Mixture Models for Community Detection

Bastien Le Chenadec bastien.le-chenadec@eleves.enpc.fr
Ecole des Ponts Paristech

Sofiane Ezzehi sofiane.ezzehi@eleves.enpc.fr
Ecole des Ponts Paristech

Theïlo Terrisse theilo.terrisse@eleves.enpc.fr
Ecole des Ponts Paristech

Quelques remarques:

- hésitez pas à utiliser Zotero pour exporter d'éventuelles références bibliographiques, histoire d'avoir un remplissage uniforme e de gagner du temps
- Une fois une abbréviation introduite (ex: SBM), si possible toujours utiliser l'abbréviation et ne pas ré-écrire l'expression entière
- Je propose de préférer le terme de "cluster" à "community". (Il faudra que je change ce que j'ai écrit du coup)

Contribution Statement

TODO

1. Introduction

Community detection is the problem of revealing communities within a graph, also referred to as "graph clustering" [3], where a community within a graph is often defined as a set of nodes that are more densely connected between them than with the rest of the network. Among existing methods to tackle this problem, model-based methods fit a mathematical model capable of explaining the observed connectivity, and from which descriptive properties of a graph maybe extracted or computed. The Stochastic Block Model (SBM) is a renowned example of graph modelling. In [2], J.-J. Daudin *et al.* explore this model from a frequentist point of view and propose a variational EM algorithm to fit this model. In this report, we quickly review some literature on community detection in Section 2. In Section **??**, we detail the method introduced in [2] as well as a variant from the literature and discuss the problem of overlapping communities. In Sections 4 and 5, our experiments are presented and discussed. Lastly, we conclude in Section 6.

2. Context and Related Work

2.1 What is Community Detection?

Community Detection has tremendous applications in the study of real networks such as social [1] or biological [4] networks. This problem is of importance to perform a meso-scopic study of graphs, where communities can be interpreted as meta-nodes that reveal new interactions associated to possibly interpretable functions within a complex system represented as a graph.

Although the definition of a "community" remains somewhat sloppy in the literature, a common definition describes it as a group of nodes with an internal density (defined as

the ratio of edges between nodes of the group that are actually present over the number of such possible edges) that is larger than its inter-community density (similarly defined, but between nodes of the group and nodes outside the group). This definition is actually specific to homophilious communities, which we will focus on, while heterophily would take an opposite definition.

2.2 Common community detection approaches

Community detection is an NP-hard problem, as testing all possible partitions of a graph has complexity $\mathcal{O}(2^n)$ with n the number of nodes. Instead, common approaches [...]

[Mention types of methods, but focus on model-based approaches]

2.3 A word on overlapping community detection approaches

3. Proposed method

In "A mixture model for random graphs" [2], the authors propose a bayesian model for graphs, and an Expectation-Maximization approach to fit this model to a given graph.

3.1 A mixture model for random graphs

We will follow the same notations as in [2]. We consider an undirected graph with n nodes and no self-loops. We denote X the adjacency matrix of this graph. As such $X_{ij} \in \{0, 1\}$ denotes the existence of an edge between nodes i and j, and $\forall i \in \{1, \dots, n\}, X_{ii} = 0$.

We consider a mixture model that spreads the vertices in Q classes, with an a priori repartition $\{\alpha_1, \dots, \alpha_Q\}$ between classes. We introduce the random variables $Z_{iq} \in \{0, 1\}$ for $i \in \{1, \dots, n\}$ and $q \in \{1, \dots, Q\}$, that represent the appartenance of node i to class q. We have the following prior distribution on Z:

$$\forall i \in \{1, \dots, n\}, \quad \sum_{q=1}^{Q} Z_{iq} = 1 \quad \text{and} \quad \forall q \in \{1, \dots, Q\}, \quad \mathbb{P}(Z_{iq} = 1) = \alpha_q \tag{1}$$

We introduce priors on the existence of edges between nodes of different classes. We denote π_{ql} the probability of an edge between a node of class q and a node of class l. Because the graph is undirected, we have $\pi_{ql} = \pi_{lq}$. Finally we suppose the following prior distribution on the existence of edges between nodes :

$$\forall q, l \in \{1, \dots, Q\}, \quad \forall i \neq j \in \{1, \dots, n\}, \quad \mathbb{P}(X_{ij} = 1 | Z_{iq} = 1, Z_{jl} = 1) = \pi_{ql} \tag{2}$$

Figure 3.1 shows the graphical model of this mixture model.

