Community Detection in Social Networks

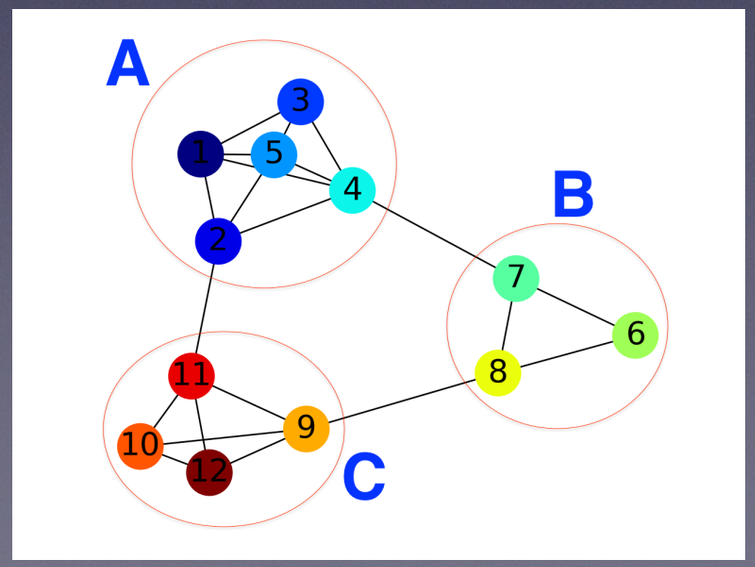
Yawen Sun
1836786

Outline

- Background
- Motivation and Objectives
- 4 types of Algorithms
- Results and Evaluation

Background

- Communities in Social Networks
- Features of communities
  - Nodes inside are communicated frequently
  - Nodes outside are connected loosely
- A simple network with 3 communities



Motivations and Objectives

Community detection in distinct fields

- Sociology
  - Helps in users behaviours analysis
- Biology
  - Helps in identification of a new population
- Computer Science
  - Can be used in recommender systems

Community Detection Algorithms

- GN Algorithm
  - A divisive algorithm based on graph theory
- Louvain
  - An agglomerative algorithm using a greedy techniques
- Spectral Clustering
  - An algorithm based on clustering techniques
- GA-Net
  - An algorithm based on genetic algorithm inspired from nature

GN Algorithm

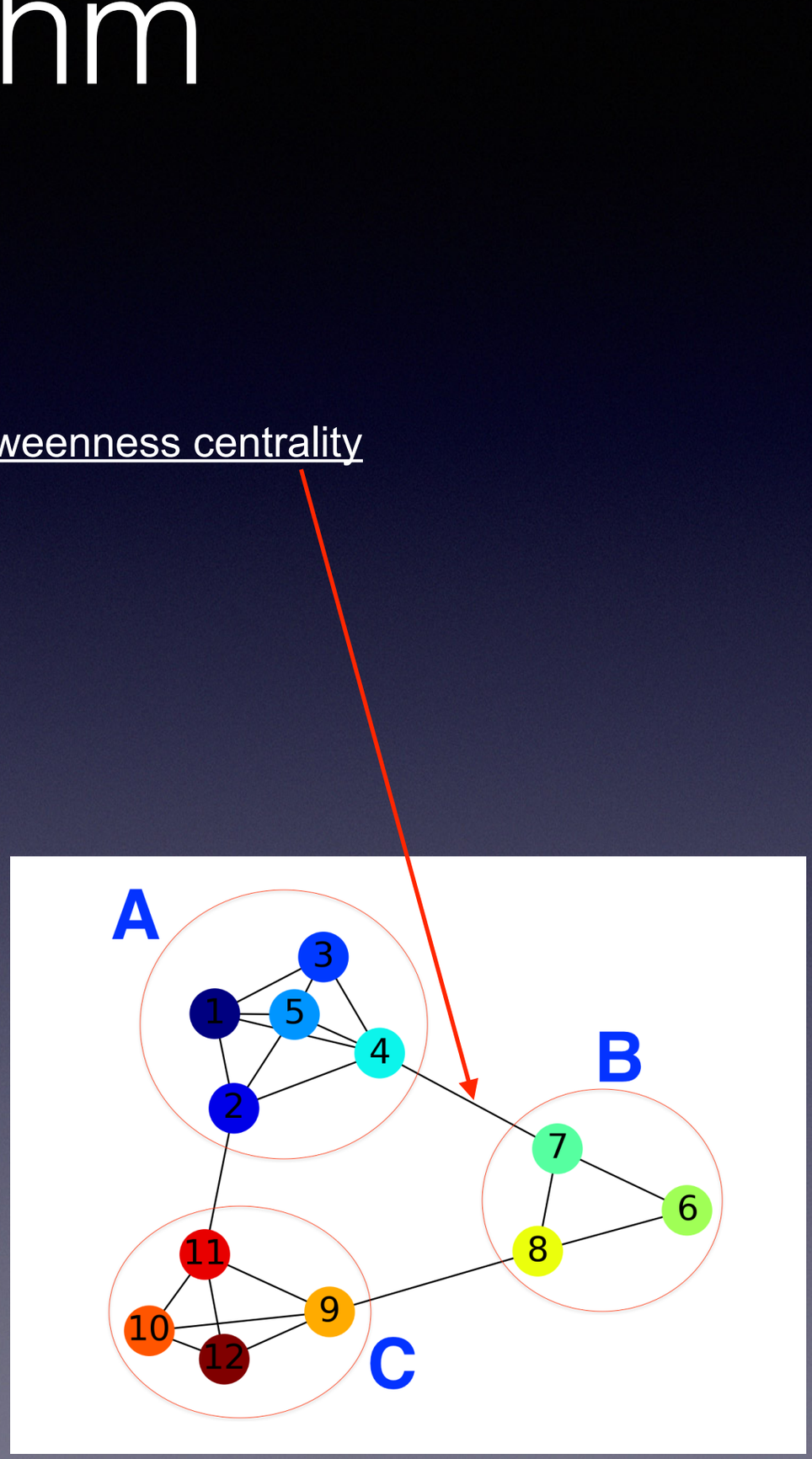
Input: the network G=(V,E)
while there is no edge:
    for e∈E do
        calculating its edge betweenness centrality
    remove an edge with the highest value of edge betweenness centrality
    if there is a tie, remove an edge in the tie set
    calculating the modularity
Output: the partition with the highest modularity

