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

<table>
<thead>
<tr>
<th></th>
<th>S1: Methods</th>
<th>S2: Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>G1: Patterns</strong></td>
<td>Nat. Sci. Rep. 13</td>
<td>CACM 15</td>
</tr>
</tbody>
</table>
| **G2: Analysis/Modeling** | ASONAM 13  
|                  | SNAM 14                      | JOI 15                           |
|                  | SocialCom 13                 | CACM 15                          |
|                  | COMSNETS 14                  | SIGKDD 15                        |
| **G3: Algorithms/Predictions** | SIGKDD 14  
|                  | IEEE TKDE (submitted)        | JCDL 14                          |
|                  | ACM TKDD (submitted)         | ICDE 15                          |
Organization of the Thesis

Chapter 1: Constant Communities in Networks
Chapter 2: Permanence and Community Structure
Chapter 3: Analyzing Real-world Communities
Chapter 4: Community-based Applications
Chapter 1: Constant Communities in Networks
Vertex Ordering

C1
A B C

C2
X Y Z

Iteration - 1
Vertex Ordering

Iteration - 2
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

\[ \Omega(v) = \frac{\text{Ind}(v)}{\text{Deg}(v)} \times \sum_{i=1}^{\text{ENG}_i(v)} \frac{1}{\text{EN}(v)} \]

- \( \Omega(v) \) = Relative permanence
- \( \text{Ind}(v) \) = # of internal neighbor
- \( \text{Deg}(v) \) = Degree of \( v \)
- \( K \) = # of external neighbor comm
- \( \text{EN}(v) \) = # of external neighbors of \( v \)
- \( \text{ENG}_i(v) \) = Connections to \( i \)th external neighbor community

\[ \Omega(A) = \frac{2}{8} \times \frac{1}{4} + \frac{1}{2} = \frac{1}{32} \]
\[ \Omega(B) = \frac{2}{8} \times \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = \frac{1}{16} \]

Discussion
Distribution of Relative Permanence

constant vertices
Improving Community Detection Algorithms

1. Identify CC
2. Construct super network
3. Apply community detection algorithm
4. Unfold super network
5. Final community structure

Collapse CC
# Modularity (Q) Improvement on Real Networks

<table>
<thead>
<tr>
<th>Networks</th>
<th>Improvemnt of Q (%)</th>
<th>Variance of Q (- CC)</th>
<th>Variance of Q (+ CC)</th>
<th>Improvemnt of Q (%)</th>
<th>Variance of Q (- CC)</th>
<th>Variance of Q (+ CC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polbook</td>
<td>3.34</td>
<td>1.74e-5</td>
<td>1.2e-32</td>
<td>1.20</td>
<td>2.25e-5</td>
<td>0</td>
</tr>
<tr>
<td>Dolphin</td>
<td>1.30</td>
<td>1.76e-5</td>
<td>0</td>
<td>1.90</td>
<td>0.9e-10</td>
<td>0</td>
</tr>
<tr>
<td>Football</td>
<td>2.45</td>
<td>2.01e-5</td>
<td>0</td>
<td>3.05</td>
<td>7.25e-8</td>
<td>6.4e-10</td>
</tr>
<tr>
<td>Email</td>
<td>4.80</td>
<td>6.89e-5</td>
<td>0.9e-12</td>
<td>5.80</td>
<td>1.7e-8</td>
<td>1.36e-12</td>
</tr>
</tbody>
</table>
Chapter 2: Permanence and Community Structure
Modularity

$$Q = \frac{1}{m} \sum_{c=1}^{n_c} \left( l_c - \frac{d_c^2}{4m} \right)$$

- **Actual edges**
- **Expected edges**

*Global Measure*
- Total internal connections
- Total external connections

*Notations:
- $m = \# \text{ edges}$
- $n_c = \# \text{ communities}$
- $l_c = \# \text{ internal edges in community } c$
- $d_c = \text{ sum of degrees of all nodes in } c$
Our Perspective of a Community
Heuristic I

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

Modularity, Conductance, Cut-ratio consider total external connections
Drug

Shoplifting

Drug
Heuristic II

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

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

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

- \( I(v) \): internal deg of \( v \)
- \( D(v) \): degree of \( v \)
- \( E_{\text{max}}(v) \): Max connection to an external neighbor
- \( C_{\text{in}}(v) \): clustering coefficient of internal neighbors

