On the Permanence of Vertices in Network Communities

Tanmoy Chakraborty
Google India PhD Fellow
IIT Kharagpur, India

20th ACM SIGKDD, New York City, Aug 24-27, 2014

Tanmoy Chakraborty

Niloy Ganguly

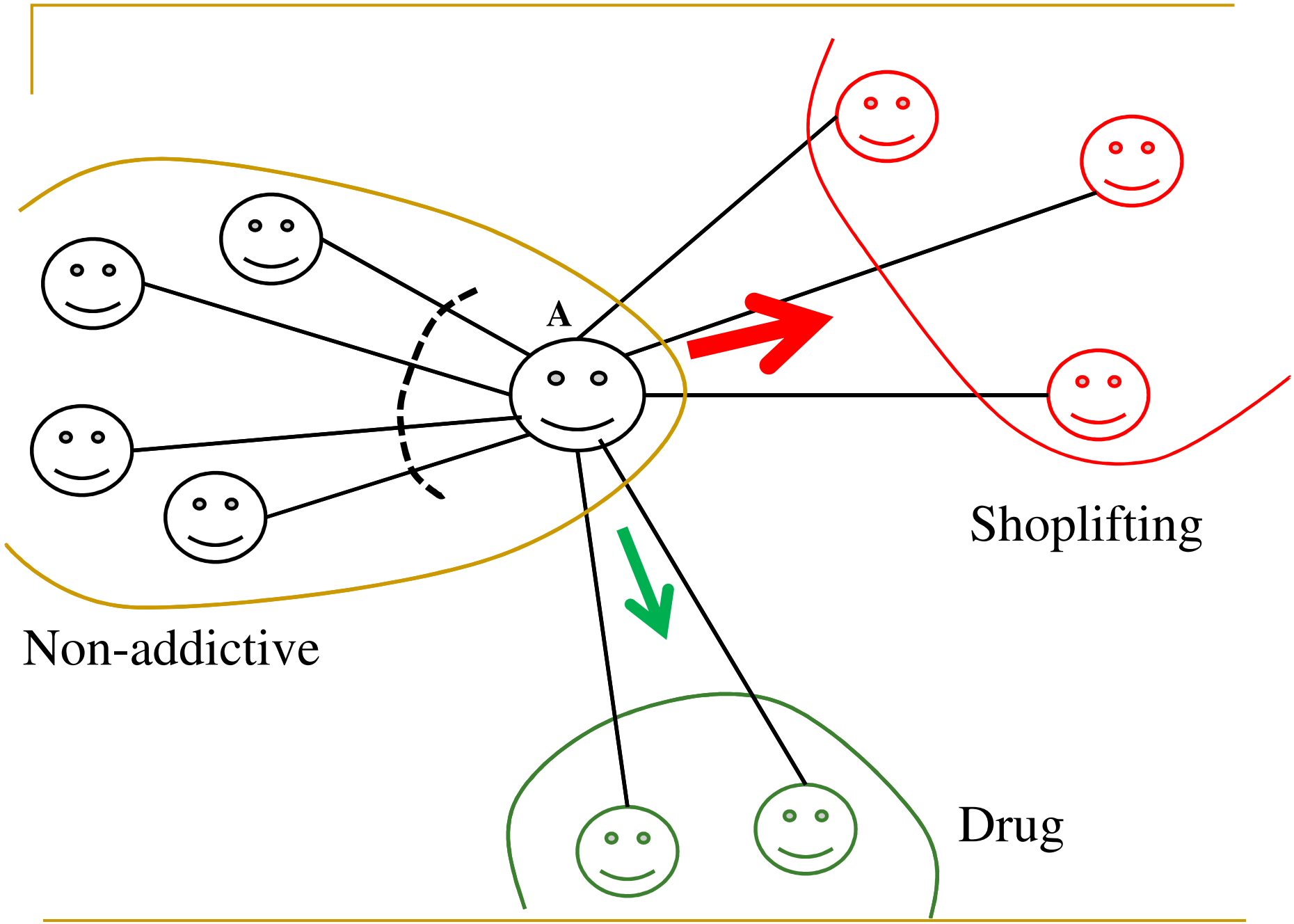
Animesh Mukherjee

IIT Kharagpur, India

Sriram Srinivasan

Sanjukta Bhowmick

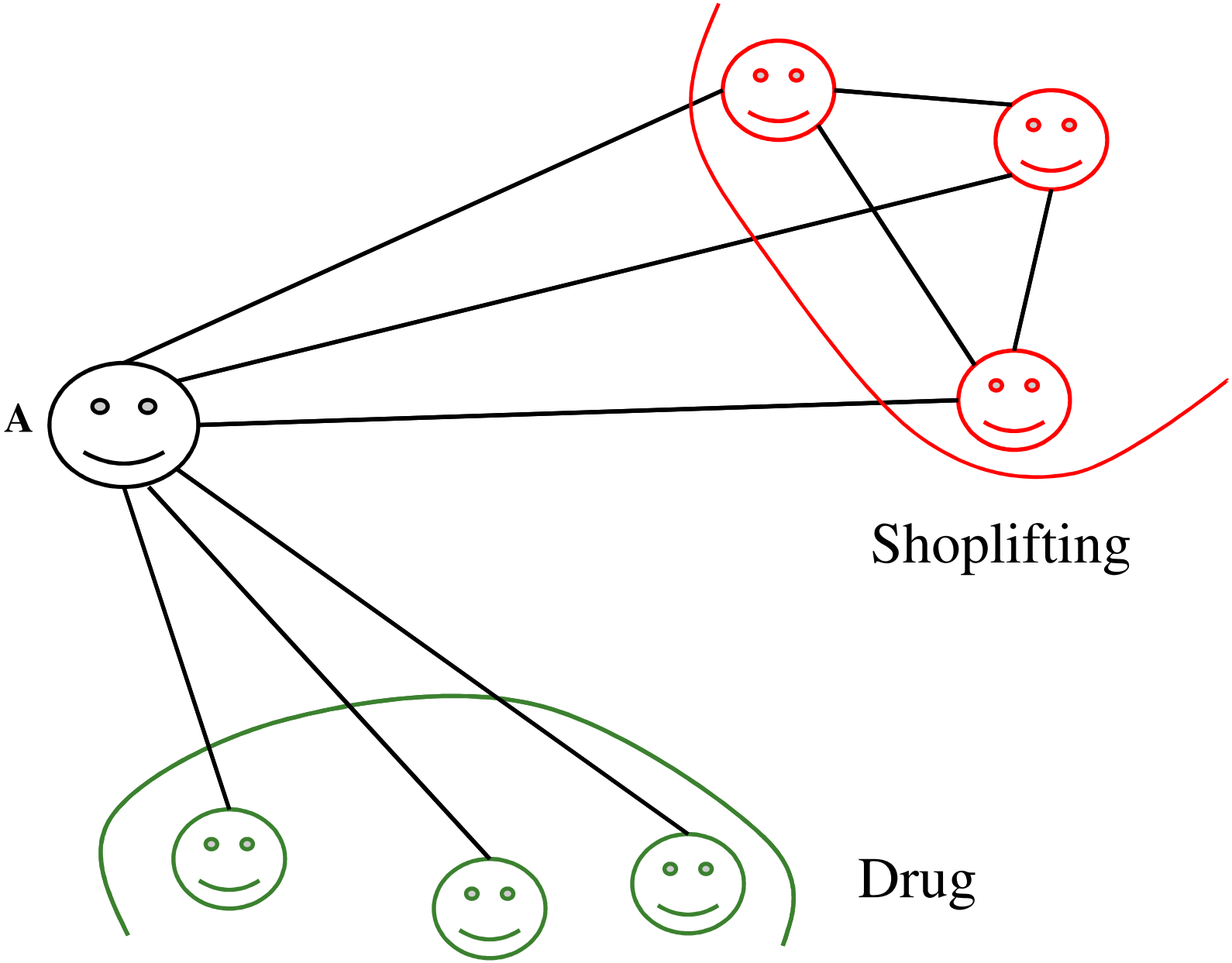
University of Nebraska, Omaha


A
Shoplifting
Non-addictive
Drug

Heuristic I

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

Modularity, Conductance, Cut-ration consider total external connections


A
Shoplifting
Drug

Heuristic II

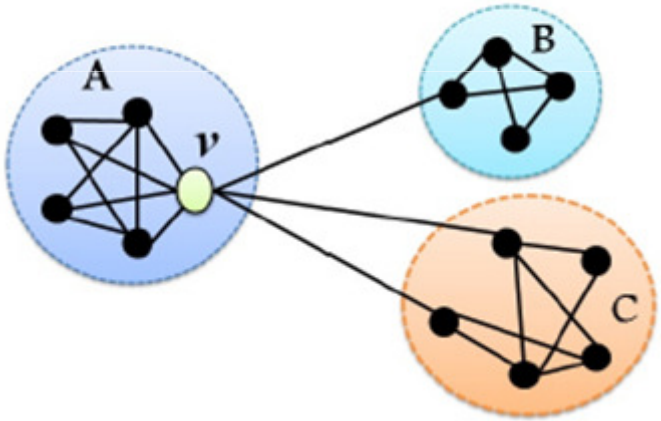
Internal neighbors should be highly connected => high clustering coefficient among internal neighbors

