Community Detection under Stochastic Block Model Likelihood Optimization via Tabu Search –Fuzzy C-Mean Method for Social Network Data

Ali Falah Yaqoob*, Basad Al-Sarray
Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq

Received: 8/7/2019 Accepted: 4/4/2020

Abstract

Structure of network, which is known as community detection in networks, has received a great attention in diverse topics, including social sciences, biological studies, politics, etc. There are a large number of studies and practical approaches that were designed to solve the problem of finding the structure of the network. The definition of complex network model based on clustering is a non-deterministic polynomial-time hardness (NP-hard) problem. There are no ideal techniques to define the clustering. Here, we present a statistical approach based on using the likelihood function of a Stochastic Block Model (SBM). The objective is to define the general model and select the best model with high quality. Therefore, integrating the Tabu Search method with Fuzzy c-Mean (FCM) is implemented in different settings. The experiments are designed to find the best structure for different types of networks by maximizing the objective functions. SBM selections are computed by applying two types of criteria, namely Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). The results show the ability of the proposed method to find the best community of the given networks.

Keywords: Community detection, Stochastic Block Model, FCM, Likelihood function, BIC, AIC.

استكشاف المجتمع في الشبكات المعقدة باستخدام دالة الإمكان النماذج الكتل العشوائية

علي فلاح يعقوب*، بسعاد علي السراي
قسم علوم الحاسوب، كلية العلوم، جامعة بغداد، بغداد، العراق

الخلاصة

هيكلية الشبكة التي تعرف باكتشاف المجتمع في الشبكات لها أهمية واسعة في مواضيع متعددة ، مثل العلوم الاجتماعية والدراسة البيولوجية والسياسة والعمل. هناك دراسات كبيرة ومتنوعة من الأساليب العملية لحول مشكلة إيجاد هيكلية الشبكة. يعرف نموذج الشبكة المعقدة على أساس السجلات هو مشكلة معقدة غير محلولة (NP-hard). لا يوجد تقنيات محلية تحديد التعدد الملائم للنموذج المعقدة. وبالتالي فإن الهدف هو تحديد النموذج المعقد (SBM) على استخدام دالة الإمكان الأعظم للنموذج الكتل العشوائية مع عدم وجود معايير. فحص أفضل النماذج بعوضة عالية، وذلك يتم تطبيق دمج أسلوب (Tabu search) مع (Fuzzy C-mean) في مختلف الأعداد. تم تطبيق اختبارات النماذج من خلال تطبيق معياري من معايير

*Email: ali.f.yaqoob94@gmail.com
1- Introduction

A complex network has a lot of communities with significant topological features common in real-world networks (biological and social networks). Community detection in networks is a key investigative tool with applications in a set of parts, ranging from finding communities in social and biological networks to identifying link farms in the World Wide Web [1], [2]. Modularity is a frequently used term in information technology and computer science. Modularity refers to the concept of making multiple modules first and then linking and combining them to form a complete system. The proposed idea is getting the best partition based on the maximization of Modularity and Likelihood functions of Stochastic Block Models (SBM). SBM is a generative tool that inclines to produce networks containing communities, with subsets that are considered by being connected with one another with particular edge densities [3]. SBM is important in statistics, machine learning, and network science, where it serves as a useful benchmark for the task of recovering community structure in graph data. In statistics, likelihood is a function of the parameters of a statistical model derived for a given data. The likelihood is used after data are available to describe plausibility of a parameter value [4]. The problems encountered are related to estimating the latent block memberships and model parameters, including modularity and likelihood maximization. Abbe and Sandon (2015) presented community detection in general SBM. They studied the partial and exact recovery of communities in the general SBM in the constant and logarithmic degree regimes and generalized the results to tackle overlapping communities [5]. Decelle et al. (2011) studied the asymptotic analysis of the SBM for modular networks. They presented the belief propagation method to find the structure of a network from its topology, where the results were applied on the generative model for social and biological networks [6]. Come and Latouche (2015) studied the problem of model selection and clustering in SBM in the case of the integrated-complete data log likelihood. They used a greedy inference method in their computations [7]. Yan et al. (2014) studied a model selection for degree-corrected block models. They applied a belief propagation method for log-likelihood of two types of the models [8]. Qin and Rohe (2013) presented the regularized spectral under the degree corrected SBM and worked on the spectral clustering in a high dimensional SBM model [9].