Betweenness centrality: $C_B(e) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(e)}{\sigma_{st}}$

A fraction of the shortest paths between two nodes.

Modularity: $Q = \sum_{i}^{n} (e_{ii} - a_i^2) \ , \ a_i = \sum_{j}^{n} e_{ij}$

A measure of how dense in communities.



M. Girvan and M. E. J. Newman (2002) Community structure in social and biological networks.

Louvain

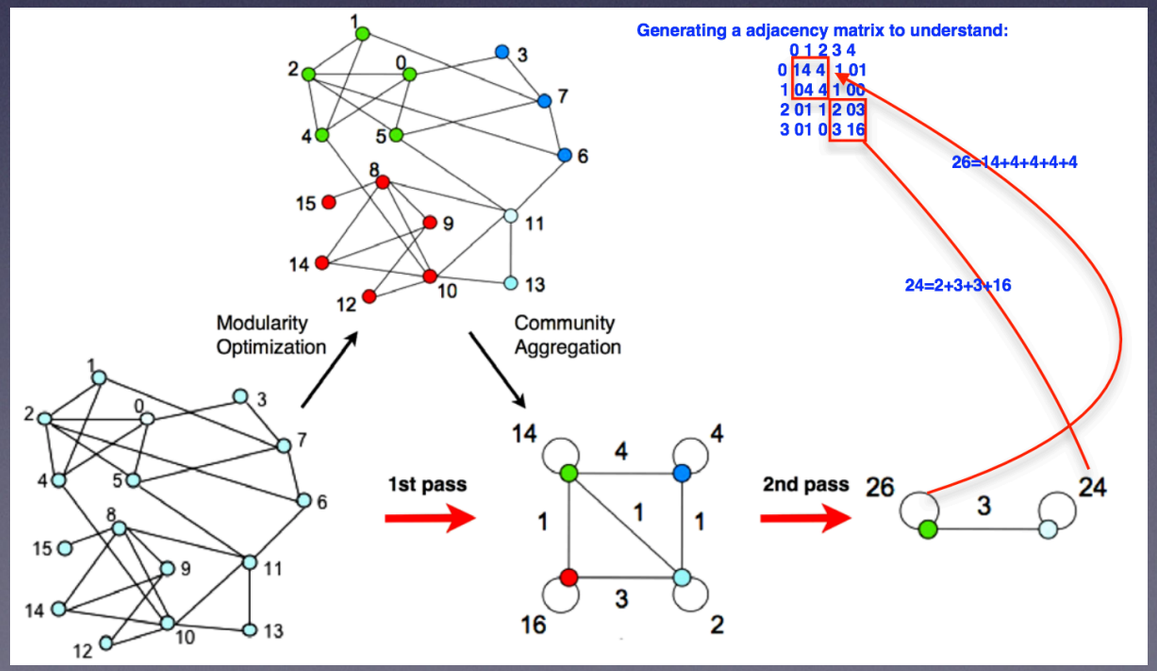
1. Assign every node as a community.
2. Combine:
    a. combine every node with its neighbour community and calculate ΔQ
    b. put this node into its neighbour whose value of ΔQ is the highest
3. Set new nodes:
    a. set one community as a new “node”
    b. change the weights of links
4. Repeat 1. until Modularity is not changed.

