Estimation in a Binomial Stochastic Blockmodel for a Weighted Graph by a Variational Expectation Maximization Algorithm

Abir El Haj, Yousri Slaoui, Pierre-Yves Louis, Zaher Khraibani

Estimation in a Binomial Stochastic Blockmodel for a Weighted Graph by a Variational Expectation Maximization Algorithm

Abir El Haj[1,2]*, Yousri Slaoui[1], Pierre-Yves Louis[1], Zaher Khraibani[2]
[1]Laboratoire de Mathématiques et Applications, Université de Poitiers, 11 Boulevard Marie et Pierre Curie, Bâtiment H3 - TSA 61125, 86073 POITIERS CEDEX 9
[2]Faculté de Sciences, Université Libanaise, Beyrouth, Liban

Abstract

Stochastic blockmodels have been widely proposed as a probabilistic random graph model for the analysis of networks data as well as for detecting community structure in these networks. In a number of real-world networks, not all ties among nodes have the same weight. Ties among networks nodes are often associated with weights that differentiate them in terms of their strength, intensity, or capacity. In this paper, we provide an inference method through a variational expectation maximization algorithm to estimate the parameters in binomial stochastic blockmodels for weighted networks. To prove the validity of the method and to highlight its main features, we set some applications of the proposed approach by using some simulated data and then some real data sets. Stochastic blockmodels belong to latent classes models. Classes defines a node's clustering. We compare the clustering found through binomial stochastic blockmodels with the ones found fitting a stochastic blockmodel with Poisson distributed edges. Inferred Poisson and binomial stochastic blockmodels mainly differs. Moreover, in our examples, the statistical error is lower for binomial stochastic blockmodels.

Keywords Binomial stochastic blockmodel; Clustering; Poisson stochastic blockmodel; Text mining; Variational inference; Weighted networks.

1 Introduction

Digital transformation challenges statistics. In many contexts, text mining is becoming a standard useful tool to find patterns of interest. This is a rising interest in particular in digital humanities and social sciences.

Beyond elementary descriptive statistics and models counting words, co-citations networks may be easily built. It means data are represented as a graph whose nodes are words and edges between two words are weighted according to the number of texts in the considered corpus citing simultaneously this pair of words. A general question of interest is to find clusters of nodes/words more closely related. Lots of community detection methods were developed in order to tackle this issue. Some probabilistic random graph models like Erdös-Renyi or the Stochastic Blockmodel (`SBM`) family can be used as statistical parametric models where the unknown cluster are latent classes. Beyond binary `SBM` (whose edges are present/absent), `SBM` with a more general distribution for the value of an edge between nodes belonging to the same class are of increasing interest and usefulness.

In this paper, we want to consider `SBM` with a binomial distribution on edges. This question is motivated by the study of co-citation networks in a text mining context where there is a maximal weight m possible for an edge corresponding to the number of documents included in the corpus. Beside developing and implementing the estimation procedure of a binomial SBM, this paper aims at comparing binomial `SBM` and `SBM` with a Poisson distributed weight. Due to the well known closeness between binomial and Poisson distributions in certain regimes of parameters, are the estimation procedures for these two models equivalent? For instance, would be a large corpus be better modeled through a Poisson `SBM` or through a binomial one? What is the number of clusters found? How is the statistical error in these cases? Following a known procedure through a Variational Expectation Maximization (`VEM`)

*e-mail adress: abir.el.haj@math.univ-poitiers.fr

algorithm (Blei et al.(2017), we develop and implement the method on simulated datasets (to validate the procedure) as well as benchmark real datasets: one in a co-citation text mining context ($m = 20$), the other one in a social networks context ($m = 14$).

The Stochastic Block Model (SBM) proposed by Anderson et al. (1992) and Holland et al. (1983) is a probabilistic random graph model which aims to produce classes, called blocks, or more generally clusters in networks. It have been used in several fields such as network and biological sciences (Fortunato 2010,P) and statistics and machine learning (Goldenberg et al. 2010). It's a generalization of the Erdõs-Reyni model proposed by Erdõs and Rényi (1959) using a latent structure on the nodes. In this model, the nodes of the network are divided into disjoint blocks such that the nodes belonging to the same block have the same inter connection probability and the same intra connection probability with nodes in others blocks. The probability of an edge between two nodes just depends on which blocks they are in, and is independent across edges.

Mariadassou et al. (2010) proposed some generalization of the SBM model to deal with random weighted graphs. Jernite et al. (2014) have treated the model with categorical edges, Airoldi et al. (2008) and Latouche et al. (2011) have focused on the SBM model with overlapping clusters. More recently, Yang et al. (2011), Xu and Hero (2013), Zreik et al. (2017) and Matias and Miel (2017) have extended the model to deal with dynamic networks where networks evolve over time.

Several authors focus on the estimation of the parameters in the SBM model. First, Snijders and Nowicki (1997) have proposed a maximum likelihood inference based on the Expectation Maximization (EM) algorithm for estimating the connection probabilities between nodes and for predicting the blocks in the model with only two blocks. Then, Nowicki and Snijders (2001) have generalized the previous work to deal with SBM model with an arbitrary block number using a Bayesian approach based on Gibbs sampling. Since the EM algorithm requires the computation of the distribution of the labels Z conditionally on the observations X which is usually intractable since the edges in the network are not independent, Daudin et al. (2008) and Jaakola (2000) have introduced approximate methods based on variational approach to estimate the parameters and to classify the clusters. They used the Variational Expectation Maximization (VEM) algorithm. This approach is known to be consistent under the SBM model according to Celisse et al. (2012). Furthermore, Latouche et al. (2012) used a variational Bayesian inference based on variational Bayes EM algorithm (VBEM), while Nowicki and Snijders used the Gibbs sampling algorithm.

In most of the methods already treated in this context, the SBM is restricted to binary networks, in which edges are unweighted. Since the most of networks are weighted, Thomas and Blitzstein (2011) have proposed to apply a threshold to a weighted relationship. This method is not effective since it produces binary graphs that only a fraction of the relevant information will be kept, and the others will be destroyed. However, Mariadassou et al. (2010), Karrer and Newman (2011), and Ball et al. (2011) have been interested in the case of weighted SBM without thresholding. They treated the case of SBM with Poisson distributed edge's weights.

