ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413940

CENTRALITY BASED NUMBER OF CLUSTER ESTIMATION IN GRAPH CLUSTERING

Mahdi Shamsi, Soosan Beheshti

Ryerson University
Department of Electrical, Computer, and Biomedical Engineering
Toronto, Canada

ABSTRACT

Graph clustering algorithms require the number of clusters as an input. However, in many real-world practical applications, the correct number of clusters is unknown. Determining the optimal number of clusters for graph clustering algorithms is an essential and challenging task, which is a form of model order selection. Here, we propose a new algorithm for estimating the number of clusters in a graph using the centrality measure. In graph theory, the centrality measure is used for determining the most important and most influential nodes within a graph. The proposed centrality based number of cluster estimation (CB-NCE) method considers minimizing the probabilistic bounds on the average central error of centrality. The desired criterion represents an information theoretic distance measure in the form of description length of centrality. The simulation results show the superior performance of the proposed algorithm among other existing methods, in terms of clustering performance metrics such as normalized mutual information, Rand index, and F-measure.

Index Terms— Graph clustering, Community detection, Model order selection, Centrality measure

1. INTRODUCTION

Graphs are generic data domains that represent complex data structures. These complex data structures can be found in many different fields such as computer networks, online social networks, chemistry, social science, image processing, and linguistics. In signal processing and computer science applications of graphs, the nodes represent the data samples and the edges show affinity or similarity or some relation between the nodes of the graph. In graph clustering, the goal is to cluster the nodes of the graph in a way that maximizes the similarity of the nodes within one cluster while maintaining the dissimilarity between different clusters. The applications of graph clustering include but not limited to community detection in social networks [1], graph signal processing [2], network assessment [3] and image segmentation [4].

One of the major challenges in graph clustering algorithms is determining the number of clusters beforehand. To address this problem, a different model order selection method, has been proposed in this context. In the model order selection method, different numbers of clusters are considered for clustering, and then the optimal cluster number is determined by optimizing some objective function. For example, in [5] maximizing the gap between the two largest eigenvalues of the adjacency matrix is considered as the objective function. In [6, 7] propose graph clustering based on eigenvector decomposition of the nonbacktracking matrix. In this method, the number of clusters can be estimated by calculating the number of real eigenvlues which their magnitude is larger than the square root of the largest eigenvalue. In [8] the degree corrected stochastic block model [9, 10] and Monte Carlo sampling method are implemented for graph clustering. In [11] a greedy algorithm for modularity maximization is used to find the optimum number of graph clusters. The K-MACE algorithm proposed in [12], obtains the correct number of clusters in k-means clustering using the available cluster compactness. Inspired by the K-MACE algorithm, a new method is proposed for estimating the optimal number of graph clusters. The proposed method denoted by (CB-NCE), concentrates on the statistical learning of the centrality measure for a range of possible number of clusters and chooses the one that minimizes this criterion.

2. PROBLEM STATEMENT AND NOTATIONS

An undirected graph $G(V, E)$ with N nodes where $V = \{v(1), \ldots, v(N)\}$ is the nodes of the graph and E is the edges of the graph, can be fully described by it's adjacency matrix $\mathbf{A}$ with size $N \times N$. $[\mathbf{A}]_{uv} = 1$ if there is an edge connecting the u and v nodes and otherwise $[\mathbf{A}]_{uv} = 0$. Due to the fact that the graph is undirected, the adjacency matrix $\mathbf{A}$ is a symmetric matrix.

Considering the true number of clusters in the graph as K^*, the size of each cluster is N_k where $k \in \{1, 2, 3, \ldots, K^*\}$. The matrix $\mathbf{A}_k$ with size $N_k \times N_k$ denotes the adjacency matrix for internal edge connection in cluster k. In this problem formulation, the total number of clusters, K^*, is unknown and estimating the number of clusters can be considered as a model order selection problem. In this paper, inspired by K-MACE algorithm [12], we propose a new approach for

Authorized licensed use limited to: IEEE Xplore. Downloaded on May 24,2021 at 02:35:47 UTC from IEEE Xplore. Restrictions apply.

finding the correct number of clusters in graph clustering.