\( Perm(v) = 0.12 \)

\( I(v) = 4, \ D(v) = 7, \ E_{\text{max}}(v) = 2 \)

\( C_{\text{in}}(v) = 5/6 \)

Discussion
Permanence

Permanence $\sim 1$

Permanence $= 0$

Permanence $\sim -1$

Wrong vertex-to-community assignment
MaxPerm: Non-overlapping Community Detection Algorithm
Major Limitations

- Limitations of optimization algorithms
  - Resolution limit \textit{(Fortunato & Barthelemy, PNAS, 07)}
  - Degeneracy of solutions \textit{(Good et al., PRE, 10)}
  - Asymptotic growth \textit{(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 $\geq 3$
## Experimental Results

<table>
<thead>
<tr>
<th>Algo</th>
<th>LFR (µ=0.1)</th>
<th>LFR (µ=0.3)</th>
<th>Football</th>
<th>Railway</th>
<th>Coauthorship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Louvain</td>
<td>0.02</td>
<td>0.00</td>
<td>0.02</td>
<td>0.14</td>
<td>0.00</td>
</tr>
<tr>
<td>FastGrdy</td>
<td>0.00</td>
<td><strong>0.87</strong></td>
<td>0.01</td>
<td><strong>0.37</strong></td>
<td>0.14</td>
</tr>
<tr>
<td>CNM</td>
<td><strong>0.14</strong></td>
<td><strong>0.40</strong></td>
<td>0.30</td>
<td>0.00</td>
<td><strong>0.05</strong></td>
</tr>
<tr>
<td>WalkTrap</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>Infomod</td>
<td>0.06</td>
<td>0.08</td>
<td>0.19</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>Infomap</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
</tr>
</tbody>
</table>

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 **independent** of the **symmetric growth** of network size and the number of communities.

**Analytical proofs:** [http://cnerg.org/permanence](http://cnerg.org/permanence)
Metric for Overlapping Communities
Overlapping Permanence (OPerm)

\[ P_{ov}^c(v) = \frac{I_c(v)}{E_{max}(v)} \times \frac{1}{D(v)} - (1 - c_{in}(v)) \cdot \frac{I_c(v)}{I(v)} \]

- \(D(v)\) = degree of \(v\)
- \(E_{max}(v)\) = Max connection to an external community
- \(C_{in}(v)\) = clustering coeff. of internal neighbors of \(v\) in \(c\)
- \(I(v)\) = \# of internal neighbors of \(v\)

\[ I_c(v) = \sum_{e \in \Gamma_c^e} \frac{1}{\chi_e} \]

\(\Gamma_c^e\) = internal edges of \(v\) in community
\(\chi_e\) = \# of communities edge \(e\) shares

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

Diagram:

- \(D(v) = 8, I(v) = 5\)
- \(P_{ov}^{c_1}(v) = \frac{1+1+\frac{1}{2}}{2 \times 8} - (1 - \frac{2}{3}) \times \frac{1+1+\frac{1}{2}}{5} = -0.01\)
- \(P_{ov}^{c_2}(v) = -0.18\)
- \(P_{ov}(v) = P_{ov}^{c_1}(v) + P_{ov}^{c_2}(v) = -0.19\)
Inference from OPerm Values

Core-periphery Structure within Communities

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

<table>
<thead>
<tr>
<th>Assortativity</th>
<th>LFR (0.1)</th>
<th>LiveJournal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree-based</td>
<td>-0.045</td>
<td>0.037</td>
</tr>
<tr>
<td>OPerm-based</td>
<td>0.645</td>
<td>0.465</td>
</tr>
</tbody>
</table>
Layers within a Community

Layer 1
Innermost layer

Layer 2
Middle layer

Layer 3
Outermost layer

Intra-layer edge
Inter-layer edge
MaxOPerm: Overlapping Community Detection Algorithm
MaxOPerm: Framework

Edge-based seed community

Combining vertices to gain OPerm

Expanding community boundary

Final community structure
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

<table>
<thead>
<tr>
<th>Basic Statistics of papers from 1960-2010</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of valid entries</td>
<td>3,473,171</td>
</tr>
<tr>
<td>Number of authors</td>
<td>1,186,412</td>
</tr>
<tr>
<td>Number of unique venues</td>
<td>6,143</td>
</tr>
<tr>
<td>Avg. number of papers per author</td>
<td>5.18</td>
</tr>
<tr>
<td>Avg. number of authors per paper</td>
<td>2.49</td>
</tr>
</tbody>
</table>
## Available Metadata

