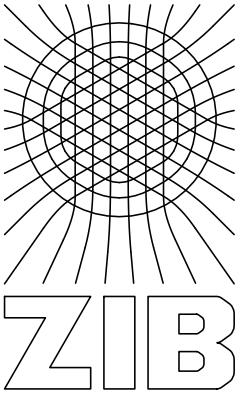
Zuse Institute Berlin



Takustraße 7
D-14195 Berlin-Dahlem
Germany

CHRISTIAN TOBIAS WILLENBOCKEL[1] AND
CHRISTOF SCHÜTTE[1,2]

[1] *Department of Mathematics and Computer Science, Freie Universität Berlin, Germany*

[2] *Zuse Institute Berlin, Germany*

Variational Bayesian Inference and Model Selection for the Stochastic Block Model with Irrelevant Vertices

ZIB Report 16-01 (January 2016)

Zuse Institute Berlin
Takustraße 7
D-14195 Berlin-Dahlem

Telefon: 030-84185-0
Telefax: 030-84185-125

e-mail: bibliothek@zib.de
URL: http://www.zib.de

ZIB-Report (Print) ISSN 1438-0064
ZIB-Report (Internet) ISSN 2192-7782

Variational Bayesian Inference and Model Selection for the Stochastic Block Model with Irrelevant Vertices

Christian Tobias Willenbockel * †[a] and Christof Schütte ‡[a,b]

[a] *Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 6, D-14195 Berlin, Germany*
[b] *Konrad–Zuse–Zentrum für Informationstechnik Berlin (ZIB), Takustraße 7, D-14195 Berlin, Germany*

Abstract

Real World networks often exhibit a significant number of vertices which are sparsely and irregularly connected to other vertices in the network. For clustering theses networks with a model based algorithm, we propose the Stochastic Block Model with Irrelevant Vertices (SBMIV) for weighted networks. We propose an original Variational Bayesian Expectation Maximization inference algorithm for the SBMIV which is an advanced version of our Blockloading algorithm for the Stochastic Block Model. We introduce a model selection criterion for the number of clusters of the SBMIV which is based on the lower variational bound of the model likelihood. We propose a fully Bayesian inference process, based on plausible informative priors, which is independent of other algorithms for preprocessing start values for the cluster assignment of vertices. Our inference methods allow for a multi level identification of irrelevant vertices which are hard to cluster reliably according to the SBM. We demonstrate that our methods improve on the normal Stochastic Block model by applying it to to Earthquake Networks which are an example of networks with a large number of sparsely and irregularly connected vertices.

Keywords: *Clustering*, *Variational Bayes EM*, *Model Selection*, *Stochastic Block Model*, *Networks*, *unsupervised classification*, *Noise*
MSC: Primary 62H30; Secondary 62H12

*corresponding author
†Email: willenbo@mi.fu-berlin.de
‡Email: schuette@zib.de

1 Introduction

Networks arise in different scientific areas like Protein–Protein interaction networks in Biology [1] or actor based networks in Sociology [2, 3]. The clustering of networks with a model based approach gives insight in the topology and formation process of the network and allows the prediction of edges or links. The Stochastic Block Model (SBM), introduced in [4], is a well established and widely used model for the clustering of networks. In the SBM, the vertices of the network are grouped in clusters (or blocks) based on the edge connection profile of the vertices. The results of the SBM are easily interpretable and link prediction is easy [2, 3].
Often real world networks are given without a known ground truth of the cluster assignment of vertices. In this situation, the task is to infer the optimal hidden cluster assignment of vertices together with the optimal number of clusters and the optimal parameters of the model.
The analysis of the statistics of many real world networks shows, that most of the vertices are sparsely and irregularly connected to other vertices of the network. If the network is weighted, e.g. the edges of the network have different weights, there can also be a huge variance of possible weights. Networks which exhibit such a connection behaviour usually have a heavy tails distribution of vertex degrees. If these edge connection properties are present in a network for many vertices, the process of inferring a SBM for the network is difficult to impossible because the huge component of irregularly and sparsely connected vertices cannot be clustered with clear results according to a SBM. In the literature , these vertices are called *irrelevant* or *noisy* vertices [5, 6]. Moreover, the irrelevant vertices can disturb the inference process of the relevant vertices which can be clustered according to a SBM [7]. So, we would prefer to identify these irregular vertices before or during the inference process to avoid biased results.
The normal SBM offers no dedicated mechanism to model the irrelevant vertices. The best we can hope for is to group these vertices in one cluster and keep them in this cluster during the optimisation. This requires an inference mechanism which locks these vertices in one cluster. An inference algorithm which has this property for the SBM was proposed in [7, 8] with the Blockloading algorithm. Nevertheless, the irrelevant vertices are not modelled explicitly by the SBM and the danger of over–or under fitting the number of clusters remains, which can lead to inferior results.
We propose the Weighted Stochastic Block Model with Irrelevant Vertices (SBMIV) to address these limitations of the SBM. The SBMIV builds on the Subset Infinite Relationals Model (SIRM) introduced by [6]. The SIRM an extension of the Infinite Relational Model (IRM) introduced by [9] as a variant of the SBM with an unlimited number of clusters. The SIRM builds on work presented in [10], [11] and [5], [6]. Vertices with irregular and sparse edge connections, which are hard to cluster according to the IRM, are considered as irrelevant in the SIRM, whereas vertices which could be clustered according to the IRM are considered as relevant. A hidden variable R_i is introduced for each vertex, with $R_i = 1$ if the vertex is

relevant and $R_i = 0$ if it is irrelevant. The edge connections of the irrelevant vertices with all other vertices of the network are generated with the same parameter which is distributed according to a Beta prior distribution. The SIRM is a model for networks with simple and unweighted edges. Contrary to the SBM, a Chinese Restaurant (CRP) Prior [12] is set for the proportions of the number of vertices in the clusters in the IRM. It was noted in [6], that the use of the CRP prior for the proportions of clusters sizes in the IRM favours the emergence of minute clusters and also can lead to biased results of the cluster assignment when irrelevant vertices are present in the network. On the other hand, non–informative priors are set on the size proportions in the Bayesian variant of the SBM [13, 7].
For inference of the SIRM, a Gibbs Sampling algorithm was proposed by [6], which samples the cluster and relevance assignment of vertices together.
We adopt an extension of our Blockloading algorithm for inference of the SBMIV which we call *Relevance Blockloading*. The Blockloading algorithm builds on an adopted Variational Bayesian Expectation Maximization (VBEM) algorithm and it was shown in [7] that it outperforms Spectral Clustering of [14], collapsed Gibbs Sampling of [15] and greedy algorithms [16].
We propose an algorithmic framework which allows the use of the Integrated Likelihood Varitational Bayes (ILVB) criterion of [17, 13] (see eqn. 108 in appendix B.1 as a model selection criterion for the number of clusters of the SBMIV. Our Relevance Blockloading algorithm offers a fully Bayesian inference process based on the use of informative prior parameters, which is independent of other algorithms for finding a start cluster assignment of the vertices. We propose an original way for the choice of informative priors on the relevance of the vertices, which allows us to calculate the relevance of vertices in the first iteration of the algorithm. This procedure for finding the relevant vertices as a first step is only dependent on the choice of the informative prior parameters and will yield the same result when initialised with the same informative priors. So, for this filtering of vertices restarts with different initial relevance assignments are unnecessary. An approach for the selection of relevant features with informative priors in Variational Bayesian framework was also proposed in [5].
In the literature it is differentiated between three main algorithmic frameworks for determining the relevant vertices [5]: There is the filtering approach where the inference of relevant vertices where the inference of the relevance of the vertices is calculated in a separate step from the assignment of relevant vertices to the clusters. Second there are embedded algorithms where the relevance classification and the cluster assignment are combined in one step, the subset clustering methods of [11, 5, 6] are examples of the embedded method. Lastly, there are wrapper methods which are feature selection algorithms which 'wrap feature search around the learning algorithms that will ultimately be applied' [5].
We will propose both a filtering and an embedded variant of the Relevance Blockloading algorithm and compare them in numerical tests. Both of our algorithmic variants allow for a step by step expansion of the number of relevant vertices in a network partition which is growing during the inference process. This algorithmic

procedures makes the Relevance Blockloading algorithm more efficient than previous variational methods.
We developed the SBMIV and its Relevance Blockloading inference framework with special regard to the model based clustering of Earthquake Networks.
Earthquake Networks, which were introduced in [18], are an example of the aforementioned real world networks with no known ground truth but a large number of sparsely and irregular connected vertices which renders the reliable inference with a model based clustering approach difficult. We will show, that our Relevance Blockloading algorithm of the SBMIV outperforms the best existing variational inference method for the clustering of networks, the Blockloading algorithm of [7].

2 Model

2.1 Stochastic Block Model

We shortly review the Stochastic Block Model (SBM) for graphs with discrete positive edge weights. The SBM was introduced in [4]. Following [2, 3] a Variational Bayesian algorithm to solve this model was proposed by [13]. A variant of the SBM with positive and discrete edge weights together with a Variational Expectation Maximization algorithm was proposed in [19]. This variant of the SBM was introduced in [19] and discussed from a Bayesian perspective in [7]. A graph $G = (V,E)$ consists of a set V of N vertices or vertices and a set of (directed) edges E connecting the vertices. The edges connecting the vertices are given by an adjacency matrix $\boldsymbol{A}$. If there is an edge from vertex i to vertex j it is $A_{ij} = w$, where $w \in (0,1,2,\dots)$ is a discrete valued weight. If there is no edge from vertex i to vertex j, it is $A_{ij} = 0$. In this paper we well consider directed and weighted graphs unless otherwise stated.
The following Stochastic Block Model (SBM) was introduced in [19] as an algorithm for generating graphs and builds on the simple edge version of the SBM of [2]. We assume that $\boldsymbol{A}$ was generated by the SBM. The SBM assigns the vertices V of the graph depending on their connection probability patterns to clusters. The SBM consists of K clusters. To each vertex i, the SBM assigns a unique cluster membership. A vertex belongs to cluster k with probability π_k with $\sum_{k=1}^{K} \pi_k = 1$. The cluster membership is given by the random variable $\boldsymbol{Z}_i \in R^{1\times K}$, with $Z_{ik} = 1$ if i is an element of cluster k and $Z_{ik} = 0$ otherwise. $\boldsymbol{Z}$ is the $N \times K$ cluster indicator matrix with matrix rows $\boldsymbol{Z}_i$ for $i \in \{1,\dots,N\}$. An edge exists within each cluster k with a weight according to the rate λ_{kk} and between cluster k and l with the rate λ_{kl}. So, the weighted Poisson SBM is generated in the following way ([17, 7]):
(i) Roll a k – sided dice with $p(i \in k|Z_{ik} = 1) = \pi_k$ for side k for each vertex i, to determine the unique cluster membership of the vertex.

(ii) Draw a realization from

$$f(\cdot;\lambda_{kl}) = \frac{\lambda_{kl}^{A_{ij}}}{A_{ij}!}\exp(-\lambda_{kl}), \tag{1}$$

for the edge A_{ij} from vertex i to vertex j, with $i \in k$ and $j \in l$. Then, the joint probability for directed graphs is:

$$p(\boldsymbol{A},\boldsymbol{Z},\boldsymbol{\pi},\boldsymbol{\lambda},K) = \prod_{i\neq j}^{N}\prod_{k,l}^{K} f(A_{ij};\lambda_{kl})^{Z_{ik}Z_{jl}} \prod_{i=1}^{N}\prod_{k=1}^{K}\pi_k^{Z_{ik}}. \tag{2}$$

The results of the clustering are easily interpretable. The prediction of new edges with this model follows naturally from the estimated parameters. Variants of the SBM for simple graphs exist [4, 2]. For example it is possible to replace the Poisson distribution in (2) with a Bernoulli distribution [2, 1]. Using the Poisson distribution also works for unweigthed graphs. In the following, we call this SBM the Poisson SBM contrary to the Bernoulli SBM of [2].

2.2 The Stochastic Block Model with Irrelevant Vertices

We propose the weighted Stochastic Block Model with irrelevant vertices (WSBMIV). We build on the Subset Infinite Relational Model (SIRM) proposed in [6] for networks with simple edges and following [19] we use the Poisson distribution to generate weighted edges in the SBM. We consider a network with N vertices. Following [6], each vertex i is considered relevant with the probability of $\phi_i \in [0,1]$. Thus, the relevance, $\boldsymbol{R}$, of the vertices is generated according to

$$p(R_i|\phi_i) = \prod_{i=1}^{N}\phi_i^{R_i}(1-\phi_i)^{(1-R_i)}. \tag{3}$$

If $R_i = 1$, the cluster membership Z_i of vertex i is determined acceding to

$$p(\boldsymbol{Z}|\pi,\boldsymbol{R}) = \prod_{i=1}^{N}\left(\prod_{k=1}^{K}\pi_k^{R_i Z_{ik}}\right). \tag{4}$$

Otherwise the vertex is not considered as cluster able and there is no cluster assignment for that vertex. The cluster membership is a $\boldsymbol{Z}_i \in \mathbb{R}^{K\times 1}$ vector. For all pairs of vertices $(i,j) \in \{1,\ldots,N\}^2$, where $R_i = 1$ and $R_j = 1$ holds, we generate the edges between those vertices dependant on their cluster assignments, $\boldsymbol{Z}_i$ and $\boldsymbol{Z}_j$, according to

$$p(\boldsymbol{A}|\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\lambda}) = \prod_{\substack{i,j\\ i\neq j}}^{N}\prod_{k,l}^{K}\left(\frac{\lambda_{kl}^{A_{ij}}}{A_{ij}!}\exp(-\lambda_{kl})\right)^{Z_{ik}Z_{jl}R_iR_j}. \tag{5}$$

If one vertex is or both vertices of the vertex pair (i, j) are irrelevant, the edges connecting this pair of vertices are generated with the same rate γ, so

$$p(\boldsymbol{A}|\boldsymbol{R},\gamma) = \prod_{\substack{i,j \\ i\neq j}}^{N} \left(\frac{\gamma^{A_{ij}}}{A_{ij}!} \exp(-\gamma) \right)^{\left(1-R_iR_j\right)}. \tag{6}$$

These considerations lead to the complete likelihood of the WSBMIV, for better readability we define $\vartheta = (\boldsymbol{\lambda}, \boldsymbol{\pi})$:

$$\begin{aligned} p(\boldsymbol{A}|\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\vartheta},\boldsymbol{\phi},\gamma) = \prod_{\substack{i,j \\ i\neq j}}^{N} \prod_{k,l=1}^{K} \left(\left(\frac{\lambda_{kl}^{A_{ij}}}{A_{ij}!} \right)^{R_iZ_{ik}R_jZ_{jl}} \exp\left(-R_iR_jZ_{ik}Z_{jl}\lambda_{kl} \right) \right) \prod_{i=1}^{N} \left(\prod_{k=1}^{K} \pi_k^{R_iZ_{ik}} \right) \\ \prod_{i\neq j}^{N} \left(\left(\frac{\gamma^{A_{ij}}}{A_{ij}!} \right)^{(1-R_iR_j)} \exp\left(-(1-R_iR_j)\gamma\right) \right) \prod_{i=1}^{N} \left(\phi_i^{R_i}(1-\phi_i)^{(1-R_i)} \right). \end{aligned} \tag{7}$$

In the case of an undirected network, the product over all i, j is replaced with the product over $i < j$. We generate a network according to the WSBMIV in the following way (cf. [17]):

Generation of Weighted SBM with irrelevant vertices (i) Flip a biased coin for each vertex $i \in \{1,\ldots,N\}$. With the probability ϕ_i the vertex is considered relevant, $R_i = 1$, and otherwise irrelevant, $R_i = 0$.
(ii) Roll a K–sided dice with $p(i \in k|Z_{ik} = 1, R_i = 1) = \pi_k$ for side k for each vertex i to determine the exclusive cluster assignment of the vertex.
(iii) Draw a realisation from $f(\cdot,\lambda_{kl})$ for the edge A_{ij} from vertex i to vertex j for all relevant vertices ($R_i = 1$ and $R_j = 1$) and cluster memberships $i \in k,\ j \in l$.
(iv) For all vertices i, j with $R_i = R_j = 0$, $R_i = 1$ and $R_j = 0$ or $R_i = 0$ and $R_j = 1$, draw realisation from $g(\cdot|\gamma)$ for the edge A_{ij}. The increased flexibility of the SBM concerning the proportions of the cluster sizes compared to the IRM also applies to (W)SBMIV. In the next section, we will address the Bayesian view of the WSBMIV.

