2016 International Symposium on Nonlinear Theory and Its Applications,
NOLTA2016, Yugawara, Japan, November 27th-30th, 2016

# An Alternative to Basic Log-likelihood for Bayesian Network Clustering

Rei Oshino† and Koujin Takeda†

†Department of Intelligent Systems Engineering, Ibaraki University
4–12–1, Nakanarusawa Hitachi, 316-8511 Ibaraki, Japan
Email: 16nm912r@vc.ibaraki.ac.jp, koujin.takeda.kt@vc.ibaraki.ac.jp

**Abstract**—We study clustering problem of vertices on graph by Bayesian inference. In Bayesian framework of clustering, stochastic block model is a standard for construction of likelihood. Here we start with a variant of stochastic block model by Karrer and Newman for theoretical discussion. It is known that naïve log-likelihood from their model does not always give natural clustering. We discuss how to modify it by adding correction term. By numerical experiment, we verify advantage of our method.

## 1. Introduction

Clustering is one of the main topics of unsupervised machine learning. In particular, clustering of vertices on graph, which we refer to as network clustering in this article (often termed as community detection generally), is also of great importance for extracting blockwise structure of real world network.

The idea of network clustering is not clearly defined mathematically in general. Its goal is to extract natural clusters for human sense. Toward this goal, several methods have been proposed and frequently used such as maximization of modularity[1] or spectral clustering[2].

Among them, we focus on Bayesian inference. In this framework, a probabilistic model of network under given network parameter is necessary, and stochastic block model (SBM)[3, 4] is a standard for the purpose. There are many variants of SBM in the definition of network probability, and we follow the one by Karrer and Newman[5] for theoretical discussion. The advantage of their SBM is its analytical simplicity: We can optimize some model parameters analytically and remove them from the model. As a result, remaining parameters are only the numbers of edges between/within clusters and the numbers of vertices in one cluster, and we maximize log-likelihood with respect to them to extract clusters. However it is reported that such naïve approach sometimes yields unnatural clusters. One way to cope with this problem is the replacement of their SBM to another, while we must avoid overfitting due to overcomplex model with many additional model parameters.

We attempt another approach here: We make an appropriate choice of information criterion. In general Bayesian inference framework, several famous information criteria are frequently used, where correction term is added to log-likelihood. In our problem the same approach can be taken naturally. However, it should be noted that for network clustering there may be an appropriate information criterion among many possibilities.

In this article we propose a form of correction term for natural network clustering, which originates from intuitive discussion. Using log-likelihood with our correction term, we conduct numerical experiment of network clustering for real network data, and compare the result with naïve method. We also discuss the relationship of our method with other works[6, 7, 8], where other information criteria are derived and proposed by theoretical argument.

## 2. Model

### 2.1. Naïve SBM

The probability of naïve SBM in [5] is given by

$$P_{\mathrm{SBM}}(G|\boldsymbol{\omega},\boldsymbol{g}) := \prod_{i<j}\frac{(\omega_{g_i,g_j})^{A_{ij}}}{A_{ij}!}\exp(-\omega_{g_i,g_j}) \times\prod_{i}\frac{(\frac{1}{2}\omega_{g_i,g_j})^{A_{ii}/2}}{(A_{ii}/2)!}\exp\left(-\frac{1}{2}\omega_{g_i,g_j}\right). \quad (1)$$

Roman subscripts $i, j$ describe vertices on network. $A_{ij}$ is $(ij)$-element of network adjacency matrix. $g_i$ is cluster index to which vertex $i$ belongs. $\omega_{g_i,g_j}$ is the expectation value of the number of edges between clusters $g_i$ and $g_j$, under the condition that the number of edges between vertex pair is independently Poisson distributed. The symbols $G$, $\boldsymbol{\omega}$ and $\boldsymbol{g}$ on l.h.s. denote a network sample, the set of $\omega_{ij}$, and the set of $g_i$(= cluster assignment index on each vertex), respectively.

