Project Report

on

Networking

DS745

Visualization and Unstructured Data Analysis

Prepared by

Sanzida Parvin

Date: March 29, 2019

Choosing a Dataset

For this project I have chosen LHDS Dataset from UserNetR package. This data set contains the United States Local Health Department (LHD) leadership network data. The data includes a network object 'lhds' consisting of 1,283 local health departments and the communication links between their leaders. The network is undirected, and ties are present or absent (unweighted). Attributes of the network members include: the state they are located in, whether or not they conduct HIV screening programs or nutrition programs, how many people live in the department jurisdiction, and the number of years of experience the leader has.

Vertex of lhds network:

A	B	C	D	E	F	G	H	I
	hivscreen	na	nutrition	popmil	state	vertex.nam	years	
1	Y	FALSE	Y	0.00818	AK	AK005	1	
2	Y	FALSE	Y	0.17444	AL	AL002	3	
3	Y	FALSE	Y	0.05744	AL	AL005	1	
4	Y	FALSE	Y	0.03442	AL	AL009	0	
5	Y	FALSE	Y	0.01406	AL	AL012	1	
6	Y	FALSE	Y	0.0263	AL	AL013	1	
7	Y	FALSE	Y	0.01381	AL	AL014	0	
8	Y	FALSE	Y	0.0148	AL	AL015	0	
9	Y	FALSE	N	0.05466	AL	AL017	3	
10	Y	FALSE	Y	0.01307	AL	AL018	1	
11	Y	FALSE	Y	0.03749	AL	AL028	1	
12	Y	FALSE	Y	0.1033	AL	AL029	1	
13	Y	FALSE	N	0.0308	AL	AL031	3	
14	NA	FALSE	N	0.6595	AL	AL038	2	
15	Y	FALSE	N	0.08913	AL	AL040	3	
16	Y	FALSE	Y	0.02229	AL	AL045	3	

Edges of lhds network

A	B	C	D	E
	from	to	na	
1	2	10	FALSE	
2	2	11	FALSE	
3	2	19	FALSE	
4	2	20	FALSE	
5	5	1003	FALSE	
6	5	6	FALSE	
7	6	11	FALSE	
8	6	17	FALSE	
9	10	11	FALSE	
10	11	19	FALSE	
11	11	26	FALSE	
12	2	12	FALSE	

Explanation of the Dataset

The dataset consists of the following attributes.

state - State where LHD is located

popmil - LHD jurisdiction population in millions

years - Code for number of years the current LHD leader has been in their position:

0 - 1-2 years

1 - 3-5 years

2 - 6-10 years

3 - 11+ years

hivscreen - Binary code indicating whether the LHD does HIV screening (Y/N)

nutrition - Binary code indicating whether the LHD does nutrition programming (Y/N)

I was created a subset of the original graph by taking the population greater than 40 thousand (0.04 million). But the density of the original graph is less than the density of the subset. So, I didn't use the subset rather use the full dataset for my networking project.

```
> data( lhds )
> lhds
 Network attributes:
  vertices = 1283
  directed = FALSE
  hyper = FALSE
  loops = FALSE
  multiple = FALSE
  bipartite = FALSE
  title = lhds
  total edges= 2708
    missing edges= 0
    non-missing edges= 2708

 Vertex attribute names:
    hivscreen nutrition popmil state vertex.names years

 Edge attribute names not shown
```

Community Detection

Modularity: Modularity is a measure of the structure of network, specifically the extent to which nodes exhibit clustering where there is greater density within the cluster and less density between them. The modularity statistics can range from -1/2 to +1. The closer to 1, the more the network exhibits clustering with respect to the given node grouping.

After running different community detection algorithms, it was seen that the modularity ranges from 0.580 to 0.914. The plot for different community structures are given below.

