Physica A 496 (2018) 339–350

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Mixture models with entropy regularization for community detection in networks

Zhenhai Chang [a,d], Xianjun Yin [a,*], Caiyan Jia [b,c,**], Xiaoyang Wang [b,c]

[a] School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, China
[b] School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
[c] Beijing Key Lab of Traffic Data Analysis and Mining, Beijing, China
[d] School of Mathematics and Statistics, Tianshui Normal University, Gansu, China

HIGHLIGHTS

- An entropy regularized mixture model (called EMM) is proposed.
- EMM is able to infer the number of communities and meanwhile identify network structure.
- The new method is compared with several competing methods on a range of well-known synthetic and real networks.

ARTICLE INFO

Article history:
Received 27 March 2017
Received in revised form 18 November 2017
Available online 4 January 2018

MSC:
00-01
99-00

Keywords:
Complex networks
Community detection
Mixture models
Entropy

ABSTRACT

Community detection is a key exploratory tool in network analysis and has received much attention in recent years. NMM (Newman's mixture model) is one of the best models for exploring a range of network structures including community structure, bipartite and core–periphery structures, etc. However, NMM needs to know the number of communities in advance. Therefore, in this study, we have proposed an entropy regularized mixture model (called EMM), which is capable of inferring the number of communities and identifying network structure contained in a network, simultaneously. In the model, by minimizing the entropy of mixing coefficients of NMM using EM (expectation–maximization) solution, the small clusters contained little information can be discarded step by step. The empirical study on both synthetic networks and real networks has shown that the proposed model EMM is superior to the state-of-the-art methods.

1. Introduction

The study of complex networks has become an important area of multidisciplinary research involving physics, mathematics, biology, social science, informatics, and other theoretical and applied sciences [1–6]. However, it is not easy to understand the complex networks by simple observation because of their complexity. One of the most effective approaches to reduce this complexity is to uncover the structures of these real-world networks [7,8]. A large number of algorithmic methods have been proposed to explore structures of observed networks during the last several years [9–14].

The Newman's Mixture Model (NMM) [15] is one of the best tools for exploring structures. The most striking advantage of NMM lies in its ability to identify a very broad range of types of structure in networks without any prior knowledge of

* Corresponding author.
** Correspondence to: No.3 Shangyuancui, Haidian District, Beijing 100044, China.
E-mail addresses: yinxj@cufe.edu.cn (X. Yin), cyjia@bjtu.edu.cn (C. Jia).

https://doi.org/10.1016/j.physa.2018.01.002
0378-4371/

the structures [16], such as assortative and disassortative structures [17], mixture structures [18], and so forth. Another advantage of NMM is that the algorithm of NMM is capable of dealing with the directed networks, undirected networks, and weighted networks. In addition, the efficiency of the algorithm is high in terms of computation complexity [16]. However, NMM assumes that in a community the total outgoing degree must be larger than zero [16,19]. To overcome this limitation, Ramasco and Mungan (2008) [19] suggested dealing with the incoming degrees, outgoing degrees, and bidirectional degrees separately; Wang and Lai (2008) [16] solved this problem by assuming that all nodes in a community share the same prior probability to connect unidirectionally to a given node. Moreover, NMM needs to know the number of communities contained in a network in advance. This is a common problem of most existing community detection methods.

In the literature, there are several approaches to infer the number of communities and identify network structures, simultaneously. One approach is modularity optimization [20]. Modularity optimization is the best known method where the modularity function is defined to measure the quality of communities contained in a network. Optimization methods such as greedy optimization [20], extremal optimization [21], and simulated annealing [22], etc. are used to optimize the modularity function. However, the modularity has been exposed to resolution limits [23–25]. Another approach is to use Bayesian statistical inference based on stochastic generative model. In particular, some of Bayesian inference methods directly put prior distributions on the number of communities K. Chen et al. (2015, 2016) [26,27] considered a fully Bayesian framework, in which the Chinese restaurant process (CRP) [28] was chosen as a prior distribution and placed on the number of communities K. While it has been recently observed that CRPs lead to inconsistent estimation of the number of communities under certain conditions [29,30]. Newman and Reinert (2016) [8] combined empirical Bayes method and maximum-entropy prior criterion, in which uniform probability distributions were placed on the number of communities K, the community assignment probabilities π, and an exponential distribution on the edge probabilities θ. Then, the posterior probability $P(K|A)$ was calculated by Markov chain Monte Carlo importance sampling [31], where A was the adjacency matrix of the observed network. The most likely value of K was the one for which $P(K|A)$ was the largest. On the contrary, some of Bayesian inference methods did not directly assume prior distributions on the number of communities K. D Jin et al. (2016) [32] used a hierarchical Bayesian approach based on the idea of ranking node popularities within communities to find K and identify network structures, simultaneously. Exponential prior was placed on each column of the expected degree matrix $D = (d_{ik})_{n\times K}$ (instead of K), where d_{ik} was the expected degree of node i in the kth community. After compressing the columns of D and removing the irrelevant communities k whose expected degrees were zero or very close to zero, the inferred number of communities was derived.

Different from the above ideas, we attempt to put a prior entropy on mixing coefficients π contained in NMM to infer the number of communities, which needs to be known in advance in the original version NMM. According to the principle of maximum entropy, if nothing is known about a distribution, then the distribution with the largest entropy should be chosen as the least-informative default. This implies that maximum-entropy (least informative) prior probability distributions on π are uniformly random, i.e., $\pi_r = 1/K$ $(r = 1, 2, \ldots, K)$. From the view of community detection, the same mixing coefficients $\pi_r = 1/K$ $(r = 1, 2, \ldots, K)$ mean that the probabilities that nodes are assigned to each community are the same. This leads to the same size of communities. On the contrary, by minimizing the prior entropy, the sizes of some communities will become smaller, and others will become larger. If the sizes of some communities are too small (i.e., the corresponding mixing coefficients are zero or close to zero), we remove them and the number of communities goes down. Therefore, putting a prior entropy on mixing coefficients can control the sizes of all communities. A natural approach for mixture models with an unknown number of communities is to put a prior entropy on mixing coefficients [33]. In this study, based on this idea, to overcome the limitation of NMM that assumes to have the knowledge of the number of communities in advance, we propose an entropy regularized NMM method (EMM) to infer the number of communities and to detect the structure of a network, simultaneously. Roughly, our method can be divided into two stages: the first stage is to minimize the entropy for obtaining the number of communities, the second stage is to estimate the parameters of NMM for getting the community labels of nodes. Theoretical analysis and experimental tests have shown the effectiveness of EMM.

The rest of the paper is structured as follows. In Section 2, we present an entropy regularized mixture model (EMM). In Section 3, we show the performance of our method on some artificial networks and real networks. Finally, we draw the conclusions in Section 4.

2. An entropy regularized mixture model (EMM)

In this section, we first review NMM, then introduce our entropy regularized mixture model (EMM).

2.1. Newman's mixture model (NMM)

Suppose we have a network of n nodes connected by directed edges, the adjacency matrix of observed network is denoted by A with elements $A_{ij} = 1$ if there is an edge from node i to j and 0 otherwise. The likelihood of NMM is defined as follows [15].

$$\begin{aligned}\Pr(A, g|\pi, \theta) &= \Pr(A|g, \pi, \theta)\Pr(g|\pi, \theta)\\ &= \prod_{ij}\theta_{g_i,j}^{A_{ij}}\cdot\prod_i \pi_{g_i}\\ &= \prod_i\left[\pi_{g_i}\prod_j\theta_{g_i,j}^{A_{ij}}\right],\end{aligned} \tag{1}$$