ΔQ:

The difference between Modularity of before-combination community and after-combination community

$$\Delta Q = \left[\frac{\Sigma_{in}+2k_{i,in}}{2m} - \left(\frac{\Sigma_{tot}+2k_i}{2m} \right)^2 \right] - \left[\frac{\Sigma_{in}}{2m} - \left(\frac{\Sigma_{tot}}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right] = \frac{1}{2m} \left(k_{i,in} - \frac{\Sigma_{tot} k_i}{m} \right)$$

$$\text{maximising } \Delta Q = \text{maximising } \left(k_{i,in} - \frac{\Sigma_{tot} k_i}{m} \right)$$



VD Blondel, JL Guillaume and R Lambiotte (2008) Fast unfolding of communities in large networks.

Spectral Clustering

Unnormalised Spectral Clustering

Input: **the value of k,** the network

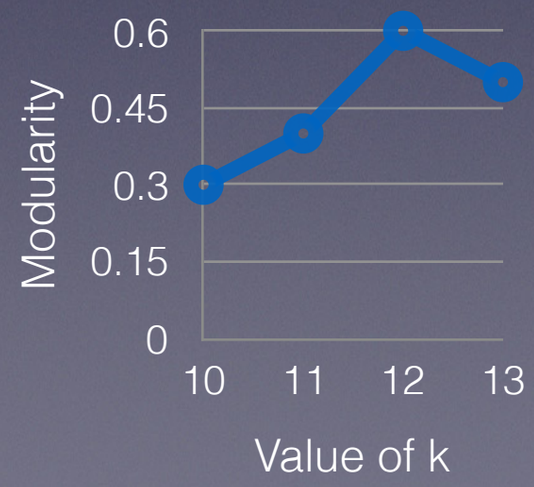
1. construct the unnormalised Laplacian matrix, L from the network
2. calculate the first k eigenvectors u1,...,uk from Laplacian matrix
3. generate matrix U (a N*k matrix) whose columns are k eigenvectors
4. apply **k-means** for N points in U with k dimensions to get k clusters

Output: k clusters

Laplacian Matrix L:

L=D(degree matrix)-A(adjacency matrix)

To find the optimal k



U Von Luxborg (2007) A tutorial on spectral clustering
AY Ng, MI Jordan and Y Weiss (2002) On spectral clustering: Analysis and an algorithm

Spectral Clustering

Normalised Spectral Clustering

Input: **the value of k**, the network

1. calculate the normalised Laplacian matrix, Lsym
2. computer first k eigenvectors u1,...,uk from normalised Laplacian matrix, Lsym
3. construct matrix U (a N*k matrix) whose columns are k eigenvectors from last step
4. generate matrix T from U through normalising, to be specific, tip
5. apply **k-means** for N points in T with k dimensions to get k clusters

Output: k clusters

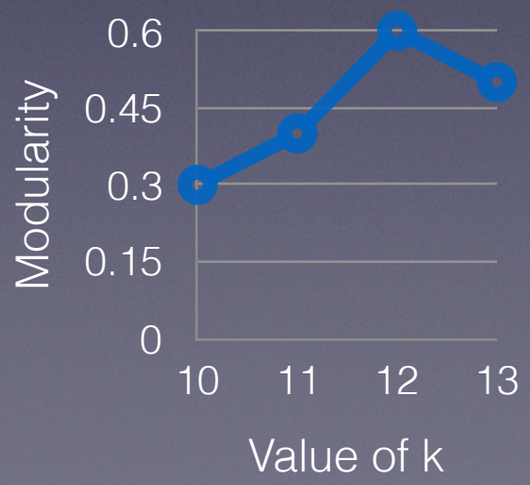
Normalised Laplacian Matrix L:

$$L_{sym} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = D^{-\frac{1}{2}}(D - A)D^{-\frac{1}{2}}$$

Matrix T:

$$t_{ij} = \frac{u_{ij}}{(\Sigma_k u_{ik}^2)^{\frac{1}{2}}}$$

To find the optimal k



U Von Luxborg (2007) A tutorial on spectral clustering
AY Ng, MI Jordan and Y Weiss (2002) On spectral clustering: Analysis and an algorithm

GA-Net

1. Initialisation:
   intialise population in the first generation *“the locus-based adjacency representation” (left part)*
2. compute fitness value for individuals
3. Selection:
   select two individuals through **roulette selection method**
4. Uniform Crossover:
   according to a generated binary string to crossover *(right part)*
5. Mutation:
   change value of the gene randomly if the changed one is appropriate
6. replace the old generation
7. repeat from step 2 to step 6 until the terminal condition is satisfied