We are interested here in the case of weighted networks, where each edge is associated with an integer value representing the capacity of ties among nodes. We provide a SBM model with binomial distributed edge weights. The binomial distribution takes the parameters m which means the maximum weight present on the edges and the parameters π_{qr} which means the matrix of probability connection between each two clusters q and r.

In this paper, we define the proposed method in section 2. We present a description of the binomial SBM model in subsection 2.1. Then, we calculate the likelihood of the complete data in subsection 2.2. We develop the proposed VEM algorithm of resolution in subsection 2.3. In subsection 2.4, we calculate the optimal number of clusters by using Integrated Classification Likelihood (ICL) criterion method. We define in section 3, the SBM model with Poisson distributed edges weights to compare it later in section 4 to our method. We set some applications of the proposed algorithm by introducing some simulated data and then two applications on real dataset in section 4 to prove its effectiveness and highlight its main features. We compare the results obtained by the proposed method to those obtained by applying the SBM model with Poisson distributed edges weights (PSBM). The final section is a conclusion of the paper.

2 Proposed method

The dataset here is incomplete since there are some latent variables that influence the distribution of the data and the formation of the clusters within the network. We compute first the likelihood of the complete data, then we calculate the likelihood of the incomplete data. Furthermore, we develop an inference method to estimate the parameters of the model.

2.1 Mixture model with latent classes

A general weighted undirected network is represented by $G := ([n], X)$, where $[n]$ is the set of nodes $\{1, ..., n\}$ for all $n \geq 1$ and X is the symmetric edge-weighted matrix of size n which encodes the observed interactions between nodes. We assume that the nodes are not connected to themselves so that for all $i \in \{1, \ldots, n\}$, we have $X_{ii} = 0$. The number of blocks in the graph is chosen equal to Q ($Q \geq 1$).

Let $Z = (Z_i)_{i\in\{1,\ldots,n\}}$ be the latent vector of $(\{0,1\}^Q)^n$ describing the belonging of the node i to cluster q when $Z_{iq} = 1$ and not when $Z_{iq} = 0$. Since a node i can belong to only one cluster then we have $\forall i, \sum_{q=1}^{Q} Z_{iq} = 1$. Moreover, the vectors Z_i for $i \in \{1, \ldots, n\}$ are independent and sampled from a multinomial distribution as following

$$Z_i \sim \mathcal{M}(1, \alpha = (\alpha_1, \ldots, \alpha_Q)),$$

where $\alpha = (\alpha_1, \ldots, \alpha_Q)$ is the vector of class proportions such as $\sum_{q=1}^{Q} \alpha_q = 1$.

The weighted matrix $X = (X_{i,j})_{i,j\in\{1,\ldots,n\}}$ associated to the network contains the weights of the edges between each two nodes in the network so that $X_{i,j} = l$ if there is an edge joining the nodes i and j and is weighted by the value l. The variables $\{X_{ij}, i, j \in [n], i < j\}$ are independent conditionally on the sigma-field generated by $\{Z_i, i \in [n]\}$, and are sampled from a binomial distribution

$$X_{ij}|Z_{iq}Z_{jl} = 1 \sim \mathcal{B}(m, \pi_{ql}),$$

where m is the maximum weight on an edge and π is the $Q \times Q$ matrix of connection probabilities between each two q-labeled and r-labeled nodes for all $q, r \in \{1, \ldots, Q\}$. Note that the parameter m is fixed according to the context.

In the sequel, we are interested in estimating the parameter $\theta = (\alpha, \pi)$ of the model in a weighted undirected network. However, we claim that all results obtained in this paper can be extended to directed networks, with or without self-loops.

2.2 Likelihood of the complete data (X, Z)

We develop here the likelihood of the complete data to estimate the parameters of the model. So, we define the joint distribution by

$$\mathbb{P}_\theta(X, Z) = \mathbb{P}_\pi(X|Z)\mathbb{P}_\alpha(Z),$$

where the laws satisfy

$$\mathbb{P}_\pi(X|Z) = \prod_{i<j}^{n}\prod_{q,l}^{Q} \mathbb{P}_{\pi_{ql}}(X_{i,j}|Z_i = q, Z_j = l) = \prod_{i<j}^{n}\prod_{q,l}^{Q} \left(\binom{m}{X_{ij}} \pi_{ql}^{X_{ij}} (1-\pi_{ql})^{m-X_{ij}} \right)^{Z_{iq}Z_{jl}}. \quad (2.1)$$

and

$$\mathbb{P}_\alpha(Z) = \prod_{i}^{n}\prod_{q}^{Q} \mathbb{P}_{\alpha_q}(Z_i) = \prod_{i}^{n}\prod_{q}^{Q} \alpha_q^{Z_{iq}}. \quad (2.2)$$

Based on the equations (2.1) and (2.2), the log-likelihood of the complete data can be expressed as follows

$$\begin{aligned} \log \mathbb{P}_\theta(X, Z) &= \log \mathbb{P}_\alpha(Z) + \log \mathbb{P}_\pi(X|Z) \\ &= \sum_{i}\sum_{q} Z_{iq} \log(\alpha_q) + \sum_{i<j}\sum_{q,l} Z_{iq}Z_{jl}(\log C_m^{X_{ij}} + X_{ij} \log \pi_{ql} + (m - X_{ij}) \log(1 - \pi_{ql})) \quad (2.3) \end{aligned}$$

2.3 Variational Inference

The log-likelihood of the incomplete data (in the sense, without knowing the labels Z) can be obtained through the marginalization $\log \mathbb{P}_\theta(X) = \log \sum_Z \mathbb{P}_\theta(X, Z)$ which involves a summation over every possible matrix Z and thus may not be tractable except for small values of n. To tackle this issue, we introduce in this section the Expectation Maximization (`EM`) algorithm developed by Dempster et al. (1977) and McLachlan and Krishnan (2007). This iterative method allow us to maximize $\log \mathbb{P}_\theta(X)$ without calculating it. However, the E-step is devoted to calculate the probability of the latent variables Z conditionally on the observed variables X which is intractable in this context since the edges X_{ij} for $i, j \in \{1, \ldots, n\}$ are not independent.