2- Community Detection and the Stochastic Block Model

Let $N$ be an even positive integer and $G$ be a random graph. For each pair of nodes, $(i, j)$ is an edge of $G$ with probability $p$ if $i$ and $j$ are in the same set, and with probability $q$ if they are in different sets. Each edge is drawn independently ($p > q$). This is known as the SBM on two communities. The goal is to recover the original partition. The question is for which values of $p$ and $q$ is it possible to recover the partition of the graph. Let $n$ be a positive integer (the number of vertices), $k$ be a positive integer (the number of communities), $p = (p_1, ..., p_k)$ be a probability vector on $[c] := (1, ..., c)$ (the prior on the $k$ communities), and $P$ be a $k \times k$ symmetric matrix with entries in $[0, 1]$ (the connectivity probabilities). The adjacency matrix $A_{ij}$ of the graph contains zeros or ones in the diagonal (these correspond to the absence or presence, respectively, of self-loops for the graph nodes). The pair $(X, G)$ is drawn under $SBM(n, p, P)$, if $X$ is an $n$-dimensional random vector with i.i.d components distributed under $P$, and $G$ is an $n$-vertex simple graph where vertices $i$ and $j$ are connected with probability $P_{X_i X_j}$ independently of other pairs of vertices [10]. The community detection problem can be formulated as finding a disjoint partition $T_1 \cup T_2 \cup T_C$. A set of node labels $s = \{s_1, ..., s_N\}$, where $s_i$ is the label of node $i$ and takes values in $\{1, 2, ..., C\}$. For any set of label assignments $s$, let $J(s)$ be the $C \times C$ matrix defined by

$$J_C(s) = \sum_{ij} A_{ij} I\{s_i = c, s_j = l\}$$

where $I$ is the indicator function. Further, let

$$J_c(s) = \sum_l J_C(c)$$

where $c \neq 1$, $J_c(s)$ is the total number of edges between $c, l$. $J_c$ is the sum of node degrees in community $c$ and $L$ is the sum of all degrees in the network. Then $J_{cc}$ is interpreted as twice the total number of edges within the community $c$ and $L$ as twice the number of edges in the whole network.
Finally, let \( n_c(s) = \sum_i I\{s_i = c\} \) be the number of nodes in the \( c^{th} \) community, and \( f(e) = (N_1 N_N, N_2 N_N, \ldots, N_N N_N)^T \). The SBM network edges variables \( A_{ij} \) with given true node labels \( c = \{c_1, \ldots, c_n\} \in \{1, \ldots, C\} \), are independent Bernoulli random variables with

\[
E[A_{ij}|c] = P_{c_ic_j}
\]

where \( P \) is a \( C \times C \) symmetric matrix. For the case of Degree Corrected SBM, Equation(2) replaced by

\[
E[A_{ij}|c] = \theta_i \theta_j P_{c_ic_j}
\]

where \( \theta_i \) is a degree parameter associated with the node \( i \), reflecting its individual propensity to form ties. A profile likelihood can be derived by maximizing over \( \theta \) by giving the following criteria [11].

\[
Y_{DCBM}(s) = \sum_c J_{ct} \log \frac{\hat{d}_{ct}}{d_t}
\]

The profile likelihood in SBM finds the optimization overall partitions by the criteria:

\[
Y_{BM}(s) = \sum_c J_{ct} \log \frac{\hat{d}_{ct}}{n_c n_t}
\]

The general modularity criteria is

\[
\gamma(s) = \sum_{ij} [A_{ij} - P_{ij}] I\{s_i = s_j\}
\]

where \( P_{ij} \) is the probability of edges falling between \( i \) and \( j \) under the null model. In the case of \( C = 1 \), SBM reduces to Erdos-Renyi random graph, where \( P_{ij} \) is a constant estimated by \( L/N^2 \).

\[
Y_{ERBM}(s) = \sum_c (J_{cc} - \frac{n_c^2}{n^2} L).
\]

The popular Newman-Girvan modularity (NGM) is

\[
\gamma_{NGM}(s) = \sum_c (J_{cc} - \frac{\hat{d}_c^2}{\hat{d}_c^2} L).
\]

3- Model selection

There are types of Penalized-likelihood information criteria, such as AIC and BIC. The Consistent AIC and the Adjusted BIC are widely used for model selection. AIC is an estimate of a constant plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model, so that a maximum AIC value means that the model is considered to be closer to the truth [12]. Let \( k \) be the number of estimated parameters in the model. Let \( \hat{L} \) be the maximum value of the likelihood function for the model. Then the AIC value of the model is

\[
AIC = 2k - 2 \ln \hat{L}
\]

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. BIC is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup [13]. The BIC is formally defined as

\[
BIC = \ln(n) k - 2 \ln(\hat{L})
\]

where \( \hat{L} \) is the maximized value of the likelihood function of the model \( M \).