3. ESTIMATION OF NUMBER OF GRAPH CLUSTERS

In graph theory and network science, centrality is used to identify the most important nodes within a graph structure. For example, in social networks, the central nodes identify the most influential person in a community, or in analysing the spread of a disease, the central nodes determine the super-spreader [13]. The centrality measure in a graph is a real value function on the nodes of the graph which determines the rank of the importance of each node. Based on the definition of importance, different centrality measures are defined on graph structures. In this paper, we have considered three different centrality measures. The following notation is used in this context:

Existing notation:
$v(i)$: Nodes of graph $G(V, E)$, $v(i) \in \{v(1), \dots, v(N)\}$.
$d(v_i, v_j)$: The distance between nodes v_i and v_j.
$r(v(i))$: Centrality measure of node $v(i)$

Degree centrality $\boldsymbol{r^D(v(i))}$: In an undirected graph, the degree of a node is the number of edges of that node. The degree of each node shows that how many other nodes are influenced and effected by that node and therefore, it can be considered as a measure of centrality [14].

Eigenvector centrality $\boldsymbol{r^E(v(i))}$: The eigenvector centrality of a node is the measurement of the influence of that node in a graph and is defined as proportional to its neighborhood nodes importance. The Katz centrality and pagerank centrality are the variations of the eigenvector centrality [14].

Closeness centrality $\boldsymbol{r^C(v(i))}$: The closeness centrality of a node is determined by the closeness of the node to all other nodes in the graph. The closeness centrality can be calculated by the average length of the shortest path between the node and all other nodes in the graph [14].

$$r^c(v(i)) = \frac{N-1}{\sum_{j=1, i \neq j}^{N} d(v(i), v(j))} \tag{1}$$

3.1. Centrality based number of clusters estimation

To extend the k-mace algorithm to graph settings, we consider clustering with different number of clusters from a minimum number of clusters up to a maximum number, $K \in [K_{min}, K_{max}]$. The new notations used in this paper are:

Our notations:
C_{Kk}: k^{th} graph cluster when the total number of clusters is K.
N_k: Size of each cluster.
$\mathbf{A}_k$: The adjacency matrix of the k^{th} cluster with respect to the internal connection of the nodes within a cluster.
$v_{Kk}(i)$: indicates that the node $v(i)$ belongs to cluster k when the total number of the clusters is K.
$\mathbf{r}_{Kk}$: The centrality vector for k^{th} cluster when the total number of the clusters is K.

$$\mathbf{r}_{Kk} = [r(v_{Kk}(i))], i \in C_{Kk} \tag{2}$$

For simplicity and without loss of generality, we consider that there is a central node in each cluster and all cluster nodes are sampled from a normal probability distribution and distributed around this central node. The node with the highest centrality within the graph cluster can be considered as the central node. If there are more than one node with the highest centrality, we can assume the central node as a new node which is connected to the node with the highest centrality. This central node is denoted as $\bar{c}_{v_{Kk}}$.

Please note that one of the mentioned centrality measures is chosen for calculating the $\mathbf{r}_{Kk}$ vector. Based on this, when $K = K^*$, each node in the k^{th} cluster can be considered as:

$$v_{Kk}(i) = \bar{c}_{v_{Kk}(i)} + \overline{w}_{v_{Kk}} \tag{3}$$

where $\bar{c}_{v_{Kk}(i)}$ is the true cluster center for node $v_{Kk}(i)$ and $\overline{w}_{v_{Kk}}$ is a white Gaussian distribution with variance σ^2_{Kk}.

The variance σ^2_{Kk} in $\overline{w}_{v_{Kk}}$ can be calculated based on the scatterness of the nodes around the central node of the graph. Therefore we have:

$$\sigma^2_{Kk} = \text{var}(d(v_{Kk}(i), \bar{c}_{v_{Kk}(i)})),\ i \in C_{Kk} \tag{4}$$

However, the true cluster number is unknown a priori. Therefore, after applying the clustering algorithm with K cluster number, where $K \in [K_{max}, K_{max}]$, the central node of each cluster is determined based on the centrality measure for all nodes within that cluster. Therefore, the center of the k^{th} cluster is:

$$\widehat{c}_{Kk} = \underset{v(i)}{\arg\max}(\mathbf{r}_{Kk}),\ \ i \in C_{Kk} \tag{5}$$

If there is more than one node with the highest centrality in $\mathbf{r}_{Kk}$, we can assume the central node as a new node which is connected to the node with the highest centrality. Consequently, the estimated center for all clusters in K clustering is:

$$\text{Estimated centers :} \qquad [\hat{c}_{K1}\ \hat{c}_{K2}\ \dots\ \widehat{c}_{KK}] \tag{6}$$

