Community Identification

Ankur Saxena[1]

[1]02CS3018 {ankur.ankursaxena@gmail.com}

I. Spectral Bisection Method

A. Langranges Matrix

Every Graph has a Langranges Matrix

$$\text{L} = \begin{cases} d_i & \text{if } i = j, \\ -1 & \text{if i is connected to j,} \\ 0 & \text{otherwise} \end{cases}$$

$$L = \begin{pmatrix} \lceil 3 & -1 & -1 \rceil & 0 & -1 & 0 \\ |-1 & 2 & -1| & 0 & 0 & 0 \\ \lfloor -1 & -1 & 3 \rfloor & -1 & 0 & 0 \\ 0 & 0 & -1 & \lceil 3 & -1 & -1 \rceil \\ -1 & 0 & 0 & |-1 & 3 & -1| \\ 0 & 0 & 0 & \lfloor -1 & -1 & 2 \rfloor \end{pmatrix}$$

Eigen vector / Eigen Value = 0

$$\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

All the derived eigen vectors are orthogonal

$v_1.v_2 = 0$

$$L = \begin{pmatrix} 1.5 & -.8 & -.6 & 0 & -.1 & 0 \\ -.8 & 1.6 & -.8 & 0 & 0 & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \end{pmatrix}$$

$$\lambda = \begin{pmatrix} 0, 0.4, 2.2, 2.3, 2.5, 3 \end{pmatrix}$$

if difference between 2^{nd} and $3^{rd}\lambda$ is high, there is clustering.

for $\lambda = 0.4$ $\begin{pmatrix} 0.2 \\ 0.2 \\ 0.2 \\ -0.4 \\ -0.7 \\ -0.7 \end{pmatrix}$

B. Kerningham - Lin Algorithm

Kerningham - Lin treats nodes as chips on a circuit board and edges as interconnections between chips. It considers two circuit boards and minimizes the number of connections across the boards.

Q = intra board connection - across board connection

Algorithm

1) Randomly divide the nodes into two groups.
2) Calculate the benefit function.
3) Consider all points, and calculate δQ if swapped.
4) Swap the one which has the maximum $|\delta Q|$ (sign not considered to avoid local maxima)
5) Repeat the steps till all the pairs are exhausted.
6) take the Q which is highest among all.

C. Wu - Huberman

This algorithm considers the network as a circuit. In this case two nodes of the network are attached to the two ends of a 1 volt battery. A threshold is

considered. If a particular node is above the threshold then it is in one community otherwise it is in other community.

It uses Kirchoff's Laws.

$\sum_{i=1}^{n} l_i = 0$

$\sum_{i=1}^{n} \frac{V_{Di} - V_c}{R} = 0$

D_i = neighbour of C

$\mathrm{V}_c = \frac{1}{n} \sum_{i=1}^{n} V_{Di}$

Probability of Random Walker reaching C = V_c

$V_A = 1$

$V_B = 0$

$V_i = \frac{1}{k_i} \sum_{j \epsilon N(i)} V_j$

$N(i) = neighborofi$

$k_i = |N(i)|$

Now, $\mathrm{V}_i = \frac{1}{k_i} \sum_{j=1}^{N} V_j a_{ij}$

$\mathrm{V}_i = \frac{1}{k_i} \left(\sum_{j=3}^{N} V_j a_{ij} + V_1 a_{ii} \right)$

$\mathrm{V}_i = \frac{1}{k_i} \sum_{j=3}^{N} V_j a_{ij} + \frac{V_1 a_{ii}}{k_i}$

$$V = \begin{pmatrix} V_3 \\ V_4 \\ \vdots \\ V_N \end{pmatrix}$$

$$B = \begin{pmatrix} \frac{a_{33}}{k_3} & \frac{a_{34}}{k_3} & \cdots & \frac{a_{3N}}{k_3} \\ \frac{a_{43}}{k_4} & \frac{a_{44}}{k_4} & \cdots & \frac{a_{4N}}{k_4} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{a_{N3}}{k_N} & \frac{a_{N4}}{k_N} & \cdots & \frac{a_{NN}}{k_N} \end{pmatrix}$$

$$C = \begin{pmatrix} \frac{a_{31}}{k_3} \; k_4 \vdots a_{N1} k_N \\ a_{41} \end{pmatrix}$$

$V = BV + C \; (1 - B) V = C$

$V = (I - B)^{-1} C$

(1-x)$^{-1} = 1 + x + x^2 + x^3 + \cdots$

V = $\sum_{m=0}^{\infty} B^m C$

let $f(v) = B_I V + C$

$f^2(v) = B(B_I + C) + C$

$f^r(v) = B^r V_I + \sum_{i=0}^{r-1} B^i C$

let $V_I = 0$

Therefore $f^r(v) = \sum_{i=0}^{r-1} B^i C$

The problem reduces to convergence of the series.