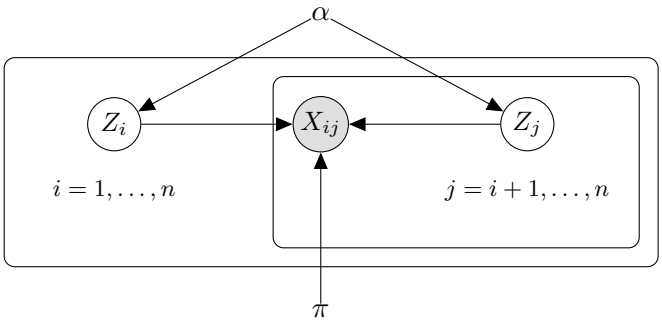

Figure 1: Graphical model of the mixture model

3.2 Variational Expectation-Maximization algorithm

The log-likelihood of the model is given by :

$$\log \mathcal{L}(X, Z) = \sum_i \sum_q Z_{iq} \log \alpha_q + \frac{1}{2} \sum_{i \neq j} \sum_{q,l} Z_{iq} Z_{jl} \times \pi_{ql}^{X_{ij}} (1 - \pi_{ql})^{1 - X_{ij}} \tag{3}$$

Because the likelihood $\mathcal{L}(X)$ is not tractable, the authors propose to use an EM algorithm to fit the model. However the E-step is not tractable either because of the posterior distribution of Z given X. Instead the authors propose to optimize (4) which is a lower bound of $\log \mathcal{L}(X)$ obtained using the Kullback-Leibler divergence between the posterior distribution of Z given X and an approximated distribution R_X.

$$\mathcal{J}(R_X) = \log \mathcal{L}(X) - \mathrm{KL}[R_X(\cdot), P(\cdot|X)] \tag{4}$$

By choosing the approximated distribution R_X to be a product of independant multinomial distributions (5), the authors obtain a fixed point relation between the parameters of the model and the parameters of the approximated distribution maximizing the lower bound $\mathcal{J}(R_X)$ (6). This fixed point relation is used in the E-step of the algorithm.

$$R_X(Z) = \prod_{i=1}^{n} h(Z_i, \tau_i) \qquad \forall i \in \{1, \ldots, n\}, h(Z_i, \tau_i) = \prod_{q=1}^{Q} \tau_{iq}^{Z_{iq}} \tag{5}$$

$$\forall i \in \{1, \ldots, n\}, \forall q \in \{1, \ldots, Q\}, \hat{\tau}_{iq} \propto \alpha_q \prod_{j \neq i} \prod_{l} \left[\pi_{ql}^{X_{ij}} (1 - \pi_{ql})^{1 - X_{ij}} \right]^{\hat{\tau}_{jl}} \tag{6}$$

This fixed point relation doesn't assure the theoretical convergence of the algorithm. We will see later that the convergence of the algorithm is not guaranteed in practice either. Finally we have the following updates for the M-step maximizing $\mathcal{J}(R_X)$:

$$\forall q, l \in \{1, \ldots, Q\}, \quad \hat{\alpha}_q = \frac{1}{n} \sum_i \hat{\tau}_{iq} \quad \hat{\pi}_{ql} = \frac{\sum_{i \neq j} \hat{\tau}_{iq} \hat{\tau}_{jl} X_{ij}}{\sum_{i \neq j} \hat{\tau}_{iq} \hat{\tau}_{jl}} \tag{7}$$

3.3 A variant

[Sofiane: présenter la variante]

Acknowledgments

Est-ce qu'on en met ?

Appendix A. plots

[...]

References

[1] P. Chunaev. Community detection in node-attributed social networks: A survey. *Computer Science Review*, 37:100286, Aug. 2020.

[2] J.-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statistics and Computing*, 18(2):173–183, June 2008.

[3] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, Feb. 2010.

[4] P. Sah, L. O. Singh, A. Clauset, and S. Bansal. Exploring community structure in biological networks with random graphs. *BMC Bioinformatics*, 15(1):220, June 2014.