2.3 Bayesian View of the SBMIV

To prepare the inference with Variational Bayesian EM methods, we state the Bayesian view of the SBMIV. The idea of the Bayesian treatment of the SBM is to set prior distributions for unknown parameters of the WSBMIV, $\boldsymbol{\Theta} = (\boldsymbol{\lambda}, \boldsymbol{\pi}, \boldsymbol{\phi}, \gamma)$. So, the parameters are treated as random variables. For the simple edge Bernoulli SBM, this idea was used in [2, 13] with conjugate prior distributions. The SIRM is also a Bayesian model [6] which allows the use of conjugate priors. Like in the SIRM, we place a Beta$(\phi_i; \zeta_i^0, \eta_i^0)$ prior distribution on the parameters ϕ_i which are

conjugate to the Bernoulli distribution and a Gamma(γ; $\alpha_\gamma^0, \beta_\gamma^0$) prior distribution on rate of the irrelevant edges, γ, which is conjugate to the Poisson distribution. Following [2, 17, 13], a Dirichlet Dir($\boldsymbol{\pi}$; $\boldsymbol{\delta}^0$) prior distribution, which is conjugate to the Multinomial distribution of the cluster assignments, is place on the parameter $\boldsymbol{\pi}$. Finally, we place a Gamma(λ_{kl}; $\alpha_{kl}^0, \beta_{kl}^0$) prior distribution on the parameters λ_{kl} [7]. We sum up the Bayesian treatment of the WSBMIV as:

$$\phi_i \sim \text{Beta}(\phi_i; \zeta_i^0, \eta_i^0) \equiv p(\phi_i), \tag{8}$$

$$\boldsymbol{\pi} \sim \text{Dir}(\boldsymbol{\pi}; \boldsymbol{\delta}^0) \equiv p(\boldsymbol{\pi}), \tag{9}$$

$$\gamma \sim \text{Gamma}(\gamma; \alpha_\gamma^0, \beta_\gamma^0) \equiv p(\gamma), \tag{10}$$

$$\lambda_{kl} \sim \text{Gamma}(\lambda_{kl}; \alpha_{kl}^0, \beta_{kl}^0) \equiv p(\lambda_{kl}). \tag{11}$$

The model generation of the Bayesian SBMIV is the same as in section 2.2, except that the model parameters have to be drawn from their respective prior distributions first, before generating the relevance assignment in (i), the cluster assignment in (ii) and the edges in steps (iii) and (iv).

3 Inference of the SBMIV

3.1 Variational Bayesian EM Inference

In the previous sections (2.2, 2.3), we explained the generation of a network according to the Poisson SBMIV. Now, we treat the inverse problem of clustering a given network according to the SBMIV. For a network given given by the adjacency matrix $\boldsymbol{A}$, we want to infer the latent variables $\boldsymbol{Z}$ and $\boldsymbol{R}$ and the unknown parameters $\boldsymbol{\Theta} = (\boldsymbol{\lambda}, \boldsymbol{\pi}, \gamma, \boldsymbol{\phi})$ of the SBMIV. We use a Variational Bayesian Expectation Maximization (VBEM) framework ([20, 21, 22, 17, 13]) to optimise the latent variables and unknown parameters of the negative log-likelihood of the SBMIV, $-\ln p(\boldsymbol{A}|K)$. The aim of the VBEM algorithm is to approximate the intractable negative log-marginal-likelihood,

$$-\ln p(\boldsymbol{A}|K) = \sum_{\boldsymbol{Z},\boldsymbol{R}} \int p(\boldsymbol{A}, \boldsymbol{\Theta}, \boldsymbol{Z}, \boldsymbol{R}) \mathrm{d}\boldsymbol{\Theta}, \tag{12}$$

with a tractable distribution $q(\cdot)$. Then, an upper variational bound of $-\ln p(\boldsymbol{A}|K)$, also called Free Energy [23, 17], dependent on the variational distribution $q(\boldsymbol{Z}, \boldsymbol{R}, \boldsymbol{\Theta})$ is derived with Jensen's inequality [23, 24, 17] .
To achieve a tractable variational distribution $q(\cdot)$ for the solution of log–likelihood of the SBMIV, we use the mean–field assumption of [17, 13] in the following way: $q(\boldsymbol{Z}, \boldsymbol{R}, \boldsymbol{\Theta}) = q(\gamma) q(\boldsymbol{\pi}) \prod_{i=1}^N q(\phi_i) \prod_{k,l}^K q(\lambda_{kl}) \prod_{i=1}^N q(\boldsymbol{Z}_i) \prod_{i=1}^N q(\boldsymbol{R}_i)$. We do not have to assume a functional form for the variational distributions $q(\cdot)$ but can infer the functional form of each distribution $q(\cdot)$ from the optimisation of the lower bound [24]. We calculate the variational bound of the negative log–likelihood,

where we omit K for the sake of brevity, in the following way:

$$-\ln\ p(\boldsymbol{A}|K) = -\ln \sum_{\boldsymbol{Z},\boldsymbol{R}} \int p(\boldsymbol{A},\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\lambda},\boldsymbol{\pi},\boldsymbol{\phi},\gamma)\mathrm{d}\boldsymbol{\Theta} \tag{13}$$

$$= -\ln \sum_{\boldsymbol{Z},\boldsymbol{R}} \int \frac{p(\boldsymbol{A},\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\Theta})}{q(\boldsymbol{Z})q(\boldsymbol{R})q(\boldsymbol{\Theta})} q(\boldsymbol{Z})q(\boldsymbol{R})q(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta} \tag{14}$$

$$\leq -\sum_{\boldsymbol{Z},\boldsymbol{R}} \int \ln\left(\frac{p(\boldsymbol{A},\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\Theta})}{q(\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\Theta})}\right) q(\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta} \tag{15}$$

$$\equiv F\left(q(\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\Theta})\right). \tag{16}$$

We provide the Free Energy of the SBMIV in Proposition 7. Now, we optimise the Free Energy (variational bound) dependent on the variational distributions $q(\cdot)$. The VBEM algorithm has an EM–like structure for the optimisation of the variational bound F with respect to the variational distribution $q(\cdot)$. The VBEM algorithm consists of two main steps: In the Expectation Step (E–step), the latent variables are optimised. In the Maximization Step (M–step), the parameters $\boldsymbol{\Theta}$ are updated.
In the case of the SBMIV, we want to infer two different types of latent variables $\boldsymbol{R}$ and $\boldsymbol{Z}$. Each variable $\boldsymbol{Z}_i$ depends on the variable R_i in the SBMIV. The same situation applies to the SIRM of [6], where a Gibbs sampling approach was proposed in which the variables R_i and $\boldsymbol{Z}_i$ are sampled together ([6]).
We propose an original algorithm to solve the SBMIV with the VBEM framework. We start with the M–step and calculate the optimal update of $q(\boldsymbol{\Theta}) = q(\boldsymbol{\lambda},\boldsymbol{\pi},\boldsymbol{\phi},\gamma)$ at step t where we keep $q^{(t)}(\boldsymbol{Z})$ and $q^{(t)}(\boldsymbol{R})$ fixed:

$$\{q^{(t+1)}(\boldsymbol{\Theta})\} = \underset{\{q(\boldsymbol{\Theta})\}}{\arg\min}\, F\left(q^{(t)}(\boldsymbol{Z}), q^{(t)}(\boldsymbol{R}), q^{(t)}(\boldsymbol{\Theta})\right). \tag{17}$$

We provide the update equations for the variational distributions of the parameters $q^{(t+1)}(\boldsymbol{\Theta})$ in the following propositions.

Proposition 1. *The optimisation of the variational bound $F[q(\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\Theta})]$ with respect to $q(\phi_i)$ shows, that $q(\phi_i)$ has the functional form of a* $\mathrm{Beta}(\phi_i;\zeta_i,\eta_i)$ *distribution. It has the same functional form as the prior distribution $p(\phi_i) = \mathrm{Beta}(\phi_i;\zeta_i^0,\eta_i^0)$. The hyperparameters $\boldsymbol{\zeta}$ and $\boldsymbol{\eta}$ for the partition vector of relevant vertices $\boldsymbol{\rho}$ are:*

$$\zeta_i = \rho_i + \zeta_i^0 \tag{18}$$

$$\eta_i = (1-\rho_i) + \eta_i^0. \tag{19}$$

Proof. See appendix A. □

Proposition 2. *The optimisation of the variational bound $F[q(\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\Theta})]$ with respect to $q(\boldsymbol{\pi})$ shows, that $q(\boldsymbol{\pi})$ has the functional form of a Dirichlet* $\mathrm{Dir}(\boldsymbol{\pi};\boldsymbol{\delta})$ *distribution. It has the same functional form as the prior distribution $p(\boldsymbol{\pi}) =$*

$\mathrm{Dir}(\boldsymbol{\pi};\boldsymbol{\delta}^0)$. *The hyper parameters* $\delta_k; k \in \{1,\ldots,K\}$ *for the relevance partition vector* $\boldsymbol{\rho}$ *and the cluster partition matrix* $\boldsymbol{Q}$ *are:*

$$\delta_k = \sum_{i=1}^{N} \rho_i Q_{ik} + \delta_k^0. \tag{20}$$

Proof. See appendix A. □

Proposition 3. *The optimisation of the Free Energy (lower variational bound)* $F[q(\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\Theta})]$ *with respect to* $q(\gamma)$ *shows, that* $q(\gamma)$ *has the functional form of a* $\Gamma(\gamma;\alpha_\gamma,\beta_\gamma)$ *distribution. It has the same functional form as the prior distribution* $p(\gamma) = \Gamma(\gamma;\alpha_\gamma^0,\beta_\gamma^0)$. *The hyperparameters* α_γ *and* β_γ *for the partition matrix* $\boldsymbol{Q}$ *are:*

$$\alpha_\gamma = \sum_{\substack{i,j \\ i\neq j}}^{N} (1-\rho_i\rho_j)A_{ij} + \alpha_\gamma^0, \tag{21}$$

$$\beta_\gamma = \sum_{\substack{i,j \\ i\neq j}}^{N} (1-\rho_i\rho_j) + \beta_\gamma^0. \tag{22}$$

Proof. See appendix A. □

Proposition 4. *The optimisation of the Free Energy (lower variational bound)* $F[q(\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\Theta})]$ *with respect to* $q(\lambda_{kl})$ *for all* $(k,l) = \{1,\ldots,K\}^2$ *shows, that* $q(\lambda_{kl})$ *has the functional form of a* $\Gamma(\lambda_{kl};\alpha_{kl},\beta_{kl})$ *distribution. It has the same functional form as the prior distribution* $p(\lambda_{kl}) = \Gamma(\lambda_{kl};\alpha_{kl}^0,\beta_{kl}^0)$. *The hyperparameters* α_{kl} *and* β_{kl}, $\forall(k,l) = \{1,\ldots,K\}^2$, *for the partition matrix* $\boldsymbol{Q}$ *are:*

$$\alpha_{kl} = \sum_{i\neq j}^{N} \rho_i\rho_j Q_{ik}Q_{jl}A_{ij} + \alpha_{kl}^0, \tag{23}$$

$$\beta_{kl} = \sum_{i\neq j}^{N} \rho_i\rho_j Q_{ik}Q_{jl} + \beta_{kl}^0. \tag{24}$$

Proof. See appendix A. □

These propositions show, that the variational distributions $q(\cdot)$ have the same functional form as the prior distributions $p(\cdot)$ of section 2.3 .
We continue the optimization of the Free Energy with respect to the latent variables in the E–step. The optimization of the Free Energy with respect to $q(\boldsymbol{Z})$ in Propostion 5 shows, that $q(\boldsymbol{Z})$ has the functional form of a Multinomial distributions.