<table>
<thead>
<tr>
<th>Metadata Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
</tr>
<tr>
<td>Unique ID</td>
</tr>
<tr>
<td>Named entity disambiguated authors’ name</td>
</tr>
<tr>
<td>Year of publication</td>
</tr>
<tr>
<td>Named entity disambiguated publication venue</td>
</tr>
<tr>
<td>Related research field(s)</td>
</tr>
<tr>
<td>References</td>
</tr>
<tr>
<td>Keywords</td>
</tr>
<tr>
<td>Abstract</td>
</tr>
</tbody>
</table>

Available @ [http://cnerg.org](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)

Peaks within 3 years from publication, then declines

Average Inward Citations
**“Impact” of a Field (Community)**

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

$$In(f^t_i) = \sum_{j \neq i} w^t_{j \rightarrow i}$$

where,

$$w^t_{j \rightarrow i} = \frac{C^t_{j \rightarrow i}}{P^t_i}$$

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

$P^t_i = \#$ of papers in field $f_i$

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

N.B.: We only consider cross-field citations

\[ \text{In}(F^t_1) = \frac{5}{3} \]
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

$RDI(X_i) = - \frac{3}{5} \log \left( \frac{3}{5} \right) - \frac{2}{5} \log \left( \frac{2}{5} \right) = 0.67$

World Wide Web (95–99)  |  NLP (95–99)

- OTH (5%)  |  OTH (5%)
- IR (16%)  |  ALGO (19%)
- HCI (13%)  |  ML (24%)
- DB (34%)  |  AI (31%)
- NETW (32%)  |  IR (20%)
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
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

Fraction of papers

Citation buckets

Application: Future Citation Count Prediction
Problem Definition

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

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

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

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

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

**Dataset is homogeneous** in terms of citation profile

Yan et al., JCDL 12
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

- **Dataset**: Training, Testing
- **Model Stage**
  - Rule-based approach
  - Categorization: PeakInt, PeakMul, PeakLate, MonDec, MonIncr, Oth
- **Query Paper**
- **Static Features**
- **SVM**
- **SVR**

Graph showing the flow of data through different stages with mappings and outputs.
Static Features

Author-centric
- Productivity (Max/Avg)
- H-index (Max/Avg)
- Versatility (Max/Avg)
- Sociality (Max/Avg)

Venue-centric
- Prestige
- Impact Factor
- Versatility

Paper-centric
- Team-size
- Reference count
- Reference diversity
- Keyword diversity
- Topic diversity
Performance Evaluation

(i) Coefficient of determination ($R^2$)
    The more, the better

(ii) Mean squared error ($\theta$)
    The less, the better

(iii) Pearson correlation coefficient ($\rho$)
    The more, the better
## Performance Evaluation

<table>
<thead>
<tr>
<th>$\Delta t$</th>
<th>$R^2$</th>
<th>$\theta$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.57</td>
<td>5.06</td>
<td>0.61</td>
</tr>
<tr>
<td>2</td>
<td>0.55</td>
<td>7.10</td>
<td>0.59</td>
</tr>
<tr>
<td>3</td>
<td>0.52</td>
<td>8.78</td>
<td>0.65</td>
</tr>
<tr>
<td>4</td>
<td>0.50</td>
<td>10.06</td>
<td>0.75</td>
</tr>
<tr>
<td>5</td>
<td>0.45</td>
<td>13.06</td>
<td>0.42</td>
</tr>
</tbody>
</table>
Application: Faceted Recommendation System for Scientific Articles
## Flat vs. Faceted Recommendation

### Flat Recommendation System

**Papers**
- A fully bayesian approach to unsupervised part-of-speech tagging
  - ACL, 2007, pp. 744–751
- Prototype-driven learning for sequence modeling
  - HLT–NAACL, 2006, pp. 320–327
- Why doesn’t EM find good HMM POS–tagger?
  - EMNLP-CoNLL, 2007, pp. 296–305
- A practical part-of-speech tagger
  - ANLC, 1992, pp. 133–140
- Minimized model for unsupervised part-of-speech tagging
  - ACL, 2009, pp. 504–512
- Constrastive estimation: training log-linear models on unlabeled data
  - ACL, 2005, pp. 354–362