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

Permanence

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

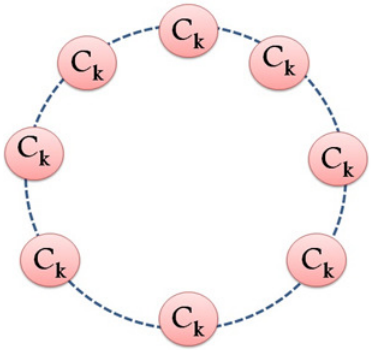
- $I(v)$=internal deg of v
- $D(v)$=degree of v
- $E_{max}(v)$=Max connection to an external neighbor
- $C_{in}(v)$=clustering coefficient of internal neighbors



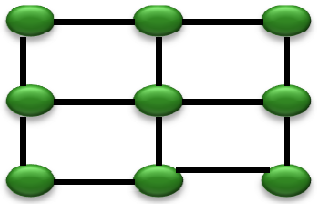
$Perm(v)$=0.12

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

Permanence



Permanence ~ 1



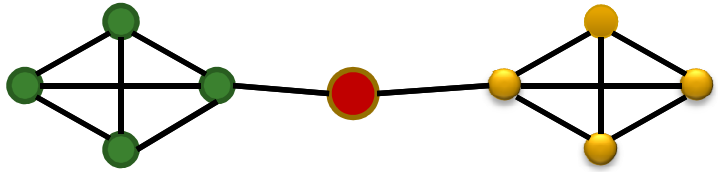
Permanence = 0

Wrong vertex-to-community assignment

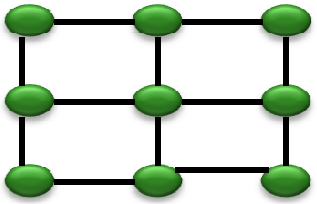
Permanence ~ -1

Research Questions

1. **Assigning every vertex** in a community is **reasonable**?



2. No one measures the **intensity of belongingness** of a vertex to a community
   - ✓ Only try to detect **best community structure** in the network
   - ✗ Never ask for whether a network possesses a strong community



Permanence answers all these questions

Test Suite of Networks

- **Synthetic Networks**

  LFR networks with different values of mixing parameter (μ)

  (Lancichinetti & Fortunato, PRE, 11)

- **Real-world Networks**
  - **Football Network**

    Nodes: teams, Edges: matches, Communities: team-conference

    (Girvan & Newman, PNAS, 02)
  - **Railway Network**

    Nodes: station, Edges: train-connections, Communities: state/provinces

    (Ghosh et al., Acta Physica, 11)
  - **Coauthorship Network**

    Nodes: authors, Edges: coauthorships, Communities: research field

    (Chakraborty et al., ASONAM, 13)

Baseline Algorithms

- **Modularity based**
  - FastGreedy *(Newman, PRE, 04)*
  - Louvain *(Blondel et al, J. Stat. Mech., 08)*
  - CNM *(Clauset et al, PRE, 04)*
- **Random-walk based**
  - WalkTrap *(Pons & Latapy, J. Graph Algo and Appln, 06)*
- **Compression based**
  - InfoMod *(Rosvall & Bergstrom, PNAS, 07)*
  - InfoMap *(Rosvall & Bergstrom, PNAS, 08)*

Permanence: A Better Community Scoring Function

Methodology

- Approach *(Steinhaeuser & Chawla, PRL, 10)*:

1. Consider a network
2. Run *N* community detection algorithms (here N=6)
3. Compute community scoring functions on these outputs
4. Rank the algos based on these values
5. Compare outputs with the ground-truth using validation metrics and rank algos again
6. Find rank-correlation

Football Network

	Mod	Per	Con	Cut	NMI	ARI	PU
FastGreedy	0.2			5	0.3		4
Louvain	0.6			2	0.5		2
CNM	0.8			1	0.8		1
WalkTrap	0.4			4	0.2		5
InfoMod	0.5			3	0.4		3
InfoMap	0.1			6	0.1		6

Intuition: Ranking of good scoring function and the validation measures should be high