Proposition 5. *The optimisation of the Free Energy (lower variational bound) with respect to* $q(\boldsymbol{Z}_i); \forall\, i = 1,\ldots,N,\ \{q^\star(\boldsymbol{Z}_i)\} = \underset{\{q(\boldsymbol{Z}_i)\}}{\arg\min}\, F\left(q(\boldsymbol{Z}),q(\boldsymbol{R}),q(\boldsymbol{\Theta})\right)$, *shows that* $q(\boldsymbol{Z}_i)$ *has the functional form of a multinomial distribution:*

$$q(\boldsymbol{Z}_i) = \mathscr{M}\left(Z_i;1,\boldsymbol{Q}_i = \{Q_{i1},\ldots,Q_{iK}\}\right). \tag{25}$$

The update equation for $\mathbb{E}(Z_{ik}) = Q_{ik}$, $\forall (i,k) \in \{1,\dots,N\} \times \{1,\dots,K\}$ *is given by:*

$$Q_{av} \propto \exp\Big(\sum_{\substack{i=1 \\ i\neq a}}^{N} \sum_{q,l}^{K} \rho_i \rho_a Q_{iq} A_{ai} \mathbb{E}(\ln \lambda_{vq}) + \sum_{\substack{i=1 \\ i\neq a}}^{N} \sum_{q,l}^{K} \rho_i \rho_a Q_{iq} A_{ia} \mathbb{E}(\ln \lambda_{qv}) \\ - \sum_{\substack{i=1 \\ i\neq a}}^{N} \sum_{q=1}^{K} \rho_a \rho_i Q_{iq} \big(\mathbb{E}(\lambda_{vq}) + \mathbb{E}(\lambda_{qv}) \big) + \rho_a \mathbb{E}(\ln \pi_v) \Big), \tag{26}$$

where $\mathbb{E}(\ln \lambda_{ql}) = \psi(\alpha_{ql}) - \ln(\beta_{ql})$, $\mathbb{E}(R_a) = \rho_a$, $\mathbb{E}(\lambda_{ql}) = \frac{\alpha_{ql}}{\beta_{ql}}$, $\mathbb{E}(\ln \pi_q) = \psi(\delta_q) - \psi\big(\sum \big(\sum_{l=1}^{K} \delta_l\big)\big)$ *and* $\psi(\cdot)$ *is the Digamma function.*

Proof. See appendix A. □

The cluster assignment is a *fuzzy*–update in eqn. 26, where a probability $Q_{ak} \in [0,1]$ is given for the cluster membership of vertex a in cluster k. We show in the next Proposition that the variational distribution $q(R_i)$ has the functional form of a Bernoulli $\text{Ber}(R_i;\rho_i)$ distribution.

Proposition 6. *The optimisation of the Free Energy (lower variational bound) with respect to* $q(R_i); \forall\, i = 1,\dots,N$, $\{q^\star(R_i)\} = \underset{\{q(R_i)\}}{\arg\min}\, F\left(q(\boldsymbol{Z}), q(\boldsymbol{R}), q(\boldsymbol{\Theta})\right)$, *shows that* $q(R_i)$ *has the functional form of a Bernoulli distribution:*

$$q(R_i) = \text{Ber}(R_i; \rho i). \tag{27}$$

The update equation for $\mathbb{E}(R_i) = \rho_i, \forall i \in \{1,\dots,N\}$ *is given by:*

$$\rho_a^\star = \frac{1}{1+\exp(-U_a)}, \tag{28}$$

with

$$U_a \equiv \sum_{\substack{i=1 \\ i\neq a}}^{N} \sum_{q,l}^{K} \rho_i Q_{iq} Q_{al} A_{ia} \mathbb{E}(\ln \lambda_{ql}) + \sum_{\substack{i=1 \\ i\neq a}}^{N} \sum_{q,l}^{K} \rho_i Q_{il} Q_{aq} A_{ai} \mathbb{E}(\ln \lambda_{ql}) \\ - \sum_{\substack{i=1 \\ i\neq a}}^{N} \sum_{q,l}^{K} \rho_i Q_{iq} Q_{al} \mathbb{E}(\lambda_{ql}) - \sum_{\substack{i=1 \\ i\neq a}}^{N} \sum_{q,l}^{K} \rho_i Q_{il} Q_{aq} \mathbb{E}(\lambda_{lq}) - \mathbb{E}(\ln \gamma) \sum_{\substack{i\neq a \\ i=1}}^{N} \rho_i (A_{ia} + A_{ai}) \\ + 2\mathbb{E}(\gamma) \sum_{\substack{i=1 \\ i\neq a}}^{N} \rho_i + \mathbb{E}(\ln \phi_a) - \mathbb{E}(\ln(1-\phi_a)) + \sum_{q=1}^{K} Q_{aq} \mathbb{E}(\ln \pi_q), \tag{29}$$

where $\mathbb{E}(\log \lambda_{vk}) = \psi(\alpha_{vk}) - \log(\beta_{vk})$, $\mathbb{E}(\lambda_{vk}) = \frac{\alpha_{vk}}{\beta_{vk}}$, $\mathbb{E}(\ln \gamma) = \psi(\alpha_\gamma) - \ln(\beta_\gamma)$, $\mathbb{E}(Z_{ik}) = Q_{ik}$, $\mathbb{E}(\pi_q) = \psi(\delta_q) - \psi\big(\sum \big(\sum_{l=1}^{K} \delta_l\big)\big) = G_q$, $\mathbb{E}(\ln \phi_a) = \psi(\zeta_a) - \psi(\zeta_a + \eta_a)$, $\mathbb{E}(\ln(1-\phi_a)) = \psi(\eta_a) - \psi(\zeta_a + \eta_a)$ *and* $\psi(\cdot)$ *is the Digamma function.*

Proof. See appendix A. □

The update of the relevance assignment of vertex a, ρ_a in eqn. 28, is also a fuzzy–update like the update of the cluster assignment, $Q_{ak}; \forall k \in \{1,\dots,K\}$, above, which gives us the expected value of the relevance of vertex i, ρ_i. The fuzziness of $\boldsymbol{\rho}$ poses a problem for the update equation of the cluster assignment (eqn. 26), because it can lead to a bias.

We introduce the following rule to get a hard assignment of the relevance of vertex i dependent on $\rho_i^\star$: If $\rho_i \geq 0.5$ we set $\rho_i = 1$ and otherwise we set $\rho_i = 0$. This rule is inspired by the Classification EM algorithm (CEM algorithm) of [25], where such a hard clustering is also used.

If we want to optimise the relevance, R_i, and cluster assignment, $\boldsymbol{Z}_i$, of vertex i, we have to deal with two cases: The first case is, that i is relevant and therefore $\rho_i = 1$ holds. In the case of $\rho_i = 1$, the update of $\boldsymbol{Q}_i$ is the same as for the normal Poisson SBM (see [7]) or appendix B.1 and we can use the update equation for the cluster assignment of relevant vertices eqn. 26 in a straightforward way.

If on the other hand $\rho_i = 0$ holds, we have to be careful with the update of $\boldsymbol{Q}_i$. In this case, it follows that $Q_{ik} = \frac{1}{K}$, $\forall k \in \{1,\dots,K\}$ which also gives biased results for the update of $\boldsymbol{Q}_a$ with $\forall a \neq i$. It also affects the update of ρ_i, which is given in eqn. 28 and 29 of Proposition 6. Moreover, from a perspective of the model (section 2.2) there is no cluster assignment for irrelevant vertices which leads to the conclusion that we should set $Q_{ik} \equiv 0$, $\forall k \in \{1,\dots,K\}$.

If we set the cluster partition matrix entries of the irrelevant vertices to zero, we can see that the update of ρ_i only depends on the last for terms of eqn. 29. Thus the update is dominated by the term $\mathbb{E}(\gamma)\sum_{i\neq a}^{N} \rho_i(A_{ia} + A_{ai})$. We found, that this leads automatically to the update $\rho_i = 1$, which is obviously wrong. So, we propose to calculate a cluster assignment $Q_{il}; \forall l \in \{1,\dots,K\}$ for the purpose of finding an unbiased relevance update first. There are different possibilities to assign the vertex i to a cluster in the case of $\rho_i = 0$. We could set $\rho_i = 1$ and calculate the updates of $Q_{il}; \forall l \in \{1,\dots,K\}$ according to eqn. 26 in Proposition 5, but with this approach, we would merge the irrelevant vertex i with a vertices in a cluster which were separated in previous iterations of the algorithm. Nevertheless this approach worked for all tests. We found that a better way is to limit the optimisation of the cluster assignment to the assignment to irrelevant status or to a newly introduced extra cluster. We provide the details in section 3.3.3. We found that this approach to fit the aim of the SBMIV model and it returned the best results.

So, in all cases, we calculate or set a preliminary cluster assignment of the vertex which is currently optimised to get a relevance assignment. Now, we can update ρ_i without the bias of missing terms because of $Q_{il} = 0; \forall l \in \{1,\dots,K\}$ or biased cluster assignments because of $Q_{il} = \frac{1}{K}; \forall l \in \{1,\dots,K\}$. If the update yields $\rho_i^\star = 1$, we keep the updated cluster assignment $Q_{il}^\star$, on the other hand, if $\rho_i^\star = 0$ holds, we set $Q_{il}^\star = 0; \forall l \in \{1,\dots,K\}$.

We conclude that we have to begin with the update of the cluster assignment $\boldsymbol{Q}_a$, with ρ_a set to one in all cases, and then we can proceed with the update of ρ_a dependent on the outcome of the update of $\boldsymbol{Q}_a$. After these two updates, we adjust $\boldsymbol{Q}_a$: If $\rho_a^\star = 1$ holds we keep the update of $\boldsymbol{Q}_a^\star$, if otherwise $\rho_a^\star = 0$ we set $Q_{al}^\star =$

$0; \forall l \in \{1, \ldots, K\}$. This way, the update of ρ_a is unbiased by missing terms of $\boldsymbol{Q}_a$ in the case of $\rho_a = 0$.
With these preparations we can state the E–step of the SBMIV, which consists of two parts, with the optimisation with respect to the cluster assignment of vertex i

$$\{q^{(t+1)}(\boldsymbol{Z}_i)\} = \arg\min_{\{q(\boldsymbol{Z}_i)\}} F\left(q^{(t)}(\boldsymbol{Z}), q^{(t)}(\boldsymbol{R}), q^{(t+1)}(\boldsymbol{\Theta})\right), \tag{30}$$

and the relevance assignment of vertex i, given by

$$\{q^{(t+1)}(\boldsymbol{R}_i)\} = \arg\min_{\{q(\boldsymbol{R}_i)\}} F\left(q^{(t+1)}(\boldsymbol{Z}), q^{(t)}(\boldsymbol{R}), q^{(t+1)}(\boldsymbol{\Theta})\right). \tag{31}$$

Our VBEM algorithm now consists of the iterations of the update equations **??** in the E– and M–step until the maximum number of iterations is reach or the Free Energy has converged,

$$F[q^{(t)}(\boldsymbol{Z}), q^{(t)}(\boldsymbol{R}), q^{(t)}(\boldsymbol{\Theta})] - F[q^{(t+1)}(\boldsymbol{Z}), q^{(t+1)}(\boldsymbol{R}), q^{(t+1)}(\boldsymbol{\vartheta})] < T, \tag{32}$$

where T is a predefined threshold. We sum up the algorithm in B in the appendix. This inference algorithm is an embedded algorithm, because the inference of the relevance and cluster assignment of each vertex is calculated together in the E–step ([5]). Our VBEM framework also allows us to calculate the relevant vertices in a separate filtering step where we skip the calculation of the cluster assignment. Then we continue the inference of the cluster assignment for the relevant vertices and skip the inference for the irrelevant vertices. This is a filtering algorithm where the inference of relevance and cluster assignment of the vertices are separated [5]. We will describe both inference schemes in detail below.
To start the inference process of the VBEM in the M–step (or alternatively in the E–step), we need a hard assignment of the relevance, ρ_i, for each vertex i and a fuzzy or hard cluster assignment, Q_i, for each relevant vertex i. We can also use randomly initialised assignments for both latent variables. In this case, we need appropriate informative prior parameters for the prior distributions. The quality of the results is highly dependent on a good choice of these start values. We will address this issue in detail in the section 3.3.2.
We also need a model selection criterion for the number of clusters. There exist three well established model selection criteria for the normal SBM: The asymptotic Integrated-completed-likelihood (ICL) of [26, 1, 19], the variational Integrated Likelihood variational Bayes (ILvb) criterion of [17, 13] and the non–asymptotic exact ICL of [16].
The ILvb is the value of the Free Energy of the SBM with non-informative priors after convergence [13]. Then the optimal number of clusters is chosen for the result with the optimal value (highest or lowest) of the Free Energy after convergence. We also tried this approach with the Free Energy of the SBMIV. We found that this approach gives biased results and leads to results where all vertices are considered irrelevant. We found an original way to use the ILvb as a model selection criterion where we use our Blockloading algorithm [7]. We will present this algorithm in detail in section 3.2.

3.2 Review of the Blockloading algorithm for the SBMIV

It can now be considered state of the art for the inference procedure of the SBM and related models, to choose subsets of the set of vertices of the network and to optimise these subsets with the cluster assignment of the other vertices kept fixed [27, 28, 7, 8]. A special place take variational algorithms which are able to optimise the number of clusters after [28] or during [7, 8] the inference process of the SBM. We will propose an inference algorithm based on our Blockloading algorithm [7] for the SBMIV in the next sections, which we call Relevance Blockloading.
We shortly review the Blockloading algorithm and its terminology (for a detailed discussion see [7, 8]). Then we propose an adaption of the Blockloading algorithm to the SBMIV.
We start with a cluster partition where all vertices are in one cluster and calculate the reference Free Energy, $F^{(ref)}$, of this cluster. We expand the cluster partition matrix to two clusters by applying the VBEM algorithm for the Poisson SBM of [7] to two clusters. If the Free Energy of the resulting partition, $F^{(trial)}$, is lower than $F^{(ref)}$, e.g. improves $F^{(ref)}$, we update the reference cluster partition matrix and the parameters with the new results for two clusters.
After the Initialisation of the algorithm, we choose an *active cluster* of the reference partition. The choice of the active cluster can severely affect the outcome of the calculation for networks with sparsely connected vertices [7]. We introduced the maximum probability strategy (**max–prob–strategy**) in [7], which lets us select the cluster $\max_l \sum_{l=1}^{K} \mathbb{E}(\lambda_{la}) + \sum_{\substack{l=1 \\ l\neq a}}^{K} \mathbb{E}(\lambda_{al}) = \sum_{l=1}^{K} \frac{\alpha_{al}}{\beta_{al}} + \sum_{\substack{l=1 \\ l\neq a}}^{K} \frac{\alpha_{la}}{\beta_{la}}, \forall l \in \{1,\ldots,K\} \equiv l_a$ as the active cluster. For other ways to select the active clusters see [7]. When we employ the max–prob–strategy, sparsely connected vertices are grouped in one cluster for the first iterations of the Blockloading algorithm [7]. The result is that sparsely connected vertices with low edge weights are kept in one extra cluster of the SBM. This property of the Blockloading algorithm leads easily to an adaption of the Blockloading inference scheme to the SBMIV, where this extra cluster of sparsely connected vertices is modelled explicitly.
In the Refinement Step, we check if vertices of the active cluster can be assigned to other cluster of the existing reference partition to improve the Free Energy. A detailed description of the Refinement Step can be found in [7].
After the Refinement Step, we determine the active cluster again. We then try to expand the active cluster into two new clusters to lower the reference Free Energy, like in the initialisation of the algorithm. If no improvement in either Expansion or Refinement Step was reached for all clusters of the existing partition, the Blockloading algorithm has converged. We sum up the Blockloading in the following overview:

Blockloading algorithm:
Input.–Adjacency matrix $\boldsymbol{A}$.
Result–Cluster partition matrix $\boldsymbol{Q}^{(ref)}$, number of clusters $K^{(ref)}$ and parameters $\vartheta^{(ref)}$.