As a result, the Average Central Error (ACE), which is denoted by z_K is defined as the error between the true center of a node and the estimated center of that node. Therefore, when K number of clusters is considered for the clustering algorithm, the central error for the k^{th} cluster is calculated by:

Authorized licensed use limited to: IEEE Xplore. Downloaded on May 24,2021 at 02:35:47 UTC from IEEE Xplore. Restrictions apply.

$$z_{Kk} = \sum_{i \in C_{Kk}} d(\overline{c}_{v_{Kk}(i)}, \hat{c}_{v_{Kk}(i)})^2 \tag{7}$$

where $\overline{c}_{v_{Kk}(i)}$ is the true center and $\hat{c}_{v_{Kk}(i)}$ is the estimated center of the node i in the k^{th} cluster. Consequently, the ACE can be computed by averaging over all central errors from all clusters:

$$z_K = \frac{1}{N} \sum_{K}^{k=1} z_{Kk} \tag{8}$$

Minimizing the ACE leads to determining the correct number of graph clusters. In [12] it has been shown that, the ACE is a sample of random variables Z_K. Minimizing the ACE is equivalent to minimizing the noiseless description length of the nodes in the graph clusters [15, 16].

$$\text{MNDL}(\overline{c}_{v_{Kk}(i)}, \hat{c}_{v_{Kk}(i)}) = \log_2 \sqrt{2\pi\sigma_{Kk}^2} + \gamma z_{Kk} \tag{9}$$

Although z_K cannot be calculated directly, in [12, 17] it has been shown that the probabilistic bounds on this error can be calculated using the available cluster compactness measure and the invented probabilistic stochastic learning method. Cluster compactness is the error between the samples of the cluster and the estimated cluster centers. Consequently, the overall cluster compactness for K clusters can be calculated by averaging over all cluster compactness from all clusters. Therefore, for k^{th} cluster we have:

$$y_{Kk} = \sum_{i \in C_{Kk}} d(v_{Kk}(i), \hat{c}_{v_{Kk}(i)})^2 \tag{10}$$

$$y_K = \frac{1}{N} \sum_{k=1}^{K} y_{Kk} \tag{11}$$

The cluster compactness y_K is also a sample of a random variable Y_K [12]. Utilizing the available cluster compactness, probabilistic bounds can be calculated for the ACE [12]:

$$\underline{z_K} \leq z_K \leq \overline{z_K} \tag{12}$$

The details of the calculation of the lower bound and upper bound of ACE in (12) are presented in [12]. Consequently, minimizing the upper bound of ACE leads to the optimal estimation of the correct number of graph clusters.

$$\hat{K} = \arg\max_K(\overline{z_K}) \tag{13}$$

Based on this, the centrality based number of cluster estimation (CB-NCE) algorithm is purposed. CB-NCE algorithm takes the graph adjacency matrix and the minimum and maximum number of clusters, $[K_{\min}, K_{\max}]$ as input and outputs the clustered graph with the optimum number of clusters. Algorithm 1 shows the pseudo code of the proposed CB-NCE method.

Algorithm 1 CB-NCE Algorithm

```
Input: Graph Data set $G(V, E)$, range of $K \in [K_{min}, K_{max}]$
Output: Estimated number of clusters $\hat{K}$, the clustering solution $[\hat{C}_1, \hat{C}_2, .., \hat{C}_{\hat{K}}]$
for ($K = K_{\min}; K \leq K_{\max}; K_{++}$) do
    Perform the clustering algorithm for $K$ and obtain: $[C_{K1}, C_{K2}, .., C_{Kj}]$
    for each cluster $C_{Kk}$, ($k = 1, .., K$) do
        Solve cluster compactness $y_{Kk}$ of cluster $C_{Kk}$
    end for
    Solve for total cluster compactness $Y_{Kk}$
end for
for ($K = K_{\min}; K \leq K_{\max}; K_{++}$) do
    Calculate the variance of each cluster $K$ based on the proposed method
end for
Estimate the optimum $\hat{K} = \arg\max_K(\overline{z_K})$ using the cluster variance and the calculated $Y_{Kk}$
Determine the graph clusters by: $[\hat{C}_1, \hat{C}_2, .., \hat{C}_{\hat{K}}]$
```

Table 1: Clustering performance metrics averaged over 100 generated random geometric graph with seven graph clusters

method	NMI	RI	F	C	NC
Louvian	0.63	0.72	0.54	0.215	0.265
NB	0.77	0.82	0.79	0.072	**0.02**
NR	067	0.78	0.57	0.061	0.840
CB-NCE	**0.81**	**0.85**	**0.82**	**0.051**	**0.02**

4. SIMULATION RESULTS

Experiment 1: To validate the proposed centrality based number of cluster estimation (CB-NCE) method, seven random geometric graph with 150 nodes with a distance threshold value of 0.125 was generated. These seven random geometric graphs then were connected together with a distance threshold value of 0.125. Then the proposed CB-NCE method was implemented for graph clustering. Figure 1 shows the clustered graph with the proposed CB-NCE algorithm. As the figure shows, the proposed CB-NCE method has detected seven existing clusters and performed the clustering algorithm with the correct number of clusters.