To tackle this issue, we introduce the Variational Expectation Maximization (`VEM`) algorithm developed by Jordan et al. (1999) and Jaakkola and Jordan (2000). This is an approximation maximization likelihood strategy based on variational approach such as in Daudin et al. (2008). This method overcomes the issue by maximizing a lower bound of the log-likelihood based on an approximation of the true conditional distribution of Z given X.

We rely on a variational decomposition of the incomplete log-likelihood as following

$$\log \mathbb{P}_\theta(X) = J_\theta(R_X(.)) + \mathrm{KL}(R_X(.) \parallel \mathbb{P}_\theta(.|X)), \tag{2.4}$$

where $\mathbb{P}_\theta(Z|X)$ is the true conditional distribution of Z given Y, $R_X(Z)$ is an approximate distribution of $\mathbb{P}_\theta(Z|X)$ and KL is the Kullback-Leibler divergence between $\mathbb{P}_\theta(Z|X)$ and $R_X(Z)$ defined by

$$\mathrm{KL}(R(.) \parallel \mathbb{P}_\theta(.|Z)) = -\sum_Z R_X(Z) \log \frac{\mathbb{P}_\theta(Z|X)}{R_X(Z)}.$$

It measures the closeness of the two distributions $\mathbb{P}_\theta(Z|X)$ and $R_X(Z)$. Furthermore, it is a non-negative measure:

$$\mathrm{KL}(R(.) \parallel \mathbb{P}_\theta(.|Z)) \geq 0. \tag{2.5}$$

We can underline that the equality is reached when $R_X(Z) = \mathbb{P}_\theta(Z|X)$.

The term $J(.)$ of the equation (2.4) is of the form

$$\begin{aligned} J_\theta(R_X(.)) &= \sum_Z R_X(Z) \log \frac{\mathbb{P}_\theta(X,Z)}{R_X(Z)} \\ &= \mathbb{E}_{R_X}[\log(\mathbb{P}_\theta(X,Z))] - \mathbb{E}_{R_X}[\log R_X(Z)], \end{aligned} \tag{2.6}$$

where $\mathbb{E}_{R_X}$ denotes the expectation with respect to R_X.

The combination of (2.4) and (2.5) ensure that

$$\log \mathbb{P}_\theta(X) \geq J_\theta(R_X).$$

Therefore, $J_\theta(R_X)$ is a lower bound of $\log \mathbb{P}_\theta(X)$.

Moreover, since $\mathbb{P}_\theta(Z|X)$ is not tractable, the classical property of KL which states that the lower bound $J_\theta(R_X)$ has a unique maximum $\mathbb{P}_\theta(X)$ reached for $R_X(Z) = \mathbb{P}(Z|X)$ is not helpful. So, we maximize $J_\theta(R_X)$ with respect to R_X and θ. By using the equations (2.6) and the log-likelihood of the complete data equation (2.3), the lower bound $J_\theta(R_X)$ can be written as follows

$$\begin{aligned} J_\theta(R_X) &= H(R_X) + \sum_i \sum_q \mathbb{E}_{R_X}(Z_{iq}) \log \alpha_q + \sum_{i<j} \sum_{q,l} \mathbb{E}_{R_X}(Z_{iq}, Z_{jl})(\log C_m^{X_{ij}} + X_{ij} \log \pi_{ql} \\ &\quad + (m - X_{ij}) \log(1 - \pi_{ql})), \end{aligned} \tag{2.7}$$

where $H(R_X) = -\sum_i \sum_q \mathbb{E}_{R_X}(Z_{iq}) \log \mathbb{E}_{R_X}(Z_{iq})$.

The E-step of the `EM` algorithm becomes tractable when we assume that the distribution $R_X(Z)$ can be factorized over the latent variable Z as follows

$$R_X(Z) = \prod_{i=1}^n R_{X,i}(Z_i) = \prod_{i=1}^n h(Z_i; \tau_i), \tag{2.8}$$

where $\{\tau_i \in [0,1]^Q, i = 1, \ldots, n\}$ are the variational parameters associated with $\{Z_i, i = 1, \ldots, n\}$ such as $\sum_q \tau_{iq} = 1, \ \forall i \in \{1, \ldots, n\}$ and h is the multinomial distribution with parameters τ_i. We have

$$\tau_{iq} = \mathbb{P}(R_X(Z_{iq} = 1)) = \mathbb{E}(R_X(Z_{iq})) = \mathbb{E}_{R_X}(Z_{iq}) \tag{2.9}$$

and

$$\tau_{iq}\tau_{jl} = \mathbb{P}(R_X(Z_{iq} = 1, Z_{jl} = 1)) = \mathbb{E}(R_X(Z_{iq}, Z_{jl})) = \mathbb{E}_{R_X}(Z_{iq}, Z_{jl}). \tag{2.10}$$

By using (2.8), (2.9), (3.2) and by developing the equation (2.7), we obtain that $J_\theta(R_X)$ can be written as follows

$$\begin{aligned} J_\theta(R_X) &= -\sum_i \sum_q \tau_{iq} \log \tau_{iq} + \sum_i \sum_q \tau_{iq} \log \alpha_q + \sum_{i<j} \sum_{q,l} \tau_{iq}\tau_{jl}(\log C_m^{X_{ij}} + X_{ij} \log \pi_{ql} \\ &\quad + (m - X_{ij}) \log(1 - \pi_{ql})). \end{aligned}$$

During the variational E-step, the parameters of the model are fixed. By maximizing the lower bound $J_\theta(R_X)$ with respect to τ, we obtain the estimate of τ by the following fixed point relation

$$\hat{\tau}_{iq} \propto \alpha_q \prod_j \prod_l \left(C_m^{X_{ij}} \pi_{ql}^{X_{ij}} (1-\pi_{ql})^{m-X_{ij}} \right)^{\hat{\tau}_{jl}}. \tag{2.11}$$

The estimation of τ is obtained from (2.11) by iterating a fixed point algorithm until convergence.

Conversely, during the M-step, the parameter τ is fixed. By maximizing the lower bound $J_\theta(R_X)$ with respect to α and under the condition $\sum_q \alpha_q = 1$, we obtain the estimate of α_q

$$\hat{\alpha}_q = \frac{1}{n} \sum_i \tau_{iq}.$$

By maximizing the lower bound $J_\theta(R_X)$ with respect to π, we obtain the estimate of π_{ql}

$$\hat{\pi}_{ql} = \frac{\sum_{i<j} \tau_{iq}\tau_{jl} X_{ij}}{m \sum_{i<j} \tau_{iq}\tau_{jl}}.$$

2.3.1 Algorithm of resolution

We present here the algorithm of resolution used to estimate the parameters of the model. We denote by t the current index for iterations in the algorithm and by ε a fixed threshold of convergence.