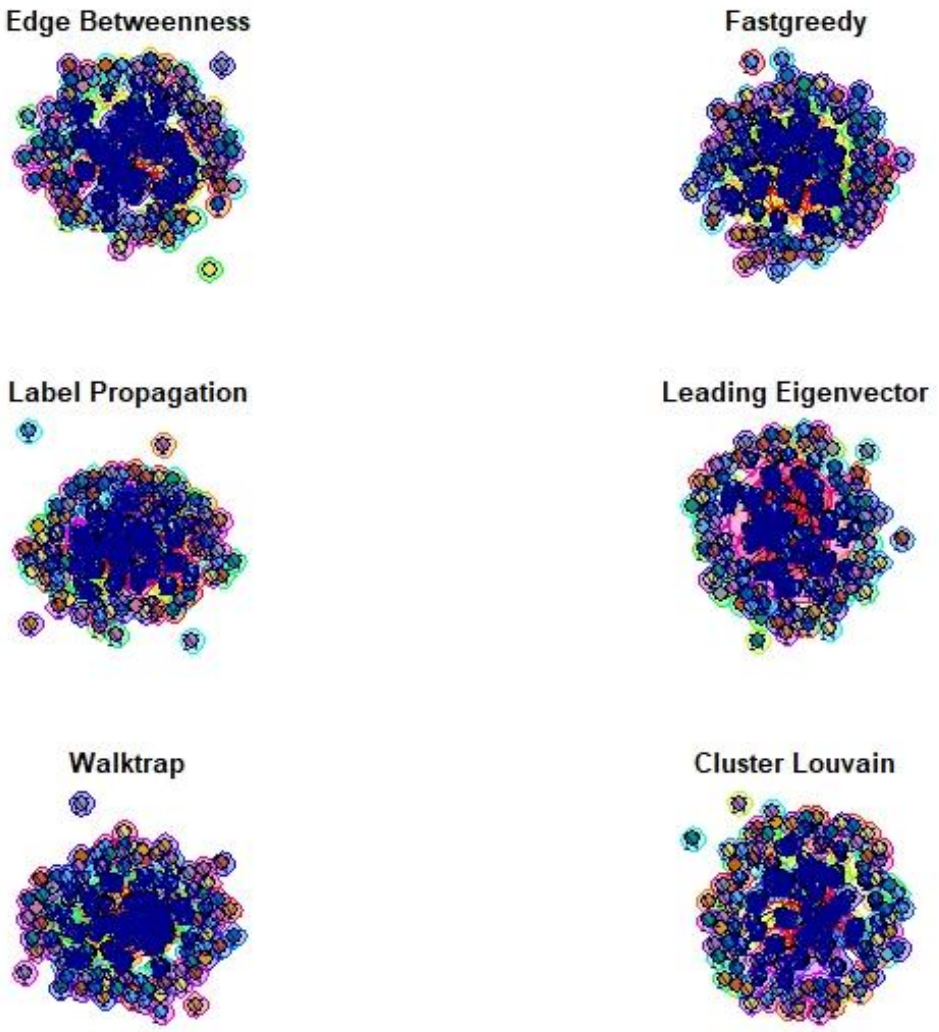


Fig: Different community structures

From the modularity of different community detection algorithms, the most closest value to 1 is for Edge Betweenness. So, this one exhibits more clustering with respect to the given node grouping.

Network Modeling

To find out the best fit network model, here I created and examined different models with different attributes. The first model was a null model, with only the 'edges' attributes.

null <- ergm(lhds ~ edges)

The second model was created after adding 'population size in millions' along with the 'edges'

popeffects <- ergm(lhds ~ edges + nodecov('popmil'))

And the third model was created using all of the attributes of the network.

```
diffhomophily2 <- ergm( lhds ~ edges +
                        nodecov( 'popmil' ) +
                        nodefactor( 'years' ) +
                        nodematch('hivscreen', diff=T, keep=2) +
                        nodematch('nutrition', diff=T, keep=2) +
                        nodematch('state') )
```

Here the "nodecov" and "nodematch" terms were used to add continuous variables such as "population size" (other example would be income, height, weight etc.) and categorical variables such as "years" (hivscreen, nutrition) respectively.