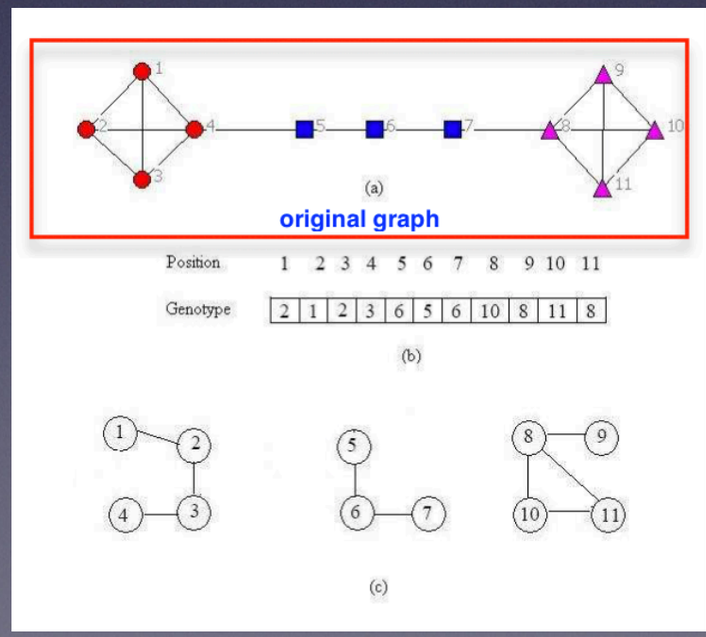
Objective Function:

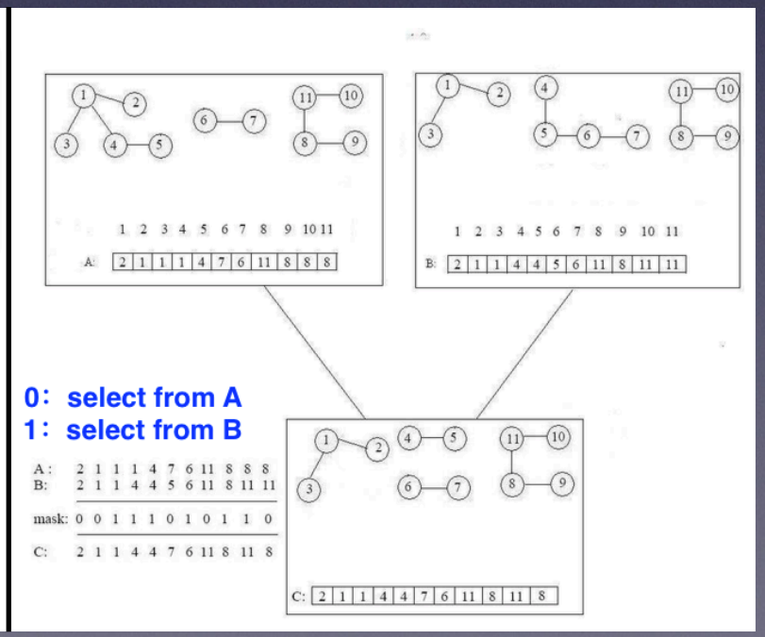
Community Score (CS)

$$CS(r) = \sum_{i}^{k} Q(Si, r) = \sum_{i}^{k} M(Si, r) * V_s \quad (10)$$

$$M(S, r) = \frac{\Sigma(a_{iJ})^r}{|I|} \quad (10.1), \quad V_s = \Sigma_{i \in I, j \in J} A_{ij} \quad (10.2)$$

$$a_{iJ} = \frac{1}{|J|} \Sigma_{j \in J} A_{ij} \quad (10.3), \quad a_{Ij} = \frac{1}{|I|} \Sigma_{i \in I} A_{ij} \quad (10.4)$$



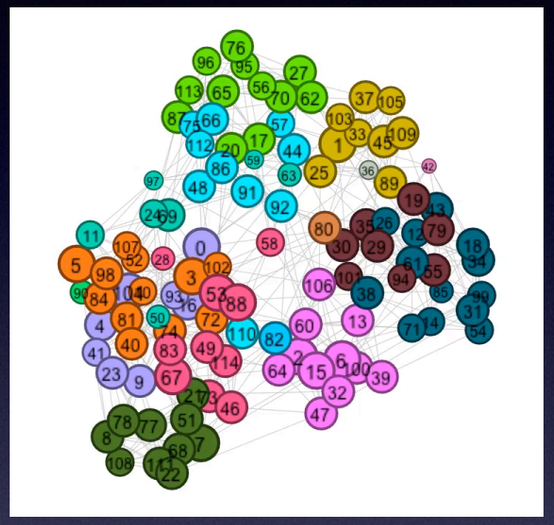


C Pizzuti (2008) Ga-net: A genetic algorithm for community detection in social networks

