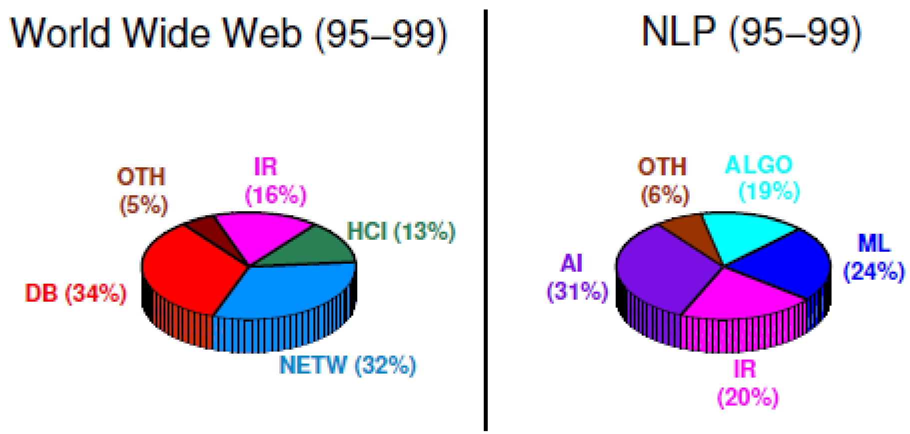
RDI of a paper X_i = $RDI(X_i) = -\sum_j p_j \log p_j$

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

More RDI, more interdisciplinarity



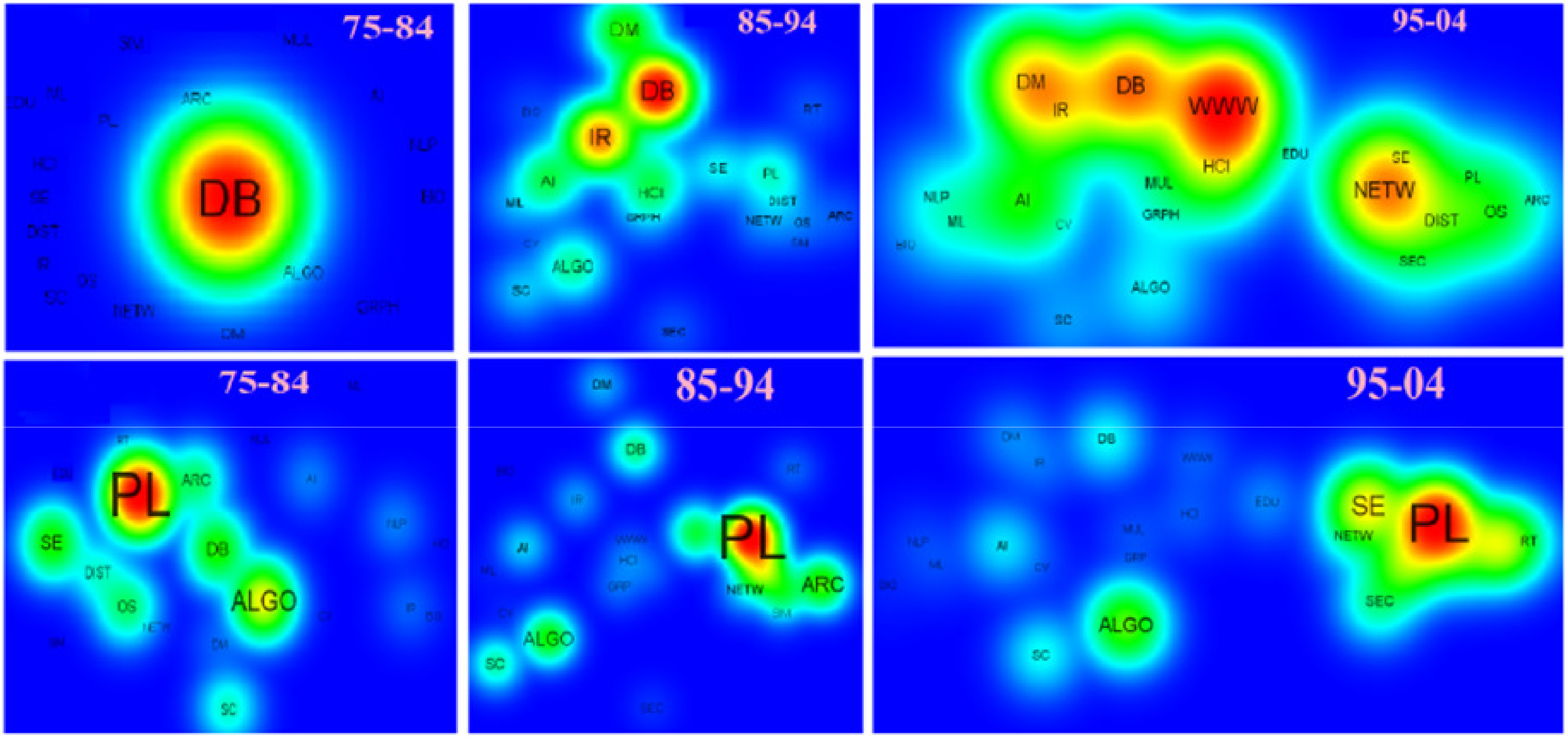
$RDI(X_i)$ = - 3/5 log (3/5) – 2/5 log (2/5)
= 0.67



Other Indicators

- Citation Diversity Index (CDI)
  - Citation based measure
- Membership Diversity Index (MDI)
  - Community based measure
- Attraction Index
  - Propensity of new researchers joining to a field

