DOI: 10.24996/ijs.2020.61.10.27

Iraqi Journal of Science

ISSN: 0067-2904

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.

اكتشاف المجتمع في الشبكات المعقدة باستعمال داله الإمكان الأعظم لنماذج 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 اهتمام واسع في مواضيع متنوعة ، مثل العلوم الاجتماعية والدراسة البيولوجية والسياسة والعمل. هناك دراسات كبيرة ويتم إعطاء الكثير من الأساليب العملية لحل مشكلة إيجاد هيكلية الشبكة. يعرف نموذج الشبكة المعقدة على أساس المجموعات هو مشكلة معقده غير محددة (NP–hard). لا توجد تقنيات مثالية لتحديد التعنقد الامثل هنا نقدم طريقة إحصائية تعتمد على استخدام دالة الامكان الاعظم لنموذج الكتلة العشوائية (SBM). وبالتالي فإن الهدف هو تحديد النموذج العام ، واختيار أفضل نموذج بجودة عالية. لذلك ، يتم تطبيق دمج أسلوب (Tabu search) مع Fuzzy C–mean في مختلف الاعدادات. تم تصميم تجارب متعددة لإيجاد أفضل هيكلية لأنواع مختلفة من الشبكات من خلال تعظيم دالة الامكان الاعظم. ويتم تحديد اختيارات النماذج من خلال تطبيق نوعين من معايير

*Email: ali.f.yaqoob94@gmail.com

اختيار النماذج مثل AIC و BIC. توضح نتائج هذا البحث قدرة الطريقة المقترحة على إيجاد أفضل هيكلية مجتمعية للشبكات المحددة.

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 will be 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), $p = (p_1, \dots, p_k)$ be a probability vector on $[c] := (1, \dots, c_g)$ (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 Λ_{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$P_{X_iX_j}$, 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 $s = \{s_1, \dots, s_N\}$, where s_i is the label of node i and takes values in $\{1,2, \dots, C\}$. For any set of label assignments s, let $J(s)$ be the $C \times C$ matrix defined by

$$J_{cl}(s) = \sum_{ij} \Lambda_{ij} \, I \, \{s_i = c \, , s_j = l\} \qquad (1)$$

where I is the indicator function. Further, let

$$J_c(s) = \sum_l J_{cl}(e) \ , \quad L = \sum_{ij} \Lambda_{ij} \qquad (2)$$

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 $\Im_{cc}$ 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 $f(e) = (\frac{N_1}{N}, \frac{N_2}{N}, \ldots, \frac{N_c}{N})^T$. The SBM network edges variables Λ_{ij} with given true node labels $c = \{c_1, \ldots, c_n\} \in \{1, \ldots, C\}$, are independent Bernoulli random variables with

$$E[\Lambda_{ij}|c] = P_{c_i c_j} \tag{3}$$

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

$$E\left[\Lambda_{ij}\middle|c\right] = \theta_i\, \theta_j\, P_{c_i c_j} \tag{4}$$

where θ_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 θ by giving the following criteria [11].

$$\Upsilon_{DCBM}(s) = \sum_{cl} \mathcal{J}_{cl} \log \frac{\mathcal{J}_{cl}}{\mathcal{J}_c\, \mathcal{J}_l} \tag{5}$$

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

$$\Upsilon_{BM}(s) = \sum_{cl} \mathcal{J}_{cl} \log \frac{\mathcal{J}_{cl}}{N_c\, N_l} \tag{6}$$

The general modularity criteria is

$$\Upsilon(s) = \sum_{ij}\left[\Lambda_{ij} - P_{ij}\right] I\,(s_i = s_j) \tag{7}$$

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.

$$\Upsilon_{ERM}(s) = \sum_c (\mathcal{J}_{cc} - \frac{N_c^2}{N^2} L) . \tag{8}$$

The popular Newman-Girvan modularity (NGM) is

$$\Upsilon_{NGM}(s) = \sum_c (\mathcal{J}_{cc} - \frac{\mathcal{J}_c^2}{L^2} L) . \tag{9}$$