(i) Blockloading Initialization.
Main Loop.
(ii) Refinement Step.
(iii) Expansion Step.
(iv) Check for Convergence of all clusters.

One of the advantages of Blockloading compared to other variational methods is, that the existing optimal partition for lower number of clusters beginning with one cluster is reused as start value partition in the following iterations of expansion and refinement. Therefore local optima are inferred one by one by the algorithm. An additional reason for the greatly improved performance is the max–prob–strategy for the choice of the active cluster mentioned above.

3.3 Relevance Blockloading Algorithm

To expand our Blockloading algorithm to the SBMIV we need a model selection criterion to evaluate the outcome of the calculation of the Initialization, Expansion and Refinement Step. A model selection for the SBMIV has to take into account, that vertices can enter or leave the relevant part of the cluster partition during the inference process. This fluctuation of vertices between relevant and irrelevant state also affects the calculation of the variational bound of the SBMIV and renders it inconsistent as a model selection criterion for the SBMIV. We discuss and propose a model selection criterion in section 3.3.1. We also need an algorithm to initialise the cluster of irrelevant vertices. This initialisation should be done early in the inference process to save computational time. We present the Initialisation of the Relevance Blockloading algorithm in section 3.3.2.
After the convergence of the Blockloading algorithm for the clusters of relevant vertices in the SBMIV, we check if the set of relevant vertices can be increased by changing the status of irrelevant vertices to relevant. We propose the relevance Expansion Step in section 3.3.2 for the Relevance Blockloading algorithm where we check if the the cluster or irrelevant vertices can be divided into an additional relevant cluster and an irrelevant cluster with a diminished number of vertices.

3.3.1 Model Selection for the SBMIV

We described in section 3.3 that we can not use the Free Energy of the SBMIV because the optimal value of this Free Energy is achieved through a partition where all vertices are considered as irrelevant. We found an original way to use the ILvb of [17, 13] for the vertex partitions returned by the VBEM inference with Blockloading. We consider the set of irrelevant vertices, $\rho_i = 0; \forall i \in \{1,\ldots,N\}$ as special cluster and build the **combined cluster partition matrix** with the cluster assignments of the relevant vertices and a vector $\mathscr{R}$ which indicates the irrelevant vertices with $\mathscr{R}_i = 1 - \rho_i; \forall i \in \{1,\ldots,N\}$. This combined cluster partition matrix is a $\boldsymbol{Q}^{(c)} \in \mathbb{R}^{N \times K^{(ref)}+1}$ matrix.

Then we calculate the Poisson ILvb [7], which is repeated for convenience in the BlockVB algorithm in appendix B.1, for the combined cluster partition matrix $\boldsymbol{Q}^{(c)}$. This Free Energy is the Reference Free Energy, $F^{(ref)}$. The Blockloading framework is now used to check for the possibility of Refinement and Expansion of the combined cluster partition $\boldsymbol{Q}^{(c)}$ measured by the Free Energy of the combined cluster partition. This is also true if we do the Expansion Step of the cluster of the irrelevant vertices (see section 3.3.2). Before we do the expansion step for the cluster of irrelevant vertices, we calculate the Free Energy of the current combined partition, $\boldsymbol{Q}^{(c)}$, then expand the cluster of the irrelevant vertices if possible with the help of the embedded BlockVB algorithm. We remark that the irrelevant vertices influence the model selection criterion.

3.3.2 Initialisation and Start Values for Relevance Blockloading

We need start values for the expected relevance of the vertices, $\boldsymbol{\rho}$, and a start cluster assignment of the relevant vertices $\boldsymbol{Q}^{(start)}$. In [7], we demonstrated that the optimisation starting with the clusters with the overall highest density connections leads to the best results for network with a high variance of edge connection probabilities. We want to transfer this approach to the inference of the SBMIV and therefore exclude irrelevant vertices at the beginning of the inference process, preferably in the first iteration of the inference algorithm.
We start the Blockloading algorithm with all vertices in one cluster and all vertices are set relevant, e.g. $\rho_i = 1; \forall i \in \{1,\ldots,N\}$. We calculate the reference Free Energy, $F^{(ref)}$, of this partition, which for this case is the Poisson ILvb repeated for convenience in appendix B.1 (see also [7]).
We recall that we aim to identify sparsely connected vertices with an uniform connection to the relevant part of the network. These irrelevant vertices are modelled by a SBM where the irrelevant vertices are connected with the same rate to all other vertices of the network (see section 2.2). We want to set special prior parameters for the Gamma$(\gamma;\alpha_\gamma^0,\beta_\gamma^0)$ prior distribution to model this edge connection profile of irrelevant vertices. We note that informative priors to set the variance of a Beta distribution for identifying relevant local and global features in a Variational Bayesian framework was used in [5] for their feature selection algorithm. We have to consider, that in most real world networks like the Earthquake Network we will present in section 4, we found that the vertices with sparse connection behaviour to have of course some variance of the probabilities for edge existence. Therefore we set the prior parameter $\alpha_\gamma^0 = 1$. A Gamma$(\gamma;1,\beta_\gamma^0)$–distribution has the form of an exponential distribution [24], which covers a wide possibility of possible values for the parameter γ. To link this Gamma prior distribution to the constant rate parameter of the irrelevant part of the SBMIV Graph, we calculate the parameter β_γ^0 so that the expectation value of γ, $\mathbb{E}[\gamma]$, is equal to the expected value of the same parameter for an SBM with all vertices in one cluster. So, we calculate the SBM with all vertices in one cluster which yields the parameters of the edge rate, α_{ER} and β_{ER}, which we calculate according to proposition 9. The Poisson SBM

with all vertices in one cluster is a special case of the Erdös–Rényi–Graph (ER–Graph) [29], so we gave the parameters the suffix ER. With the help of these two parameters we see that

$$\frac{1}{\beta_\gamma^0} = \frac{\alpha_{ER}}{\beta_{ER}} \Rightarrow \beta_\gamma^0 = \frac{\beta_{ER}}{\alpha_{ER}}. \tag{33}$$

Thus, we have a $\gamma \sim \text{Gamma}(\gamma; 1, \frac{\beta_{ER}}{\alpha_{ER}})$–prior–distribution which has the form of an exponential distribution and $\mathbb{E}[\gamma] = \frac{\beta_{ER}}{\alpha_{ER}}$.
Now that we have calculated the prior distributions for the irrelevant vertices, we can run our Relevance BlockVB algorithm for the SBMIV (see appendix B) for all vertices set to relevant and and all vertices in one cluster with the Relevance-priors $\left(1, \beta_\gamma^0\right)$ calculated above.
We note that because all vertices are in one cluster we do not need to calculate a cluster assignment. Our Relevance–prior parameters play the role of start values. The Relevance Initialisation returns the relevant vertices, which are grouped together in one cluster and the irrelevant vertices.
Now, we calculate the Free Energy, $F^{(trial)}$, of the combined cluster partition matrix, $\boldsymbol{Q}^{(c)} \in \mathbb{R}^{N\times 2}$, which is set up according to section 3.3.1. If $F^{(trial)} < F^{(ref)}$ holds, we apply the Blockloading algorithm to the relevant vertices returned by our Relevance BlockVB algorithm.
Another algorithmic approach is to apply the normal BlockVB algorithm for the Poisson SBM of [7] (see also appendix B.1 to the active clusters of the relevant vertices. This algorithmic variant leads to the Filtering Relevance Blockloading algorithm.
We emphasise that we do not need repeated initialisations for different start values with this inference. We remark that this is a tremendous advantage compared to all other variational algorithms we are aware of, which all need repeated initialisations to some extent. This is especially important for large networks. We sum up our initialisation in algorithm:

Relevance Initialisation
Input.–Adjacency matrix $\boldsymbol{A}$. Cluster partition matrix with all vertices in one cluster.
Result–Initialisation of relevant and irrelevant vertices.
(i) Calculate the ER parameters for all vertices in one cluster.
(ii) Calculate the Relevance Priors.
(iii) Apply the Relevance BlockVB algorithm of appendix B with the Relevance priors.

We always calculate the reference Free Energy, $F^{(ref)}$, and the trial Free Energy, $F^{(trial)}$, of the combined partition matrix, $\boldsymbol{Q}^{(c)} \in \mathbb{R}^{N\times K^{(ref)}+1}$. When the Blockloading algorithm for the relevant vertices has converged, we can check if the set of relevant vertices can be expanded. To do this we use our newly introduced *Relevance Expansion Step* in section 3.3.3.

3.3.3 Relevance Expansion Step and Convergence

After the convergence of all active clusters with relevant vertices we proceed by setting the irrelevant vertices to active. For the expansion of the set of relevant vertices, we use the same framework as in the Relevance Initialisation Step with some adaptions. The main idea stays the same of setting back all vertices back to relevant status and grouping them in a new cluster.Therefore we build the combined partition matrix, $\boldsymbol{Q}^{(c)}$, and set all vertices to relevant. Then we use the current reference parameters of the irrelevant cluster, $\left(\alpha_\gamma^{(ref)}, \beta_\gamma^{(ref)}\right)$, returned by the last iteration of the Blockloading inference for the relevant vertices, to determine the Relevance–prior parameters following eqn. 33.
We set all vertices in the irrelevant clusters to active and start the inference with the Relevance BlockVB algorithm for the combined partition matrix, $\boldsymbol{Q}^{(c)}$ and the Relevance priors.
We only set irrelevant vertices to the added irrelevant cluster of the combined cluster partition matrix, $\boldsymbol{Q}^{(c)}$, in the E – step and don't do a full optimisation of the cluster partition matrix in the E–Step. We found that this procedure has better separation properties which means that less vertices are re–labeled from relevant to irrelevant, than doing a full optimisation.
A full E–step where all vertices in the active cluster could be assigned to any of the relevant clusters can lead to the merging of clusters which were separated before. This clustering is not optimal but due to the limited number of clusters compared to the expansion/influx by/of the new relevant vertices. Of course a full E–Step or other procedures are possible.
Like in the initialisation step, we calculate the trial Free Energy, $F^{(trial)}$, for the returned combined trial cluster partition matrix, $\boldsymbol{Q}^{(c)}_{(trial)}$. If the reference Free Energy was improved, e.g. if $F^{(trial)} < F^{(ref)}$ holds, we update all parameters and hidden variables, $\left(\boldsymbol{\rho}^{(ref)}, \boldsymbol{Q}^{(ref)}, \boldsymbol{\Theta}^{(ref)}\right)$. Then we restart the Relevance Blockloading algorithm for the now updated set of relevant vertices. We remark, that a relevant vertex may be found as irrelevant during the optimization of relevant clusters, following the embedded E–Step of section 3.1, but a vertex can only enter the set of relevant vertices from irrelevant status during the relevance Expansion Step where the set of irrelevant vertices is active.
If otherwise $F^{(trial)} \geq F^{(ref)}$ holds, the Relevance Blockloading algorithm has converged. We sum up the whole algorithm:

Embedded Relevance Blockloading algorithm
Input.–Adjacency matrix $\boldsymbol{A}$.
Result–Cluster partition matrix, $\boldsymbol{Q}^{(ref)}$, number of clusters, $K^{(ref)}$, parameters, $\Theta^{(ref)}$ and the relevance assignment of vertices, $\boldsymbol{\rho}$.
(i) Relevance Initialization Step.
Main Loop.
(ii) Embedded Refinement Step.

(iii) Embedded Expansion Step.
(iv) Check for Convergence of relevant clusters.
(v) Relevance Expansion Step.
(vi) Check for Convergence of the irrelevant cluster.

For the Embedded Relevance Blockloading algorithm we use the Relevance BlockVB algorithm presented in appendix B in all cases. The irrelevant vertices do not influence the inference process of the cluster assignment of the relevant vertices at all. This feature of Embedded Relevance Blockloading is comparable to the Gibbs sampling procedure in [6]. So we say that the ERB is an algorithm **without noise influence**.
For the Filtering variant of the Relevance Blockloading algorithm, we replace steps *(ii)* and *(iii)* with the BlockVB Refinement and Expansion Step proposed in [7]. The BlockVB algorithm is repeated in appendix B.1. We sum up the Filtering Relevance Blockloading (FRB) algorithm:

Filtering Relevance Blockloading algorithm
Input.–Adjacency matrix $\boldsymbol{A}$.
Result–Cluster partition matrix, $\boldsymbol{Q}^{(ref)}$, number of clusters, $K^{(ref)}$, parameters, $\Theta^{(ref)}$ and the relevance assignment of vertices, $\boldsymbol{\rho}$.
(i) Relevance Initialization Step with Relevance BlockVB.
Main Loop.
(ii) Refinement Step with BlockVB.
(iii) Expansion Step with BlockVB.
(iv) Check for Convergence of relevant clusters.
(v) Relevance Expansion Step with Relevance BlockVB.
(vi) Check for Convergence of the irrelevant cluster.

The Refinement and Expansion Step for the relevant vertices of the FRB algorithm are applied to the combined cluster partition matrix $\boldsymbol{Q}^{(c)}$. The difference to the Blockloading algorithm of [7] is, that the cluster of irrelevant vertices is only active in the Relevance Expansion Step (RE–Step).
We remark, that in the Refinement Step, vertices can leave the set of relevant vertices and become irrelevant but a vertex can only become relevant in the RE–Step. The presence of the irrelevant vertices in a separate cluster influences the inference process for the relevant vertices. This is separates our FRB algorithm from the algorithm proposed in [6]. Therefore the FRB is an algorithm **with noise influence**.
We will compare both the Filtering and the Embedded Relevance Blockloading algorithm with numerical tests in section 4.