Evolutionary Landscape

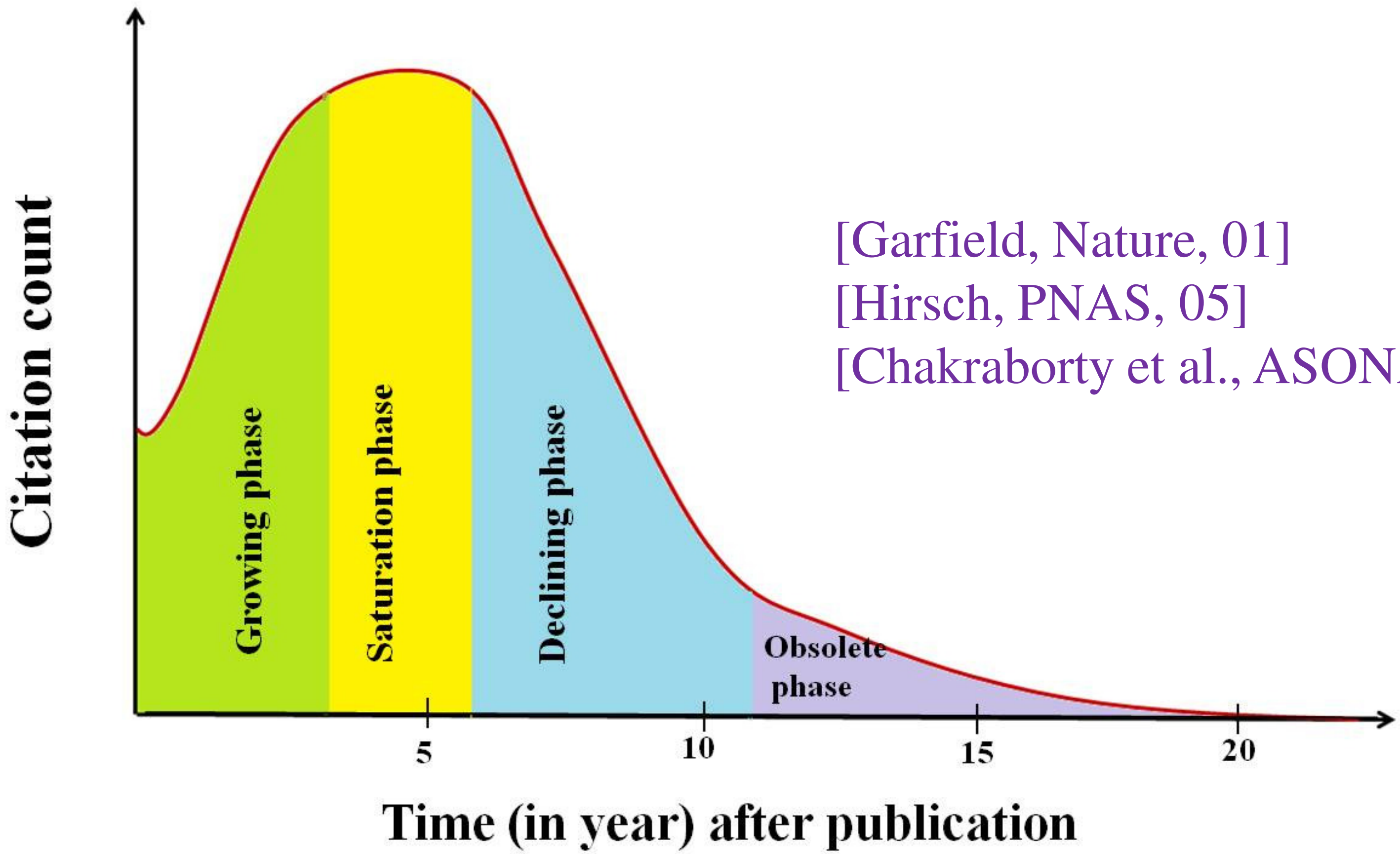


- Fields are grouped based on the **connection proximity**
- The **size of the font** indicates the **relative importance** (# of incoming citations) of a field

Chapter 4: Community-based Applications

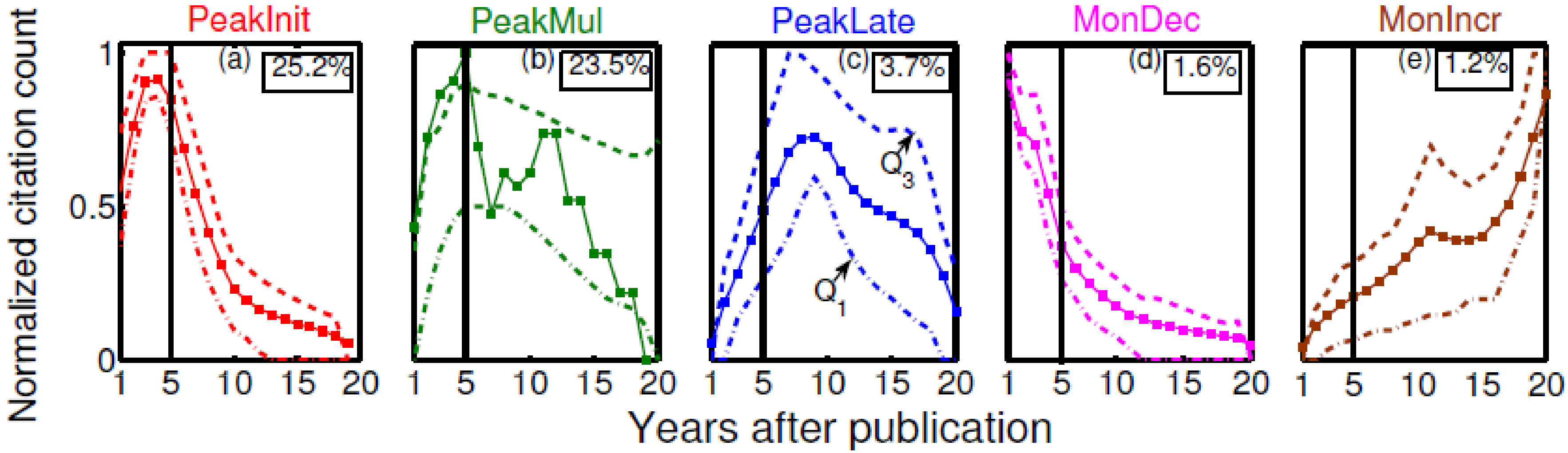
Citation Profile of an Article

Common consensus about the growth of citation count of a paper over time after publication



[Garfield, Nature, 01]
[Hirsch, PNAS, 05]
[Chakraborty et al., ASONAM, 13]

