Identification of Neutral Members in Social Networks using a Distance and Fuzzy based Model

A. Galappaththi
Postgraduate Institute of Science
University of Peradeniya
Peradeniya

M. A. P. Chamikara
Postgraduate Institute of Science
University of Peradeniya
Peradeniya

Y.P.R. D. Yapa
Department of Statistics and Computer Science
University of Peradeniya
Peradeniya

S. R. Kodituwakku
Department of Statistics and Computer Science
University of Peradeniya
Peradeniya

J. Gunatilake
Department of Geology
University of Peradeniya
Peradeniya

ABSTRACT

Understanding community structure helps to interpret the role of actors in a social network. Actor has close ties to actors within a community than actors outside of its community. Community structure reveals important information such as central members in communities and bridges members who connect communities. Clustering algorithms like hierarchical clustering, affinity propagation, modularity and spectral graph clustering had been applied in social network clustering to identify community structures in it. This study proposes a novel method for distance measurement between nodes and centroids. Distance is measured based on the shortest path length and number of common nearest neighbors with one path length. This measure, "Proportional closeness" is used to assign nodes to the closest centroid. A fuzzy system is also applied to find the closest centroid to a node when similar proportional closeness values are present for multiple centroids. The method has been applied to two artificial networks and one real world network data to test its accuracy on membership identification. The results revealed that the method successfully assigns members to its nearest centroid and leave neutral members aside without assigning to any centroid.

General Terms

Membership identification, Closeness

Keywords

eigenvector centrality, centroids, fuzzy system, proportional closeness, fuzzy closeness

1. INTRODUCTION

This study is focused on membership of entities in sub-communities of networks. In social networks, nodes represent actors and edges represent interaction between them [1]. Edges can be directed or undirected, depending on the representation of the interaction among actors [1]. For example, a network displays telephone calls between actors, directed edges show who initiate the calls. On the other hand, an undirected edge in a social media network represents simply a friendship between two actors. Though self-loops are also possible in graphs and networks, they are seldom considered in social networks.

The network structure shows sub communities inside the network and their interactions. In general, it can be observed that link densities are comparatively high within the groups than in-between the groups [2].The variation in link densities in the network shows the varied interests of individual actors and communities that they are part of. Different community structures can be observed in sparse networks and it is very interesting to study them compared to dense networks. For example, in a fully connected network each actor has direct edges to all others in the network. Therefore, every actor plays a similar role in the network. Some cases, actors in the network may be crucial to the information flow of the network. For example, information flow in a bowtie network with seven actors as shown in Figure 1 and it strictly depends on the participation of actor 4. Though symmetric networks like bowtie or hundred percent identical actors are not commonly available in real word [3]. The purpose of this study is to find identical behaviours of actors and find the actors who fall within the boundaries of communities in a network.

1.1 Related Work

Graph clustering algorithms are applied to find communities in many social network analysis studies. Hierarchical clustering [4, 5], affinity propagation [6] and spectral graph clustering [7] are popular methods for clustering of networks. Modularity of a network, proposed by Newman and Girvan [8] finds optimal community structure for networks based on the network topology. The model compares the edge placement with the null model of the graph to find the modularity. Then the network is partitioned into communities as it maximizes the modularity. Modularity also contains similar properties of spectral graph clustering as rows and columns of the modularity matrix and the Laplacian matrix sum up to zero [2]. Hierarchical clustering [4, 5] requires the measuring of similarity between nodes presented in the network. Hierarchical agglomerative clustering method was applied instead of divisive clustering method, where each node is a cluster at the beginning and it is merged till all become one large cluster. Number of clusters can be determined by choosing a cut point in the dendrogram. However, this algorithm doesn't provide good results on real world network as presented by Newman in [9] for Zachary's karate club network [10]. Community detection by edge removal requires to find edge betweenness and remove the edge with the highest betweenness value. However, betweenness need to be recalculated because once an edge is removed, the network topology is changed [9]. Clustering by message passing proposed by Frey and Dueck [6] which also requires a similarity measurement between nodes in the network and an additional preference value for each node. The preference value has a direct effect on the number of clusters returned by the algorithm.

One of the limitations of the methods mentioned above is the failure to leave neutral members in network without assigning them to a centroid. Therefore, a novel method is required to assign nodes to central members while leaving neutral members intact.

