Communications of the Korean Statistical Society
2011, Vol. 18, No. 2, 213–217

DOI: 10.5351/CKSS.2011.18.2.213

Local Centers of the Social Network

Myung-Hoe Huh[1,a]

[a]Department of Statistics, Korea University

Abstract

For the social network of n nodes, one might be interested in finding k nodes to disseminate the information as quickly as possible or to identify key nodes of high "local centrality". I propose two algorithms for determining k "local centers" of the network and work on a real case.

Keywords: Social network analysis, neighborhood, closeness, local centers.

1. Background and Aim

Social network is a social structure made up of individuals called nodes, which are connected by one or more specific types of interdependency, such as friendship, kinship, common interest, financial exchange, dislike, sexual relationships, or relationships of beliefs, knowledge or prestige (*Wikipedia* "social network", retrieved 2010/11/05). In most social networks, a number of nodes are more central than the others in various respects such as the closeness, the betweenness and the power (Wasserman and Faust, 1994). Huh's (2010) booklet is a readable introduction to the social network analysis.

Rather than detecting a batch of central nodes which are often clustered together, one may be interested in finding out several nodes that are "locally" central. For instance, Figure 1 shows the network in which five stars of five vertices are connected to a star of five vertices. In this simple case, the local central nodes are 7, 8, 9, 10 and 11. The aim of this paper is to identify the local central nodes of a somewhat complex network.

The usages of local centers are 1) the identification of the overall structure of the network and 2) the speedy delivery of the message or the goods to the most part of the network.

In Section 2, I propose two algorithms for determining k local centers. In Section 3, I work on a real case of the network of 439 nodes. In Section 4, I remark on the issue of choosing k and the merit of local centers, compared to the use of global centers.

2. Two Algorithms for Determining k Local Centers

Here, I will propose two algorithms for determining k local centers of the network, of which the nodes are labeled by $1, \ldots, n$.

2.1. Maxmin algorithm

As a modification of K-means clustering algorithm, k local centers of the network of n nodes can be determined by the following algorithm:

This research was supported by a Korea University Grant during the academic year 2010.

[1] Professor, Department of Statistics, Korea University, Anam-Dong, Sungbuk-Gu, Seoul 136-701, Korea.
E-mail: stat420@korea.ac.kr

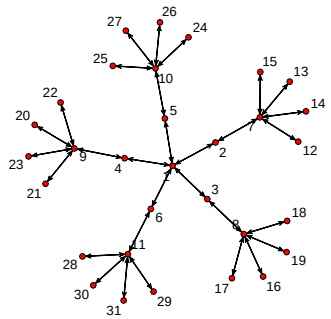

Figure 1: *Star-of-stars network*

0) Select k nodes for initial local centers $L_1, \ldots, L_k$ by the maxmin algorithm, which will be written shortly.

1) For each node $i\,(= 1, \ldots, n)$, find the closest local center $c(i)$ by searching $L_1, \ldots, L_k$ for

$$\min \{d(L_1, i), \ldots, d(L_k, i)\}.$$

Then n nodes are partitioned into k subgroups $G_1, \ldots, G_k$ of which the member nodes are attracted to $L_1, \ldots, L_k$, respectively.

2) For each subgroup $G_j\,(j = 1, \ldots, k)$, update the local center L_j by the node c closest to all member nodes of G_j, in the sense that

$$\max_{c \in G_j} \sum_{i \in G_j,\, i \neq c} \frac{1}{d(c, i)}.$$

3) Repeat Step 1 and Step 2 until the list of k local centers does not change.

4) Compute the harmonic average of interdistances within subgroups by

$$D_{avrg} = \left[\frac{1}{n-k} \sum_{i \neq L_1, \ldots, L_k} \frac{1}{d(c(i), i)} \right]^{-1}.$$

For Step 0 of the list of initial local centers, I use the "maxmin" algorithm:

1) Choose one node at random. Denote this node by L_1.

2) Find the node L_2 that is located farthest from L_1. Thus L_2 is set to c by

$$\max_c d(L_1, c).$$

3) For $j = 3, \ldots, k$, find the node c sequentially by

$$\max_c \min \{d(L_1, c), \ldots, d(L_{j-1}, c)\},$$

and set $L_j = c$.

When the algorithm is applied to the network of Figure 1 with $k = 5$, the nodes 7, 8, 9, 10, 11 (or 1, 7, 8, 9, 10) appear as the local centers with $D_{avrg} = 1.02$ (or 1.14), depending on the first seed that is chosen randomly by the algorithm.

2.2. δ-Neighborhood algorithm

Suppose that the message from the initial node can be reached to its connected nodes provided the distance between two nodes is less than or equal to δ. The objective is to find k nodes to maximize the spread of message to the whole network. I propose the following algorithm:

1) Select the first node L_1 among the nodes c which achieves

$$\max_c \sum_{i=1}^{n} I\,[\,d(c,i) \le \delta\,].$$

Discard the nodes that can be reached from L_1 within distance δ and retain the other nodes of the network. In subsequent steps, the step number is denoted by j and the number of nodes retained is denoted by n_j. Set the step number j equal to 2.