Algorithm 1 Variational Expectation Maximization algorithm for inference in SBM

Initialization: Initialize τ^0 with the `k-means` algorithm.

1: Update the parameters τ and θ iteratively

$$\theta^{(t+1)} = \arg\max_\theta J_\theta(R_X; \tau^{(t)}) \qquad \text{M-step}$$

$$\tau^{(t+1)} = \arg\max_\tau J_{\theta^{(t+1)}}(R_X; \tau) \qquad \text{VE-step}$$

2: Repeat Step 1 until $\|\theta^{(t+1)} - \theta^{(t)}\| < \varepsilon$.

2.4 Integrated Classification Likelihood (ICL)

In the sections above, we estimated the parameters of the model by fixing the number of blocks Q since the `SBM` model function requires number of latent groups Q as an input argument. We are interested here in choosing the number of clusters $\hat{Q}$ that will optimally fit the data. One of the proposed method consists in iterating the `SBM` model with different values of Q and then choosing the optimal number of clusters by evaluating goodness of fit for each group sizes. This method is expensive in terms of time and computing since we evaluate the goodness of fit for all the groups.

Another approach consists in using the Bayesian Information Criterion (BIC). The optimal number of clusters is obtained by running the model for different values of Q and then by choosing the one which provides the higher value of `BIC`. We have

$$BIC(G) = \log \mathbb{P}_\theta(X) - \frac{V_Q}{2} \log n,$$

where V_Q is the number of parameters of the model for the Q groups. However, This method involves the computation of the log-likelihood of the given data X which is intractable.

Thus, Daudin et al. (2008) proposed the Integrated Classification Likelihood (ICL) criterion to estimate Q in a `SBM` model. This method is an approximation of the complete data likelihood. The ICL is of the form

$$\begin{aligned} ICL(Q) &= \sum_i \sum_q \hat{\tau}_{iq} \log \hat{\alpha}_q + \sum_{i<j} \sum_{q,l} \hat{\tau}_{iq}\hat{\tau}_{jl} (\log C_m^{X_{ij}} + X_{ij} \log \hat{\pi}_{ql} + (m - X_{ij}) \log(1-\hat{\pi}_{ql})) \\ &\quad - \frac{1}{2} \left(\frac{Q(Q+1)}{2} \log \frac{n(n-1)}{2} - (Q-1) \log n \right). \end{aligned}$$

The `VEM` algorithm is run for different values of Q and $\hat{Q}$ is chosen such that ICL is maximized.

3 SBM with Poisson distributed edges weights

In this section, we define the `SBM` with Poisson distributed edges weights method in order to compare it later to our method. We assume that the vectors Z_i for $i \in \{1, \ldots, n\}$ are independent and are drawn from a multinomial distribution such as:

$$Z_i \sim \mathcal{M}(1, \alpha = (\alpha_1, \ldots, \alpha_Q))$$

and that the edges $\{X_{ij}, i, j \in [n], i < j\}$ are independent conditionally on the sigma-field generated by $\{Z_i, i \in [n]\}$ and are drawn from a Poisson distribution such as:

$$X_{ij}|Z_{iq}.Z_{jl} = 1 \sim \mathcal{P}(\lambda_{ql}).$$

The parameter λ_{ql} is $Q \times Q$ matrix of mean connection probabilities between the latent groups.

According to Mariadassou et al. (2010) and following the same steps as before, the lower bound of the log-likelihood can be expressed as follows:

$$\begin{aligned} J_\theta(R_X) &= \sum_i \sum_q \mathbb{E}_{R_X}(Z_{iq}) \log \alpha_q + \sum_{i<j} \sum_{q,l} \mathbb{E}_{R_X}(Z_{iq} Z_{jl})(-\lambda_{ql} + X_{ij} \log(\lambda_{ql}) - X_{ij}!) \\ &\quad - \sum_i \sum_q \mathbb{E}_{R_X}(Z_{iq}) \log \mathbb{E}_{R_X}(Z_{iq}). \end{aligned} \tag{3.1}$$

The estimation of the model parameters α and λ can be calculated directly by fixing the parameter τ and then by maximizing (3.1) with respect to α and λ respectively. They can be expressed as follows

$$\hat{\alpha}_q = \frac{1}{n} \sum_i \tau_{iq} \quad \text{and} \quad \hat{\lambda}_{ql} = \frac{\sum_{i<j} \tau_{iq} \tau_{jl} X_{ij}}{\sum_{i<j} \tau_{iq} \tau_{jl}}.$$

By maximizing (3.1) with respect to τ and by fixing the parameters λ and α, we can obtain the estimated parameter $\hat{\tau}$ by the following fixed point relation

$$\hat{\tau}_{iq} \propto \alpha_q \prod_j \prod_l \left(\frac{e^{-\lambda_{ql}} \lambda_{ql}^{X_{ij}}}{X_{ij}!} \right)^{\hat{\tau}_{jl}}. \tag{3.2}$$

The estimation of τ is obtained from (3.2) by iterating a fixed point algorithm until convergence.

4 Numerical experiments

This section aims at highlighting the main features of the proposed inference algorithm and at proving its validity by introducing two simulated data and then by applying our algorithm on a real dataset. Furthermore, numerical comparisons with the Poisson `SBM` is performed to prove the effectiveness of the proposed approach.

4.1 Simulated data

First, we perform the stochastic blockmodel using simulated data with a binomial output distribution. The graph has $n = 20$ vertices. We choose for this simulation a fixed number of clusters Q equal to three. We use in the simulation the following parameters:

$$\bar{\alpha} = (0.2, 0.5, 0.3) \quad \text{and} \quad \bar{\pi} = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.2 & 0.5 & 0.3 \\ 0.1 & 0.3 & 0.6 \end{bmatrix}.$$

We visualize the network in Figure 1 using `Gephi` software with the layout algorithm Force Atlas. Furthermore, in all the applications below, the width of the lines used to represent the edges is proportional to its weights.