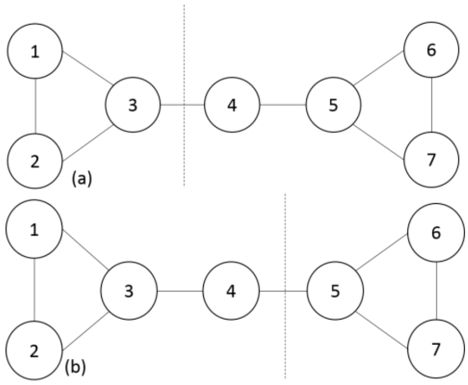

Figure 1: Bowtie network with equally weighted edges partitioned into two communities (a) community structure given by spectral graph clustering (b) community structure identified with the maximum modularity

2. METHOD PROPOSED

To describe the problem, three graph partitioning algorithms were applied on a bowtie network with seven nodes and equally weighted eight edges to identify possible community structures. To demonstrate the above limitation of the existing clustering algorithms; modularity, spectral graph clustering, hierarchical clustering, community detection by edge removing were applied to artificial networks shown in Figure 1 and Figure 2. Modularity method [2] founds first four nodes into one community and the rest to the other community while spectral graph clustering [7] portioned the graph's first three nodes into one and the rest to the other. Community detection by removing edges [8] also requires to choose which edge should be removed since both 3-4 and 4-5 edges received the similar values for edge betweenness, which received result similar to modularity of spectral graph clustering based on the selection of the edge to remove. Complete link hierarchical clustering algorithms produced results similar to the spectral graph clustering algorithm, partitioning nodes 1 to 3 and nodes 4 to 7 into separate communities. The distances for hierarchical clustering was computed by taking the reciprocal of shortest path length among edges. Affinity propagation also clustered node 4 with nodes 1, 2 and 3 while nodes 5, 6 and 7 were selected to other cluster.

It was also found that with weighted edges, nodes are attracted to the community with the highest weight. Therefore, it is required to find a way to interpret the role of actors who has symmetric interactions with communities.

To address the inability in methods mentioned above, in finding neutral members in a network, additional feature "proportional closeness" is extracted from the network in addition to the basic centrality measures: degree; betweenness and closeness [1]. Given a network, it is required to find central members of the network and calculate the distance from each central member to every node of the network. Then this will provide a measure indicating how close each node to the central members of the network. Eigenvector centrality measure has been used to find the centrality of each node in the network. The threshold has been determined by the maximum slope of the sorted eigenvector centrality measures to separate most of the central members of the network. For the Zachary's karate club network [10], nodes 1, 3, 33 and 34 have been found as the most central members of the network. For the bowtie network, nodes 3 and 5 have been found as the central members. To find the strength of the proposed method, an artificial network has been created by adding three more nodes to the fourth node in bowtie network (Figure 2). Shortest path length of each node from central nodes, betweenness, closeness, degree and eccentricity have been extracted from the network as features for the model. Then a new proportional closeness value has been calculated to each centroid by the following model.

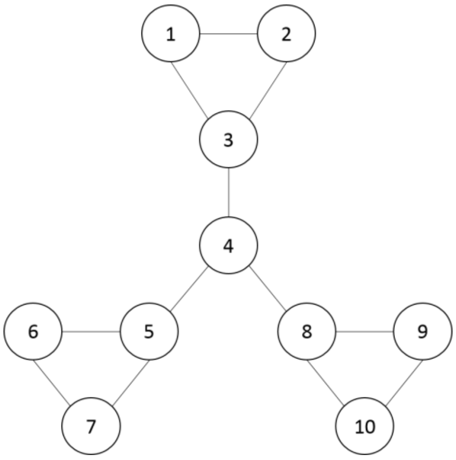

Figure 2: Artificial network created by adding three more nodes to fourth node in bowtie network

2.1 Proportional Closeness

$$D_n^{(i)} = \frac{1}{d_n}\left(N^1_{l\in\{i,n\}} - \frac{1}{c-1}\sum_{j\neq i} N^1_{l\in\{j,n\}}\right) - \left(L_n^{(i)} - \frac{1}{c-1}\sum_{j\neq i} L_n^{(j)}\right) \quad (1)$$

Table 1. Symbols and notations

Symbol	Description
d_n	Degree of node n
$N^1_{l\in\{i,n\}}$	Number of common nearest neighbors for node n and centroid i with 1 path length away
c	Number of centroids
$L_n^{(i)}$	Shortest path length from node n to centroid i
$D_n^{(i)}$	Proportional closeness from node n to centroid i