The simulation statistics of the given models are shown below

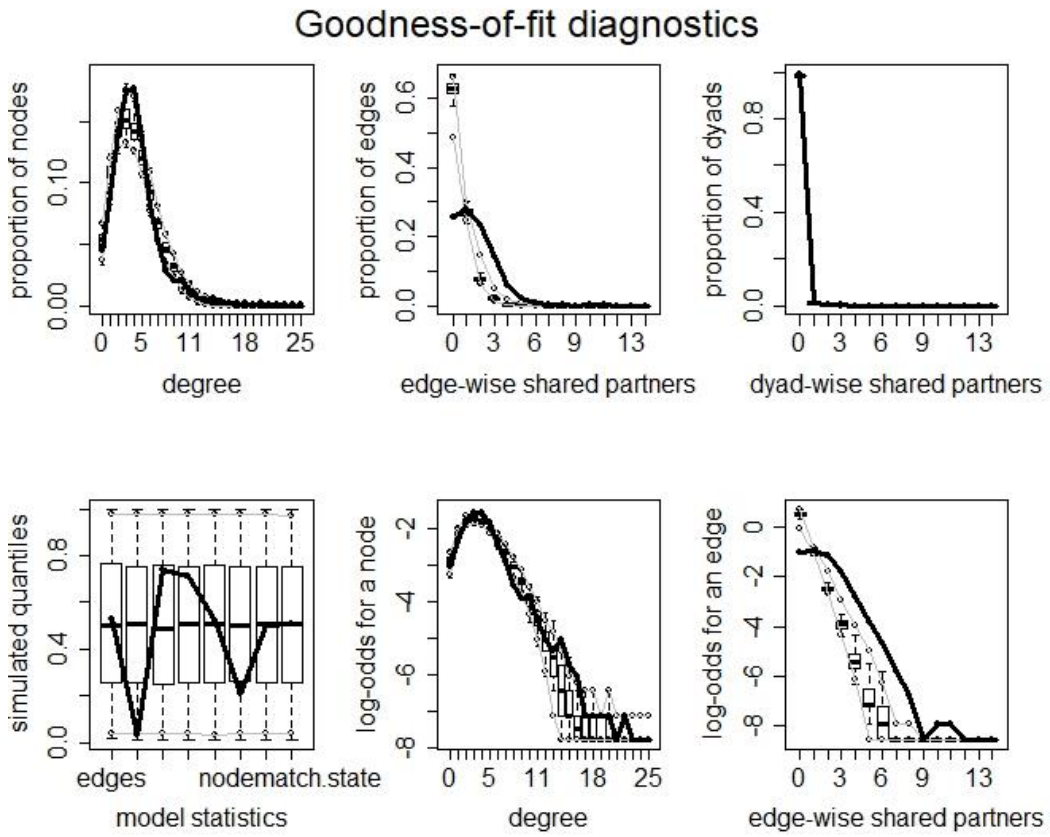
```
                edges degree0 degree1 degree2 degree3 degree4 degree5 triangle
lhds             2708      58     117     182     223     226     172     1437
Null             2628      21      86     163     253     281     200        9
Main effects     2686      12      92     174     224     251     227       19
Diff homophily 2 2679      42     119     195     222     234     171     1123
> |
```

Simulation statistics shows that the diff homophily2 model performed much better than other two models so far. Overall edges, and number of nodes with one degree, two-degree, three degree, four

and also five degree were well matched with the original network. Number of triangles is also very similar to the original network. Also, the lower AIC value of diff homophily2 model indicates the better fit for this model.

Goodness of Fit Plots

The dark line needs to be within the gray lines to indicate goodness of fit. Gray lines indicate the confidence interval. The plot shows that degree terms have been modeled well, but there are some issues with edge-wise shared partners, and dyad-wise shared partners.



References:

1. Luke, D. A. (2016). *A User's Guide to Network Analysis in R*. Springer.

2. Course materials of DS745 – Visualization and Unstructured Data Analysis

3. Google and piazza discussion board