We can show in Figure 1 the structure of the simulated data graph. There are three apparent communities. By applying our algorithm implemented in `R` programming language, we obtained that the vertices of the network are grouped into three clusters as shown in Table 1.

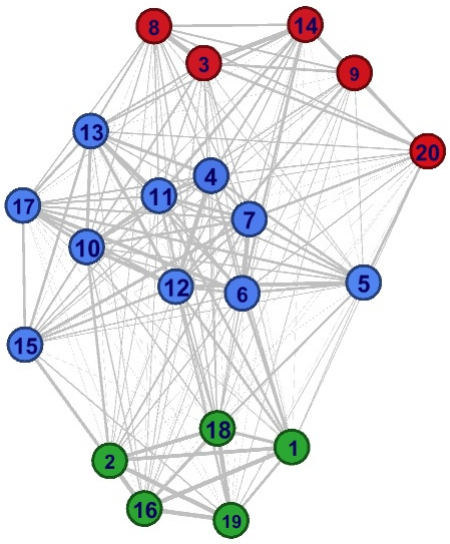

Figure 1: First simulated data graph visualization with `Gephi`.

Clusters	Vertices
1	3 8 9 14 20
2	4 5 6 7 10 11 12 13 15 17
3	1 2 16 18 19

Table 1: Grouping first simulated graph vertices into clusters

Table 1 shows clearly that the nodes of the first simulated graph are split into three clusters which are the same as the three clusters shown in the Figure 1 which confirm the effectiveness of our method. The time of convergence of the algorithm is 0.22second (CPU Corel3 - 4GB RAM) which is so satisfying.

We sample now $S = 8$ random graphs according to the same mixture model. Then we calculated in Table 2 and Table 3, for each parameter, the estimated Root Mean Squares Error (`RMSE`) defined by:

$$RMSE(\bar{\alpha}_q) = \sqrt{\frac{1}{S}\sum_{s=1}^{S}(\hat{\alpha}_q^{(s)} - \bar{\alpha}_q)^2} \quad \text{and} \quad RMSE(\bar{\pi}_{qr}) = \sqrt{\frac{1}{S}\sum_{s=1}^{S}(\hat{\pi}_{qr}^{(s)} - \bar{\pi}_{qr})^2},$$

where the superscript s labels the estimates obtained in simulation s.

	`RMSE`$(\bar{\alpha}_1)$	`RMSE`$(\bar{\alpha}_2)$	`RMSE`$(\bar{\alpha}_3)$
`SBM` with binomial distribution on edges	0.05	0.14	0.12
`SBM` with Poisson distribution on edges	0.33	0.24	0.15

Table 2: Root Mean Squares Error of the parameter $\bar{\alpha}_q$ for the first simulated data using the binomial `SBM` model and the Poisson `SBM` model.

`RMSE`	$\bar{\pi}_{1.}$	$\bar{\pi}_{2.}$	$\bar{\pi}_{3.}$
$\bar{\pi}_{.1}$	0.01	0.06	0.14
$\bar{\pi}_{.2}$	0.06	0.12	0.07
$\bar{\pi}_{.3}$	0.14	0.07	0.13

Table 3: Root Mean Squares Error of the parameter $\bar{\pi}_{qr}$ for the first simulated data using the binomial `SBM` model.

`RMSE`	$\bar{\pi}_{1.}$	$\bar{\pi}_{2.}$	$\bar{\pi}_{3.}$
$\bar{\pi}_{.1}$	0.3	0.8	0.88
$\bar{\pi}_{.2}$	0.8	0.5	0.68
$\bar{\pi}_{.3}$	0.88	0.68	0.4

Table 4: Root Mean Squares Error of the parameter $\bar{\pi}_{qr}$ for the first simulated data using the Poisson `SBM` model.

Applying the `SBM` model with Poisson distributed edges weights to this simulated data, we find the same three clusters of the model. We give in Table 2, 3 and 4, the `RMSE` of the parameters $\bar{\alpha}$ and $\bar{\pi}$ obtained by our algorithm and those find by the Poisson `SBM` model. Results show that the `RMSE` of

the parameters of the proposed model are closer to 0 and then the results is more satisfying than those obtained by the Poisson SBM model.

Since we get better results using our method, we apply on the next simulated data example only the proposed algorithm (related to the binomial SBM).

We introduce now a graph with a larger number of vertices to confirm that the proposed algorithm is valid for larger weighted networks. The observed graph here has $n = 70$ vertices and a fixed number of clusters Q equal to 5. The parameters used are:

$$\bar{\alpha} = (0.2, 0.1, 0.3, 0.35, 0.05) \quad \text{and} \quad \bar{\pi} = \begin{bmatrix} 0.5 & 0.1 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.4 & 0.2 & 0.1 & 0.2 \\ 0.1 & 0.2 & 0.6 & 0.05 & 0.05 \\ 0.1 & 0.1 & 0.05 & 0.4 & 0.35 \\ 0.2 & 0.2 & 0.05 & 0.35 & 0.2 \end{bmatrix}.$$

We visualize in Figure 2 the network using Gephi software with the layout algorithm Force Atlas.

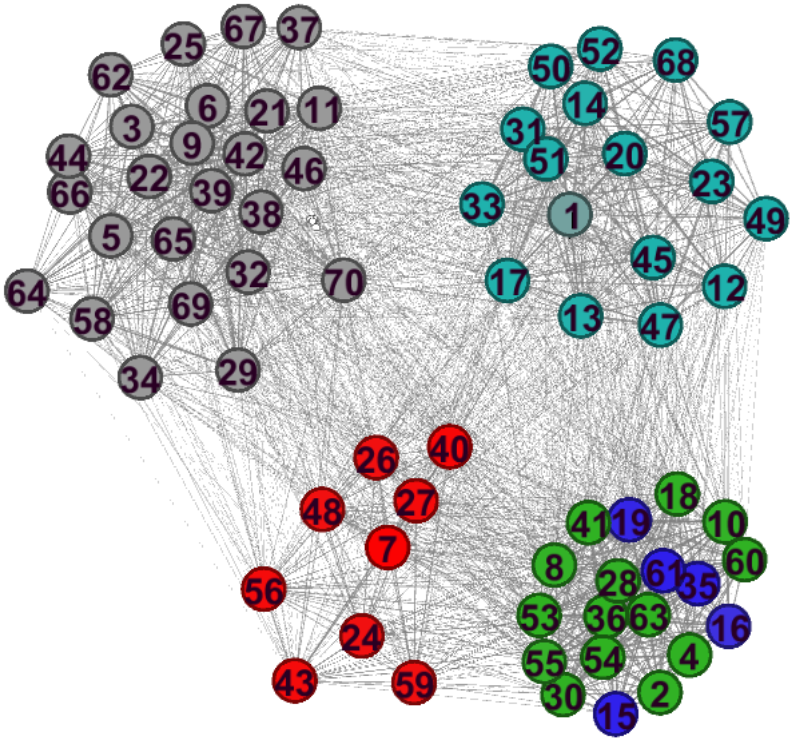

Figure 2: Second simulated graph visualization with Gephi.