Equation 1 measure the proportional closeness and Table 1 describes symbols involved in the equation. The proportional distance for node *n* to centroid *i* is calculated by considering the normalized shortest path length to each centroid and number of nearest neighbours common to both node n and centroids which are in one path length distance. When considering the shortest path length (where is the shortest path length from node n to centroid *i*), to remove the effect of other centroids *j*, the average shortest path length of all *j* from (c is number of centroids) the shortest path length to the centroid *i* is subtracted. Then we consider the number of common neighbours with one path length away (where is count of common nearest neighbours *l* to both *i* and *n*, which are one path length away) from both actor and the centroid. Similarly we remove the effect of common neighbours with actor to other centroids by taking average of counts. To remove the effect of degree to the count, it was divided by the degree. When number of nearest neighbors for a centroid and a node is higher compared to other centroids that implies the particular node is close to the particular centroid compared to others. Therefore, count of nearest neighbors positively effect to the proportional closeness. However, shortest path length is negatively affected since a node is more close to a centroid it has a shorter length compared to length to other centroids. Then *n* is assigned to centroid *i* for maximum value of proportional closeness.

2.2 Fuzzy Closeness

"Proportional closeness" measurement sometimes generates equal values for the same node for several centroids. This raises an ambiguity in assigning the corresponding node into a particular centroid. To overcome this, a fuzzy model was generated using the Matlab fuzzy logic toolbox [11], in which "Proportional closeness" and the centrality of the corresponding centroid are taken as the input parameters. The model then generates a generalized value for the closeness as marks (fuzzy closeness) as in Equation 2.

$$Fuzzy\ Closeness = \frac{\int_0^1 \mu_z\left(\mu_D(x) * \mu_{Centrality}(y)\right) zdz}{\int_0^1 \mu_z(z)dz} \qquad (2)$$

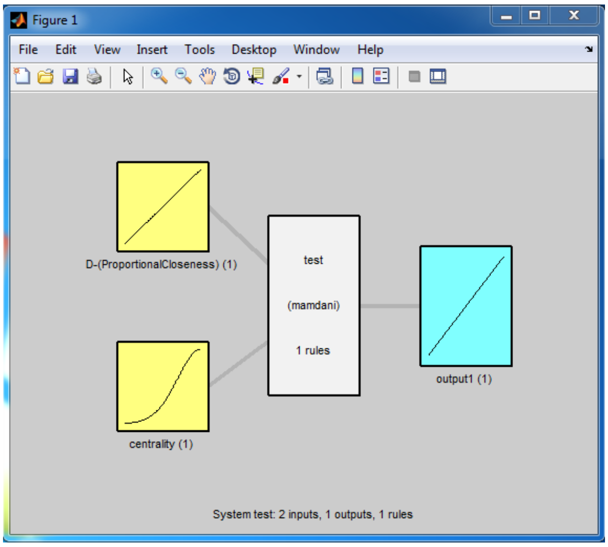
Figure 1
File Edit View Insert Tools Desktop Window Help
D-(ProportionalCloseness) (1)
test
(mamdani)
1 rules
output1 (1)
centrality (1)

System test: 2 inputs, 1 outputs, 1 rules

Figure 3: The model takes two inputs; D (proportional closeness) and centrality measurement. The output generates a fuzzy closeness value as marks. The model uses Mamdani fuzzy inference to generate the output.

Figure 3 shows the developed fuzzy model in which Mamdani [12] method is used as the fuzzy inference technique. Input "D" is modelled as a linear membership function while centrality is modelled as a "Sigmoid right" membership function. The output is modelled as a linear membership function.

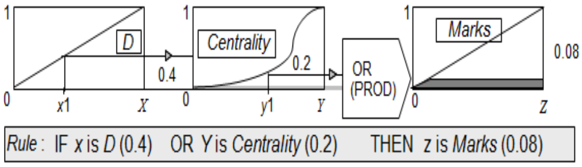
1
D
0.4
0
x1
X
1
Centrality
0.2
0
y1
Y
OR
(PROD)
1
Marks
0.08
0
Z
Rule : IF x is D (0.4) OR Y is Centrality (0.2) THEN z is Marks (0.08)

Figure 4: Fuzzy rule evaluation by the model. The instance shows x_1 = 0.3, y_1 = 0.6

The one and only one rule of the rule base basically provides the "proportional closeness", a biasness towards the centroids with higher centrality values as a "fuzzy closeness" measurement. Figure 4 shows the rule evaluation of the fuzzy model.