Naïve Bayesian network clustering is based on the maximization of the log-likelihood with respect to $\boldsymbol{\omega}, \boldsymbol{g}$ under a given network sample $G$.

$$\mathcal{L}_{\mathrm{SBM}}(G|\boldsymbol{\omega},\boldsymbol{g}) := \log P_{\mathrm{SBM}}(G|\boldsymbol{\omega},\boldsymbol{g}). \quad (2)$$

After maximization with respect to $\boldsymbol{\omega}$ analytically, the maximum log-likelihood is expressed as

$$\mathcal{L}_{\mathrm{SBM}}(G|\boldsymbol{g}) := \max_{\boldsymbol{\omega}}\{\log P_{\mathrm{SBM}}(G|\boldsymbol{\omega},\boldsymbol{g})\} = \sum_{\alpha\beta} m_{\alpha\beta}\log\frac{m_{\alpha\beta}}{n_\alpha n_\beta}, \quad (3)$$

where

$$m_{\alpha\beta} := \sum_{ij} A_{ij}\delta_{g_i,\alpha}\delta_{g_j,\beta}, \quad (4)$$

is the number of edges between clusters indexed by Greek letters $\alpha, \beta$, and $n_\alpha$ is the number of vertices in cluster $\alpha$. Then, our task is to maximize log-likelihood in (3) for a given $G$ with respect to cluster assignment $\boldsymbol{g}$, or equivalently the set of variables $\{m_{\alpha\beta}, n_\alpha\}$.

Naïve SBM in [5] is suitable for theoretical analysis or argument, however it generates the network without natural clusters. Therefore this model must be corrected.

### 2.2. Degree-corrected SBM

Degree-corrected SBM is also proposed in [5] as a variant of naïve SBM, where the probability of network $G$ is replaced by

$$P_{\mathrm{DC-SBM}}(G|\boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{g}) := \prod_{i<j} \frac{(\theta_i \theta_j \omega_{g_i,g_j})^{A_{ij}}}{A_{ij}!} \exp(-\theta_i \theta_j \omega_{g_i,g_j}) \times \prod_i \frac{(\frac{1}{2}\theta_i^2 \omega_{g_i,g_j})^{A_{ii}/2}}{(A_{ii}/2)!} \exp\left(-\frac{1}{2}\theta_i^2 \omega_{g_i,g_j}\right). \tag{5}$$

$\theta_i$ is the expected value of degree(= the number of connected edges) for vertex $i$, and $\boldsymbol{\theta}$ is the set of $\theta_i$. After maximization with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$ analytically, log-likelihood becomes

$$\begin{aligned} \mathcal{L}_{\mathrm{DC-SBM}}(G|\boldsymbol{g}) &:= \max_{\boldsymbol{\omega},\boldsymbol{\theta}} \{\log P_{\mathrm{DC-SBM}}(G|\boldsymbol{\omega}, \boldsymbol{\theta}, \boldsymbol{g})\} \\ &= \sum_{\alpha\beta} m_{\alpha\beta} \log \frac{m_{\alpha\beta}}{\kappa_\alpha \kappa_\beta}, \end{aligned} \tag{6}$$

where $\kappa_\alpha$ is the sum of all vertex degrees in cluster $\alpha$. Although we introduce novel parameter set $\boldsymbol{\theta}$, the final expression of log-likelihood is as simple as the original naïve model: $n_\alpha$ is just replaced by $\kappa_\alpha$. Therefore, degree-correlated SBM is regarded as the simplest variant of naïve SBM.

With this modification, we can extract more natural clusters in comparison with naïve SBM, however it still fails to extract clusters in some graphs.

## 3. Method

We start our discussion with log-likelihoods of naïve SBM (3) or degree-correlated SBM (6). For appropriate information criteria, we add correction terms to these log-likelihoods as follows,

$$\mathcal{L}^*_{\mathrm{SBM}}(G|\boldsymbol{g}) := \mathcal{L}_{\mathrm{SBM}}(G|\boldsymbol{g}) + \sum_\alpha m_{\alpha\alpha} - \sum_{\alpha<\beta} m_{\alpha\beta}. \tag{7}$$

$$\mathcal{L}^*_{\mathrm{DC-SBM}}(G|\boldsymbol{g}) := \mathcal{L}_{\mathrm{DC-SBM}}(G|\boldsymbol{g}) + \sum_\alpha m_{\alpha\alpha} - \sum_{\alpha<\beta} m_{\alpha\beta}. \tag{8}$$

Intuitively, the first correction term enhances the density of edges inside single cluster, and the second correction term penalizes the intertwining edges between clusters. By maximizing the log-likelihood with these correlations, extraction of more natural clusters is expected.