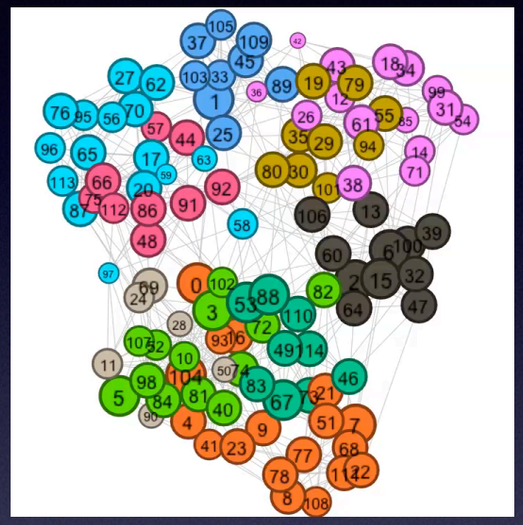
Results and Evaluation

Results of the football team dataset

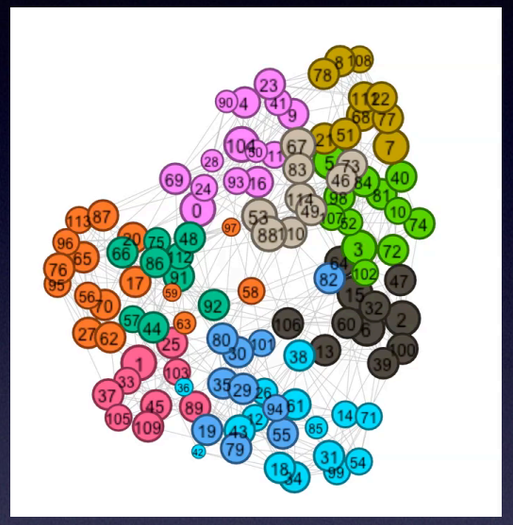
Orignal network



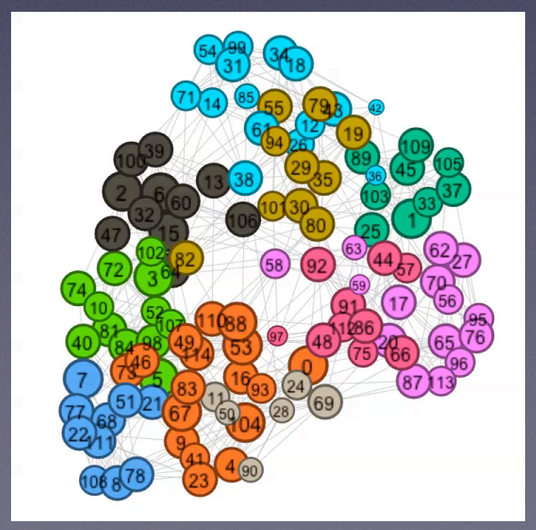
GN Algorithm



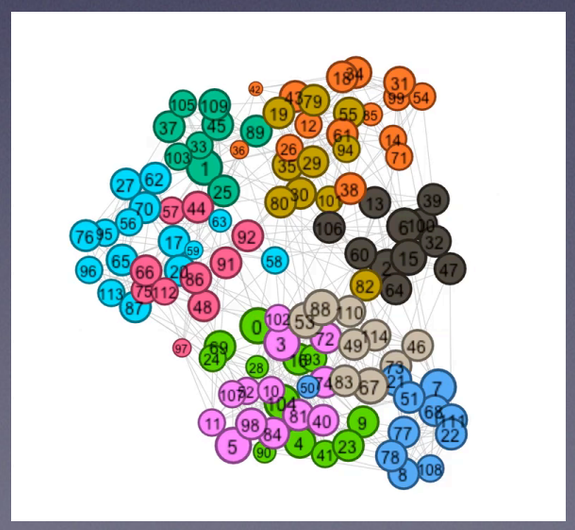
Louvain



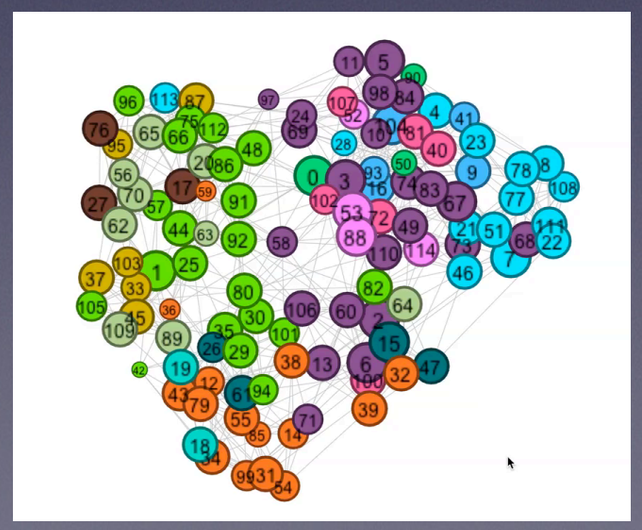
Unnormalised Spectral Clustering



Normalised Spectral Clustering



GA-Net



Results and Evaluation

Evaluation using Metrics in the football team dataset

- Internal Metrics
  - Modularity

$$Q = \sum_{i}^{n} (e_{ii} - a_i^2) \ , \ a_i = \sum_{j}^{n} e_{ij}$$

  - Internal Edge Density

$$IED = \frac{m_S}{n_S(n_S-1)/2}$$

  - Inverse Conductance