3.3.4 Successive Filtering

The relevance Expansion Step can be applied to a given adjacency matrix repeatedly without inference of the relevant part of the model. We propose the following

Filtering procedure to cluster a network independently of different start values. We call this algorithm the *Successive Filtering* algorithm. It allows us to divide the cluster partition matrix into different macro–clusters consisting of several clusters which then can be further refined expanded in parallel.
For a given adjacency matrix, we start with the relevance Initialisation Step of section 3.3.2 where we start with a cluster partition of all vertices in one cluster and set to relevant. We proceed by calculating the reference Free Energy,$F^{(ref)}$, and the Relevance priors. The relevance Expansion Step (RE–Step) yields a cluster or relevant vertices and the cluster of irrelevant vertices. We calculate the trial Free Energy of the combined matrix.
As described in section 3.3.2, we check if the trial Free Energy is lower than the reference Free Energy. If the $F^{(ref)}$ could be improved, we apply the relevance Expansion Step to the new cluster of irrelevant vertices. We continue this procedure as long as $F^{(ref)}$ can be improved.
This algorithm does not need several re–initialisation with different vertices and converges very fast. It provides us with separated parts of the network. With the help of this algorithm we can extract subnetworks which consist of clusters with homogeneous connection profiles compared with the rest of the network. On each of the returned subnetwork we perform the Relevance Expansion Step (RE–Step).

4 Numerical Experiments

4.1 Earthquake Network

The Earthquake Network which was introduced in [18], maps the spatial and temporal succession of earthquakes of a chosen region to a network. For convenience, we repeat our short exposition of the construction of the Earthquake Network (EN) and some important facts about the dataset presented in [7]. Important statistical properties of earthquake catalogue data are inherited by the EN [18, 30, 31].
One of the important findings presented in [30, 31] is, that the degree distribution of the Earthquake Networks under survey follows a heavy tails power law distribution. The consequence is, that the majority of vertices is sparsely and irregularly connected to other vertices of networks and there are vertices which have In [7] we clustered an example network of the Southern California Area (details presented below) with the Blockloading algorithm according to the SBM.

Table 1: Results of the Fully Bayesian Filtering Relevance Blockloading algorithm with noise influence for the Poisson SBMIV for the weighted Earthquake Network. Normalized Mutual Information (NMI) calculated in comparison to the best result of all tests for the combined matrix $\boldsymbol{Q}^c$ and for the irrelevant vertices (IV). Results were ordered according to the difference to the reference Free Energy ΔF_{ref}. Number of clusters K.

ΔF_{ref}	0	61	88	92	133	172	210	302	406	480
NMI $\boldsymbol{Q}^c$	1	0.94	0.93	0.93	0.89	0.92	0.93	0.91	0.93	0.94
NMI IV	1	0.94	0.92	0.95	0.93	0.89	0.92	0.91	0.95	0.89
K	52	50	51	52	50	51	51	50	51	50
no. of IV	1051	1049	1053	1053	1103	1059	1068	1064	1051	1063

Table 2: Results of the Filtering Relevance Blockloading algorithm with noise influence for the Poisson SBMIV for the weighted Earthquake Network. Normalized Mutual Information (NMI) calculated in comparison to the best result of all tests for the combined matrix $\boldsymbol{Q}^c$ and for the irrelevant vertices (IV). Results were ordered according to the difference to the reference Free Energy ΔF_{ref}. Number of clusters K.

ΔF_{ref}	0	130	272	299	301	333	523	537	617	635
ΔF_{ref}^{best}	36	166	308	335	337	369	559	573	654	672
NMI $\boldsymbol{Q}^c$	1	0.94	0.95	0.93	0.93	0.96	0.95	0.96	0.94	0.95
NMI IV	1	0.93	0.93	0.94	0.93	0.96	0.95	0.96	0.94	0.95
NMI best $\boldsymbol{Q}^c$	0.95	0.94	0.93	0.94	0.92	0.93	0.92	0.93	0.93	0.92
NMI best IV	0.91	0.94	0.94	0.95	0.95	0.9	0.92	0.95	0.93	0.89
K	49	47	47	46	45	48	45	45	46	45
no. of IV	1020	1039	1034	1038	1040	1016	1024	1037	1030	1010

Table 3: Results of the Embedded Relevance Blockloading algorithm without noise influence for the Poisson SBMIV for the weighted Earthquake Network. Normalized Mutual Information (NMI) calculated in comparison to the best result of all tests for the combined matrix $\boldsymbol{Q}^c$ and for the irrelevant vertices (IV). Results were ordered according to the difference to the reference Free Energy ΔF_{ref}. Number of clusters K.

ΔF_{ref}	0	134	175	242	382	412	438	687	875	1026
ΔF_{ref}^{best}	287	422	462	529	670	699	726	974	1162	1313
NMI $\boldsymbol{Q}^c$	1	0.83	0.81	0.82	0.77	0.82	0.95	0.8	0.77	0.93
NMI IV	1	0.66	0.4	0.7	0.4	0.66	0.95	0.65	0.4	0.95
NMI best $\boldsymbol{Q}^c$	0.82	0.85	0.79	0.85	0.8	0.83	0.82	0.84	0.78	0.81
NMI best IV	0.63	0.76	0.56	0.73	0.56	0.77	0.64	0.77	0.56	0.63
K	50	48	49	47	47	47	47	47	46	46
no. of IV	837	1023	1350	997	1348	1025	841	1034	1345	834

Table 4: Results of the Blockloading algorithm with non–informative priors for the Poisson SBM applied to the weighted Earthquake Network. Normalized Mutual Information (NMI) calculated in comparison to the best result of all tests for the combined matrix $\boldsymbol{Q}^c$ and for the proxy cluster of irrelevant vertices (proxy IV). Results were ordered according to the difference to the reference Free Energy ΔF_{ref}. Number of clusters K. Comparison with the relevance matrix, ΔF_{ref} and the combined matrix of the best result of all tests measured by the Free Energy.

ΔF_{ref}	0	20	132	163	198	228	276	304	346	517
ΔF_{ref}^{best}	322	342	454	486	520	550	598	626	668	839
NMI $\boldsymbol{Q}^c$	1	0.95	0.91	0.9	0.91	0.92	0.92	0.81	0.92	0.92
NMI proxy IV	1	0.96	0.97	0.95	0.91	0.94	0.99	0.63	0.98	0.98
NMI best $\boldsymbol{Q}^c$	0.85	0.86	0.84	0.85	0.86	0.86	0.85	0.87	0.84	0.85
NMI best IV	0.76	0.78	0.78	0.78	0.81	0.79	0.77	0.75	0.76	0.76
K	46	46	46	48	48	48	48	45	44	46
no. of IV	1157	1144	1149	1146	1120	1140	1156	931	1161	1163

The EN is constructed for a chosen geographical area and time span. A square grid is put on the area of interest [32]. The EN unfolds in the following way:
(i) Place a vertex in the first square where seismic activity occurs at the start of the observation interval.
(ii) Place a second vertex where the next time seismic activity occurs and place a (directed) edge between the last two vertices of seismic activity pointing to the latest vertex of activity.
(iii) Continue until the end of observation.

We constructed the Earthquake Network of the Southern California area (32s, 37n; 122w, 114w) for the time interval from January 1, 1984 to December 31, 2013. We chose a square length of $10\,km$ for the grid and did not include depth information of the earthquake catalog contrary to [18]. This results in 4256 squares. We used the earthquake catalogue data from the Southern California Earthquake Data center (SCDEC) [33].
Earthquake catalogues have a minimum magnitude of completeness (see e.g. [34]). The earthquake catalogue is expected to list every earthquake with magnitude equal or higher than the magnitude of completeness. It was shown in [34] that the SCDEC catalogue is complete for a magnitude of $M \geq 1.8$ on the Richter Scale from January 1, 1984 onwards. We used only earthquakes with magnitude $M \geq 1.8$ for the construction of the EN.
We set the entries on the diagonal of the adjacency matrix of the EN to zero. These entries represent aftershocks in the EN. The resulting adjacency matrix of the EN has $N = 2324$ vertices and 58718 edges. The highest edge weight of the EN was 240 and the lowest 1 (and 0 if there is no edge between the two vertices).
We evaluate and compare the numerical tests with the same principles as in [7],

so to compare the values of the Free Energy F in the following tests, we take the best of all values of the Free Energy F, F_{ref}, and calculate the difference $\Delta F_{ref} = F - F_{ref} \geq 0$.
To compare different cluster partitions $\boldsymbol{Q}$ and $\boldsymbol{Q}_0$ we use the Normalized Mutual Information (NMI) ([35]). A NMI($\boldsymbol{Q}_0$,$\boldsymbol{Q}$) of 1 means that both partitions $\boldsymbol{Q}$ and $\boldsymbol{Q}_0$ are identical. The NMI is zero when no information about $\boldsymbol{Q}_0$ can be deduced from $\boldsymbol{Q}$.
We tested the three versions of our Relevance Blockloading algorithm for the SBMIV presented in section 3, namely the Filtering Relevance Blockloading algorithm with noise influence and non–informative priors for the relevant vertices, the fully Bayesian Filtering Relevance Blockloading algorithm with noise influence and the Embedded Relevance Blockloading algorithm without noise influence and non–informative priors for the relevant vertices.
All algorithms were initialised for ten times with different start values. We used the Relevance priors in all algorithms for the Initialisation and the Relevance Expansion Step presented in section 3.3.2. The best result, measured by the Free Energy, was returned by the fully Bayesian version of the Filtering Relevance Blockloading algorithm without noise influence with a Free Energy of $F_{ref}^{best} = 133414$ and $K_{ref} = 52$ clusters. The fully Bayesian Blockloading algorithm is explained in [8]. The results of the tests for this algorithm are presented in table 1.
We compare the combined cluster partition matrices $\boldsymbol{Q}^c$ by calculating the NMI of each combined cluster partition matrix with the combined cluster partition matrix of the best result measured by the Free Energy. This measure shows how reliably an algorithm finds the same combined cluster partition for different start values.
To calculate the reliability of the inference of the relevant and irrelevant vertices we build a partition matrix where all relevant vertices are assigned to one column of the matrix and the irrelevant vertices to the other column. Then we compare these relevance partition matrix with the relevance partition matrix of the best result by calculating the NMI of theses matrices.
We repeat these calculations of the NMI with respective to the overall best result, measured by the Free Energy, of all tested algorithms. A close second best result was returned by the Filtering Relevance Blockloading algorithm without noise influence and non-informative priors for the relevant part of the model (mixed approach, see also [8]), with a Free Energy of $F_{ref} = 133449$ ($\Delta F_{ref} = 36$) and $K = 49$ clusters. The variance of the Free Energy for the mixed approach Filtering Blockloading algorithm was higher where the result with the highest Free Energy had $\Delta F_{ref} = 672$ and $K = 45$ clusters. In tables 1 and 2 we can see that the difference of the Free Energy ΔF_{ref} differs less for the full Bayesian version of the Filtering Relevance algorithm (table 1), but that in general all results have a high degree of similarity and for both algorithms. The mixed approach version seems to be bit more reliable whereas the full Bayesian version returns the best results of all tests measured by the Free Energy with non-informative priors. Both algorithms identify mostly the same vertices as irrelevant.
The results of the Embedded Relevance Blockloading algorithm without noise in-

fluence are less reliable than the those of the Filtering Blockloading algorithms measured by the similarity of the combined cluster partition matrices, relevance partition matrices and ΔF_{ref}. The best Free Energy of the Embedded Relevance algorithm was $\Delta F_{ref} = 287$ with $K =$ clusters and the worst $\Delta F_{ref} = 672$ with $K =$ clusters. We also see in table 3 that the Embedded Relevance algorithm returns vastly differing numbers of irrelevant vertices and the relevance matrices have a sub par NMI with respect to reliability and the overall best result of all algorithms. These results show, that for the Variational Bayesian Blockloading algorithm the Filtering approach works better contrary to the findings for the Gibbs Sampling inference applied to the closely related SIRM in [6], where the joint sampling of the cluster and relevance assignment were proposed as the best inference method for such a model with irrelevant vertices.

For reference we give the results for the normal Blockloading algorithm with the mixed approach for the Poisson SBM in table 4. We remark that these results differ to those presented in [7] because we repeated the tests with an updated version of our code and we found a minor bug in our code for the calculation of the NMI. The Blockloading algorithm for the SBM does not explicitly model irrelevant vertices. We take the cluster with the lowest summed expected edge existence rates (cf. section 3.2) as a proxy for the irrelevant vertices. In all tests, this proxy irrelevant clusters also was the cluster with highest number of vertices.

This observation and the built–in noise suppression with the max-prob-strategy for the choice of the active cluster presented in section 3.2 justify the choice of these clusters as a proxy for the cluster of irrelevant vertices.

The best result of the Blockloading algorithm for the SBM was $\Delta F_{ref} = 322$ and $K = 46$ clusters and the worst result was $\Delta F_{ref} = 839$. The Blockloading algorithm for the SBM is reliable but less than the Filtering Blockloading algorithm with non–informative clusters. We show in table 4 that the proxy relevance cluster has a similarity of less than $NMI = 0.8$ compared to the relevance cluster of Filtering Blockloading.

We conclude that the best choice for an inference algorithm for the SBMIV of all tested algorithms is the full Bayesian Filtering Relevance Blockloading algorithm with noise influence which also improves on the results of the Blockloading algorithm of [7] (repeated and extended in table 4) for the the normal SBM. The Filtering Relevance Blockloading algorithm with non-informative priors for the relevant priors is also a viable choice and takes a close second place in the quality of the best results and a first in general reliability. Both, the fully Bayesian and the non–informative Bayesian Filtering Relevance Blockloading algorithms clearly return better results than the Blockloading algorithm for the normal SBM of [7].

5 Conclusion

We introduced the Stochastic Block Model with Irrelevant Vertices (SBMIV) for weighted networks. We proposed the Relevance Blockloading algorithm for the inference of the SBMIV. We showed that the Relevance Blockloading algorithm can be employed as a filtering algorithm, where the determination of the relevance of vertices is separated from the cluster assignment and as an embedded algorithm where the relevance and the cluster assignment of vertices are done in the same Expectation Step. We introduced a new model selection criterion for the SBMIV, based on the Integrated Likelihood Variational Bayes (ILVB) criterion and the algorithmic framework of the (Relevance) Blockloading algorithm. We showed that the algorithmic framework of the Blockloading algorithm facilitates the identification of irrelevant vertices at the beginning and during the inference process of the model. Our new relevance informative priors for the identification of the cluster of the irrelevant vertices make the inference of relevant vertices independent of other algorithms for the initialisation of start values. We demonstrated that our filtering Relevance Blockloading algorithm together with the SBMIV improves the results for earthquake networks when compared to existing variational inference methods with the same model criterion.

Acknowledgements

C.T.W. acknowledges the funding through a GeoSim fellowship.