The structure of network graph in Figure 2 shows clearly five apparent communities. By applying our algorithm implemented in the software R, we obtain that the optimal number of clusters is five and that the nodes are grouped into these five clusters as shown in Table 5.

Clusters	Vertices	Cardinality
1	1 12 13 14 17 20 23 31 33 45 47 49 50 51 52 57 68	17
2	3 5 6 9 11 21 22 25 29 32 34 37 38 39 42 44 46 58 62 64 65 66 67 69 70	25
3	2 4 8 10 18 28 30 36 41 53 54 55 60 63	14
4	15 16 19 35 61	5
5	7 24 26 27 40 43 48 56 59	9

Table 5: Grouping second simulated graph vertices into clusters.

The nodes of the graph are clearly grouped into the same five clusters shown in Figure 2.
We calculate in the Table 6 and Table 7 the RMSE of the parameters $\bar{\alpha}$ and $\bar{\pi}$.

RMSE($\bar{\alpha}_1$)	RMSE($\bar{\alpha}_2$)	RMSE($\bar{\alpha}_3$)	RMSE($\bar{\alpha}_4$)	RMSE($\bar{\alpha}_5$)
0.07	0.03	0.18	0.23	0.09

Table 6: Root Mean Squares Error of the parameter $\bar{\alpha}_q$ for the second simulated SBM with binomial output.

RMSE	$\bar{\pi}_{1.}$	$\bar{\pi}_{2.}$	$\bar{\pi}_{3.}$	$\bar{\pi}_{4.}$	$\bar{\pi}_{5.}$
$\bar{\pi}_{.1}$	0.1	0.01	0.01	0.003	0.39
$\bar{\pi}_{.2}$	0.01	0.1	0.1	0.41	0.09
$\bar{\pi}_{.3}$	0.01	0.1	0.31	0.05	0.04
$\bar{\pi}_{.4}$	0.003	0.41	0.05	0.07	0.26
$\bar{\pi}_{.5}$	0.39	0.09	0.04	0.26	0.38

Table 7: Root Mean Squares Error of the parameter $\bar{\pi}_{qr}$ for the second simulated SBM with binomial output on edges.

All the obtained values are close to zero which demonstrates that the estimated parameters are closer to the true parameters. The time of convergence of the algorithm is 0.47second (CPU Corel3 - 4GB RAM) which is so satisfying.

4.2 Text mining through terms co-occurence network (Reuters-21578 data set)

The Reuters-21578 data set contains a collection of documents that appeared on Reuters newswire in 1987. The data set is available online at http://kdd.ics.uci.edu/databases/reuters21578/reuters 21578.html. For more explanation about this data, we refer the reader to (Lewis 1997). We are interested in this example in 20 exemplary news articles from the Reuters-21578 data set of topic crude. The data is available in the package tm (Feinerer et al. 2008) of the software R under the name of crude data where all documents belong to the topic crude dealing with crude oil. We build a term-by-document matrix of the corpus crude by doing a text mining treatment. We interpret a term as important according to a simple counting of frequencies, we chose the frequent terms that co-occur at least six times in the documents. Then, we compute the correlations between them in the term-by-document matrix and we chose those out higher than 0.5. The figure visualizing the correlation between these terms is available in (Feinerer et al. 2008).

We transform the term-by-document matrix into a one mode matrix which is the term-by-term matrix. The network associated to this matrix is an undirected network of 21 vertices and 97 edges, where each vertex is a term and there is an edge between a pair of terms if they co-occur together at least one time in the documents. The edge weights are represented in the obtained matrix where each cell indicates the number of documents where both the row and the column terms co-occur.

The graph associated with this network is visualized in Figure 3 using Gephi software with the layout algorithm Force Atlas.

We present some global characteristics of the structure of the associated graph with "non weighted" edges: the assortativity coefficient is equal to 0.23 which is a positive value, that means that the terms presented in the documents of the reuters-21578 corpus tends to occur with other terms that have equally high or equally low number of occurrence. The average clustering coefficient (transitivity) is equal to 0.84 which shows the completeness of the neighborhood of the vertices in the network. The density of the graph is equal to 0.51 which indicates that the graph of the network is dense. Note that the transitivity and the density value are close which means that the graph is not highly clustered.

We apply our algorithm. We obtain that the terms are grouped into four clusters as presented in Table 8. Table 8 shows the distribution of the network's terms into groups which means that the terms of each group are frequently co-occuring together in the documents.

We apply now the SBM with a Poisson distributed edge's weight. We obtain that the terms are grouped into two clusters as presented in the follwong table 9

We define the total variation distance between two probability distributions μ and v on the sample space Ω by $d_{TV}(\mu, v) = \frac{1}{2}\sum_{x\in\Omega}|\mu(x) - v(x)|$. The mean of the total variation distance between the binomial and the Poisson distribution is equal to $md_{TV} = 6.5$ which means that the two approaches are not close and then the two fitted model are so different.