$$Inverse\ Conductance = 1 - \frac{c_S}{2m_S+c_S}$$

- External Metrics
  - Nomalised Mutual Information

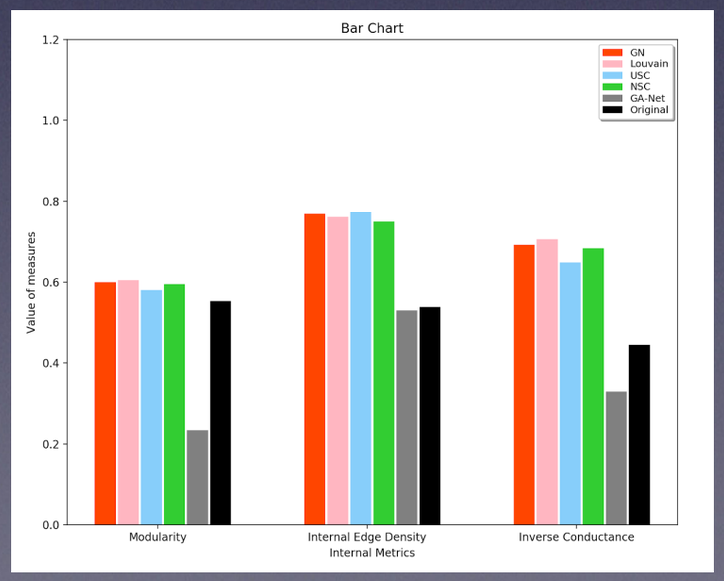
$$I(A,B) = \frac{-2\sum_{i=1}^{c_A}\sum_{j=1}^{c_B} N_{ij}\log(N_{ij}N/N_{i.}N_{.j})}{\sum_{i=1}^{c_A} N_{i.}\log(N_{i.}/N) + \sum_{j=1}^{c_B} N_{.j}\log(N_{.j}/N)}$$

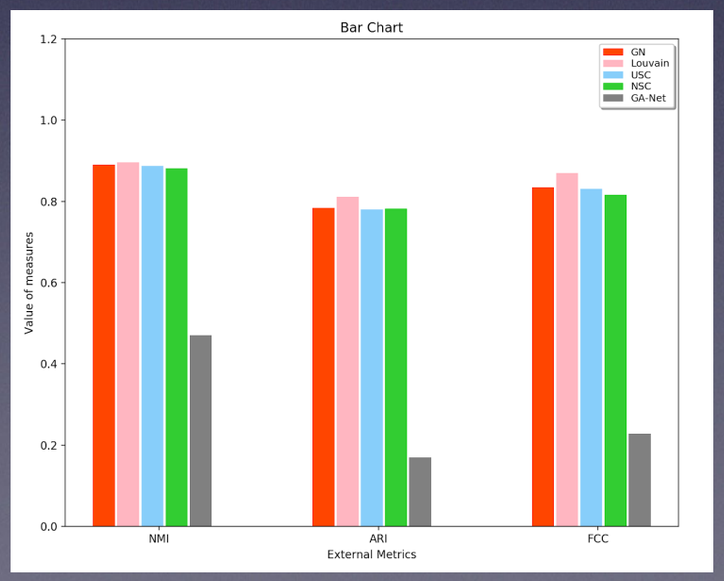
  - Adjusted Rand Index

$$ARI = \frac{\binom{n}{2}\sum_{r=1}^{R}\sum_{c=1}^{C}\binom{t_{rc}}{2} - \left[\sum_{r=1}^{R}\binom{t_{r.}}{2}\sum_{c=1}^{C}\binom{t_{.c}}{2}\right]}{\frac{1}{2}\binom{n}{2}\left[\sum_{r=1}^{R}\binom{t_{r.}}{2} + \sum_{c=1}^{C}\binom{t_{.c}}{2}\right] - \left[\sum_{r=1}^{R}\binom{t_{r.}}{2}\sum_{c=1}^{C}\binom{t_{.c}}{2}\right]}$$