A Proofs and Propositions

Proof of **Proposition 1** in section 3.1:

Proof. The terms of the lower bound F dependent on $q(\phi_i)$ are:

$$F[q(\phi_i)] = -\mathbb{E}_{\boldsymbol{R},\boldsymbol{\phi}}[\ln p(\boldsymbol{R}|\boldsymbol{\phi})] - \mathbb{E}_{\boldsymbol{\phi}}(\ln p(\boldsymbol{\phi})) + \mathbb{E}_{\boldsymbol{\phi}}(\ln q(\boldsymbol{\phi})) + \text{const.}. \tag{34}$$

$$\begin{aligned} &= -\sum_{i=1}^{N} \left(\rho_i \mathbb{E}_{\phi}(\ln \phi_i) + (1-\rho_i)\mathbb{E}_{\phi}(\ln(1-\phi_i))\right) \\ &\quad - \mathbb{E}_{\boldsymbol{\phi}}(\ln p(\boldsymbol{\phi})) + \mathbb{E}_{\boldsymbol{\phi}}(\ln q(\boldsymbol{\phi})) + \text{const.}. \end{aligned} \tag{35}$$

We use Variational Bayesian optimisation of F with respect to $q(\phi_i)$ which yields:

$$\begin{aligned} \frac{\delta F[q(\cdot)]}{\delta q(\phi_i)} &= -(\zeta_i^0 - 1)\ln(\phi_i) - (\eta_i^0 - 1)\ln(1-\phi_i) + 1 + \ln q(\phi_i) \\ &\quad - \rho_i \ln \phi_i - (1-\rho_i)\ln(1-\phi_i) + \text{const.}. \end{aligned} \tag{36}$$

It follows that

$$q(\phi_i) \propto \exp\left(\left(\rho_i + \zeta_i^0 - 1\right)\ln \phi_i + \left((1-\rho_i) + \eta_i^0 - 1\right)\right). \tag{37}$$

Equation 37 shows that $q(\phi_i)$ has the functional form of a
$\text{Beta}(\rho_i + \zeta_i^0, (1-\rho_i) + \eta_i^0)$ Beta distribution, so after normalisation it holds that $q(\phi_i) = \text{Beta}(\phi_i; \zeta_i, \eta_i)$, where

$$\zeta_i = \rho_i + \zeta_i^0, \tag{38}$$

$$\eta_i = (1-\rho_i) + \eta_i^0. \tag{39}$$

for $\forall i \in \{1,\ldots,N\}$. □

Proof of **Proposition 2** in section 3.1:

Proof. The terms of the Free Energy F which depend on $q(\boldsymbol{\pi})$ are:

$$F[q(\boldsymbol{\pi})] = -\mathbb{E}_{\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\pi}}[\ln p(\boldsymbol{Z}|\boldsymbol{\pi},\boldsymbol{R})] - \mathbb{E}_{\boldsymbol{\pi}}[\ln p(\boldsymbol{\pi})] + \mathbb{E}_{\boldsymbol{\pi}}[\ln q(\boldsymbol{\pi})] + \text{const.} \tag{40}$$

$$= -\sum_{q=1}^{K}\sum_{i=1}^{N} \rho_i Q_{iq} \mathbb{E}_{\boldsymbol{\pi}}[\ln \pi_q] - \mathbb{E}_{\boldsymbol{\pi}}[\ln p(\boldsymbol{\pi})] + \mathbb{E}_{\boldsymbol{\pi}}[\ln q(\boldsymbol{\pi})] + \text{const.}. \tag{41}$$

Variational Bayesian optimisation of the Free Energy F yields:

$$\frac{\delta F[q(\boldsymbol{\pi})]}{\delta q(\boldsymbol{\pi})} = -\sum_{q=1}^{K}\sum_{i=1}^{N} \rho_i Q_{iq} \ln \pi_q - \sum_{q=1}^{K} (\delta_q^0 - 1) \ln \pi_q + \ln q(\boldsymbol{\pi}) + \text{const.} \tag{42}$$

$$\Rightarrow\ q(\boldsymbol{\pi}) \propto \exp\left(\sum_{q=1}^{K}\sum_{i=1}^{N} \left(\rho_i Q_{iq} + \delta_q^0 - 1\right) \ln \pi_q \right). \tag{43}$$

Normalisation of eqn. (43) shows that $q(\boldsymbol{\pi})$ has the functional form of a $\text{Dir}(\boldsymbol{\pi};\boldsymbol{\delta})$ Dirichlet distribution with the update equations

$$\delta_k = \sum_{i=1}^{N} \rho_i Q_{ik} + \delta_k^0;\ \forall k \in \{1,\ldots,K\}, \tag{44}$$

for the parameters. □

Proof of **Proposition 3** in section 3.1:

Proof. The terms of the lower bound F dependent on $q(\gamma)$ are:

$$F[q(\gamma)] = \sum_{\substack{i,j \\ i\neq j}}^{N} \Big(-(1-\rho_i\rho_j A_{ij}) \mathbb{E}_\gamma(\ln \gamma) + (1-\rho_i\rho_j) \mathbb{E}_\gamma(\gamma) \Big)$$
$$- \mathbb{E}_\gamma(\ln p(\gamma)) + \mathbb{E}_\gamma(\ln q(\gamma)) + \text{const.}. \tag{45}$$

We use Variational Bayesian for the optimisation of F with respect to $q(\gamma)$ which yields:

$$\frac{\delta F[q(\cdot)]}{\delta q(\gamma)} = \sum_{\substack{i,j \\ i\neq j}}^{N} \Big(-(1-\rho_i\rho_j) A_{ij} \ln(\gamma) + (1-\rho_i\rho_j)\gamma$$
$$- (\alpha_\gamma^0 - 1)\ln(\gamma) + \beta_\gamma^0 \gamma + 1 + \ln q(\gamma) \Big) + \text{const.}. \tag{46}$$

It follows that

$$q(\gamma) \propto \exp\left(\left(\left(\sum_{\substack{i,j \\ i\neq j}}^{N}(1-\rho_i\rho_j)A_{ij}\right)+\alpha_\gamma^0-1\right)\ln(\gamma)-\left(\sum_{\substack{i,j \\ i\neq j}}^{N}(1-\rho_i\rho_j)+\beta_\gamma^0\right)\gamma\right). \tag{47}$$

Equation 47 shows that $q(\lambda_{ql})$ has the functional form of a $\Gamma(\sum_{\substack{i,j \\ i\neq j}}^{N}(1-\rho_i\rho_j)A_{ij}+\alpha_\gamma^0, \sum_{\substack{i,j \\ i\neq j}}^{N}(1-\rho_i\rho_j)+\beta_\gamma^0)$ Gamma distribution.

□

Proof of **Proposition 4** in section 3.1:

Proof. The terms of the lower bound F dependent on $\boldsymbol{\lambda}$ are:

$$\begin{aligned} F\left[q(\boldsymbol{\lambda})\right] =& -\mathbb{E}_{\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\lambda},\gamma}[\ln p(\boldsymbol{A}|\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\lambda},\gamma)]+\mathbb{E}_{\boldsymbol{\lambda}}[\ln q(\boldsymbol{\lambda})] \\ & -\mathbb{E}_{\boldsymbol{\lambda}}[\ln p(\boldsymbol{\lambda})]+\text{const.} \qquad (48) \\ =& \sum_{q,l}^{K}\sum_{\substack{i,j \\ i\neq j}}^{N}\left(-\rho_i Q_{iq}\rho_j Q_{jl}A_{ij}\mathbb{E}_{\boldsymbol{\lambda}}\left[\ln\lambda_{ql}\right]+\rho_i\rho_j Q_{iq}Q_{jl}\mathbb{E}_{\boldsymbol{\lambda}}\left[\lambda_{ql}\right]\right) \\ & -\mathbb{E}_{\boldsymbol{\lambda}}\left[\ln p(\boldsymbol{\lambda})\right]+\mathbb{E}_{\boldsymbol{\lambda}}\left[\ln q(\boldsymbol{\lambda})\right]+\text{const..} \qquad (49) \end{aligned}$$

Variational Bayesian optimisation of F with respect to $q(\lambda_{ql})$ yields:

$$\begin{aligned} \frac{\delta F\left[q(\cdot)\right]}{\delta q(\lambda_{ql})} =& \sum_{\substack{i,j \\ i\neq j}}^{N}\left(-\rho_i\rho_j Q_{iq}Q_{jl}A_{ij}\ln\lambda_{ql}+\rho_i\rho_j Q_{iq}Q_{jl}\lambda_{ql}\right)+\ln q(\lambda_{ql}) \\ & -\left(\alpha_{ql}^0\ln(\beta_{ql}^0)-\ln\Gamma(\alpha_{ql}^0)+(\alpha_{ql}^0-1)\ln\lambda_{ql}-\beta_{ql}^0\lambda_{ql}\right)+\text{const..} \qquad (50) \end{aligned}$$

It follows that

$$q(\lambda_{ql}) \propto \exp\Big(\big(\sum_{\substack{i,j \\ i\neq j}}^{N}\rho_i\rho_j Q_{iq}Q_{jl}A_{ij}+\alpha_{ql}^0-1\big)\ln\lambda_{ql}-\big(\sum_{\substack{i,j \\ i\neq j}}^{N}\rho_i\rho_j Q_{iq}Q_{jl}+\beta_{ql}^0\big)\lambda_{ql}\Big). \tag{51}$$

Equation (51) shows that $q(\lambda_{ql})$ has the functional form of a $\Gamma(\sum_{i\neq j}^{N}\rho_i\rho_j Q_{ik}Q_{jl}A_{ij}+\alpha_{kl}^0, \sum_{i\neq j}^{N}\rho_i\rho_j Q_{ik}Q_{jl}+\beta_{kl}^0)$ Gamma distribution.

□

Proof of **Proposition 5** in section 3.1:

Proof. We collect the terms of the Free Energy F which depend on $q(\boldsymbol{Z})$:

$$F[q(\boldsymbol{Z})] = -\mathbb{E}_{\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\Theta}}[\ln p(\boldsymbol{A},\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\lambda},\gamma,\boldsymbol{\pi},\boldsymbol{\phi})] + \mathbb{E}_{\boldsymbol{Z}}[\ln q(\boldsymbol{Z})] + \text{const.} \tag{52}$$

$$= -\mathbb{E}_{\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\lambda},\gamma}[\ln p(\boldsymbol{A}|\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\lambda},\gamma)] - \mathbb{E}_{\boldsymbol{R},\phi}[\ln p(\boldsymbol{R}|\boldsymbol{\phi})] - \mathbb{E}_{\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\pi}}[\ln p(\boldsymbol{Z}|\boldsymbol{\pi},\boldsymbol{R})]$$
$$+ \mathbb{E}_{\boldsymbol{Z}}[\ln q(\boldsymbol{Z})] + \text{const.} \tag{53}$$

$$= \mathbb{E}_{\boldsymbol{Z}}\Big(-\sum_{\substack{i,j \\ i\neq j}}^{N}\sum_{q,l}^{K} \rho_i\rho_j Z_{iq}Z_{jl}A_{ij}\mathbb{E}(\ln\lambda_{ql}) + \sum_{\substack{i,j \\ i\neq j}}^{N}\sum_{q,l}^{K} \rho_i\rho_j Z_{ik}Z_{jl}\mathbb{E}(\lambda_{ql})$$
$$- \sum_{i=1}^{N}\sum_{q=1}^{K} \rho_i Z_{iq}\mathbb{E}(\ln\pi_q)\Big) + \mathbb{E}_{\boldsymbol{Z}}\Big(\sum_{i=1}^{N} q(\boldsymbol{Z}_i)\ln q(\boldsymbol{Z}_i)\Big) + \text{const.}. \tag{54}$$

Variational Bayesian Optimization of F with respect to $q(\boldsymbol{Z}_a)$ yields:

$$\ln q(\boldsymbol{Z}_a) \propto \sum_{v=1}^{K} Z_{av}\Big(\sum_{\substack{i=1 \\ i\neq a}}^{N}\sum_{q=1}^{K} \rho_a\rho_i Q_{iq}A_{ai}\mathbb{E}(\ln\lambda_{vq}) + \sum_{\substack{i=1 \\ i\neq a}}^{N}\sum_{q=1}^{K} \rho_a\rho_i Q_{iq}A_{ia}\mathbb{E}(\ln\lambda_{qv})$$
$$- \sum_{\substack{i=1 \\ i\neq a}}^{N}\sum_{q=1}^{K} \rho_a\rho_i Q_{iq}\left(\mathbb{E}(\lambda_{vq}) + \mathbb{E}(\lambda_{qv})\right) + \rho_a\mathbb{E}(\ln\pi_v)\Big). \tag{55}$$

Taking the exponential of eqn. 55 leads to:

$$q(\boldsymbol{Z}_a) \propto \exp\Big(\sum_{v=1}^{K} Z_{av}\Big(\sum_{\substack{i=1 \\ i\neq a}}^{N}\sum_{q=1}^{K} \rho_a\rho_i Q_{iq}A_{ai}\Big(\psi(\alpha_{vq}) - \ln(\beta_{vq})\Big) + \sum_{\substack{i=1 \\ i\neq a}}^{N}\sum_{q=1}^{K} \rho_a\rho_i Q_{iq}A_{ia}\Big(\psi(\alpha_{qv}) - \ln(\beta_{qv})\Big)$$
$$- \sum_{\substack{i=1 \\ i\neq a}}^{N}\sum_{q=1}^{K} \rho_a\rho_i Q_{iq}\left(\frac{\alpha_{vq}}{\beta_{vq}} + \frac{\alpha_{qv}}{\beta_{qv}}\right) + \rho_a\Big(\psi(\delta_q) - \psi(\sum\left(\sum_{l=1}^{K}\delta_l\right))\Big)\Big)\Big), \tag{56}$$

where $\psi(\cdot)$ is the Digamma function. After normalisation of eqn. 56, we see that $q(Z_a)$ has the functional form of a $\mathscr{M}(\boldsymbol{Z}_a; 1, \boldsymbol{Q}_a = \{Q_{a1},\ldots,Q_{aK}\})$ Multinomial distribution. □

Proof of **Proposition 6** in section 3.1:

Proof. The terms of the lower bound F dependent on $q(\boldsymbol{R})$ are:

$$
\begin{aligned}
F\left[q(\boldsymbol{R})\right] &= -\mathbb{E}_{\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\Theta}}[\ln p(\boldsymbol{A},\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\lambda},\gamma,\boldsymbol{\pi},\boldsymbol{\phi})] + \mathbb{E}_{\boldsymbol{R}}[\ln q(\boldsymbol{R})] + \text{const.} && (57)\\
&= -\mathbb{E}_{\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\lambda},\gamma}[\ln p(\boldsymbol{A}|\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\lambda},\gamma)] - \mathbb{E}_{\boldsymbol{R},\phi}[\ln p(\boldsymbol{R}|\boldsymbol{\phi})] - \mathbb{E}_{\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\pi}}[\ln p(\boldsymbol{Z}|\boldsymbol{\pi},\boldsymbol{R})]\\
&\quad + \mathbb{E}_{\boldsymbol{R}}[\ln q(\boldsymbol{R})] + \text{const.} && (58)\\
&= \mathbb{E}_{\boldsymbol{R}}\Big[\sum_{q,l}^{K}\sum_{\substack{i,j\\ i\neq j}}^{N}\left(-R_iR_jQ_{iq}Q_{jl}A_{ij}\mathbb{E}[\ln\lambda_{ql}] + R_iR_jQ_{iq}Q_{jl}\mathbb{E}[\lambda_{ql}]\right)\\
&\quad - \left(\sum_{\substack{i,j\\ i\neq j}}^{N}(1-R_iR_j)A_{ij}\mathbb{E}[\ln\gamma] - (1-R_iR_j)\mathbb{E}[\gamma]\right)\\
&\quad + \sum_{i=1}^{N} -R_i\mathbb{E}[\ln\phi_i] - (1-R_i)\mathbb{E}[\ln(1-\phi_i)]\\
&\quad - \sum_{q=1}^{K}\sum_{i=1}^{N}R_iQ_{iq}\mathbb{E}[\ln\pi_q]\Big] + \mathbb{E}_{\boldsymbol{R}}\Big(\sum_{i=1}^{N}q(R_i)\ln q(R_i)\Big) + \text{const..} && (59)
\end{aligned}
$$

Variational optimisation of F with respect to $q(R_a)$ leads to:

$$
\begin{aligned}
\ln q(R_a) =& R_a\Big[\left(\sum_{\substack{i=1\\ i\neq a}}^{N}\sum_{q,l}^{K}\rho_iQ_{iq}Q_{al}A_{ia}\mathbb{E}(\ln\lambda_{ql}) - \rho_iQ_{iq}Q_{al}\mathbb{E}(\lambda_{ql})\right)\\
&+ \left(\sum_{\substack{i=1\\ i\neq a}}^{N}\sum_{q,l}^{K}\rho_iQ_{il}Q_{aq}A_{ai}\mathbb{E}(\ln\lambda_{ql}) - \rho_iQ_{il}Q_{aq}\mathbb{E}(\lambda_{ql})\right)\\
&- \sum_{\substack{i=1\\ i\neq a}}^{N}\rho_i(A_{ia}+A_{ai})\mathbb{E}(\ln\gamma) + 2\mathbb{E}(\gamma)\sum_{\substack{i=1\\ i\neq a}}^{N}\rho_i + \mathbb{E}(\ln\phi_a)\\
&- \mathbb{E}[\ln(1-\phi_a)] + \sum_{q=1}^{K}Q_{aq}\mathbb{E}(\ln\pi_q)\Big] + \text{const..} \qquad (60)
\end{aligned}
$$

To see that, by eqn.(60), $q(R_a)$ has the functional form of the logarithm of a Bernoulli $\text{Ber}(R_a;\rho_a)$ distribution, we use the following observation [36]:

$$
\ln\text{Ber}(R_a;\rho_a) = R_a\ln\rho_a + (1-R_a)\ln(1-\rho_a) = R_a\ln\left(\frac{\rho_a}{1-\rho_a}\right) + \text{const.}, \quad (61)
$$

now we set $U_a = \ln\left(\frac{\rho_a}{1-\rho_a}\right)$ which leads to

$$
\rho_a = \exp(U_a)(1-\rho_a) \Rightarrow \rho_a = \frac{1}{1+\exp(-U_a)}. \quad (62)
$$

So, we set

$$
\begin{aligned}
U_a \equiv & \left(\sum_{\substack{i=1 \\ i \neq a}}^{N} \sum_{q,l}^{K} \rho_i Q_{iq} Q_{al} A_{ia} \mathbb{E}(\ln \lambda_{ql}) - \rho_i Q_{iq} Q_{al} \mathbb{E}(\lambda_{ql}) \right) \\
& + \left(\sum_{\substack{i=1 \\ i \neq a}}^{N} \sum_{q,l}^{K} \rho_i Q_{il} Q_{aq} A_{ai} \mathbb{E}(\ln \lambda_{ql}) - \rho_i Q_{il} Q_{aq} \mathbb{E}(\lambda_{ql}) \right) \\
& - \sum_{\substack{i=1 \\ i \neq a}}^{N} \rho_i (A_{ia} + A_{ai}) \mathbb{E}(\ln \gamma) + 2\mathbb{E}(\gamma) \sum_{\substack{i=1 \\ i \neq a}}^{N} \rho_i + \mathbb{E}(\ln \phi_a) \\
& - \mathbb{E}[\ln(1-\phi_a)] + \sum_{q=1}^{K} Q_{aq} \mathbb{E}(\ln \pi_q) \qquad (63)
\end{aligned}
$$

which gives us together with eqn. (62) the optimal update, $\rho_a^\star$. □

Proposition 7. *The Free Energy after convergence (Integrated Likelihood variational bound) for the Poisson Stochastic Block Model with irrelevant vertices for K clusters, is given by*

$$
\begin{aligned}
F[q(Z,R,\boldsymbol{\Theta})] = & \sum_{i=1}^{N} \left(\ln \left(\frac{\Gamma(\zeta_i + \eta_i)}{\Gamma(\zeta_i) + \Gamma(\eta_i)} \right) - \ln \left(\frac{\Gamma(\zeta_i^0 + \eta_i^0)}{\Gamma(\zeta_i^0) + \Gamma(\eta_i^0)} \right) \right) \\
& - \alpha_\gamma^0 \ln(\beta_\gamma^0) + \ln \Gamma(\alpha_\gamma^0) + \alpha_\gamma \ln(\beta_\gamma) - \ln \Gamma(\alpha_\gamma) \\
& - \sum_{q,l}^{K} \alpha_{q,l}^0 \ln(\beta_{ql}^0) + \ln \Gamma(\alpha_{ql}^0) + \sum_{q,l}^{K} \alpha_{ql} \ln(\beta_{ql}) - \ln \Gamma(\alpha_{ql}) \\
& - \ln \left(\Gamma \left(\sum_{q=1}^{K} \delta_q^0 \right) \right) + \sum_{q=1}^{K} \ln(\Gamma(\delta_q^0)) + \ln \left(\Gamma \left(\sum_{q=1}^{K} \delta_q \right) \right) - \sum_{q=1}^{K} \ln(\Gamma(\delta_q)) \\
& + \sum_{q=1}^{K} \sum_{i=1}^{N} Q_{iq} \ln Q_{iq}. \qquad (64)
\end{aligned}
$$

where $\boldsymbol{Q}$ *is the cluster partition matrix,* $\boldsymbol{A}$ *the adjacency matrix,* $\boldsymbol{R}$ *the relevant vertices,* $\boldsymbol{\Theta} = (\boldsymbol{\lambda}, \boldsymbol{\pi}, \gamma, \boldsymbol{\phi})$ *the model parameters and* $\boldsymbol{\vartheta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\alpha}_\gamma, \boldsymbol{\beta}_\gamma, \zeta_i, \eta_i)$ *the hyper parameters.*

Proof.

$$F[q(\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\Theta})] = -\sum_{\boldsymbol{Z},\boldsymbol{R}} \int \ln\left(\frac{p(\boldsymbol{A},\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\Theta})}{q(\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\Theta})}\right) q(\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta} \tag{65}$$

$$\begin{aligned}
&= -\mathbb{E}_{\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\Theta}}[\ln p(\boldsymbol{A}|\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\Theta})] - \mathbb{E}_{\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\phi}}[\ln p(\boldsymbol{Z}|\boldsymbol{\pi},\boldsymbol{R})] - \mathbb{E}_{\boldsymbol{\lambda}}[\ln p(\boldsymbol{\lambda})] \\
&\quad - \mathbb{E}_{\gamma}[\ln p(\gamma)] - \mathbb{E}_{\boldsymbol{\pi}}[\ln p(\boldsymbol{\pi})] - \mathbb{E}_{\boldsymbol{R},\boldsymbol{\phi}}[\ln p(\boldsymbol{R}|\boldsymbol{\phi})] - \mathbb{E}_{\boldsymbol{\phi}}[\ln p(\boldsymbol{\phi})] \\
&\quad + \sum_{i=1}^{N} \mathbb{E}_{Z}[\ln q(Z_i)] + \mathbb{E}_{\boldsymbol{\pi}}[\ln q(\boldsymbol{\pi})] + \mathbb{E}_{\boldsymbol{\phi}}[\ln q(\boldsymbol{\phi})] \\
&\quad + \mathbb{E}_{\boldsymbol{\lambda}}[\ln q(\boldsymbol{\lambda})] + \mathbb{E}_{\gamma}[\ln q(\gamma)]
\end{aligned} \tag{66}$$

$$\begin{aligned}
&= \sum_{q,l}^{K}\sum_{\substack{i,j \\ i\neq j}}^{N} \left(-\rho_i\rho_j Q_{iq}Q_{jl}A_{ij}\left(\psi(\alpha_{ql}) - \ln(\beta_{ql})\right) + \rho_i\rho_j Q_{iq}Q_{jl}\left(\frac{\alpha_{ql}}{\beta_{ql}}\right)\right) \\
&\quad + \sum_{q,l}^{K} -(\alpha_{ql}^0 - 1)(\psi(\alpha_{ql}) - \ln(\beta_{ql})) + \ln\left(\Gamma(\alpha_{ql}^0)\right) \\
&\quad + \sum_{q,l}^{K} \beta_{ql}^0\left(\frac{\alpha_{ql}}{\beta_{ql}}\right) - \alpha_{ql}^0\ln(\beta_{ql}^0) + \alpha_{ql}\ln(\beta_{ql}) - \ln\left(\Gamma(\alpha_{ql})\right) \\
&\quad + \sum_{q,l}^{K} (\alpha_{ql} - 1)\left(\psi(\alpha_{ql}) - \ln(\beta_{ql})\right) - \beta_{ql}\left(\frac{\alpha_{ql}}{\beta_{ql}}\right) \\
&\quad - \left(\sum_{\substack{i,j \\ i\neq j}}^{N} (1-\rho_i\rho_j)A_{ij}\left(\psi(\alpha_\gamma) - \ln(\beta_\gamma)\right) - (1-\rho_i\rho_j)\frac{\alpha_\gamma}{\beta_\gamma}\right) \\
&\quad - \alpha_\gamma^0\ln(\beta_\gamma^0) + \ln\Gamma(\alpha_\gamma^0) - (\alpha_\gamma^0 - 1)\left(\psi(\alpha_\gamma) - \ln(\beta_\gamma)\right) - \beta_\gamma\frac{\alpha_\gamma}{\beta_\gamma} \\
&\quad + \sum_{i=1}^{N} \Big(-\rho_i\left(\psi(\zeta_i) - \psi(\zeta_i+\eta_i)\right) - (1-\rho_i)\left(\psi(\eta_i) - \psi(\zeta_i+\eta_i)\right)\Big) \\
&\quad - \left(\sum_{i=1}^{N} (\zeta_i^0 - 1)\left(\psi(\zeta_i) - \psi(\zeta_i+\eta_i)\right) + (\eta_i^0 - 1)\left(\psi(\eta_i) - \psi(\zeta_i+\eta_i)\right)\right) \\
&\quad - \left(\sum_{i=1}^{N} (\zeta_i - 1)\left(\psi(\zeta_i) - \psi(\zeta_i+\eta_i)\right) + (\eta_i - 1)\left(\psi(\eta_i) - \psi(\zeta_i+\eta_i)\right)\right) \\
&\quad - \sum_{i=1}^{N} \left(\ln(\Gamma(\zeta_i^0+\eta_i^0)) - \ln(\Gamma(\zeta_i^0)) - \ln(\Gamma(\eta_i^0))\right) \\
&\quad - \sum_{i=1}^{N} \left(\ln(\Gamma(\zeta_i+\eta_i)) - \ln(\Gamma(\zeta_i)) - \ln(\Gamma(\eta_i))\right)
\end{aligned}$$

$$
\begin{aligned}
&- \sum_{q=1}^{K} \sum_{i=1}^{N} \rho_i Q_{iq} \big(\psi(\delta_q) - \psi \left(\sum_{l=1}^{K} \delta_q \right)) - \ln \left(\Gamma \left(\sum_{l=1}^{K} \delta_l^0 \right) \right) \\
&+ \sum_{q=1}^{K} \ln \left(\Gamma(\delta_q^0) \right) - \sum_{q=1}^{K} (\delta_q^0 - 1) \left(\psi(\delta_q) - \psi \left(\sum_{l=1}^{K} \delta_l \right) \right) \\
&+ \ln \left(\Gamma \left(\sum_{q=1}^{K} \delta_q \right) \right) - \sum_{q=1}^{K} \ln \left(\Gamma(\delta_q) \right) \\
&+ \sum_{q=1}^{K} (\delta_q - 1) \left(\psi(\delta_q) - \psi \left(\sum_{l=1}^{K} \delta_l \right) \right) + \sum_{i=1}^{N} \sum_{k=1}^{K} Q_{ik} \ln Q_{ik}. \qquad (67)
\end{aligned}
$$

We insert the update equations for the hyper parameters $\boldsymbol{\vartheta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\alpha}_\gamma, \boldsymbol{\beta}_\gamma, \zeta_i, \eta_i)$ of Propositions 1, 2, 3 and 4 into eqn. (67). This yields eqn. (64). □

B Relevance Poisson BlockVB algorithm

Input: Combined Cluster Partition matrix $Q^{(c)}$, adjacency matrix $\boldsymbol{A}$, vector of relevant vertices $\boldsymbol{\rho}$, relevance prior hyperparameters α_{ER} and β_{ER}, indices of active vertices I.
Initialisation: Set non informative [37] priors for the Gamma$(\lambda_{kl};\alpha_{kl},\beta_{kl})$ distributions: $\alpha_{kl}^0 = \frac{1}{3}$; $\beta_{kl}^0 = \frac{1}{1000}$ $\forall k,l \in \{1,\ldots,K\}$. Set non informative priors of the Dirichlet prior distribution of cluster assignments: $\delta_k^0 = 1, \forall k \in \{1,\ldots,K\}$ [38, 13]. Set the informative relevance priors for the Gamma$(\gamma;\alpha_\gamma,\beta_\gamma)$ distributions of the irrelevant cluster:

$$\alpha_\gamma^0 = 1; \tag{68}$$

$$\beta_\gamma^0 = \frac{1}{\left(\frac{\alpha_{ER}}{\beta_{ER}}\right)}. \tag{69}$$