Results

Lighter color is better

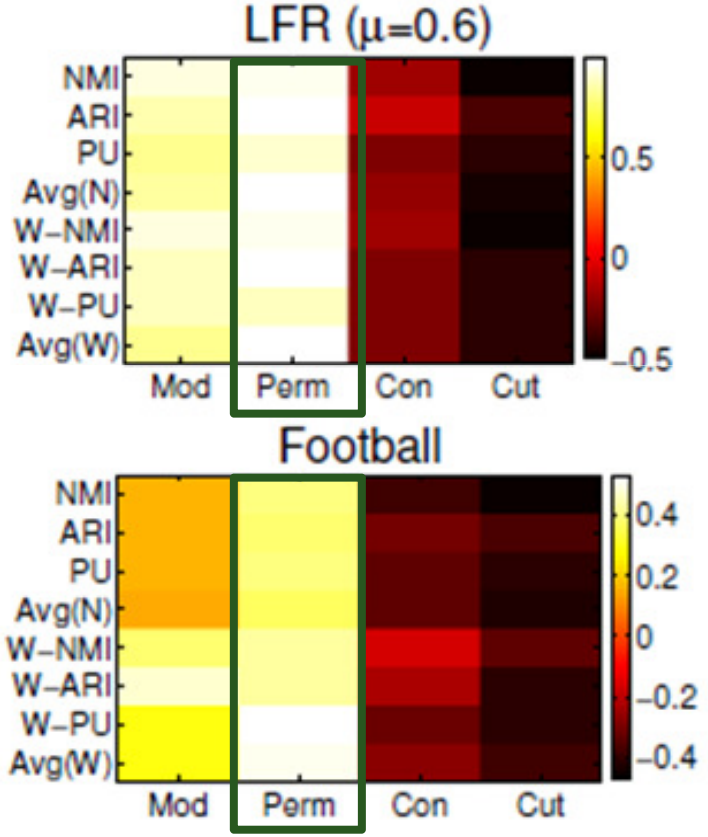


Fig. : Heat maps depicting pairwise Spearman's rank correlation

Networks	Modularity	Permanence	Conductance	Cut
LFR(μ=0.1)	**0.88**	**0.88**	**0.88**	0.02
LFR(μ=0.3)	0.61	**0.74**	0.72	0.28
LFR(μ=0.6)	0.87	**0.96**	-0.18	-0.44
Football	0.25	**0.43**	-0.29	-0.41
Railway	0.43	**0.46**	0.08	-0.48
Coauthorship	**0.92**	**0.92**	0.76	0.86

Table : Performance of the community scoring functions averaged over all the validation measures

Developing 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)*

Community Detection Based on Maximizing Permanence

- Follow similar strategy used in Louvain algorithm (a greedy modularity maximization) *(Blondel et al., J. Stat. Mech, 07)*
- Selecting seed nodes helps converge the process faster
- We only consider those communities having size >=3
- Communities having size<3 remain as singleton

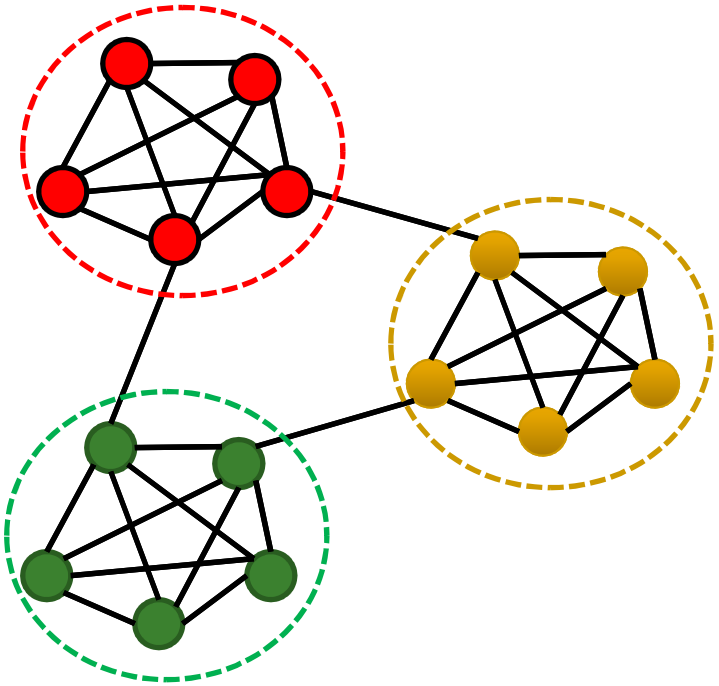
Experimental Results

Why ????