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

$$BIC = \ln(n)\; k - 2 \ln(\widehat{L})$$

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
              $If\ sol1.Cost \geq BestSol1.Cost\ and\ sol2.Cost \geq BestSol2{:}Cost\ then$
                          $BestSol1 = sol1;\ BestSol2 = sol2;$
Step 3.6. Save Best Quality Ever Found
Step 3.7. $Cost.Best(iter, 1) = BestSol1.Cost;$
$Cost.Best(iter, 2) = BestSol2.Cost.$
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 Tabu search 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 nAction1Update $Tabu.List$ 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
              $If\ sol1.Cost \geq BestSol1.Cost\ and\ sol2.Cost \geq BestSol2{:}Cost\ then$
                          $BestSol1 = sol1;\ BestSol2 = sol2;$
Step 4.7. Save Best Quality Ever Found
Step 4.8. $Cost.Best(iter, 1) = BestSol1.Cost;\ Cost.Best(iter, 2) = BestSol2.Cost$
```

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]

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

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

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

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

Network	**AIC (TS-FCM)**	**BIC (TS-FCM)**	**No. of Best Cluster**	**AIC (FCM-TS)**	**BIC (FCM-TS)**	**No. of Best Cluster**
Karate	2224×10^{3}	2224×10^{3}	16	2225×10^{3}	2225×10^{3}	16
Dolphin	1395×10^{4}	1395×10^{4}	29	1394×10^{4}	1394×10^{4}	30
AFC	5463×10^{4}	5463×10^{46}	52	5464×10^{4}	5464×10^{4}	55
Facebook	2228×10^{6}	2235×10^{6}	1545	2228×10^{6}	2235×10^{6}	1545
Political	2734×10^{12}	2735×10^{12}	586	2066×10^{5}	2069×10^{5}	747
Protein	2076×10^{12}	2078×10^{12}	1205	9070×10^{4}	9150×10^{4}	1346
Internet	1991×10^{13}	1993×10^{13}	3108	2073×10^{5}	2084×10^{5}	3476
Chesapeake	1076×10^{7}	1077×10^{7}	12	2219×10^{7}	2231×10^{7}	15
Delaunay	1809×10^{11}	1809×10^{11}	498	2891×10^{4}	2915×10^{4}	512
Twitter	6846×10^{14}	6857×10^{14}	897	6846×10^{14}	6857×10^{14}	1015

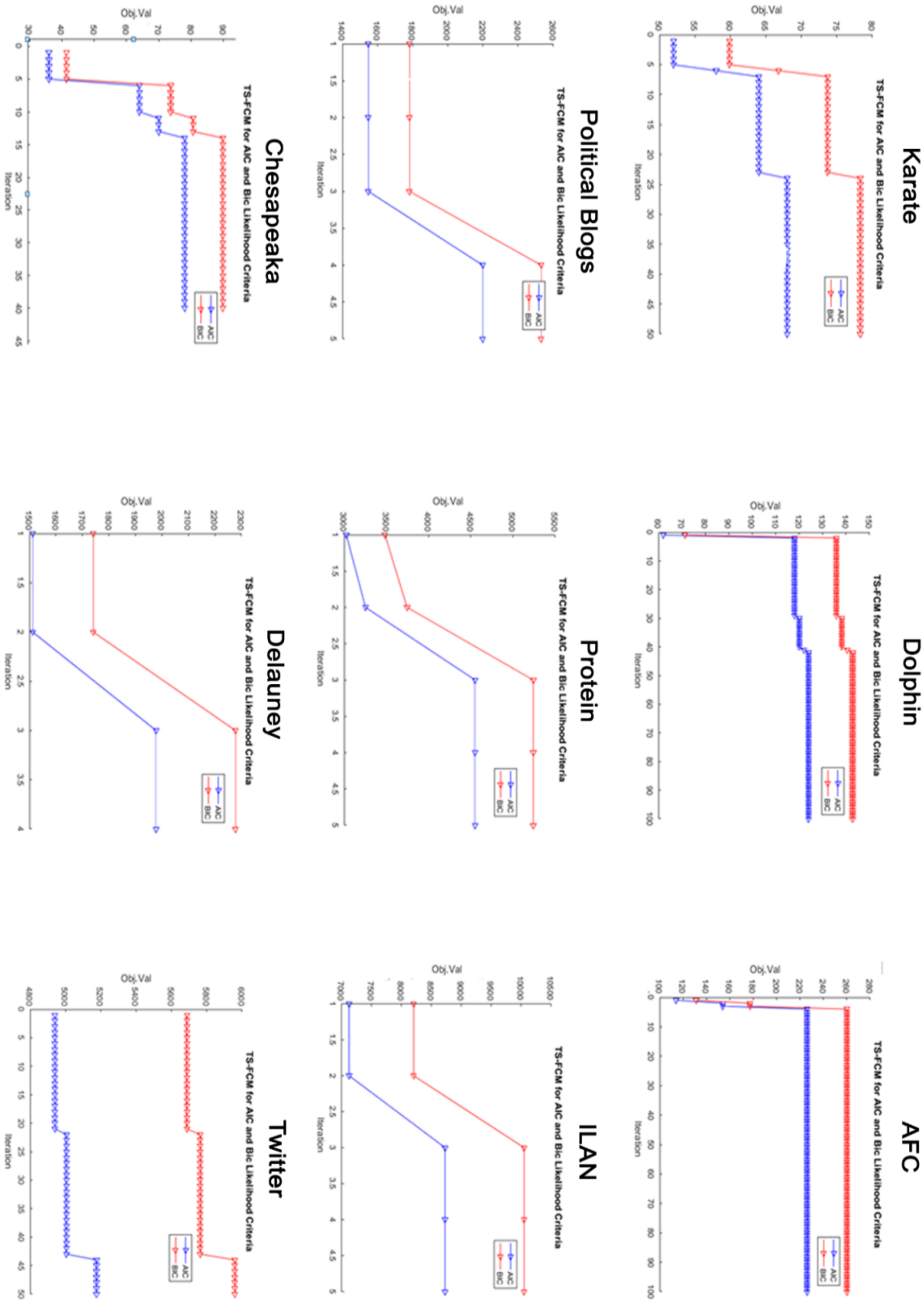

Figure 1-SBM AIC, BIC values using TS-FCM.

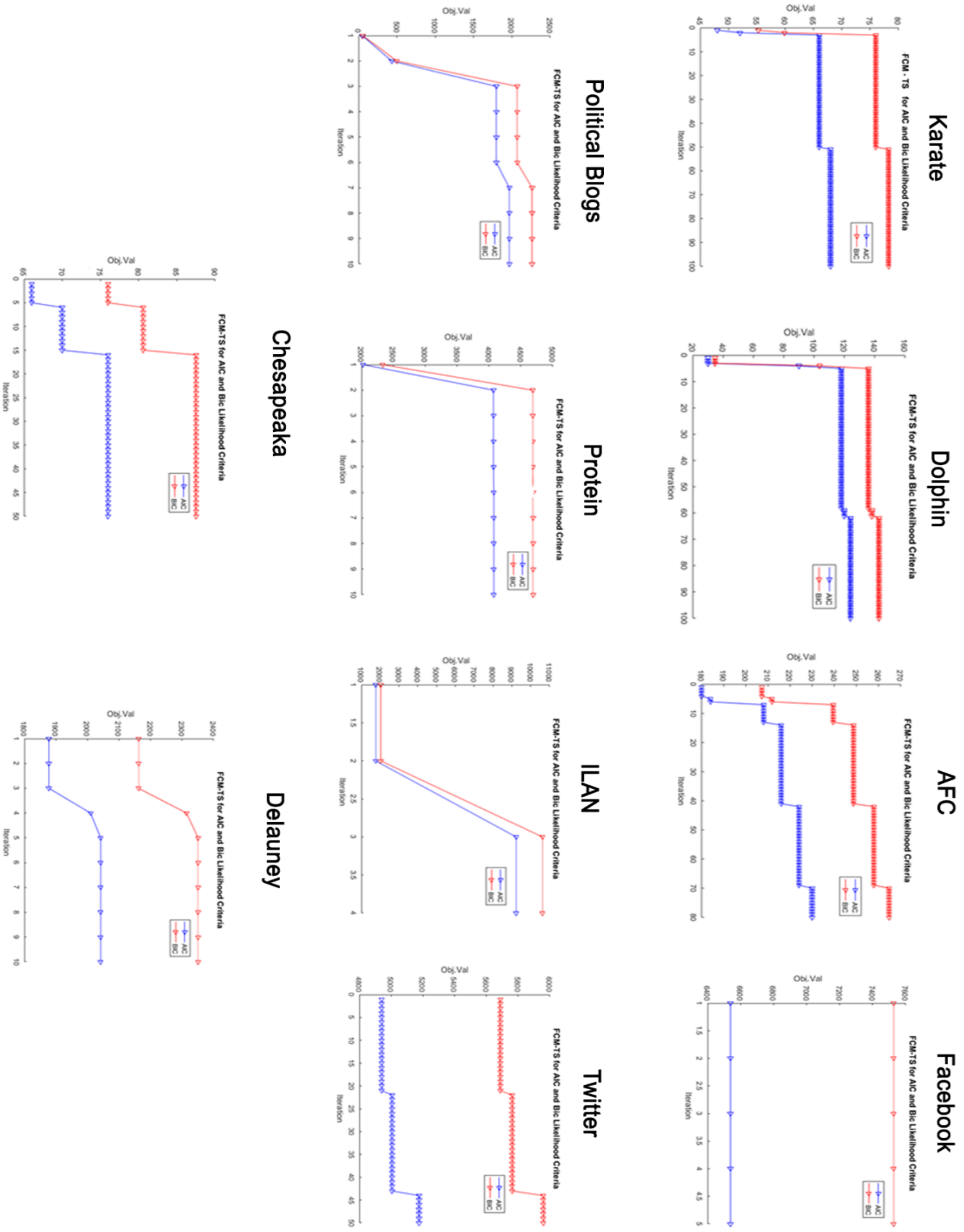