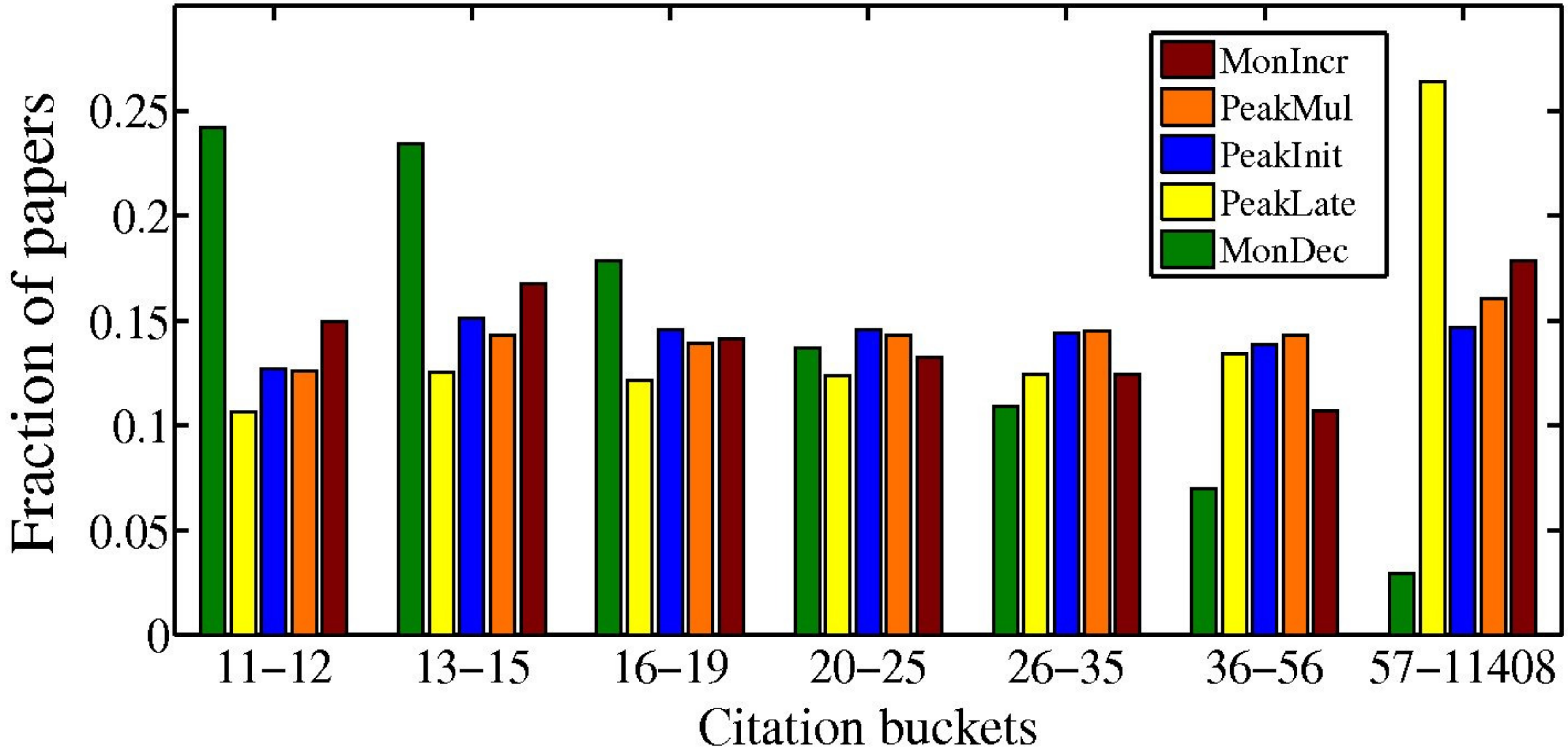
Six Universal Citation Profiles



Q1 and Q3 represent the first and third quartiles of the data points respectively.

Another category: **'Oth'** => having less than one citation (on avg) per year

More on the Categories



Contribution of papers from each category in different citation buckets

Application:
Future Citation Count Prediction

Problem Definition

Citation counts:
Given the set of scientific articles D, the citation counts $(C_T(.))$ of an article $d \in D$ is defined as:

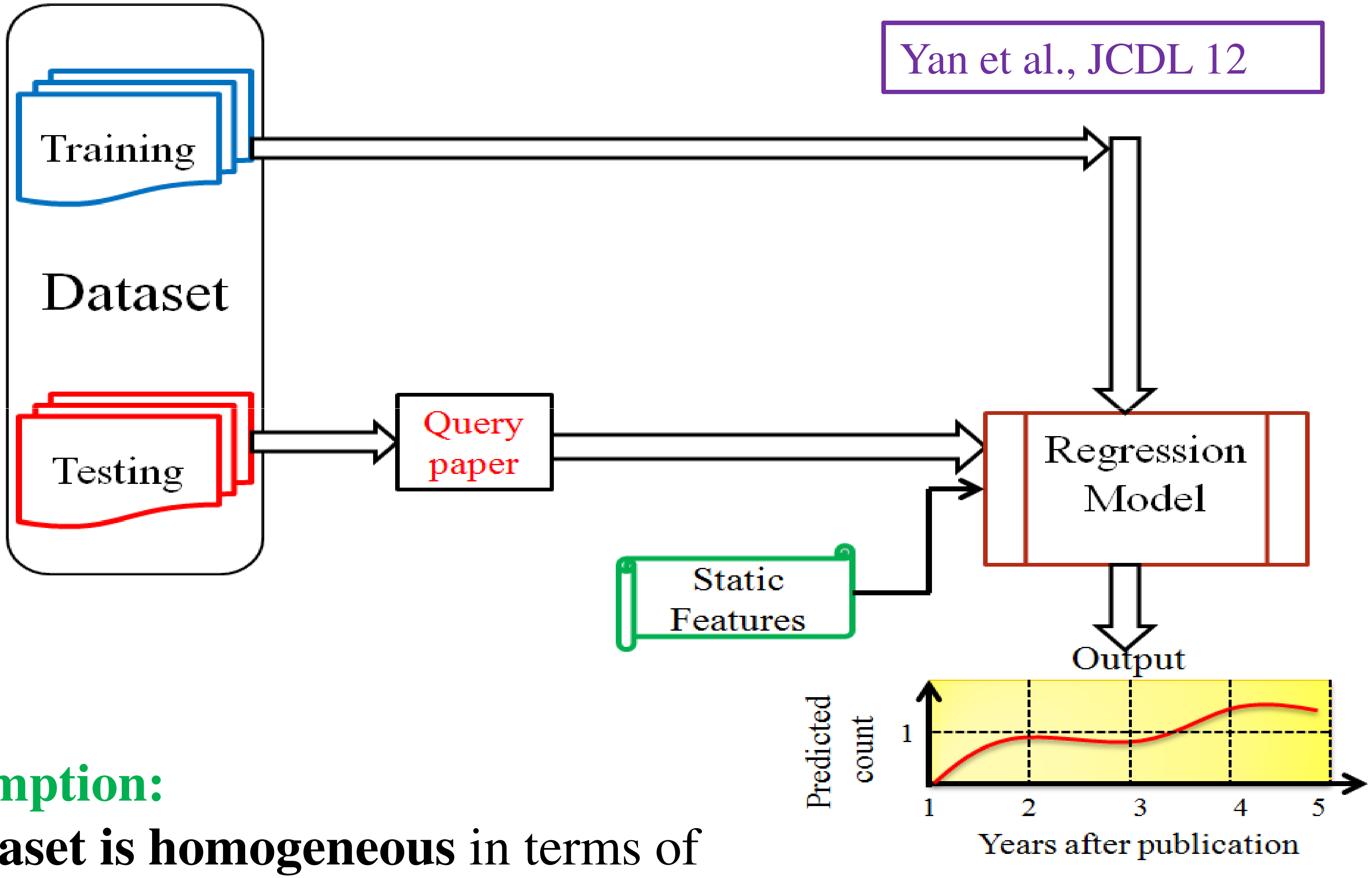
$$citing(d) = \{d' \in D : d' \ cites \ d\}$$
$$C_T(d) = |citing(d)|$$

Learning task: Given a set of features $F = \{f_1, f_2, ..., f_n\}$, our goal is to learn a predictive function ψ to predict the citation counts of an article d after a give time period Δt of its publication. Formally, this can be written as:

$$\psi(d|F, \Delta t) \rightarrow C_T(d|\Delta t)$$

we consider $\Delta t \in \{1, 5\}$

Traditional Framework

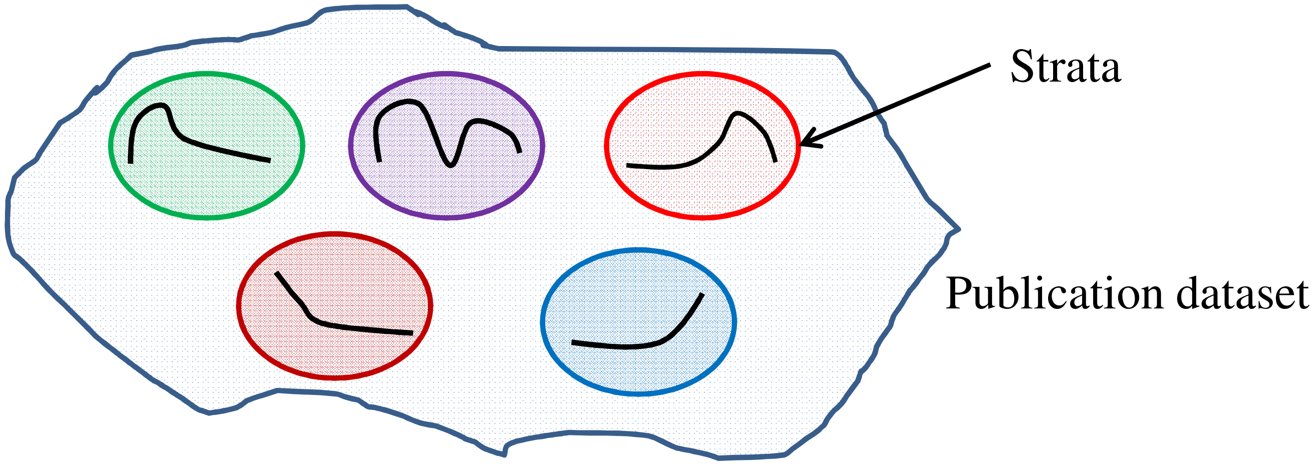


Assumption:

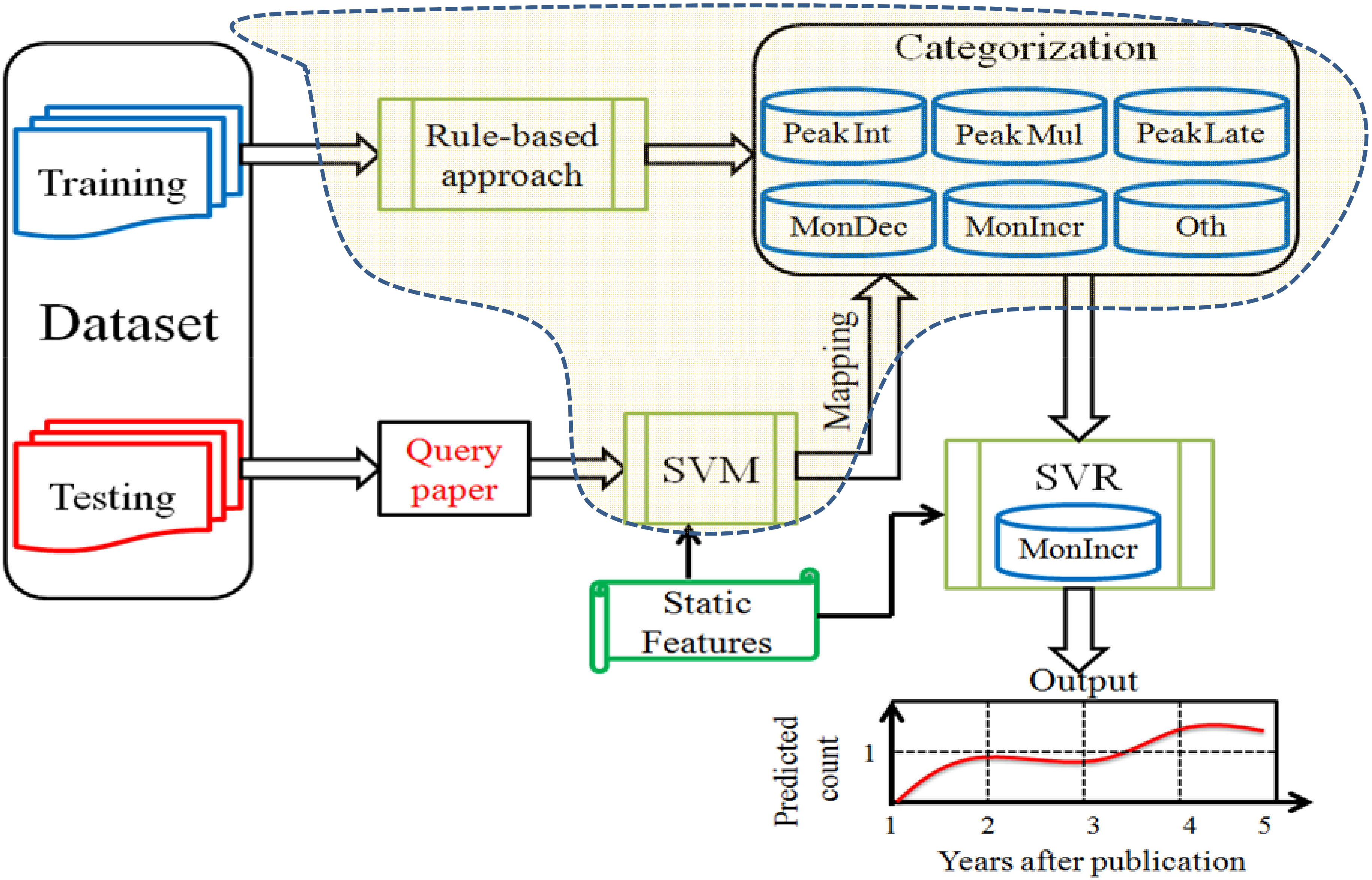
Dataset is homogeneous in terms of citation profile

Stratified Learning

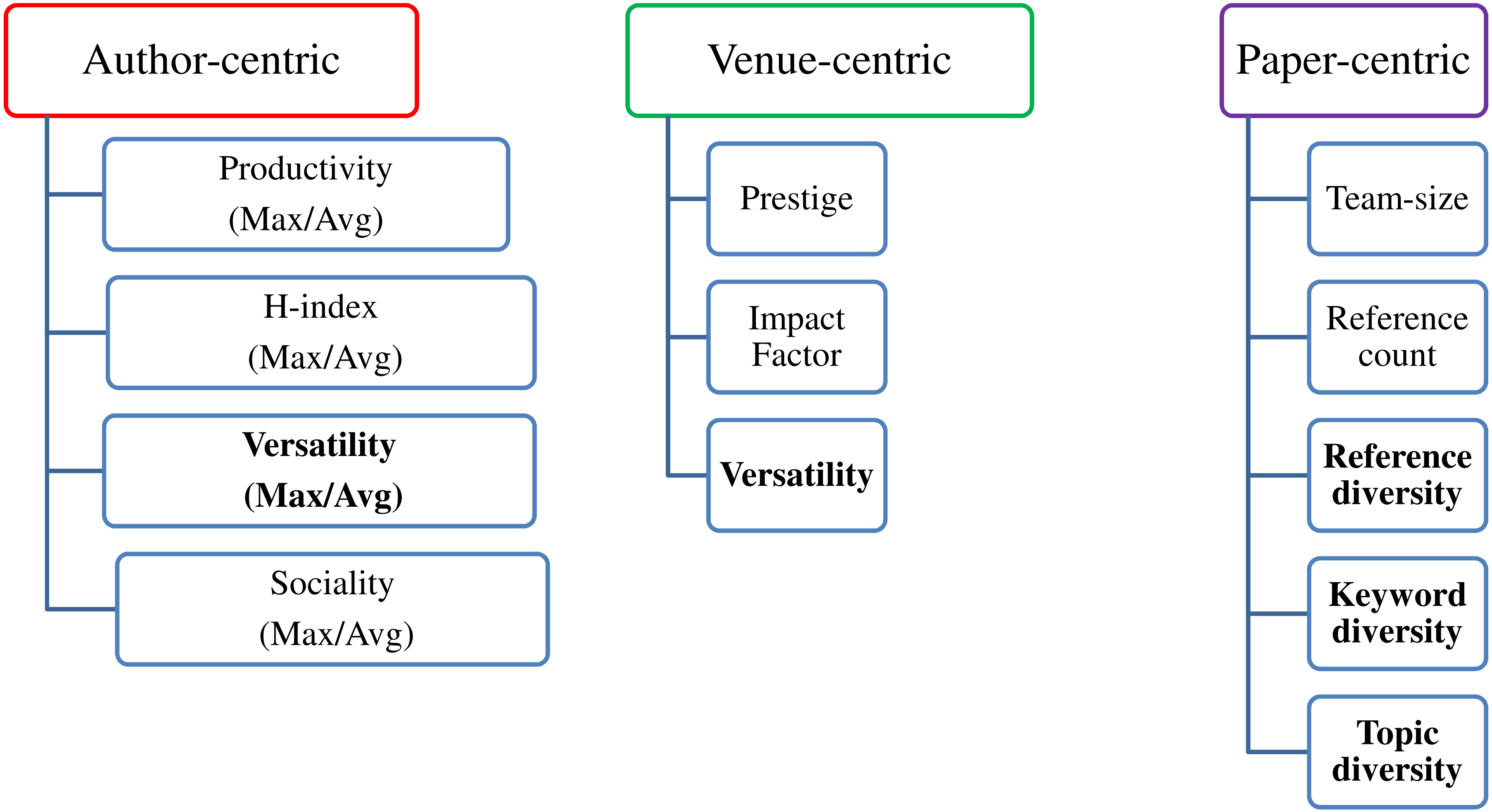
- **Stratification** is the process of dividing members of the population into **homogeneous subgroups** before sampling.
- The **strata** should be mutually exclusive
  - Every element in the population must be assigned to only one stratum