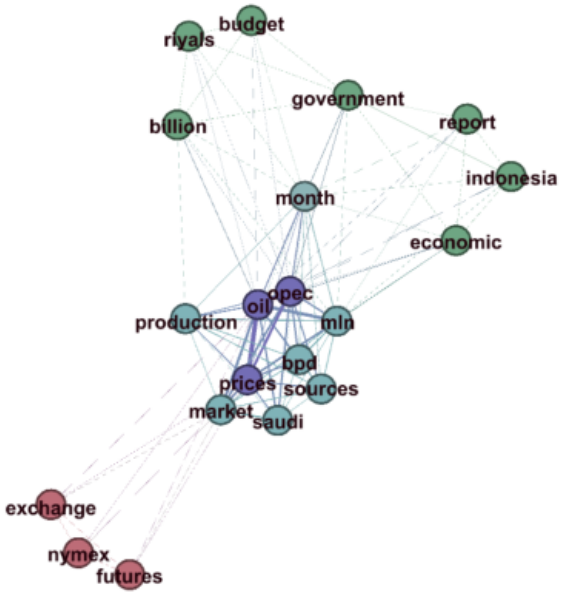

Figure 3: Network of terms of the the Reuters-21578 corpus visualization with `Gephi`.

Clusters	Vertices
1	oil opec prices
2	mln bpd month sources production saudi market
3	billion budget riyals government economics indonesia report
4	exchange nymex futures Kuwait

Table 8: Grouping the terms of the network of terms of the Reuters-21578 corpus into clusters using binomial `SBM`.

Clusters	Vertices
1	oil opec prices mln bpd sources production saudi market Kuwait
2	billion budget exchange futures riyals government economics indonesia month nymex report

Table 9: Grouping the terms of the network of terms of the Reuters-21578 corpus into clusters using Poisson `SBM`.

4.3 Social network: a benchmark dataset (Deep South network)

The data was collected by Davies et al. (1941) in the Southern United State 1930s in order to report a comparative study of social in black and in white society. They are interested in the percentage of the contacts between individuals which have approximately the same class levels so they collect the deep South data which represents the participation of 18 white women in a series of 14 informal social events over a nine-month period. The data is available in the package `manet` in software `R` under the name `deepsouth` http://cran.r-projet.org/web/packages/manet/manet.pdf. For more explanation about this data, we refer the reader to (Linton 2003). This data is considered as a benchmark in comparing social network analysis method. The authors focus on the analysis of two-mode data which means the women-by-event matrix data.

We transform the data into a single mode matrix which is the women-by-women matrix by multiplying the data matrix by its transpose. the network associated to this matrix is an undirected network of 18 vertices and 139 edges, where each vertex represents a Southern women among the 18 and there is an edge between a pair of women if they participate together in one of the 14 events a least. The edge weights are represented in the obtained matrix where each cell indicates the number of events co-attended by both the row and the column women.

The graph associated with the obtained network is visualized in Figure 4 using `Gephi` software with the layout algorithm Force Atlas. We present some global characteristics of the structure of the associated graph with "non weighted" edges: the assortativity coefficient is equal to 0.11 which is a positive value,

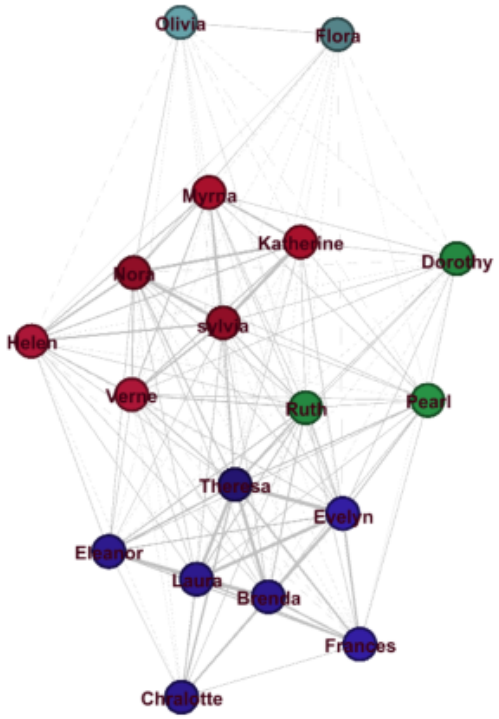

Figure 4: Deep South network visualization with `Gephi`.

that means that the Southern's women tends to participate to social events with other women that have equally high or equally low number of participation in the events. The average clustering coefficient (transitivity) is equal to 0.93 which shows the completeness of the neighborhood of the vertices in the network. The density of the graph is equal to 0.9 which indicates that the graph of the network is dense. Note that the transitivity and the density value are close which means that the graph is not highly clustered.

We apply our algorithm on the network to cluster the women into groups based on their occurrence in the events. The results are shown in Table (10). In table 10, each cluster represent the women which

Clusters	Vertices
1	Olivia Flora
2	Evelyn Laura Theresa Brenda Charlotte Frances Eleanor
3	Verne Myrna katherine Sylvia Nora Helen
4	Pearl Ruth Dorothy

Table 10: Grouping the women of deep South network into clusters.

are frequently met together in the informal social events.

We compare in the following the results obtained by the proposed method to several already existing methods: `BGR74` proposed by Breiger (1974) and is based on algebraic approaches, `FRE92`, `FRE193` and `FR293` proposed by Freeman (1993) and Freeman (1994) and is based on various algorithms to search for an optimal partition and `OSB00` proposed by Osbourn and Martinez (1995) and is based on the algorithm `VERI`. then, we compare it with the Poisson `SBM`.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
BGR74	W	W	W	W	W	W	W		W	W	W	W	W	WW	WW		W	W
FRE92	W	W	W	W	W	W	W		W	W	W	W	W	W	W	W		
FR193	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
FR293	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
OSB00	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
BSBM	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
PSBM	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W

Table 11: Clustering the women of the deep South network by different methods.

Table 11 shows the clusters obtained by different methods. At each line, The symbol "W" corresponds to women and all the W of the same color correspond to the women in the same cluster. The `BSBM` line corresponds to our method while the `PSBM` line corresponds to the Poisson `SBM`.

The mean of the total variation distance between the binomial and the Poisson distribution is equal to $md_{TV} = 3.4$ which means that the two model are different.