Lets assume the series converges, with constant number of steps the precision is high.

II. Random Walk Model

Divisive Clustering

A. Newman Garvan

Initial cluster - entire graph.

1) Calculate edge betweenness of each node
2) Remove the edge with highest edge betweenness
3) Recalculate

Edge Betweenness

- no of shortest path though a edge.
- amount of current passing though that edge
- Expected no of random walkers passing though that edge.

$$E = \begin{pmatrix} e_{11} & e_{12} & \cdots & e_{1m} \\ & \ddots & \vdots & \vdots \\ & & & e_{mm} \end{pmatrix}$$

Let there be m clusters

e_{ij} = no of edges between c_i and c_j

Q = $\sum e_{ii} - \delta_{random} = \sum e_{ii} - \sum e_{ii}^{rand}$

E^2 is sufficiently random

Therefore $Q = \sum e_{ii} - \sum_i \sum_k e_{ik} e_{ki}$

Calculating edge betweenness is a hard problem.

B. Radicchi

Defines the clustering coefficient of an edge.

- Remove edge with lowerst C_{ij}

$\mathrm{C}_{ij} = \frac{Z_{ij}+1}{min(k_i - 1, k_j - 1)}$

Z_{ij} = no of triangles.

$min(k_i - 1, k_j - 1)$ = min excess degree = max triangles possible.

C. Shortest Path Edge Betweenness

Algorithm

- Initialize vertex(s) $d_s = 0, w_s = 0$
- Each vertex i adjacent d_i = 1 w_i = 1
- Each vertex j adjacent to i
  1) If j has not been assigned distance then d_j = d_i + 1 , w_j = w_i
  2) If j has already been assigned distance and d_j = d_i + 1 then w_j = w_j + w_i
  3) d_i + 1 $\geq$d$_j$ then do nothing
- Repeat until finished.

Algorithm for edge betweenness

1) Find every leaf t and assign $\frac{w_i}{w_t}$ to edge (t,i)
2) For all (j,i) where j is closer to source

W$_{ji} = \frac{w_j}{w_i} + \sum_{\forall k} W_{ik}$

D. Random Walker

P = $\frac{1}{k_j}$

k_j = degree of j

P = $\frac{A_{ij}}{k_j}$

A = connection matrix

M = AD^{-1}

M = transition probability

D = diagonal matrix $D_{ii} = k_i$

$$M = \begin{pmatrix} \frac{A_{11}}{k_1} & \frac{A_{12}}{k_2} & \cdots \\ \frac{A_{21}}{k_1} & \vdots & \\ \vdots & & \end{pmatrix}$$

The path of the random walker passes through the edge u to v.

$s \rightarrow \cdots \rightarrow u \rightarrow v \rightarrow \cdots \rightarrow t$

s = start

t = terminate

M^2 = transition probability at 2 steps

M^3 = transition probability at 3 steps

t is an absorbing node

therefore remove t from matrix

M_t = matrix with t removed

M_t^n = n hops

$[M_t^n]_{us}$ = probability from s to u in n hops.

Therefore the total probability will be

P = $\frac{1}{k_u}[M_t^0 + M_t^1 + M_t^2 + \cdots + M_t^\infty]$

P = $\frac{1}{k_u}[I - M_t]^{-1}$

P = $D^{-1}(I - M_t)^{-1}S$ for all pairs

S = start node

P$_{u \rightarrow v} = [\frac{1}{k_u}(I - M_t)^{-1}]_{us} - [\frac{1}{k_v}(I - M_t)^{-1}]_{vs}$

P = forward - backward

E. Modularity

$$\begin{pmatrix} e_{11} & & & e_{14} \\ & e_{22} & & \\ & & e_{33} & \\ & & & e_{44} \end{pmatrix}$$

e_{ii} = edges within a cluster

e_{ij} = edges between 2 clusters

a_i = $\sum_{\forall k} e_{ik}$

$if(e_{ii} = a_i^2)$

P(ii) = P(i).P(i)

goodness of a cluster

$Q = \sum(e_{ii} - a_i^2)$