Our Framework: 2-stage Model



Static Features



Performance Evaluation

(i) Coefficient of determination (R^2)

The more, the better

(ii) Mean squared error (θ)

The less, the better

(iii) Pearson correlation coefficient (ρ)

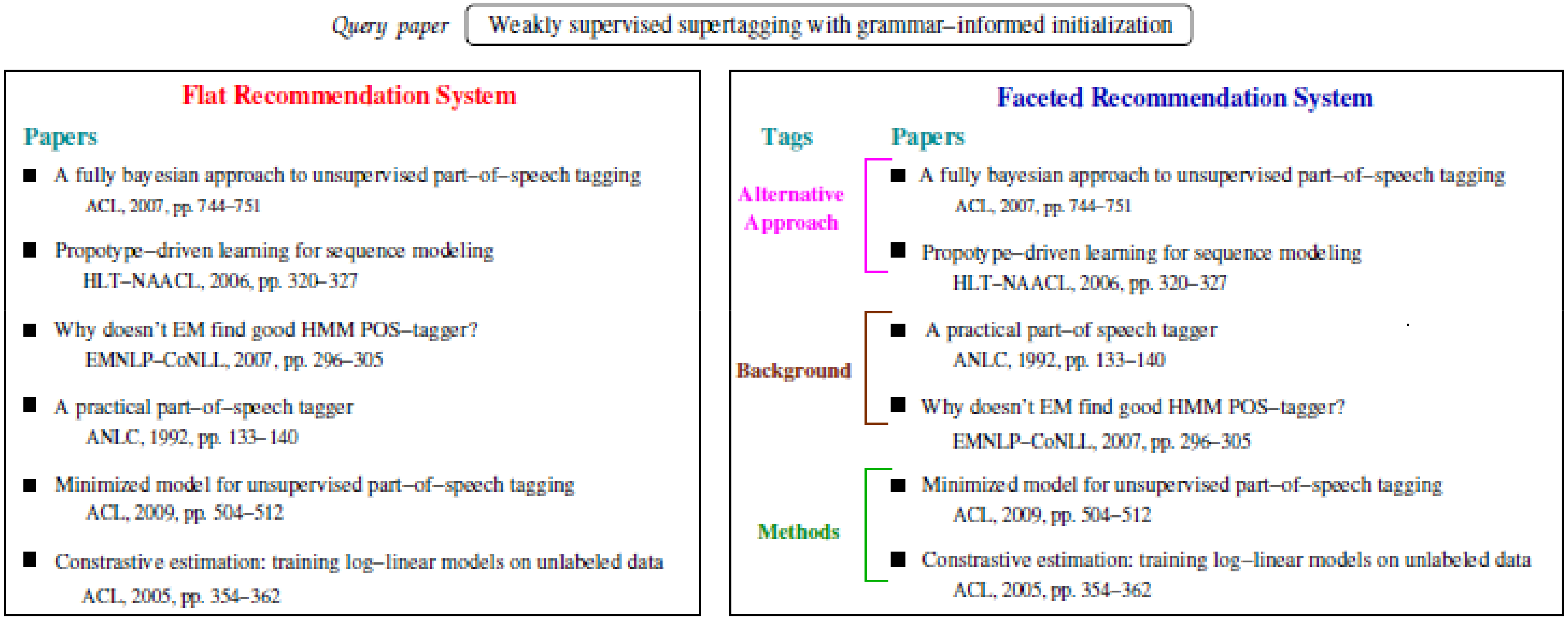
The more, the better

Performance Evaluation

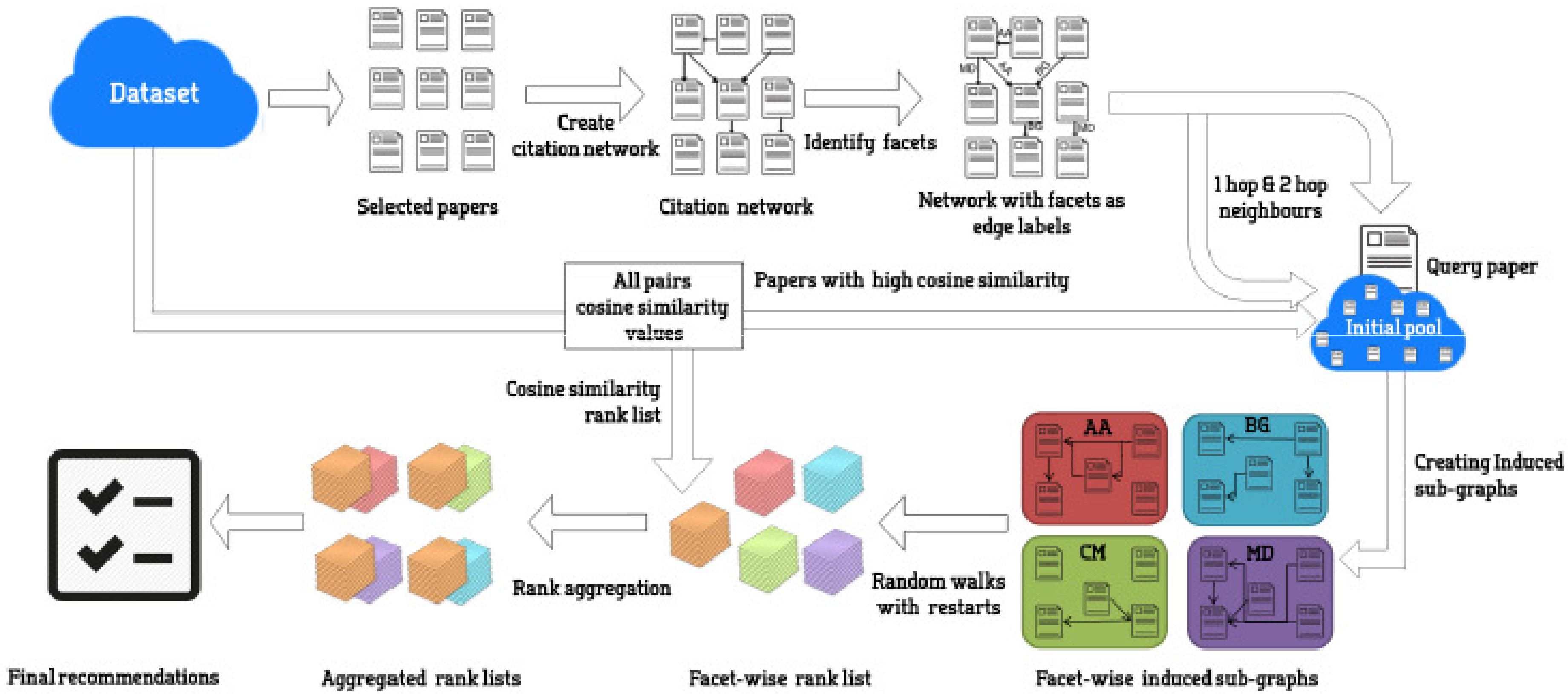
	Baseline		
	R^2	θ	ρ
Δt=1	0.57	5.06	0.61
Δt=2	0.55	7.10	0.59
Δt=3	0.52	8.78	0.65
Δt=4	0.50	10.06	0.75
Δt=5	0.45	13.06	0.42