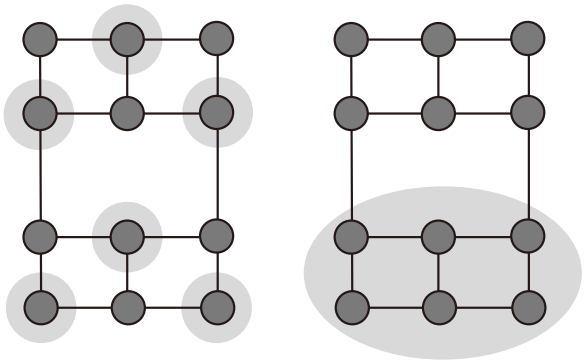

Figure 1: The result of clustering for a designed network. Left: Clusters by maximization of $\mathcal{L}_{\mathrm{DC-SBM}}(G|\boldsymbol{g})$. Right: Clusters by maximization of $\mathcal{L}^*_{\mathrm{DC-SBM}}(G|\boldsymbol{g})$. Vertices on shaded background belong to the same cluster.

The advantage of such correction terms is that we do not need to introduce novel network parameters. We only use $m_{\alpha\beta}$ in correction terms, which has already been in the original log-likelihood. This is desirable for avoiding over-complex modeling. On the other hand, we should keep in mind that arbitrary multiplication factors can be introduced to correction terms. We set them unity in this article.

## 4. Numerical Experiment

### 4.1. Clustering of artificial network

First, we prepare several artificial small-size networks, in which clusters are obvious. An example of artificial network is shown in Figure 1. We extract the clusters by maximizing $\mathcal{L}_{\mathrm{DC-SBM}}$ or $\mathcal{L}^*_{\mathrm{DC-SBM}}$, where the number of clusters(= two) is known in advance. For maximization, we conduct exhaustive search in this experiment.

The result is depicted in the same figure. Log-likelihood with correlation $\mathcal{L}^*_{\mathrm{DC-SBM}}$ yields natural clusters, whereas naïve log-likelihood $\mathcal{L}_{\mathrm{DC-SBM}}$ gives unnatural ones. We also attempted other artificial networks having similar network/cluster structure, and found that $\mathcal{L}^*_{\mathrm{DC-SBM}}$ is also more successful. Hence we conclude that our method can cure the defect in the naïve model.

In Figure 1, vertices in the same cluster are not mutually intertwined by many edges. By naïve SBM, such networks will be generated only with very low probability, therefore we cannot extract natural clusters. By incorporating correction terms, we can avoid unnatural results.

### 4.2. Clustering of small size network dataset

Next we apply our method to real network data of relatively small size (the number of vertices is $10^1 \sim 10^2$).

| | (A) $\mathcal{L}_{\mathrm{DC-SBM}}(G\|\boldsymbol{g})$ | (B) $\mathcal{L}^*_{\mathrm{DC-SBM}}(G\|\boldsymbol{g})$ | (C) both |
|---|---|---|---|
| "Karate Club" | 0.06 | 0.27 | 0.04 |
| "Dolphin" | 0.01 | 0.16 | 0.04 |

Table 1: Clustering result of Zachary's "Karate club" dataset and "Dolphin" dataset. We search global maximum of $\mathcal{L}_{\mathrm{DC-SBM}}(G|\boldsymbol{g})$ or $\mathcal{L}^*_{\mathrm{DC-SBM}}(G|\boldsymbol{g})$ with 100 random initial conditions. We show the successful ratio of three cases: (A) Global maximum is found only for $\mathcal{L}_{\mathrm{DC-SBM}}(G|\boldsymbol{g})$. (B) Only for $\mathcal{L}^*_{\mathrm{DC-SBM}}(G|\boldsymbol{g})$. (C) For both.

We use Zachary's "Karate club" dataset[9] (34 vertices), and "Dolphin" dataset[10] (62 vertices). There is the correct answer of clustering only for "Karate club" dataset, where two clusters exist. We extract the clusters on these networks by maximizing $\mathcal{L}_{\mathrm{DC-SBM}}$ or $\mathcal{L}^*_{\mathrm{DC-SBM}}$, and by assuming again the number of clusters is two and known, although "Dolphin" dataset is thought to have more communities (i.e. we focus on the largest two clusters).

For extraction, we use Kernighan-Lin algorithm[11]. In this algorithm we select the most relevant vertex to increase $\mathcal{L}_{\mathrm{DC-SBM}}(G|\boldsymbol{g})$ or $\mathcal{L}^*_{\mathrm{DC-SBM}}(G|\boldsymbol{g})$ by the change of its cluster assignment, then actually move it to another cluster. We repeat it until reaching local (or sometimes global) maximum of $\mathcal{L}_{\mathrm{DC-SBM}}(G|\boldsymbol{g})$ or $\mathcal{L}^*_{\mathrm{DC-SBM}}(G|\boldsymbol{g})$. We perform this algorithm under 100 random initial cluster assignments, and among 100 final resulting assignments we regard the largest $\mathcal{L}_{\mathrm{DC-SBM}}(G|\boldsymbol{g})$ or $\mathcal{L}^*_{\mathrm{DC-SBM}}(G|\boldsymbol{g})$ as global maximum, which appears under several initial conditions.