### Faceted Recommendation System

#### Tags
- A fully bayesian approach to unsupervised part-of-speech tagging
  - ACL, 2007, pp. 744–751
- Prototype-driven learning for sequence modeling
  - HLT–NAACL, 2006, pp. 320–327

#### Background
- A practical part-of-speech tagger
  - ANLC, 1992, pp. 133–140
- Why doesn’t EM find good HMM POS–tagger?
  - EMNLP-CoNLL, 2007, pp. 296–305

#### Methods
- Minimized model for unsupervised part-of-speech tagging
  - ACL, 2009, pp. 504–512
- Constrastive estimation: training log-linear models on unlabeled data
  - ACL, 2005, pp. 354–362
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

<table>
<thead>
<tr>
<th>Facets</th>
<th>VanillaPR</th>
<th>FeRoSA-CS</th>
<th>FeRoSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>BG</td>
<td>0.65</td>
<td>0.51</td>
<td>0.79</td>
</tr>
<tr>
<td>AA</td>
<td>0.48</td>
<td>0.34</td>
<td>0.56</td>
</tr>
<tr>
<td>MD</td>
<td>0.62</td>
<td>0.39</td>
<td>0.62</td>
</tr>
<tr>
<td>CM</td>
<td>0.44</td>
<td>0.38</td>
<td>0.62</td>
</tr>
<tr>
<td>Average</td>
<td>0.55</td>
<td>0.40</td>
<td>0.65</td>
</tr>
</tbody>
</table>
Evaluation by the Original Authors

Hi XXX,

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

1. **Background paper (BG)**
   - Lattice Minimum Bayes–Risk Decoding for Statistical Machine Translation
   - R & F

2. **Alternative Approach (AA)**
   - Structural and Topical Dimensions in Multi-Task Patent Translation
   - R & Not F

3. **Comparison (CM)**
   - Sequential Labeling with Latent Variables, An Exact Inference Algorithm and its Efficient Approximation
   - Not R & Not F

4. **Method (MD)**
   - Feature-Rich Translation by Quasi-Synchronous Lattice Parsing
   - Not R & Not F

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

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

<table>
<thead>
<tr>
<th>Systems</th>
<th>OI@3</th>
<th>OP@3</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS</td>
<td>0.27</td>
<td>0.61</td>
</tr>
<tr>
<td>MAS</td>
<td>0.17</td>
<td>0.45</td>
</tr>
<tr>
<td>LLQ</td>
<td>0.13</td>
<td>0.41</td>
</tr>
<tr>
<td>f-FeRoSA</td>
<td>0.43</td>
<td>0.79</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OP</th>
<th>f-FeRoSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>OP@3</td>
<td>0.79</td>
</tr>
<tr>
<td>OP@5</td>
<td>0.78</td>
</tr>
<tr>
<td>OP@10</td>
<td>0.71</td>
</tr>
</tbody>
</table>
Towards an ACL Anthology Corpus with Logical Document Structure. An Overview of the ACL 2012 Contributed Task

- **Comparison.** Integrating User-Generated Content in the ACL Anthology
- **Comparison.** Predicting a Scientific Community’s Response to an Article
- **Method. Background.** Repurposing Theoretical Linguistic Data for Tool Development and Search
- **Comparison.** Scientific Paper Summarization Using Citation Structure
- **Method. Background.** Rediscovering ACL Discoveries Through the Lens of Global Lesk Similarity
- **Method.** Towards High-Quality Text Stream Extraction from PDF Technical Reports: Method 1
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
“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**: \( f(x) = a \cdot x^{-k} \), \( 2 \leq k \leq 3 \)
- **Stretched exponential**: \( f_k(x) = e^{-x^k} \), \( 0 \leq k \leq 1 \)
- **Log-normal**: If \( x \) is log normal, \( y = \ln(x) \) follows a normal dist
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., 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., 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., Rising Popularity of Interdisciplinary Research - an Analysis of Citation Networks, Workshop on Social Networks, COMSNETS, Bangalore, 2014.


- 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, 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


- 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.


- 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