Application:
Faceted Recommendation System for Scientific Articles

Flat vs. Faceted Recommendation



FeRoSA: Workflow Diagram



Four facets:

Background, Alternative Approach, Methods, Comparison

Experimental Setup

Baseline: Flat Recommendation Systems:

- Google Scholar (GS), Microsoft Academic Search (MAS) and LLQ (Liang et al., 11)

Baseline: Faceted Recommendation Systems:

- VanillaPR and FeRoSA-CS

Ground-truth Generation:

- Number of query papers = 30 (30 recommendations per query)
- Number of experts in NLP = 8

Metrics:

- Overall Precision (OP)
- Overall Impression (OI)
- Faceted Evaluation: Faceted Precision (TP)

Faceted Evaluation based on Ground-truth

Facets	**VanillaPR**	**FeRoSA-CS**	**FeRoSA**
BG	0.65	0.51	**0.79**
AA	0.48	0.34	**0.56**
MD	**0.62**	0.39	**0.62**
CM	0.44	0.38	**0.62**
Average	0.55	0.40	**0.65**

Evaluation by the Original Authors

FeRoSA: Evaluation by the Authors

Hi XXX,

For your paper, "**Topic Segmentation with Hybrid Document Indexing**", would you recommend the following as:

	R & F	R & Not F	Not R & Not F
1. Background paper (BG) Lattice Minimum Bayes-Risk Decoding for Statistical Machine Translation	●	○	○
2. Alternative Approach (AA) Structural and Topical Dimensions in Multi-Task Patent Translation	○	●	
3. Comparison (CM) Sequential Labeling with Latent Variables: An Exact Inference Algorithm and its Efficient Approximation	●	○	
4. Method (MD) Feature-Rich Translation by Quasi-Synchronous Lattice Parsing	○	○	

R: Recommendation is relevant; F: Facet is correct

You may find more recommendations for your paper @ http://www.ferosa.org/beta/D07-1037.html

Remarks: The system made a few mistakes but overall I think it is a good approach.

- **12 authors responded**
- **75% cases, the recommendation is marked as relevant**
- **BG:0.49, AA: 0.42 MD: 0.52, CM: 0.59**

Flat Evaluation

(a)

Systems	OI@3	OP@3
GS	0.27	0.61
MAS	0.17	0.45
LLQ	0.13	0.41
f-FeRoSA	**0.43**	**0.79**

(b)

OP	f-FeRoSA
OP@3	0.79
OP@5	0.78
OP@10	0.71

Table : (a) Flat evaluation of the competing systems; (b) overall precision of f-FeRoSA at different number of recommendations.

www.ferosa.org

Conclusions
&
Future Work

Takeaways

- Community: mesoscopic view of a network
- Constant community captures the invariant substructure of a network
- Permanence measures the belongingness of a node in a community
- Real-world community in citation network reveals scientific paradigms of Computer Science domain
- Applications such as search and recommendation systems perform significantly well

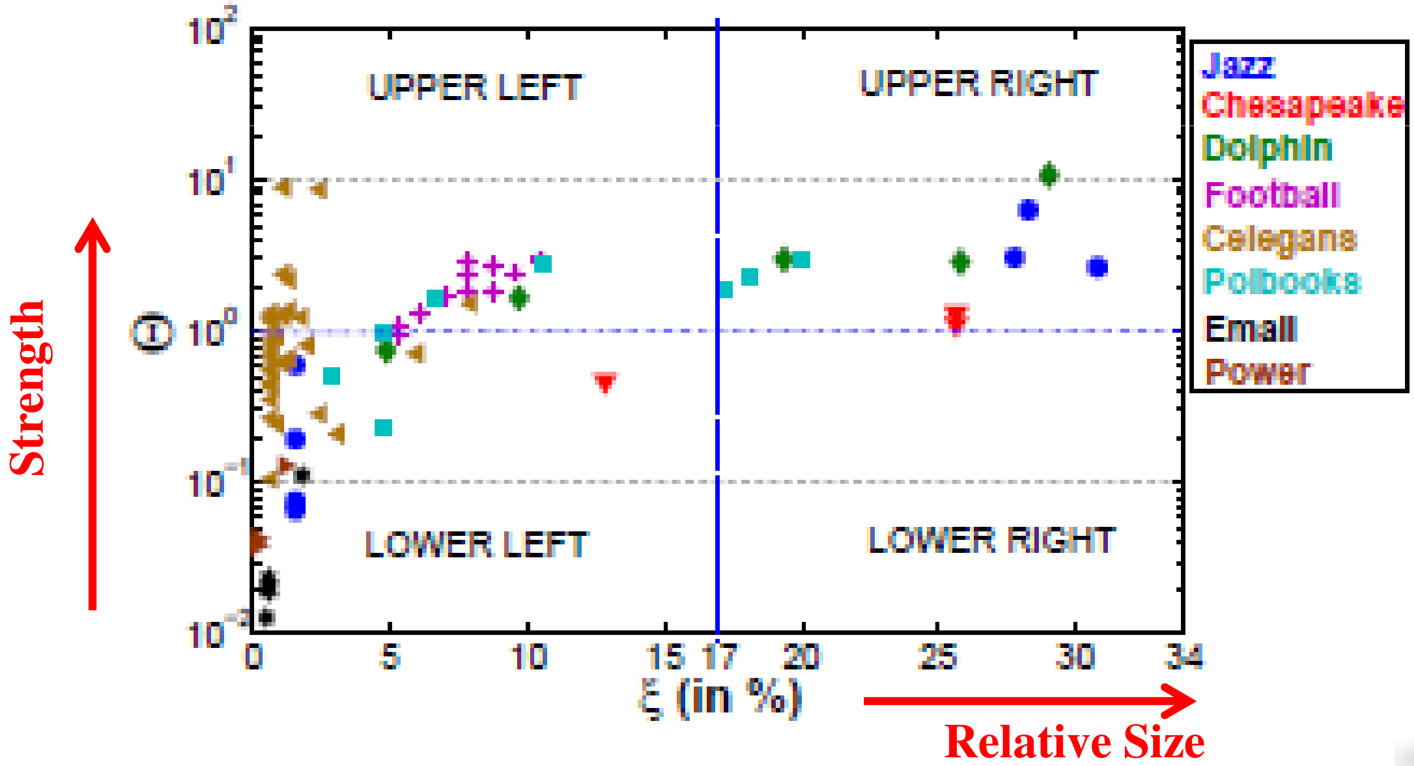
Future Work

- Local and dynamic community detection
- Explore more on the core-periphery structure within a community
- Citation categorization of individual authors
- Evaluate the real systems on larger datasets to show the Scalability and robustness