2) Select the next node L_j among the nodes c which achieves

$$\max_c \sum_{i=1}^{n_{j-1}} I\,[\,d(c,i) \le \delta\,].$$

Discard the nodes that can be reached from L_j within radius δ and retain the other nodes of the network. Increase the step number j by 1.

3) Repeat Step 2 while the step number j is less than or equal to k and $n_{j-1} \ge 1$.

4) Output the number of nodes finally retained in the network and compute the coverage which is defined by the ratio of the number of nodes reached from $L_1, \ldots, L_k$ with the distance less than or equal to δ over the total number of nodes in the network.

When the algorithm is applied to the network of Figure 1 with $\delta = 1$ and $k = 5$, the nodes 7, 8, 9, 10, 11 appear as the local centers with the number of unreached nodes n_k equal to 1.

3. The Case of Faux Magnolia High Network

The Faux Magnolia High(FMH) Network is obtained by surveying friend relationship among 1461 students of a high school located in the southern United States (Goodreau *et al.*, 2008). FMH data is available in `ergm` package of open-source statistical language `R` (Hunter *et al.*, 2008). In the network, the largest component of students communicating each other consists of 439 students. Hereafter, I will apply two algorithms of Section 2 to the subnetwork of FMH. See Figure 2. I start with $k = 10$.

3.1. Maxmin local centers

Out of ten trials of the algorithm with $k = 10$, it turns out that the students with ID's 21, 63, 122, 356, 677, 991, 1009, 1019, 1064, 1263 are local centers with $D_{avrg} = 2.6$, the harmonic mean of the distances between the local center and its membership nodes. In Figure 2, maxmin local centers are colored in red. By setting $k = 5$, the harmonic mean of interdistances within subgroups increases to 3.7. For $k = 20$, the harmonic mean of interdistances decreases to 2.0.

3.2. δ-Neighborhood local centers

For $\delta = 7$ and $k = 10$, it turns out that the students with ID's 2, 294, 479, 677, 769, 914, 1142, 1170, 1358, 1385 are local centers with the coverage 94.5%. For $\delta = 4$ and $k = 10$, the coverage drops to 73.3%. The coverage returns to 94.3% by increasing k to 20.

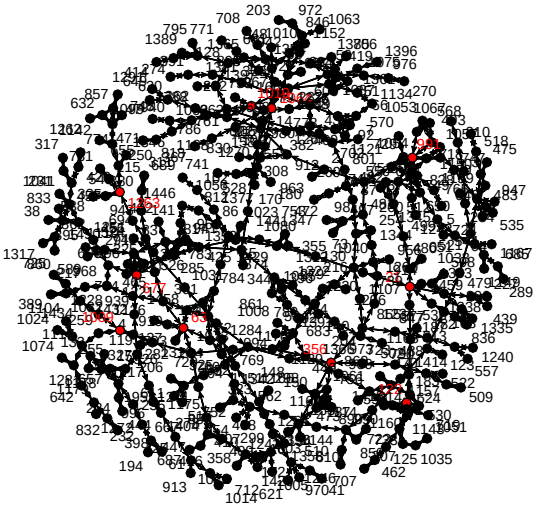

Figure 2: *The largest component of Faux Magnolia High(FMH) network*

4. Remarks

Finding k local centers in the network and k centroids in the multivariate dataset share several common features such as the maxmin seeding of initial group centers (centroids) and allocation of the nodes (observations) to the nearest group center (centroid). They are not the same, however. Although determining the number k is a very complex matter in K-means clustering, the choice of k in social networks, as seen in Section 3, is directly linked to the average interdistances between nodes within subgroups or the coverage with neighborhood radius of δ. Therefore, one may set k to keep the interdistances below a certain level or to secure the coverage above a certain level.

Conventional analyses of social networks focus on the detection of global k nodes with the largest centrality index values (Freeman, 1979). For the largest component of Faux Magnolia High network, $k(= 10)$ groups induced by the global centers inflate the average interdistance to 3.1 (larger than 2.6 of 10 subgroups by the local centers). On coverage rate, the difference is more drastic: it is 80.5% by 10 global centers, much less than 94.5% by 10 local centers.

One may adopt time-proven hierarchical clustering algorithms such as complete linkage or single linkage in clustering n nodes of the network, followed by identification of the central nodes in respective clusters. Such methods may work well in most cases, however, it should be noted that they are not addressing the issues raised in this study directly. Of course, traditional clustering methods can be used only for *undirected* networks.

References

Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification, *Social Networks*, **1**, 215–239.

Goodreau, S. M., Handcock, M. S., Hunter, D. R., Butts, C. T. and Morris, M. (2008). A statnet tutorial, *Journal of Statistical Software*, **24**, 1–26.

Huh, M. H. (2010). *Introduction to Social Network Analysis using R*, Freedom Academy, Seoul. (Written in Korean)

Hunter, D. R., Handcock, Butts, C. T., M. S., Goodreau, S. M. and Morris, M. (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks, *Journal of Statistical Software*, **24**, 1–29.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*, Cambridge University Press.

Received November 2010; Accepted January 2011