4- Integrating Tabu via FCM for SBM and Degree-Corrected Stochastic Block Model (DCSBM) likelihood maximization

Tabu search is a global optimization algorithm, the basic concept of which was described by Glover (1989), who presented it as a meta-heuristic superimposed on another heuristic. The overall approach is to avoid entrainment in cycles by forbidding or penalizing moves which take the solution, in the next iteration, to points in the solution space previously visited (hence “tabu”) [14]. The idea of the tabu method is a simulation to the human behavior which appears to operate with a random element that leads to inconsistent behavior, given similar circumstances [15]. Fuzzy clustering is a hard clustering algorithm which requires that each data point of the data set belongs to one and only one cluster. Fuzzy c-means (FCM) clustering was developed by Dunn in 1973 and improved by Bezdek in 1981 [16]. The integrating of tabu search with FCM under SBM is explained in Algorithm 1. Another integration under DCSBM is explained in Algorithm 2.
Algorithm 1. Tabu Search with Fuzzy c-mean For SBM Likelihood Function

| Input: | D dataset, max.iter maximum number of iterations |
| Output: | Cost. bst: best maximum likelihood objective values |

Step.1 Initialization: Tabu. List = [], Q = 0, TC1: Tabu counter for Model, TC2: Tabu Counter for adjacency, use Tabu counter to save best solution node or label.

Step 2. Generate random permutation for Lb1 and Lb2 to find Solution.

Step 3. For iter = 1 to max.iter do {Main TS loop}

Step 3.1. If .Cost >= Init. Cost then Init. Sol = new.sol. {Save the new values}

Step 3.2. sol = NewSol. Update Current Solution:

Step 3.3. For i=1 to nAction1 { Update Tabu. List for model}

If i == NewSol.Index, then add to Tabu. List TC1(i) = TL1; otherwise, reduce Tabu Counter

TC1(i) = max(TC1(i) − 1, 0);

Step 3.4. Update Tabu. List for adjacency

For i=1 to nAction2

If i == NewSol.Index, then add to Tabu List TC2(i) = TL2;

otherwise, reduce Tabu Counter TC2(i) = max(TC2(i) − 1, 0);

Step 3.5. Update Best Solution Ever Found

\[
\text{If } sol1.\text{Cost} \geq \text{BestSol1.\text{Cost}} \text{ and } sol2.\text{Cost} \geq \text{BestSol2.\text{Cost}} \text{ then } \\
\text{BestSol1} = sol1; \text{ BestSol2} = sol2;
\]

Step 3.6. Save Best Quality Ever Found


End

Algorithm 2. Tabu Search with Fuzzy c-mean For DCSBM Likelihood Function

| Input: | D dataset, max.iter maximum number of iterations |
| Output: | Cost. bst: best maximum likelihood objective values |

Step.1 Initialization: Tabu. List = [], Init. Q = 0, TC1: Tabu counter for Model, TC2: Tabu Counter for adjacency

Step 2. Define an anonymous function for computing Best cost from Algorithm (2) with the real parameters and the dependent parameters before the function name.

Step 3. Generate random permutation for Lb1 and Lb2 to find Solution.

Step 4. For iter = 1 to max.iter do {Main TS loop}

Step 4.1. If .Cost >= Init. Cost then Init. Sol = new.sol. Save the new values.