  - Fraction of Correctly Classified nodes

an element in estimated community is assigned correctly if at least half neighbours assigned correctly

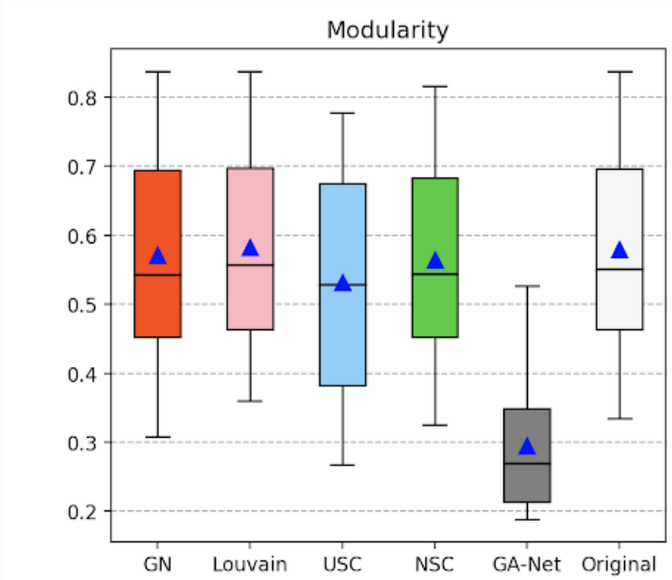


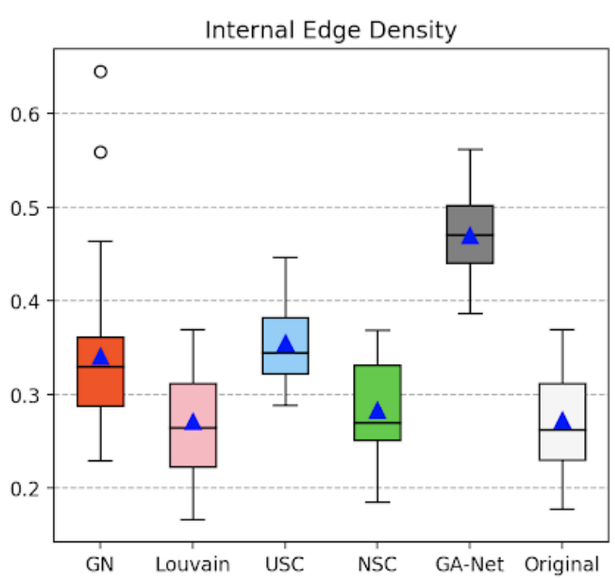


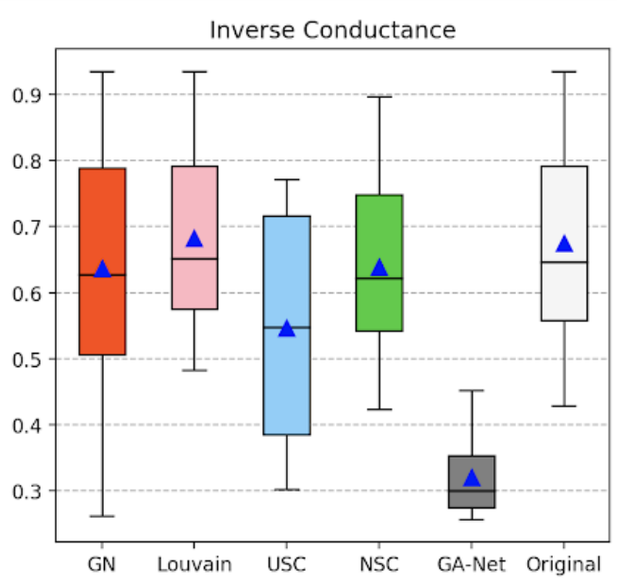
J Yang, J Leskovec (2015) Defining and evaluating network communities based on ground-truth
GK Orman, V Labatut and H Cherifi (2012) Comparative evaluation of community detection algorithms: a topological approach