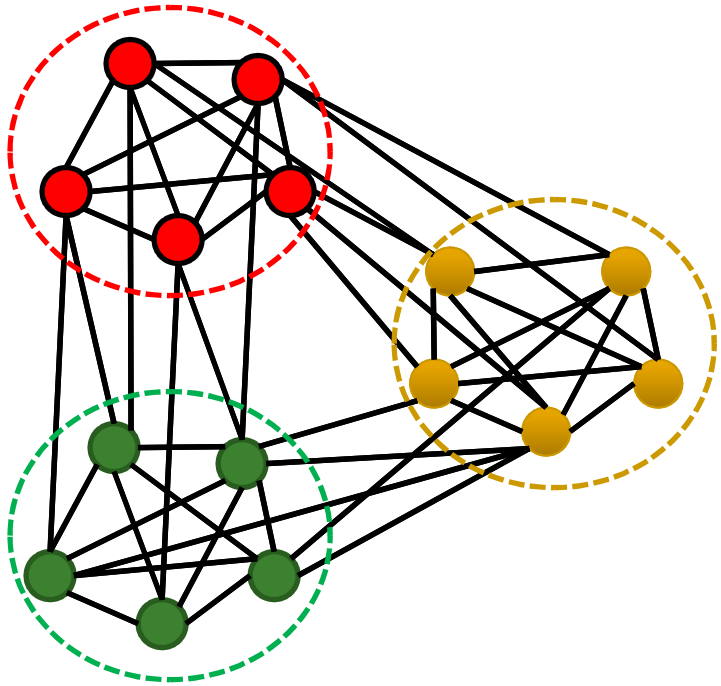
Algo	LFR (μ=0.1)	LFR (μ=0.3)	LFR (μ=0.6)	Football	Railway	Coauthorship
Louvain	0.02	0.00	-0.75	0.02	0.14	0.00
FastGrdy	0.00	0.87	0.02	0.01	0.37	0.14
CNM	0.14	0.40	-0.13	0.30	0.00	0.05
WalkTrap	0.00	0.00	-0.50	0.02	0.02	0.01
Infomod	0.06	0.08	-0.20	0.19	0.04	0.00
Infomap	0.00	0.00	-0.72	0.02	-0.02	0.03

Table: Differences of our algorithm with the other algorithms averaged over all validation measures

LFR (μ = 0.1) vs. LFR (μ = 0.6)



LFR (μ = 0.1)

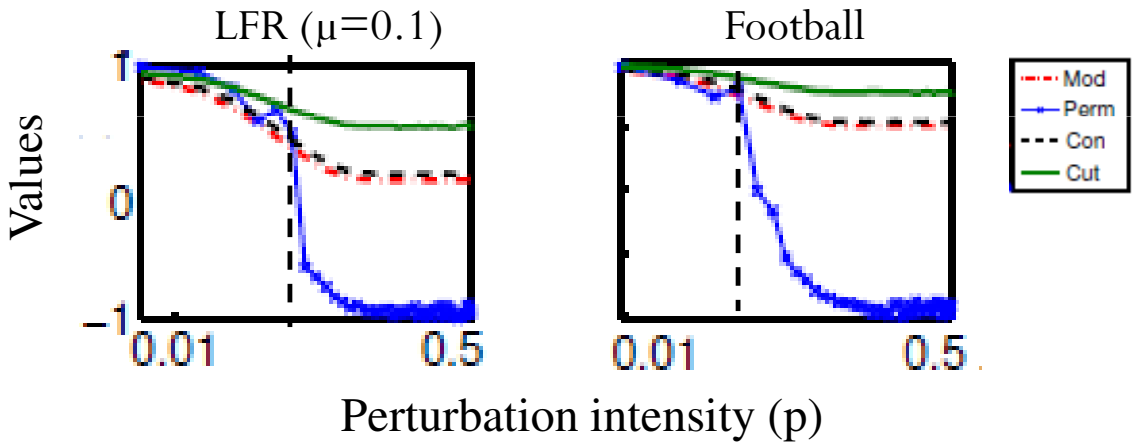


LFR (μ = 0.6)

μ = Avg**. ratio** between the number of **external connections** to **its degree**

Permanence is Nice

- **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)

Issues Related with Modularity Maximization

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

Take Away

- **Permanence**
  - a better community scoring function
  - sensitive to perturbation after a certain threshold
  - indicates the eligibility of a network for community detection

- **Maximizing permanence**
  - a better community detection algorithm
  - can detect small-size communities
  - ameliorates existing limitations

❑ **Future work**

- Recast permanence for overlapping communities
- Recast permanence for weighted graphs
- Recast permanence for dynamic community detection

Thank you

http://cnerg.org/permanence