Step 4.2. sol = NewSol. Update Current Solution:

Step 4.3. For i=1 to nAction1 Update Tabu. List for model

If i == NewSol.Index, then add to Tabu. List TC1(i) = TL1; otherwise, reduce Tabu Counter

TC1(i) = max(TC1(i) − 1, 0);

Step 4.4. Update Tabu. List for adjacency

For i=1 to nAction2

If i == NewSol.Index, then add to Tabu List TC2(i) = TL2;

otherwise, reduce Tabu Counter TC2(i) = max(TC2(i) − 1, 0);

Step 4.6. Update Best Solution Ever Found

\[
\text{If } sol1.\text{Cost} \geq \text{BestSol1.\text{Cost}} \text{ and } sol2.\text{Cost} \geq \text{BestSol2.\text{Cost}} \text{ then } \\
\text{BestSol1} = sol1; \text{ BestSol2} = sol2;
\]

Step 4.7. Save Best Quality Ever Found


5- Experimental Results and Discussion

This section deals with the experimental part of this paper. The results show the ability of the proposed algorithms to find the optimal solution, the best clusters, based on the values of SBM and
DCSBM objective functions. The experiments are designed for the real networks with different topics and complicity. The details of the networks are given by Table-1.

Table 1- Details of the networks used in this work [17]

<table>
<thead>
<tr>
<th>Networks</th>
<th>No. of Nodes</th>
<th>No. of Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zackary Karate</td>
<td>34</td>
<td>78</td>
</tr>
<tr>
<td>Dolphin</td>
<td>62</td>
<td>158</td>
</tr>
<tr>
<td>American Football Collage (AFC)</td>
<td>115</td>
<td>613</td>
</tr>
<tr>
<td>Facebook</td>
<td>3958</td>
<td>84241</td>
</tr>
<tr>
<td>Protein</td>
<td>2284</td>
<td>6644</td>
</tr>
<tr>
<td>Political blogs</td>
<td>1107</td>
<td>9537</td>
</tr>
<tr>
<td>Internet Level AS Network(ILAN)</td>
<td>6444</td>
<td>11284</td>
</tr>
<tr>
<td>Chesapeake</td>
<td>39</td>
<td>170</td>
</tr>
<tr>
<td>Delaunay</td>
<td>1024</td>
<td>3056</td>
</tr>
<tr>
<td>Twitter</td>
<td>2623</td>
<td>21000</td>
</tr>
</tbody>
</table>

The values of the metrics (AIC, BIC), based on optimal values of likelihood functions of the models under study (SBM, DCSBM), are given in Tables -2 and 3, respectively. These values of the metrics are obtained by applying TS-FCM to optimize the partitions of the networks based on maximum likelihood functions. Here the number of clusters is defined on the range of 2 to n/2. The results are given in Figures- 1-4 which are divided into three parts; the first part represents the partitions of the network, the second part represents the integrating of TS with FCM objective function, and the third part represent the model selection criteria using the integration of TS with FCM for modeling the networks.

Table 2-Comparison of AIC and BIC values for SBM using TS-FCM

<table>
<thead>
<tr>
<th>Network</th>
<th>AIC (TS-FCM)</th>
<th>BIC (TS-FCM)</th>
<th>No. of Best Cluster</th>
<th>AIC (FCM-TS)</th>
<th>BIC (FCM-TS)</th>
<th>No. of Best Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Karate</td>
<td>68</td>
<td>78</td>
<td>16</td>
<td>68</td>
<td>78</td>
<td>17</td>
</tr>
<tr>
<td>Dolphin</td>
<td>123</td>
<td>143</td>
<td>29</td>
<td>123</td>
<td>143</td>
<td>31</td>
</tr>
<tr>
<td>AFC</td>
<td>225</td>
<td>260</td>
<td>52</td>
<td>229</td>
<td>263</td>
<td>54</td>
</tr>
<tr>
<td>Facebook</td>
<td>6580</td>
<td>7540</td>
<td>1545</td>
<td>6590</td>
<td>7550</td>
<td>1758</td>
</tr>
<tr>
<td>Political</td>
<td>2200</td>
<td>2570</td>
<td>586</td>
<td>1989</td>
<td>2450</td>
<td>747</td>
</tr>
<tr>
<td>Protein</td>
<td>4520</td>
<td>5250</td>
<td>1189</td>
<td>4050</td>
<td>4700</td>
<td>1346</td>
</tr>
<tr>
<td>ILAN</td>
<td>8700</td>
<td>10000</td>
<td>3108</td>
<td>9000</td>
<td>10500</td>
<td>3476</td>
</tr>
<tr>
<td>Chesapeake</td>
<td>78</td>
<td>90</td>
<td>12</td>
<td>76</td>
<td>87.5</td>
<td>15</td>
</tr>
<tr>
<td>Delaunay</td>
<td>1990</td>
<td>2285</td>
<td>498</td>
<td>2090</td>
<td>2360</td>
<td>512</td>
</tr>
<tr>
<td>Twitter</td>
<td>5192</td>
<td>5978</td>
<td>952</td>
<td>5192</td>
<td>5979</td>
<td>1024</td>
</tr>
</tbody>
</table>