Discussion on Feedback

Prof. Frank Schweitzer

❑ “Networks containing a large number of such constant communities are less likely to be affected by perturbation.” Explain to what kind of perturbation this statement should apply.



Discussion on Feedback

Prof. Frank Schweitzer

- "We aggregate these two criteria to formulate permanence of a vertex" (4.1) is certainly only one of different ways to include the given heuristics I and II. Discuss alternatives, in order to provide more evidence for your specific choice.

  Permanence

- Discuss the relation between the relative permanence and permanence of a node. What are the conceptual differences, what is the additional information provided in each of these measures?

  Relative Permanence

Discussion on Feedback

Prof. Frank Schweitzer

- Explain the meaning of a power law, and distinguish it from other types of distributions (stretched exponential, log-normal, beta etc.) . Explain methods to test distributions in general, and methods to verify the power law behavior in data, specifically.

- Power law: $f(x) = a.x^{-k}$, $2 \le k \le 3$
- Stretched exponential: $f_k(x) = e^{-x^k}$, $0 \le k \le 1$
- Log-normal: If x is log normal, $y=ln(x)$ follows a normal dist

Discussion on Feedback

Prof. Y. Narahari

Thank you very much for such inspiring comments.
I would be happy to take up any queries.

Publications from the Thesis

Journals

- Chakraborty et al., On the Categorization of Scientific Citation Profiles in Computer Sciences, ***Communications of the ACM*** (***CACM***), 58:9, pp. 82-90, 2015. (IF: 2.836)

- Chakraborty et al. Understanding and Modeling Diverse Scientific Careers of Researchers, ***Journal of Informetrics***, 9:1, ISSN 1751-1577, pp. 69-78, Jan 2015. (IF: 3.609)

- Chakraborty et al. Constant Communities in Complex Networks**,** ***Nature Scientific Reports***, 3, 1825, ISSN 2045-2322, 2013. (IF: 5.078)

- Chakraborty et al., Computer Science Fields: A Quantitative Route to the Rise and Fall of scientific Research, ***Social Network Analysis and Mining*** (***SNAM***), 4:1, Springer Vienna, ISSN 1869-5450, pp. 1-18, 2014.

Publications from the Thesis

Conferences

- Chakraborty et al., On the formation circles in co-authorship networks, ***ACM SIGKDD***, Sydney, August 10 - 13, 2015, pp. 109-118.
- Chakraborty et al., DiSCern: A Diversified Citation Recommendation System for Scientific Queries, ***ICDE***, Seoul, Korea, April 3-17, 2015, pp. 555-566.
- Chakraborty et al., Towards a Stratified Learning Approach to Predict Future Citation Counts, ***JCDL***, London, Sep 8-12, 2014, pp. 351-360.
- Chakraborty et al., On the Permanence of Vertices in Network Communities, ***ACM SIGKDD,*** New York City, Aug 24 - 27, 2014, pp. 1396-1405.
- Chakraborty et al., Rising Popularity of Interdisciplinary Research - an Analysis of Citation Networks, Workshop on Social Networks, **COMSNETS**, Bangalore, 2014.
- Chakraborty et al., Automatic Classification and Analysis of Interdisciplinary Fields in Computer Sciences, ***SocialCom***, Washington D.C., Sep 8- 14, 2013, pp. 180 – 187.
- Chakraborty et al., Computer Science Fields as Ground-truth Communities: Their Impact, Rise and Fall, ***ASONAM***, Canada, Aug 25-28, 2013, pp. 426-433.

Manuscripts under Review

- Chakraborty et al, Overlapping Permanence: A New Vertex-based Metric to Analyze Overlapping Communities, ***IEEE TKDE.***

- Chakraborty et al., FeRoSA: A Faceted Recommendation System for Scientific Articles, ***ACM TIST*** (*to be submitted*).

- Chakraborty et al., Permanence and Community Analysis in Complex Networks, ***ACM TKDD***.