Figure 2-SBM AIC, BIC values using FCM-TS.

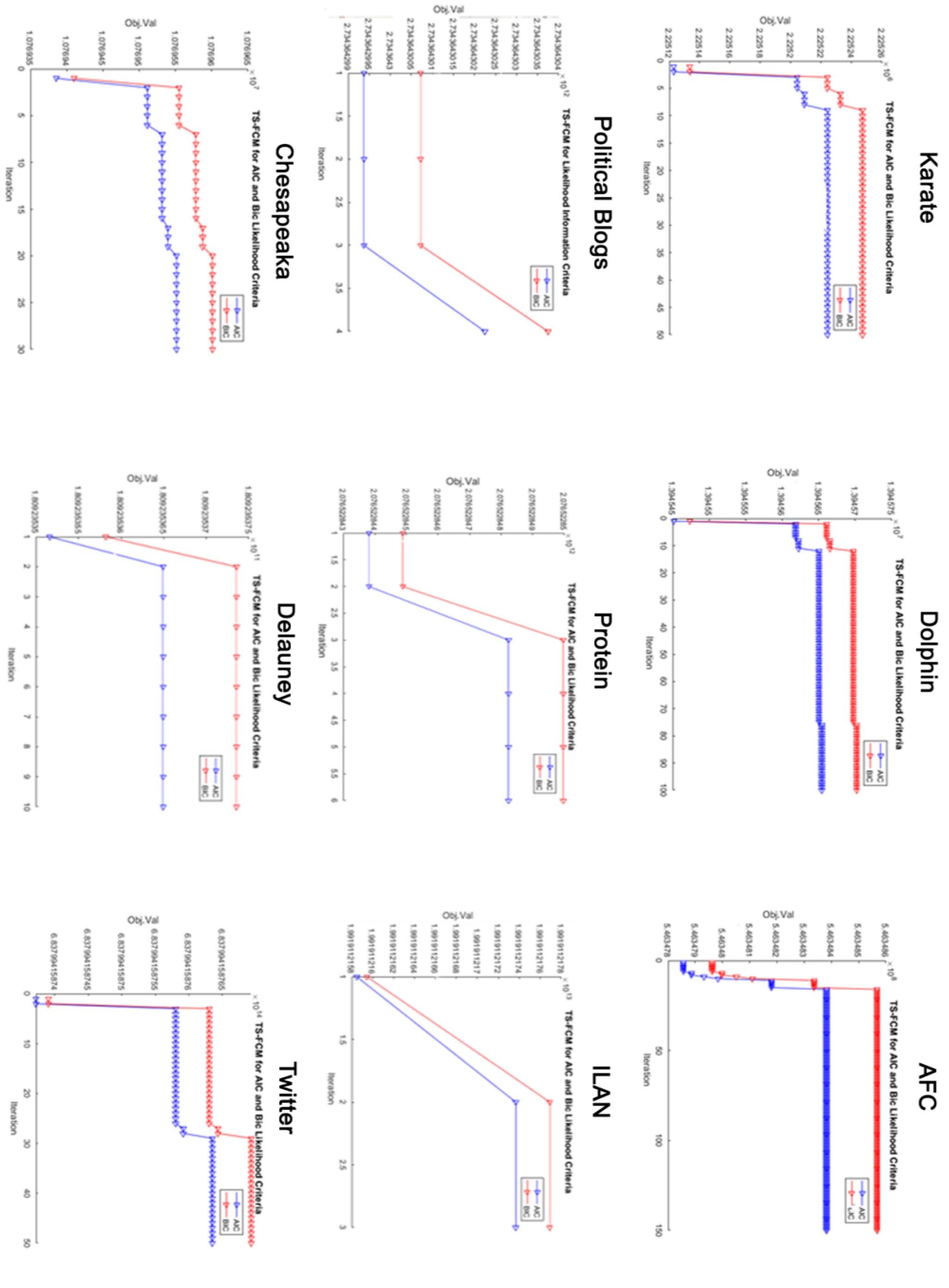

Figure 3-DCSBM AIC, BIC values using TS-FCM.

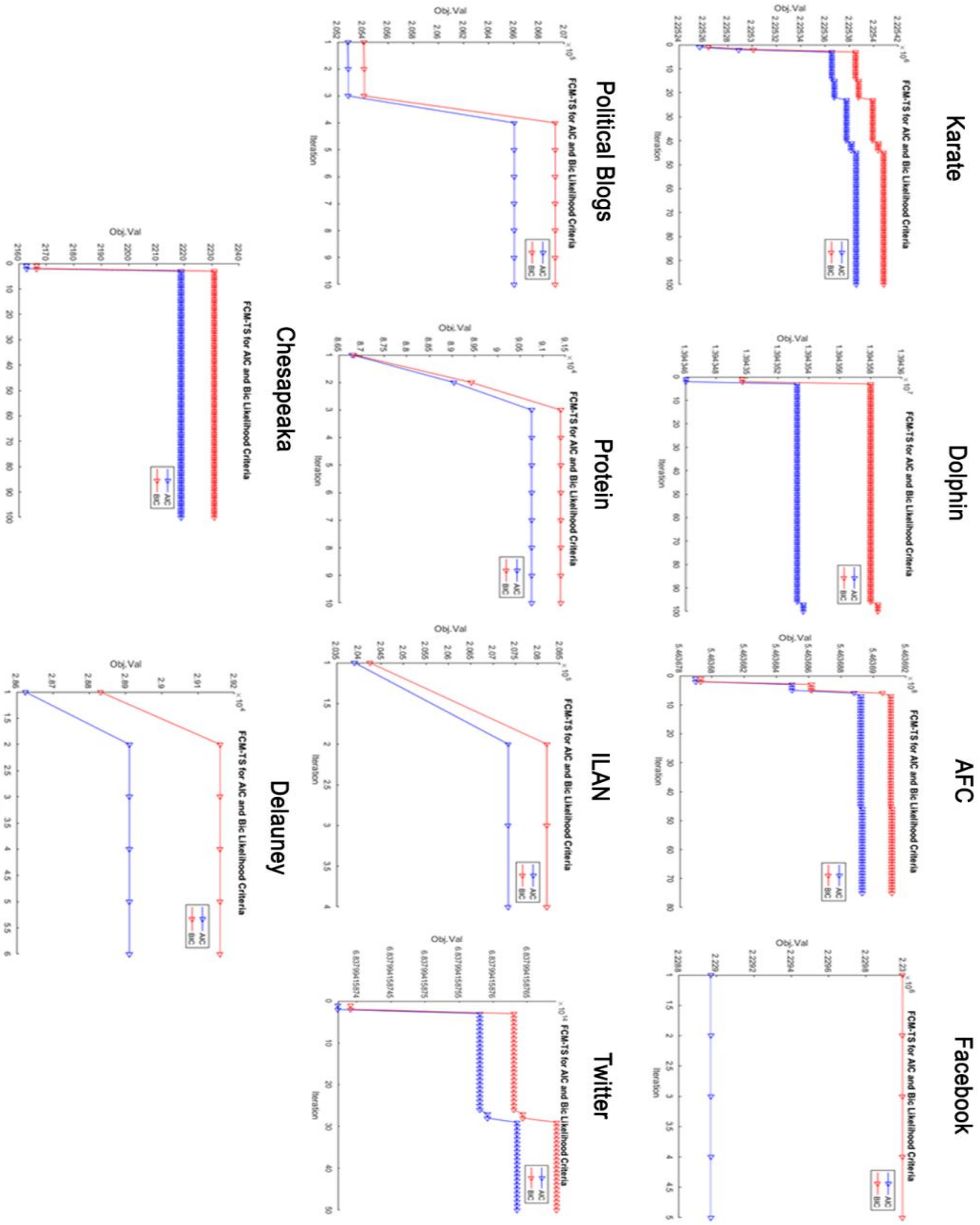

Figure 4-DCSBM AIC, BIC values using TS-FCM.

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

1. Mark EJ Newman. **2004**. "Fast algorithm for detecting community structure in networks." *Physical review E*; **69**(6): 066133.
2. Santo Fortunato and Darko Hric. **2016**. "Community detection in networks: A user guide." *Physics Reports*; **659**: 1–44.
3. Jierui Xie, Stephen Kelley, and Boleslaw K Szymanski. **2013**. "Overlapping community detection in networks: The state-of-the-art and comparative study." *Acm computing surveys* (csur); **45**(4): 43.
4. Steve Harenberg, Gonzalo Bello, L Gjeltema, Stephen Ranshous, Jitendra Harlalka, Ramona Seay, Kanchana Padmanabhan, and Nagiza Samatova. **2014**. "Community detection in large-scale networks: a survey and empirical evaluation." Wiley Interdisciplinary Reviews: *Computational Statistics*; **6**(6): 426–439.