Table 3-Comparison of AIC and BIC values for DCSBM using TS-FCM

<table>
<thead>
<tr>
<th>Network</th>
<th>AIC (TS-FCM)</th>
<th>BIC (TS-FCM)</th>
<th>No. of Best Cluster</th>
<th>AIC (FCM-TS)</th>
<th>BIC (FCM-TS)</th>
<th>No. of Best Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Karate</td>
<td>2224 × 10^4</td>
<td>2224 × 10^3</td>
<td>16</td>
<td>2225 × 10^3</td>
<td>2225 × 10^3</td>
<td>16</td>
</tr>
<tr>
<td>Dolphin</td>
<td>1395 × 10^4</td>
<td>1395 × 10^4</td>
<td>29</td>
<td>1394 × 10^4</td>
<td>1394 × 10^4</td>
<td>30</td>
</tr>
<tr>
<td>AFC</td>
<td>5463 × 10^4</td>
<td>5463 × 10^4</td>
<td>52</td>
<td>5464 × 10^4</td>
<td>5464 × 10^4</td>
<td>55</td>
</tr>
<tr>
<td>Facebook</td>
<td>2228 × 10^6</td>
<td>2235 × 10^6</td>
<td>1545</td>
<td>2228 × 10^6</td>
<td>2235 × 10^6</td>
<td>1545</td>
</tr>
<tr>
<td>Political</td>
<td>2734 × 10^12</td>
<td>2735 × 10^12</td>
<td>586</td>
<td>2066 × 10^5</td>
<td>2069 × 10^5</td>
<td>747</td>
</tr>
<tr>
<td>Protein</td>
<td>2076 × 10^12</td>
<td>2078 × 10^12</td>
<td>1205</td>
<td>9070 × 10^4</td>
<td>9150 × 10^4</td>
<td>1346</td>
</tr>
<tr>
<td>Internet</td>
<td>1991 × 10^13</td>
<td>1993 × 10^13</td>
<td>3108</td>
<td>2073 × 10^5</td>
<td>2084 × 10^5</td>
<td>3476</td>
</tr>
<tr>
<td>Chesapeake</td>
<td>1076 × 10^7</td>
<td>1077 × 10^7</td>
<td>12</td>
<td>2219 × 10^7</td>
<td>2231 × 10^7</td>
<td>15</td>
</tr>
<tr>
<td>Delaunay</td>
<td>1809 × 10^11</td>
<td>1809 × 10^11</td>
<td>498</td>
<td>2891 × 10^4</td>
<td>2915 × 10^4</td>
<td>512</td>
</tr>
<tr>
<td>Twitter</td>
<td>6846 × 10^14</td>
<td>6857 × 10^14</td>
<td>897</td>
<td>6846 × 10^14</td>
<td>6857 × 10^14</td>
<td>1015</td>
</tr>
</tbody>
</table>
Figure 1-SBM AIC, BIC values using TS-FCM.
Figure 2-SBM AIC, BIC values using FCM-TS.
Figure 3-DCSBM AIC, BIC values using TS-FCM.
Figure 4-DCSBM AIC, BIC values using TS-FCM.

6. Conclusions
In this work, we present different types of algorithms to solve problems of community detection in complex networks. The statistical models are presented here for modeling and finding the best structure of complex networks. These models are Stochastic Block Model and Degree Corrected
Stochastic Block Model. The objective functions adopted for estimating these block models are likelihood function and modularity function, by applying Integrating TS-FCM. The proposed hybrid algorithm was used for computing maximum likelihood function (SBM-DCSBM) models for different types of the complex and real networks. The experiments of this study were conducted by using different settings and metrics to select the best partitions. Here AIC and BIC were used to define the best model based on optimal values of maximum likelihood function. The results show the ability of the proposed method to find best community structure of the networks. The best values were achieved by the BIC criteria for all networks.

References