5 Conclusion

In this paper, we developed an inference method based on a variational expectation maximization (`VEM`) algorithm to estimate the parameters in a binomial stochastic blockmodel for weighted graphs. Since the log-likelihood of the incomplete data $\log \mathbb{P}_\theta(X) = \log \sum_Z \mathbb{P}_\theta(X, Z)$ is intractable except for network with small number of nodes n, we used an expectation maximization (`EM`) algorithm to tackle this issue. Since the edges of the network are not assumed to be independent, the computation of $\mathbb{P}(Z|X)$ is not possible and then the E-step of the `EM` algorithm which requires the calculation of $\mathbb{P}(Z|X)$ is intractable. We performed a variational expectation maximization (`VEM`) method to overcome this issue. This method is based on two steps. The first step consists of estimating the parameters α and π of the model by fixing the parameter τ and then by maximizing the lower bound J of the log-likelihood while the second step consists of estimating the parameter τ by fixing the model parameters and then by maximizing the lower bound J of the log-likelihood. We showed the effectiveness of our proposed method by using first, two simulated data and then on two real dataset. We compared our algorithm to the Poisson `SBM` model. Results show that our proposed method gives better results than the other method. We point out that the computational time of convergence of our proposed algorithm is so satisfying. Furthermore, this method is very easy to implement using the software R.

Moreover we found completely different fitted models by using the binomial and the Poisson distribution separately. The number of clusters can be different and clusters may differ. It is not the parameters regime where a binomial distribution could be approximated by a Poisson, or vice-versa.

We plan to make an extension of our work by proposing an algorithm using variational Bayesian inference for the binomial SBM model with weighted edges.

References

Airoldi, E., Blei, D., Fienberg, S., and Xing, E. (2008). Mixed membership stochastic blockmodels. The Journal of Machine Learning Research, **9**, 1981–2014.

Anderson, C. J., Wasserman, S., and Faust, K. (1992). Building stochastic blockmodels. Social Networks, **14**, 137–161.

Ball, B., Karrer, B., and Newman, M. E. (2011). Efficient and principled method for detecting communities in networks. Physical Review E, **84**, 036–103.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. Journal of the American Statistical Association, **112**, 859–877.

Breiger, R. L. (1974). The duality of persons and groups. Social Forces, **53**, 181–190.

Celisse, A., Daudin, J. J., and Pierre, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. Electronic Journal of Statistics, **6**, 1847–1899.

Daudin, J., Picard, F., and Robin, S. (2008). A mixture model for random graph. Statistics and Computing, **18**, 1–36.

Davies, A, Gardner, B. B., and Gardner, M. R. (1941). Deep South, Chicago: The University of Chicago Press.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B, **39**, 1–38.

Erdõs, P., and Rényi, A. (1959). On random graphs. Publicationes Mathematicae, **6**, 290–297.

Feinerer, I., Hornik, K. and Meyer, D. (2008). Text mining infrastructure in R. Journal of Statistical Software, **25**, 1–54.

Fortunato, S. (2010). Community detection in graphs. Physics Reports, **486**, 75–174.

Freeman, L. C. (1993). On the sociological concept of "group": a empirical test of two models. American Journal of Sociology, **98**, 152–166.

Freeman, L. C. (1994). Finding groups with a simple genetic algorithm. Journal of Mathematical Sociology, **17**, 227–241.

Goldenberg, A., Zheng, A. X., Fienberg, S. E., and Airoldi, E. M. (2010). A survey of statistical network models. Foundations and Trends in Machine Learning, **2**, 129–233.

Holland, P. W., Laskey, K. B., and Leinhardt S. (1983). Stochastic blockmodels: First steps. Social Networks, **5**, 109–137.

Jaakkola, T. S., and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. Statistics and Computing, **10**, 25–37.

Jaakkola. T. S. (2000). Tutorial on variational approximation methods. Advanced mean field methods: theory and practice, 129–159.

Jernite, Y., Latouche, P., Bouveyron, C., Rivera, P., Jegou, L., and Lamassé, S. (2014). The random subgraph model for the analysis of an acclesiastical network in merovingian gaul. Annals of Applied Statistics, **8**, 55–74.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. Machine learning, **37**, 183–233.

Karrer, B. and Newman, M.E.J. (2011). Stochastic blockmodels and community structure in networks. Physical Review E, **83**, 016–107.

Latouche, P., Birmelé, E., and Ambroise, C. (2011). Overlapping stochastic block models with application to the French political blogosphere. Annals of Applied Statistics. **5**, 309–336.

Latouche, P., Birmelé, E., and Ambroise, C. (2012). Variational bayesian inference and complexity control for stochastic block models. Statistical Modelling, **12**, 93–115.

Lewis, D. (1997). "Reuters-21578 Text Categorization Collection Distribution 1.0." http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html.

Linton C. (2003). Finding Social Groups: A Meta-Analysis of the Southern Women Data, In Ronald Breiger, Kathleen Carley and Philippa Pattison, eds. Dynamic Social Network Modeling and Analysis. Washington: The National Academies Press.

Mariadassou, M., Robin, S., and Vacher, C. (2010). Uncovering latent structure in valued graphs: a variational approach. Annals of Applied Statistics, **4**, 715–742.

Matias, C., and Miele, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. Journal of the Royal Statistical Society: Series B (Statistical Methodology), **79**, 1119–1141.

McLachlan, G., and Krishnan, T. (2007). The EM algorithm and extensions. John Wiley and Sons, **382**.

Nowicki, K., and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. Journal of the American Statistical Association, **96**, 1077–1087.

Osbourn, G. C., and Martinez, R. F. (1995). Empirically defined regions of influence for clustering analyses. Pattern Recognition, **28**, 1793–1806.

Porter, M. A., Onnela, J. P., and Mucha, P. J. (2009). Communities in networks. Notices of the American Mathematical Society, **56**, 1082–1097.

Snijders, T. A., and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. Journal of classification, **14**, 75–100.

Thomas, A.C. and Blitzstein, J.K. (2011). Valued ties tell fewer lies: Why not to dichotomize network edges with thresholds. arXiv:1101–0788.

Xu, K., and Hero III, A. (2013). Dynamic stochastic blockmodels: Statistical models for time-evolving networks. In Social Computing, Behavioral-Cultural Modeling and Prediction. Springer, 201–210.

Yang, T., Chi, Y., Zhu, S., Gong, Y., and Jin, R. (2011). Detecting communities and their evolutions in dynamic social networks: a bayesian approach. Machine learning, **82**, 157–189.

Zreik, R., Latouche, P., and Bouveyron, C. (2017). The dynamic random subgraph model for the clustering of evolving networks. Computational Statistics, **32**, 501–533.