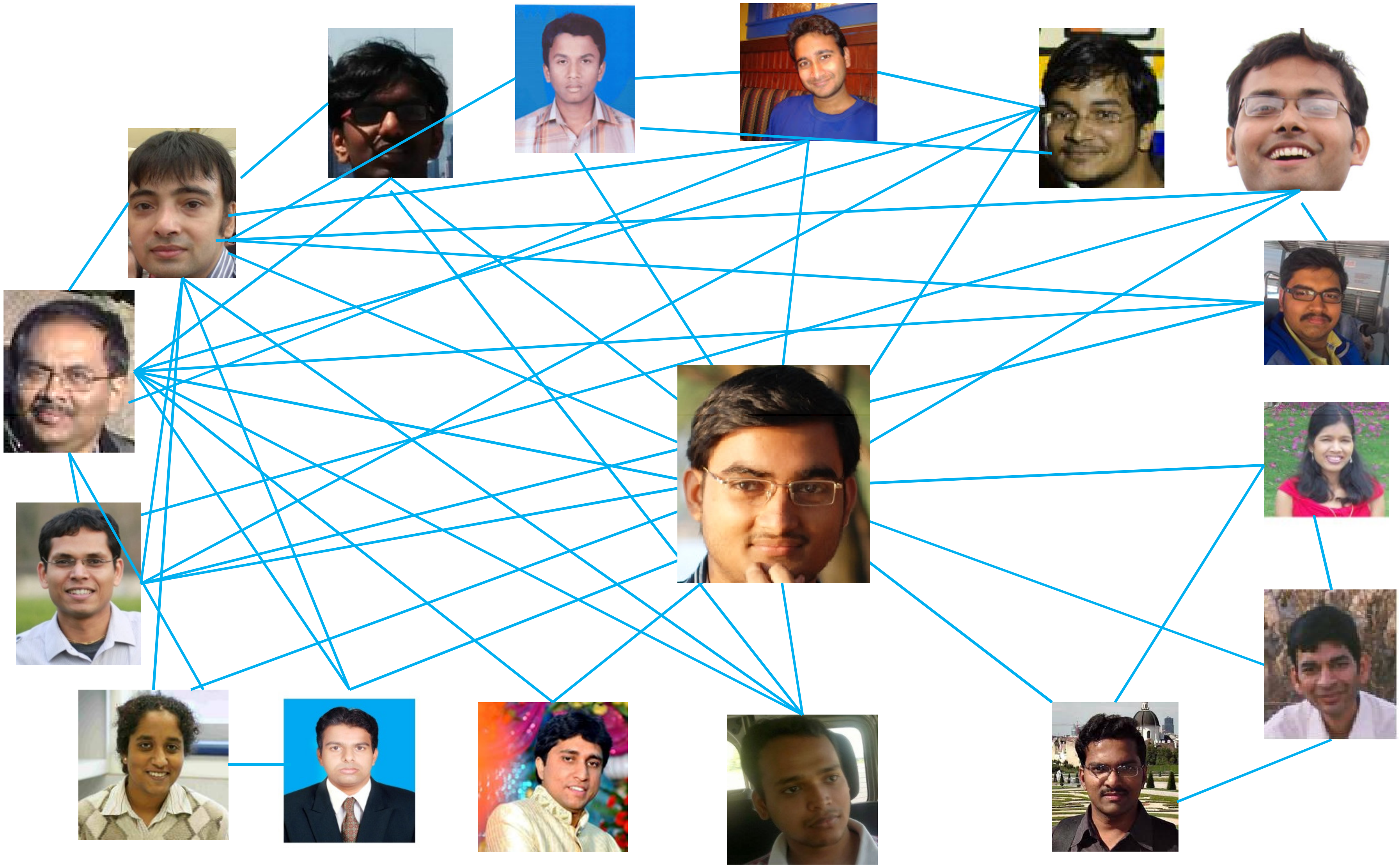
Publications outside the Thesis

- M. Singh, V. Patidar, S. Kumar, **T. Chakraborty**, A. Mukherjee, P. Goyal. The role of citation context in predicting long-term citation profiles: an experimental study based on a massive bibliographic text dataset, CIKM, 2015 (accepted).
- **T Chakraborty**, Leveraging non-overlapping communities to detect overlapping community structure, JSTAT, 5, pp. P05017, 2015.
- S. Srinivasan, **T. Chakraborty**, S. Bhowmick. Identifying Base Clusters And Their Application To Maximizing Modularity, Contemporary Mathematics. Graph partitioning and Graph Clustering. AMS-DIMACS, 2012.
- M. Singh, **T. Chakraborty**, A. Mukherjee, P. Goyal. ConfAssist: A Conflict resolution framework for assisting the categorization of Computer Science conferences, JCDL, Tennessee, USA, 2015.
- **Chakraborty** et al., Automatic Classification of Scientific Groups as Productive: An Approach based on Motif Analysis, *ASONAM* ,Beijing, 2014.
- **Chakraborty** et al., Analysis and Modeling of Lowest Unique Bid Auctions, SocialCom, Stanford, CA, USA, 2014.
- **Chakraborty** & Chakraborty, OverCite: Finding Overlapping Communities in Citation Network, BASNA, ASONAM, Canada, 2013..

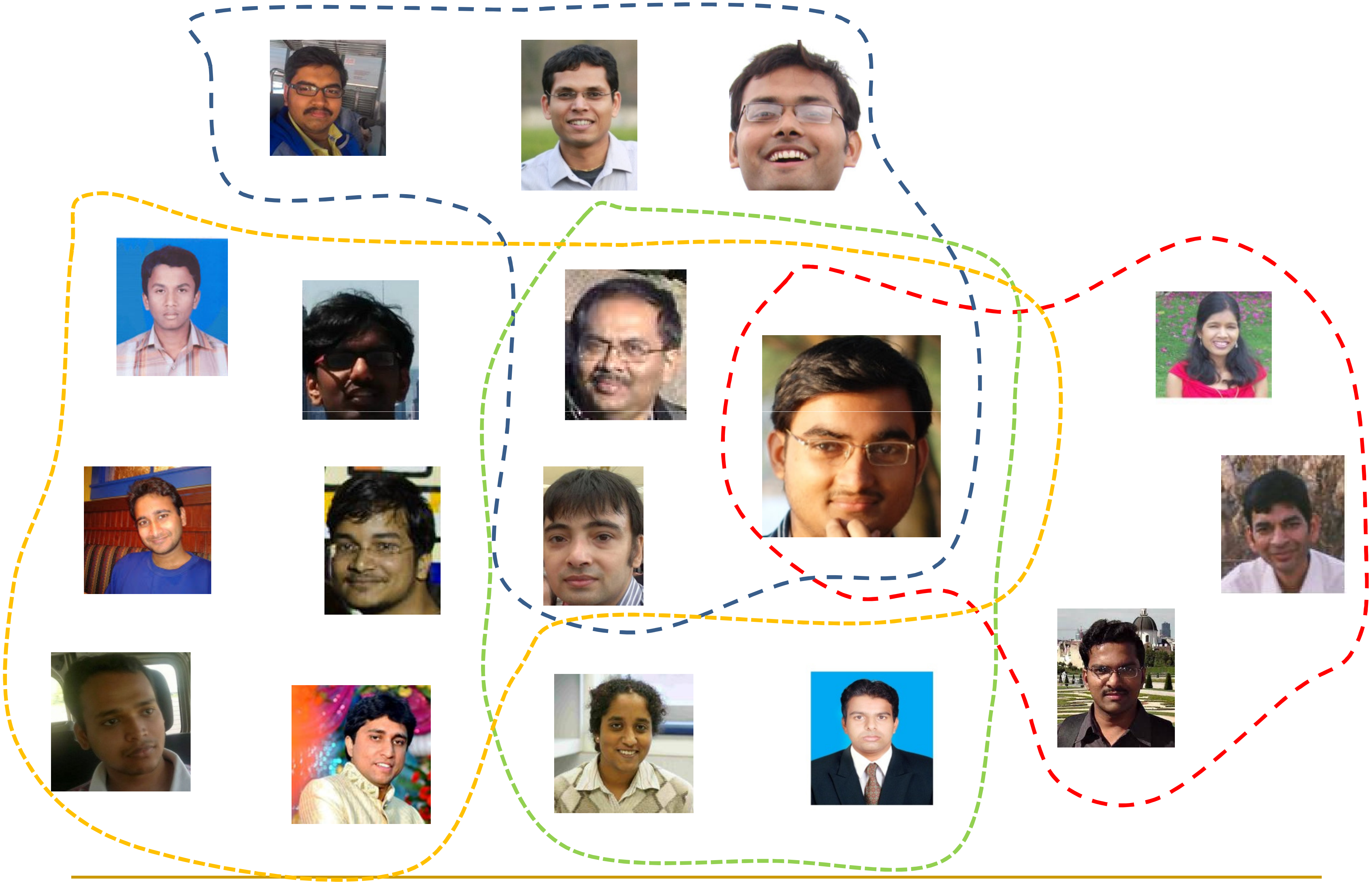
Awards & Recognitions

- Google PhD Fellowship
- First Prize in Microsoft Techvista 2015
- Best Demo Award, IBM Day, IIT Kharagpur, 2015
- Best Presentation Award, Workshop in Social Networking, COMSNETS 2014
- Honorable Mention Award, Microsoft Techvista 2013
- Best paper nomination in ASONAM, 2013
- Our KDD 2014 paper was invited as a premier paper at COMAD 2014, XRCI 2015 and ACM iKDD 2015

Thanks to the Co-authors



Thanks to the Co-authors



Thank you