As a result, we finally reach the same cluster assignments both by $\mathcal{L}_{\mathrm{DC-SBM}}(G|\boldsymbol{g})$ and by $\mathcal{L}^*_{\mathrm{DC-SBM}}(G|\boldsymbol{g})$ for "Karate club" and "Dolphin" datasets. In addition, the final cluster assignments are natural. (For "Karate club", only one vertex is misclassified.) However, the successful ratios of reaching global maximum from 100 initial assignments are different, as summarized in Table 1. This means that natural cluster assignment can be found by corrected log-likelihood $\mathcal{L}^*_{\mathrm{DC-SBM}}(G|\boldsymbol{g})$ more easily than the naïve one $\mathcal{L}_{\mathrm{DC-SBM}}(G|\boldsymbol{g})$.

### 4.3. Clustering of large size network dataset

Next we apply our method to real network data of relatively large size (the number of vertices is $10^2 \sim 10^3$). We use the datasets as follows:

- "football"[12] (114 vertices, 1224 edges)
- "euroroad"[13] (1174 vertices, 2834 edges)
- "netscience"[14](1460 vertices, 5484 edges)

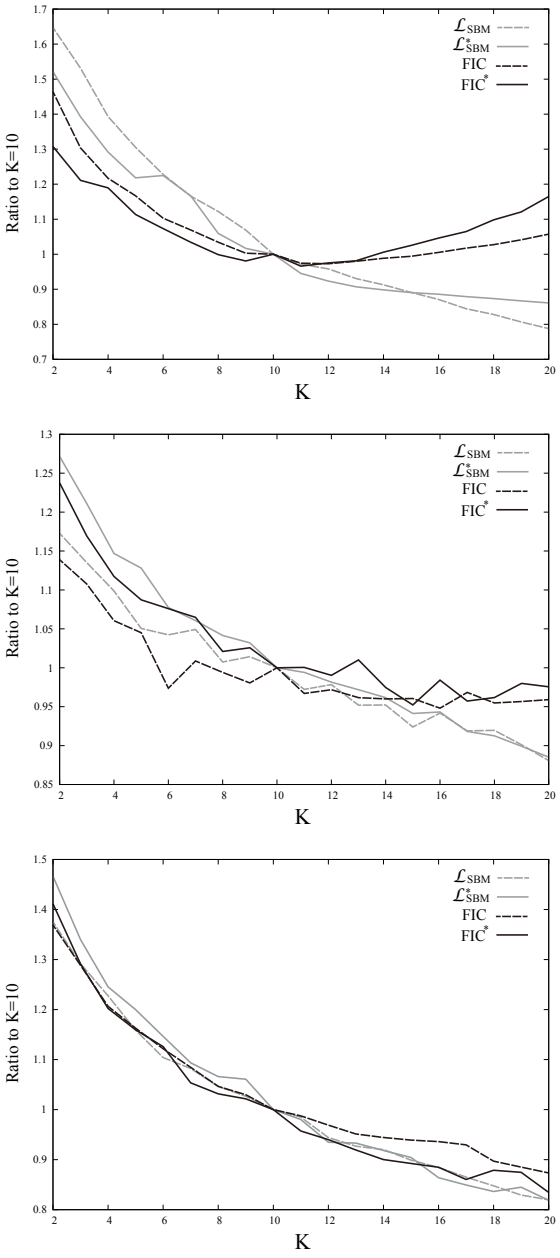


Figure 2: Clustering result of large size network dataset. Top: "football", Middle: "euroroad", Bottom: "netscience".

Here we assume the number of clusters (denoted by $K$) is unknown, which must be determined. For this purpose, we maximize the following quantities with respect to $K$:

- $\mathcal{L}_{\mathrm{SBM}}$ in (2)
- $\mathcal{L}^*_{\mathrm{SBM}}$ in (7)
- Factorized Information Criterion (FIC)[7, 8]: It is proposed for factorized asymptotic Bayesian inference and can be applied to network clustering by SBM for determination of optimal $K$.