Finally the fuzzy closeness value is obtained using the above equation which is applied on the output shape obtained from the alpha cut as shown in Fig. 4. After modeling the inference engine of the developed fuzzy model, the rule surface shown in Figure 5 is generated. The rule surface provides a clear idea of how the output value has varied against the two inputs: centrality and d (proportional closeness). It is very clear that when the two inputs have values close to 1 the effect of fuzzy closeness also comes close to one, letting the corresponding node to be biased towards the centroid in consideration.

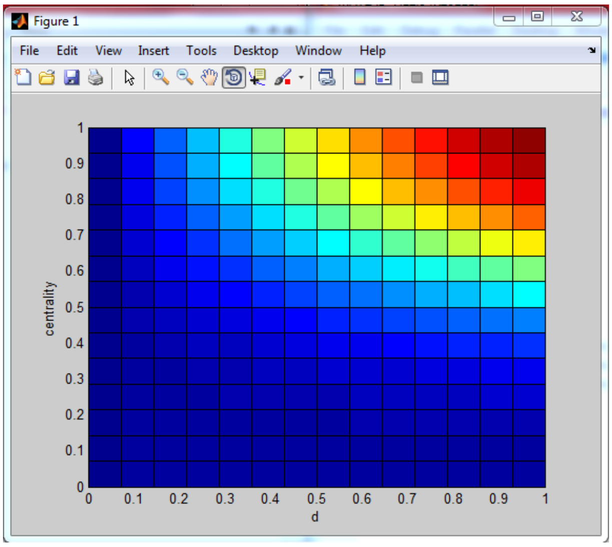

Figure 5: This is a two dimensional view of the rule surface represents the variation of the output value against the two inputs in which the output value varies between 0 to 1 from dark blue to dark brown.

3. RESULTS AND DISCUSSION

Three networks bowtie (Figure 1), extended bowtie (Figure 2) and Zachary's karate club have been analysed using the Matlab Tools for Network Analysis package [13] in Octave [14]. For the bowtie network, node 4 has no attraction to either of two centroids. For the network in Fig. 2, four central members-node 3, 4, 5 and 8 have been identified by the eigenvector centrality. When the centroid 4 was isolated, all other centroids received two directly connected nodes who are not centroids. Out of curiosity we test the network with only three centroids by removing the centroid 4. All three centroids attracted the same node sets as in test with four centroids except node 4 has no attraction to either of centroids. The karate club network has been analysed separately with four centroids and two centroids. When only actors 1 and 34 are considered for centroid all other actors correctly attracted to them as depicted in Figure 6. More comparative results are displayed in Table 2 including original faction membership and subgroup membership of each node returned by the model proposed in the paper. When all four actors: 1, 3, 33 and 34 are considered as centroids, actor 3 was isolated from actor 1 and other actors previously attracted to centroid 1. Actor 9 was attracted to centroid 33 while rest of the actors resided with centroid 34. This method returns the same value for proportional closeness to both centroids for node 4, in the bowtie network mentioned in earlier. With the fuzzy model the result of similar values is omitted. For example, the "Proportional Closeness" for the 19th instance (node) generated are equal to 0.3559, 0.49152, 0.4915 and 0.4915 for the centroids 1, 3, 33 and 34 respectively. Therefore, there was a problem of assigning the 19th node in to one of the four centroids. With the fuzzy model it generates the values 0.5123, 0.5223, 0.5218 and 0.5258 for the centroids 1, 3, 33 and 34 respectively. From the fuzzy closeness values, it is clear that the 19th node should be close to the centroid 34 in which the highest closeness value has resulted.