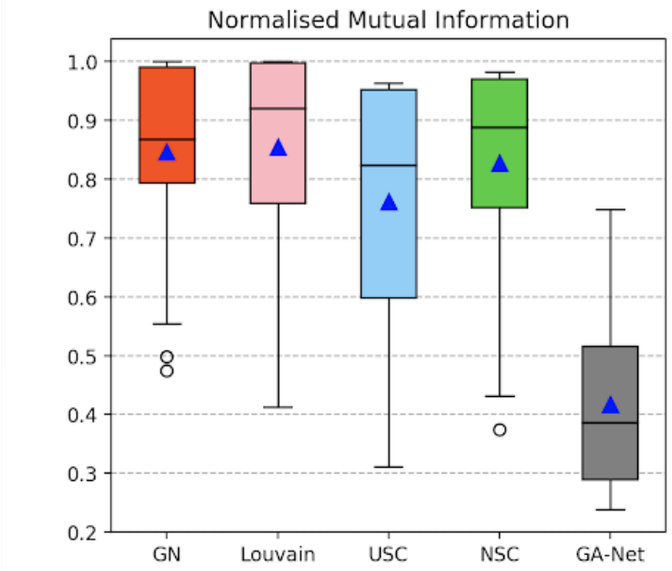
Results and Evaluation

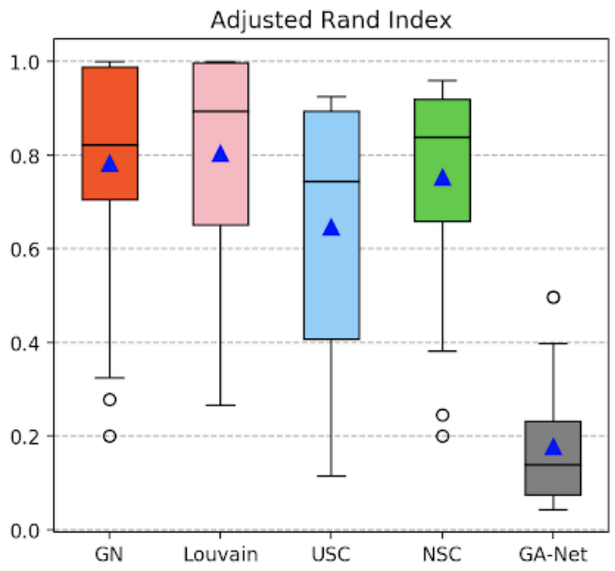
Evaluation using Metrics in several synthetic networks

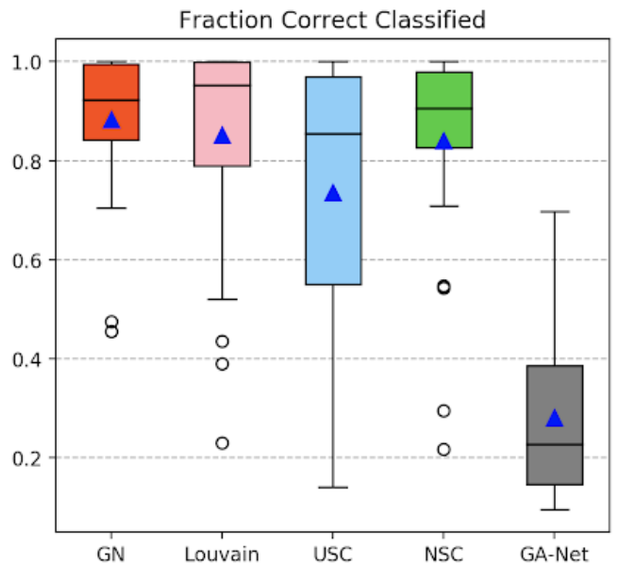












Results and Evaluation

Comparing Algorithms using plain language

Name	Advantages	Disadvantages
GN	Simple to realise; Do not need to set the number of communities	Time-consuming
Louvain	The fastest one among these; Do not need to set the number of communities	A complex algorithm
USC/NSC	Perform good, but the normalised one are better	Needs to set the number of communities
GA-Net	Do not need to set the number of communities	Has low accuracy

Results and Evaluation

Comparing Algorithms using statistical test

Oneway test (Anova, F-test)

Internal metric -Modularity

Algorithms VS Original	T values	Degree fo freedom	p-values
GN	3.8723	29	0.0005652
Louvain	-2.2519	29	0.03207
USC	9.3677	29	2.843E-10
NSC	12.686	29	2.321E-13
GA-Net	26.926	29	2.2E-16

External metric -Normalised Mutual Information

Comparison	T values	Degree fo freedom	p-values
U S C v s NSC	-6.4289	29	4.946E-07
G N v s Louvain	-0.63826	29	0.5283
Louvain vs GA	19.445	29	2.2E-16

Results from GN and Louvain are similar.

Anova for all metrics

Metrics	F values	Degree of freedom	p-values
Modularity	18.838	(5, 174)	6.027E-15
Internal Edge Density	49.872	(5, 174)	2.2E-16
Inverse Conductance	28.029	(5, 174)	2.2E-16
Normalised Mutual Information	31.271	(4, 145)	2.2E-16
Adjusted Rand Index	42.39	(4, 145)	2.2E-16
Fraction Correct Classified	44.53	(4, 145)	2.2E-16

Thanks for watching

Yawen Sun
1836786