To compare the proposed method with the louvain method [11], nonbacktracking matrix (NB) method [6,7] and Newman-

Authorized licensed use limited to: IEEE Xplore. Downloaded on May 24,2021 at 02:35:47 UTC from IEEE Xplore. Restrictions apply.

Table 2: Graph datasets.

Dataset	Nodes	Edges	Ground truth meta information
IEEE reliability test sytem [18]	73 power stations	108 power lines	3 interconnected power systems
Hibernia Internet backbone map [19]	55 cities	162 connections	city names and geographical locations
Cogent Internet backbone map [19]	197 cities	243 connections	city names and geographical locations

Table 3: Clustering Performance Comparison

Dataset	method	NMI	Rand index	F-measure	conductance	NC
IEEE reliablity test sytem [18]	Louvian	0.74	0.84	0.67	0.144	0.169
	NB	0.75	0.88	0.81	0.070	0.**100**
	NR	0.72	0.82	0.64	0.680	0.804
	CB-NCE	**0.88**	**0.93**	**0.91**	**0.043**	0.22
Hibernia Internet backbone map [19]	Louvian	0.27	0.51	0.33	0.222	0.263
	NB	0.73	0.89	0.09	**0.027**	0.053
	NR	0.73	0.89	0.90	**0.027**	0.053
	CB-NCE	**0.98**	**0.99**	**0.96**	0.030	**0.048**
Cognet Internet backbone map [19]	Louvian	0.25	0.54	0.26	0.186	0.204
	NB	0.26	0.54	0.58	**0.073**	0.109
	NR	0.48	**0.68**	0.63	0.148	**0.043**
	CB-NCE	**0.51**	**0.68**	**0.67**	0.115	0.072

Reinert (NR) method [8], internal and external clustering metrics were calculated. These clustering performance metrics are normalized mutual information (NMI) [20], Rand index (RI) [20], F-measure (F) [20], conductance (C) [21] and normalized cuts (NC) [21]. Please note that higher values of normalized mutual information (NMI), Rand index (RI) and F-measure (F) indicate the better performance while lower values for conductance (C) and normalized cuts (NC) are desirable.

Table 1 shows the simulation results which were calculated by averaging the performance indexes over 100 different generated random geometric graphs. According to table 1, the proposed CB-NCE method has superior performance metrics compared to other existing methods.

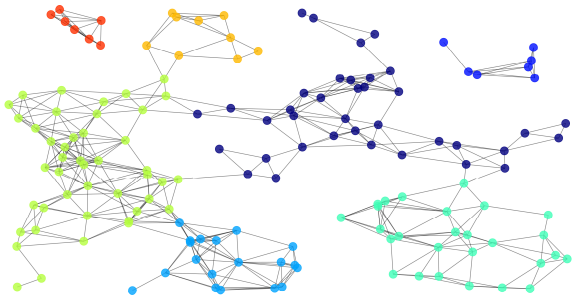

Fig. 1: An example of CB-NCE clustering method on a random geometric graph generated by 7 graph clusters.

Experiment 2: The proposed CB-NCE method was implemented on real-world graph datasets. Table 2 shows the detailed information of the datasets. Table 3 compares the proposed CB-NCE algorithm with louvain method [11], non-backtracking matrix (NB) method [6,7] and Newman-Reinert (NR) method [8] in terms of the aforementioned internal and external clustering metrics. As the table indicates, the proposed CB-NCE method outperforms other existing algorithms in terms of normalized mutual information, Rand index, and F-measure in all datasets.

5. CONCLUSION

In this paper, a new approach for determining the optimal number of clusters in graph clustering, is presented. The proposed centrality based number of cluster estimation (CB-NCE) algorithm obtains probabilistic bounds on the average central error (ACE) based on the centrality measure vector and cluster compactness. As the simulation results indicate, the proposed CB-NCE method outperforms other existing algorithms in this context in terms of normalized mutual information, Rand index, F-measure, conductance, and normalized cuts.