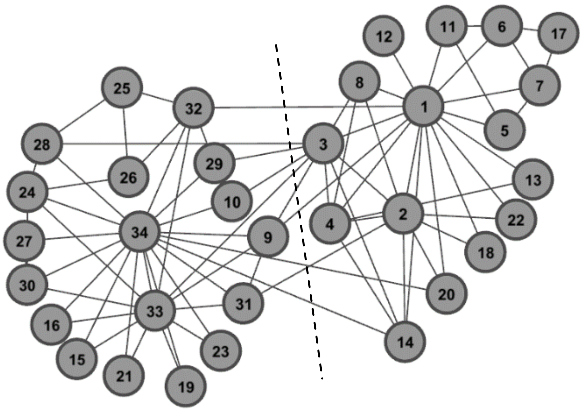

Figure 6: Zachary's karate club network. Node sizes are varied according to the degree of node. Dashed line displays the sub-groups found by proportional closeness of nodes to centroids 1 and 34

This work assumed that the networks is undirected and there are no self-loops. Further for large sparse networks, eigenvector centrality doesn't provide acceptable values to find centroids. Therefore, more general method is required to find central members prior to membership identification.

Table 2. Faction membership of karate club and subgroup membership returned from the model proposed in the paper

Faction membership from reference [10]		Results returned from the method proposed in the paper for centroids 1 and 34	
Mr. Hi	John	Node 1	Node 34
1	9	1	9
2	10	2	10
3	15	3	15
4	16	4	16
5	19	5	19
6	21	6	21
7	23	7	23
8	24	8	24
11	25	11	25
12	26	12	26
13	27	13	27
14	28	14	28
17	29	17	29
18	30	18	30
20	31	20	31
22	32	22	32
	33		33
	34		34

4. CONCLUSION

The method proposed in this study clearly identifies memberships of nodes to central member based on the interactions presented in the network. For the karate club network, the method correctly identified members attracted to two central members (manager and instructor) after the separation of the club, as presented in reference [10]. Fuzzy closeness measurement generated a generalized value for the closeness of a particular node to a particular centroid. The computational complexity of the method is depends on number of nodes in the network (n) and number of centroids (m) input to the algorithm resulting $O(nm)$ time complexity. Comparatively this is a very low value compare to algorithms run with time complexity of square of number of nodes and number of iterations to converge times the square of number of nodes. When a node is symmetric it shows no closeness to any centroid, which is the prime objective of this study.

The model accounts common nearest neighbours with only one path length. However for large networks with long network diameter, members may have more than one transitive nodes between them and the centroid. Therefore, as a future direction of the study, common nearest neighbours with more than one path length can also be considered in calculating the proportional distance.

5. ACKNOWLEDGMENTS

This work was funded by the National Research Council (NRC) of Sri Lanka [Grant No: 11-071]

6. REFERENCES

[1] Hanneman, R. A., & Riddle, M. 2005. Introduction to social network methods.

[2] Newman, M. E. 2006. Modularity and community structure in networks. Proceedings of the National Academy of Sciences, 103(23), 8577-8582.

[3] White, H. C., Boorman, S. A., & Breiger, R. L. 1976. Social structure from multiple networks. I. Blockmodels of roles and positions. American journal of sociology, 730-780.

[4] Xu, J. J., & Chen, H. 2005. CrimeNet explorer: a framework for criminal network knowledge discovery. ACM Transactions on Information Systems (TOIS), 23(2), 201-226.

[5] John Scott, & Peter J. Carrington (Eds.). (2011). The SAGE handbook of social network analysis. SAGE publications.

[6] Frey, B. J., & Dueck, D. 2007. Clustering by passing messages between data points. Science, 315(5814), 972-976.

[7] Fiedler, M. 1973. Algebraic connectivity of graphs. Czechoslovak Mathematical Journal, 23(2), 298-305.

[8] Girvan, M., & Newman, M. E. 2002. Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99(12), 7821-7826.

[9] Newman, M. E. (2004). Detecting community structure in networks. The European Physical Journal B-Condensed Matter and Complex Systems, 38(2), 321-330.

[10] Zachary, W. W. 1977. An information flow model for conflict and fission in small groups. Journal of anthropological research, 452-473.

[11] MATLAB version R. 2011aFuzzy logic toolbox version Natick, Massachusetts: The MathWorks Inc., 2010.

[12] S. N. Sivanandam, S. Sumathi and S. N. Deepa. Introduction to Fuzzy Logic using MATLAB, 2007, 120-131

[13] Gergana. 2014. Octave Networks Toolbox First Release. ZENODO. http://doi.org/10.5281/zenodo.10778

[14] John W. Eaton, David Bateman, S├©renHauberg, RikWehbring 2014. GNU Octave version 3.8.1 manual: a high-level interactive language for numerical computations. CreateSpace Independent Publishing Platform. ISBN 1441413006,URL http://www.gnu.org/software/octave/doc/interpreter/