- FIC with our correction term like (7) and (8) (denoted by FIC$^*$ in Figure 2)

In this experiment, we first fix $K$ and maximize these quantities. For maximization, we perform simulated annealing under a single random initialization of cluster assignment. Then we vary $K$ and search the maximum with respect to $K$ as well.

The result is depicted in Figure 2. The ratio of each quantity to $K = 10$ is shown on the vertical axis. It should be noted that the maximum of the original quantity is depicted as minimum, because all quantities are computed as negative.

For "football", $\mathcal{L}_{\mathrm{SBM}}$ and $\mathcal{L}^*_{\mathrm{SBM}}$ do not give the minimum. On the other hand, ICL and ICL$^*$ give the minimum around $K = 10$. The result of $K = 10$ seems quantitatively natural for the size of this dataset from the viewpoint of unsupervised learning, although we do not show the picture of clustering result here. For ICL$^*$, the curvature of the graph around minimum seems larger than ICL, therefore this implies our method helps the determination of $K$ by an arbitrary optimization algorithm.

We cannot find a clear minimum for other two datasets, however, for "euroroad" there might be a minimum in the range of $10 < K < 20$, and the experiment with high precision will be desired. For "netscience", the experiment of higher $K$ will be necessary for finding a minimum.

## 5. Summary and discussion

We proposed an alternative to naïve log-likelihood for network clustering. By the result of numerical experiment, we verified our log-likelihood with correction yields more natural result of network clustering, or enables us to find natural clusters more easily.

Information criteria for network clustering are also proposed in preceding works, and FIC is one of them. In [8], the higher order correction to FIC under Laplace approximation is calculated under sparse network structure. Their resulting criterion is termed F$^2$IC in their article, where the number of edges between clusters plays a significant role like our method. These criteria are obtained analytically, and we must discuss the relation between our method with them as a future work.


## Acknowledgments

We are thankful to T. Kawamoto for his comment on the related work[8]. This work is supported by KAKENHI Nos. 24700007, 25120013 (KT).


## References


[1] M. E. J. Newman and M. Girvan, "Finding and Evaluating Community Structure in Networks," *Phys. Rev. E*, vol.69, 026113, 2004.

[2] J. Shi and J. Malik: "Normalized Cut and Image Segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.22, pp.888-905, 2000.

[3] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic Blockmodels: First Steps," *Social Networks*, vol.5, pp.109-137, 1983.

[4] S. Wasserman and C. J. Anderson, "Stochastic a Posteriori Blockmodels: Construction and Assessment," *Social Networks*, vol.9, pp.1-36, 1987.

[5] B. Karrer and M. E. J. Newman, "Stochastic Blockmodels and Community Structure in Networks," *Phys. Rev. E*, vol.83, 016107, 2011.

[6] J. J. Daudin, F. Picard, and S. Robin, "A Mixture-model for Random Graphs," *Statistics and Computing*, vol.18, pp.173-183, 2008.

[7] R. Fujimaki and S. Morinaga, "Factorized Asymptotic Bayesian Inference for Mixture Modeling," *proc. of Artificial Intelligence and Statistics Conference*, 2012.

[8] K. Hayashi, T. Konishi, and T. Kawamoto, "A Tractable Fully Bayesian Method for the Stochastic Block Model," *preprint*, arXiv:1602.02256.

[9] W. Zachary, "An Information Flow Models for Conflict and Fission in Small Groups," *J. of Anthropological Research*, vol.33, pp.452-473, 1997.

[10] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The Bottlenose Dolphin Community of Doubtful Sound Features a Large Proportion of Long-Lasting Associations - Can Geographic Isolation Explain This Unique Trait?," *Behavioral Ecology and Sociobiology*, vol.54, pp.396-405, 2003.

[11] B. W. Kernighan and S. Lin, "An Efficient Heuristic Procedure for Partitioning Graphs," *The Bell System Technical Journal*, pp.231-307, 1970.

[12] M. Girvan and M. E. J. Newman, "Community Structure in Social and Biological Networks," *Proc. of Nat. Acad. Sci.*, vol.99, pp.7821-7826, 2002.

[13] L. Šubelj and M. Bajec, "Robust Network Community Detection Using Balanced Propagation," *Eur. Phys. J. B*, vol.81, pp.353-362, 2011.

[14] M. E. J. Newman, "Finding Community Structure in Networks Using the Eigenvectors of Matrices," *Phys. Rev. E*, vol.74, 036104, 2006.