Authorized licensed use limited to: IEEE Xplore. Downloaded on May 24,2021 at 02:35:47 UTC from IEEE Xplore. Restrictions apply.

6. REFERENCES

[1] Scott White and Padhraic Smyth, "A spectral clustering approach to finding communities in graphs," in *Proceedings of the 2005 SIAM international conference on data mining*. SIAM, 2005, pp. 274–285.

[2] Aliaksei Sandryhaila and José MF Moura, "Discrete signal processing on graphs," *IEEE transactions on signal processing*, vol. 61, no. 7, pp. 1644–1656, 2013.

[3] Pin-Yu Chen and Alfred O Hero, "Assessing and safeguarding network resilience to nodal attacks," *IEEE Communications Magazine*, vol. 52, no. 11, pp. 138–143, 2014.

[4] Noeleene Mallia-Parfitt and Georgios Giasemidis, "Graph clustering and variational image segmentation for automated firearm detection in x-ray images," *IET Image Processing*, vol. 13, no. 7, pp. 1105–1114, 2019.

[5] Marzia Polito and Pietro Perona, "Grouping and dimensionality reduction by locally linear embedding," in *Advances in neural information processing systems*, 2002, pp. 1255–1262.

[6] Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang, "Spectral redemption in clustering sparse networks," *Proceedings of the National Academy of Sciences*, vol. 110, no. 52, pp. 20935–20940, 2013.

[7] Alaa Saade, Marc Lelarge, Florent Krzakala, and Lenka Zdeborová, "Spectral detection in the censored block model," in *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2015, pp. 1184–1188.

[8] Mark EJ Newman and Gesine Reinert, "Estimating the number of communities in a network," *Physical review letters*, vol. 117, no. 7, pp. 078301, 2016.

[9] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt, "Stochastic blockmodels: First steps," *Social networks*, vol. 5, no. 2, pp. 109–137, 1983.

[10] Brian Karrer and Mark EJ Newman, "Stochastic blockmodels and community structure in networks," *Physical review E*, vol. 83, no. 1, pp. 016107, 2011.

[11] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, pp. P10008, 2008.

[12] Soosan Beheshti, Edward Nidoy, and Faizan Rahman, "K-mace and kernel k-mace clustering," *IEEE Access*, vol. 8, pp. 17390–17403, 2020.

[13] Cezary Bartosiak, Rafał Kasprzyk, and Andrzej Najgebauer, "The graph and network theory as a tool to model and simulate the dynamics of infectious diseases," *Bio-Algorithms and Med-Systems*, vol. 9, no. 1, pp. 17–28, 2013.

[14] Stephen P Borgatti and Martin G Everett, "A graph-theoretic perspective on centrality," *Social networks*, vol. 28, no. 4, pp. 466–484, 2006.

[15] Mahdi Shamsi, Faizan Rahman, and Soosan Beheshti, "Correct number of clusters (cnc) description length in arbitrary shape clustering," in *2019 16th Canadian Workshop on Information Theory (CWIT)*. IEEE, 2019, pp. 1–4.

[16] Azadeh Fakhrzadeh and Soosan Beheshti, "Minimum noiseless description length (mndl) thresholding," in *2007 IEEE Symposium on Computational Intelligence in Image and Signal Processing*. IEEE, 2007, pp. 146–150.

[17] Faizan Rahman and Soosan Beheshti, "Kernel k-mace clustering," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2018, pp. 2002–2006.

[18] Cliff Grigg, Peter Wong, Paul Albrecht, Ron Allan, Murty Bhavaraju, Roy Billinton, Quan Chen, Clement Fong, Suheil Haddad, Sastry Kuruganty, et al., "The ieee reliability test system-1996. a report prepared by the reliability test system task force of the application of probability methods subcommittee," *IEEE Transactions on power systems*, vol. 14, no. 3, pp. 1010–1020, 1999.

[19] Simon Knight, Hung X Nguyen, Nickolas Falkner, Rhys Bowden, and Matthew Roughan, "The internet topology zoo," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 9, pp. 1765–1775, 2011.

[20] Mohammed J Zaki and Wagner Meira, *Data mining and analysis: fundamental concepts and algorithms*, Cambridge University Press, 2014.

[21] Jianbo Shi and Jitendra Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

Authorized licensed use limited to: IEEE Xplore. Downloaded on May 24,2021 at 02:35:47 UTC from IEEE Xplore. Restrictions apply.