5. Emmanuel Abbe and Colin Sandon. **2015**. "Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery." In Foundations of Computer Science (FOCS), IEEE 56th Annual Symposium on, 2015; pages 670–688. IEEE.
6. Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborov´a. **2011**. "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications." *Physical Review E*; **84**(6): 066106.
7. Etienne Come and Pierre Latouche. **2015**. "Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood." *Statistical Modelling*; **15**(6): 564–589.
8. Xiaoran Yan, Cosma Shalizi, Jacob E Jensen, Florent Krzakala, Cristopher Moore, Lenka Zdeborov´a, Pan Zhang, and Yaojia Zhu. **2014**. "Model selection for degree-corrected block models." Journal of Statistical Mechanics: *Theory and Experiment*; (5):P05007.
9. Tai Qin and Karl Rohe. **2013**. "Regularized spectral clustering under the degree-corrected stochastic block model." In Advances in Neural Information Processing Systems; pages 3120–3128.
10. Emmanuel Abbe. **2017**. "Community detection and stochastic block models: recent developments". ; ArXiv: 1703.10146.
11. Zhao, Y., Levina, E., Zhu, J., et al. **2012**. "Consistency of community detection in networks under degree-corrected stochastic block models." *The Annals of Statistics*; **40**(4): 2266-2292.
12. Joseph E Cavanaugh. **1997**. "Unifying the derivations for the akaike and corrected akaike information criteria." *Statistics & Probability Letters,* **33**(2): 201–208.
13. John J Dziak, Donna L Coffman, Stephanie T Lanza, and Runze Li. 2017. "Sensitivity and specificity of information criteria." Peer J Preprints.
14. Fred Glover. **1989**. "Tabu search part I." *ORSA Journal on computing*; **1**(3): 190–206.
15. Fred Glover. **1990**. "Tabu search part II." *ORSA Journal on computing*; **2**(1): 4–32.
16. J.C. Bezdek. 2013. "Pattern Recognition with Fuzzy Objective Function Algorithms." Advanced Applications in Pattern Recognition. Springer US; ISBN 9781475704501. URL https://books.google.iq/books?id=z6XqBwAAQBAJ.
17. Stanford Large Network Dataset Collection, https://snap.stanford.edu/data/index.html.