Set the non informative prior hyperparameters [38] for the distribution of relevant and irrelevant vertices:

$$\zeta_i^0 = \frac{1}{2}\forall i \in \{1,\ldots,N\}, \tag{70}$$

$$\eta_i^0 = \frac{1}{2}\forall i \in \{1,\ldots,N\}. \tag{71}$$

Prepare M–Step: Calculate

$$S_{\alpha_{kl}} = \sum_{i\neq j}^{N} \rho_i \rho_j Q_{jl} Q_{ik} A_{ij}, \tag{72}$$

$$S_{\beta_{kl}} = \sum_{i\neq j}^{N} \rho_i \rho_j Q_{ik} Q_{jl}, \tag{73}$$

$$S_{\delta_k} = \sum_{i=1}^{N} \rho_i Q_{ik}, \tag{74}$$

and

$$\alpha_{kl} = S_{\alpha_{kl}} + \alpha_{kl}^0, \tag{75}$$

$$\beta_{kl} = S_{\beta_{kl}} + \beta_{kl}^0, \tag{76}$$

$$\delta_k = S_{\delta_k} + \delta_k^0. \tag{77}$$

Initialise parameters for the relevance assigment:

$$S_{\alpha_\gamma} = \sum_{\substack{i,j \\ i \neq j}}^{N} (1 - \rho_i \rho_j) A_{ij} \tag{78}$$

$$\alpha_\gamma = S_{\alpha_\gamma} + \alpha_\gamma^0; \tag{79}$$

$$S_{\beta_\gamma} = \sum_{\substack{i,j \\ i \neq j}}^{N} (1 - \rho_i \rho_j) \tag{80}$$

$$\beta_\gamma = S_{\beta_\gamma} + \beta_\gamma^0. \tag{81}$$

Initialise parameters for the Beta$(\phi_i; \zeta_i, \eta_i)$ distribution of relevant and irrelevant vertices:

$$\zeta_i = \rho_i + \alpha_i^0, \tag{82}$$

$$\eta_i = 1 - \rho_i + \beta_i^0. \tag{83}$$

Main Loop: Until convergence of F.
Repeat: Update active vertices $a \in I$.
E–Step for $\boldsymbol{Q}$*:*
Embedded E-Step:
Set active vertex a to relevant $(\rho_a = 1)$ and calculate cluster assignment of a for all $v = \{1, \ldots, K\}$:

$$\begin{aligned} Q_{av}^\star \propto \exp\Big(& \sum_{\substack{i=1 \\ i \neq a}}^{N} \sum_{q,l}^{K} \rho_i \rho_a Q_{iq} A_{ai} \mathbb{E}(\ln \lambda_{vq}) + \sum_{\substack{i=1 \\ i \neq a}}^{N} \sum_{q,l}^{K} \rho_i \rho_a Q_{iq} A_{ia} \mathbb{E}(\ln \lambda_{qv}) \\ & - \sum_{\substack{i=1 \\ i \neq a}}^{N} \sum_{q=1}^{K} \rho_a \rho_i Q_{iq} \big(\mathbb{E}(\lambda_{vq}) + \mathbb{E}(\lambda_{qv}) \big) + \rho_a \mathbb{E}(\ln \pi_v) \Big). \end{aligned} \tag{84}$$

Normalise all updated matrix rows.
Filtering E-Step:
(Alternative to Embedded E-step.) Set matrix entry of expansion cluster of $\boldsymbol{Q}^c$ to 1.

E–Step for $\boldsymbol{\rho}$*:* Calculate relevance assignment of vertex a:

$$\begin{aligned} U_a = & \sum_{\substack{i=1 \\ i \neq a}}^{N} \sum_{q,l}^{K} \rho_i Q_{iq} Q_{al} A_{ia} \mathbb{E}(\ln \lambda_{ql}) + \sum_{\substack{i=1 \\ i \neq a}}^{N} \sum_{q,l}^{K} \rho_i Q_{il} Q_{aq} A_{ai} \mathbb{E}(\ln \lambda_{ql}) \\ & - \sum_{\substack{i=1 \\ i \neq a}}^{N} \sum_{q,l}^{K} \rho_i Q_{iq} Q_{al} \mathbb{E}(\lambda_{ql}) - \sum_{\substack{i=1 \\ i \neq a}}^{N} \sum_{q,l}^{K} \rho_i Q_{il} Q_{aq} \mathbb{E}(\lambda_{lq}) - \mathbb{E}(\ln \gamma) \sum_{\substack{i \neq a \\ i=1}}^{N} \rho_i (A_{ia} + A_{ai}) \\ & + 2\mathbb{E}(\gamma) \sum_{\substack{i=1 \\ i \neq a}}^{N} \rho_i + \mathbb{E}(\ln \theta_a) - \mathbb{E}(\ln(1 - \theta_a)) + \sum_{q=1}^{K} Q_{aq} \mathbb{E}[\ln \pi_q] \end{aligned} \tag{85}$$

it follows that

$$\rho_a^\star = \frac{1}{1+\exp(-U_a)}. \tag{86}$$

Round ρ_a. Set

$$\rho_a^\star = \begin{cases} 1, & \text{if } \rho_a^\star \geq 0.5 \\ 0, & \text{else} \end{cases}. \tag{87}$$

Update the relevant entries of the partition matrix Q: Set matrix row a to zero if $\rho_a^\star = 0$.

M–Step: Update the parameters of the distributions of the SBM.

$$S_{\alpha_{kl}} = \sum_{\substack{i,j \\ i\neq j}}^{N} \rho_i \rho_j Q_{ik} Q_{jl} A_{ij}, \tag{88}$$

$$S_{\beta_{kl}} = \sum_{\substack{i,j \\ i\neq j}}^{N} \rho_i \rho_j Q_{ik} Q_{jl}, \tag{89}$$

$$S_{\delta_k} = \sum_{i\in I} \rho_i Q_{ik}, \tag{90}$$

and the parameters of the irrelevant vertices:

$$\alpha_\gamma = \sum_{\substack{i,j \\ i\neq j}}^{N} (1-\rho_i\rho_j) A_{ij} + \alpha_\gamma^0, \tag{91}$$

$$\beta_\gamma = \sum_{\substack{i,j \\ i\neq j}}^{N} (1-\rho_i\rho_j) + \beta_\gamma^0, \tag{92}$$

$$\zeta_i = \rho_i + \zeta_i^0, \ \forall i = \{1,\ldots,N\}, \tag{93}$$

$$\eta_i = 1 - \rho_i + \eta_i^0, \ \forall i = \{1,\ldots,N\}. \tag{94}$$

Convergence: Check the convergence of the variational lower bound, $F[q(\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\Theta})]$:

$$\begin{aligned}
F[q(\boldsymbol{Z},\boldsymbol{R},\boldsymbol{\Theta})] = &\sum_{i=1}^{N} \ln\left(\frac{\Gamma(\zeta_i+\eta_i)}{\Gamma(\zeta_i)+\Gamma(\eta_i)}\right) - \ln\left(\frac{\Gamma(\zeta_i^0+\eta_i^0)}{\Gamma(\zeta_i^0)+\Gamma(\eta_i^0)}\right) \\
&- \alpha_\gamma^0 \ln(\beta_\gamma^0) + \ln\Gamma(\alpha_\gamma^0) + \alpha_\gamma \ln(\beta_\gamma) - \ln\Gamma(\alpha_\gamma) \\
&- \sum_{q,l}^{K} \alpha_{q,l}^0 \ln(\beta_{ql}^0) + \ln\Gamma(\alpha_{ql}^0) + \sum_{q,l}^{K} \alpha_{ql} \ln(\beta_{ql}) - \ln\Gamma(\alpha_{ql}) \\
&- \ln\left(\Gamma\left(\sum_{q=1}^{K} \delta_q^0\right)\right) + \sum_{q=1}^{K} \ln(\Gamma(\delta_q^0)) + \ln\left(\Gamma\left(\sum_{q=1}^{K} \delta_q\right)\right) - \sum_{q=1}^{K} \ln\left(\Gamma(\delta_q)\right) \\
&+ \sum_{q=1}^{K}\sum_{i=1}^{N} Q_{iq} \ln Q_{iq}. \qquad (95)
\end{aligned}$$

Repeat until convergence or for the chosen number of iterations.

B.1 Poisson Block VB algorithm [7]

Input: partition matrix $Q^{(start)}$, active Cluster c and adjacency matrix A.
Initialization: Find indices I of vertices in the active cluster, $i \in c$.
Set non informative prior parameters for the Gamma prior distribution: $\alpha_{kl}^0 = \frac{1}{3}$ and $\beta_{kl}^0 = 1/1000$ for all k,l [37], and for the Dirichlet distributions $\delta_k^0 = 1\,\forall k$ [13, 38].
Initialize update formulas for the M-Step:

$$S_{\alpha_{kl}} = \sum_{i\neq j}^{N} Q_{ik} Q_{jl} A_{ij}, \qquad (96)$$

$$S_{\beta_{kl}} = \sum_{i\neq j}^{N} Q_{ik} Q_{jl}, \qquad (97)$$

$$S_{\delta_k} = \sum_{i=1}^{N} Q_{ik}. \qquad (98)$$

Prepare the M–Step with:

$$S_{\alpha_{kl}}^I = \sum_{\substack{i=1\\ i\neq j}}^{N}\sum_{\substack{j\in I\\ i\neq j}} Q_{ik} Q_{jl} A_{ij} + \sum_{\substack{i\in I\\ i\neq j}}\sum_{\substack{j\notin I\\ j=1}}^{N} Q_{ik} Q_{jl} A_{ij}, \qquad (99)$$

$$S_{\beta_{kl}}^I = \sum_{\substack{i=1\\ i\neq j}}^{N}\sum_{\substack{j\in I\\ i\neq j}} Q_{ik} Q_{jl} + \sum_{\substack{i\in I\\ i\neq j}}\sum_{\substack{j\notin I\\ j=1}}^{N} Q_{ik} Q_{jl}, \qquad (100)$$

$$S_{\delta_k}^I = \sum_{i\in I} Q_{ik}. \qquad (101)$$

Calculate the parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta})$ $\forall k, l \in \{1, \ldots, K\}$ according to Proposition 9 and 10:

$$\alpha_{kl} = S_{\alpha_{kl}} + \alpha_{kl}^0, \tag{102}$$

$$\beta_{kl} = S_{\beta_{kl}} + \beta_{kl}^0, \tag{103}$$

$$\delta_k = S_{\delta_k} + \delta_k^0. \tag{104}$$

Main Loop: Until convergence of F.
Expectation Step: Until convergence of the matrix entries Q_{ik}, $\forall i \in I$, $k = \{1, \ldots, K\}$.
Calculate updates of all matrix entries Q_{ik} $i \in I$ and $k = 1, \ldots, K$ according to Proposition 8.
Calculate the norm $Q_{ik}^\star = Q_{ik} / \left(\sum_{k=1}^K Q_{ik}\right)$ of the updates Q_{ik}.
Check for convergence of the matrix entries Q_{ik}, $\forall i \in I$, $k = 1, \ldots, K$.
Maximization Step: Update the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ of the Gamma prior distributions and the parameters $\boldsymbol{\delta}$ of the Dirichlet prior distributions. Set $S_{\alpha_{kl}}^{old} = S_{\alpha_{kl}}^I$, $S_{\beta_{kl}}^{old} = S_{\beta_{kl}}^I$ and $S_{\delta_k}^{old} = S_{\delta_k}^I$. Calculate $S_{\alpha_{kl}}^I$, $S_{\beta_{kl}}^I$ and $S_{\delta_k}^I$. Calculate M–Step Updates:

$$\alpha_{kl} = S_{\alpha_{kl}} - S_{\alpha_{kl}}^{old} + S_{\alpha_{kl}}^I + \alpha_{kl}^0, \tag{105}$$

$$\beta_{kl} = S_{\beta_{kl}} - S_{\beta_{kl}}^{old} + S_{\beta_{kl}}^I + \beta_{kl}^0, \tag{106}$$

$$\delta_k = S_{\delta_k} - S_{\delta_k}^{old} + S_{\delta_k}^I + \delta_k^0. \tag{107}$$

Calculate $F(\boldsymbol{Q}, \boldsymbol{\vartheta})$ according to

$$\begin{aligned} F[q(\cdot)] = &\sum_{k,l}^K \ln\left(\frac{\beta_{kl}^{\alpha_{kl}}\Gamma(\alpha_{kl}^0)}{\beta_{kl}^{0^{\alpha_{kl}^0}}\Gamma(\alpha_{kl})}\right) + \sum_{i=1}^N \sum_{k=1}^K Q_{ik} \ln Q_{ik} \\ &+ \ln\left(\frac{\Gamma\left(\sum_{x=1}^K \delta_x\right)\prod_{x=1}^K \Gamma(\delta_x^0)}{\Gamma\left(\sum_{x=1}^K \delta_x^0\right)\prod_{x=1}^K \Gamma(\delta_x)}\right). \end{aligned} \tag{108}$$

Check for convergence of F.

Proposition 8 ([7])**.** *The optimal estimate of the expectation of the latent variable Z_{ik}, $\mathbb{E}[z_{ik}] = Q_{ik}$ for all $i \in V, q = 1, \ldots, N$, $Q_{iv}^\star = \arg\min_{Q_{iv}} F(\boldsymbol{Q}, \boldsymbol{\vartheta})$, is given by:*

$$Q_{iv} \propto \exp\Big(\sum_{\substack{i=1 \\ i \neq j}}^N \sum_{k=1}^K A_{ai} Q_{ik} C_{vk} + \sum_{\substack{i=1 \\ i \neq j}}^N \sum_{k=1}^K A_{ia} Q_{ik} C_{kv} \tag{109}$$

$$- \sum_{\substack{i=1 \\ i \neq j}}^N \sum_{k=1}^K Q_{ik} D_{vk} + G_v\Big), \tag{110}$$

where $\mathbb{E}_{\boldsymbol{\lambda}}[\log \lambda_{vk}] = \psi(\alpha_{vk}) - \log(\beta_{vk}) = C_{vk}$, $\mathbb{E}_{\boldsymbol{\lambda}}[\lambda_{vk}] + \mathbb{E}_{\boldsymbol{\lambda}}[\lambda_{kv}] = \frac{\alpha_{vk}}{\beta_{vk}} + \frac{\alpha_{kv}}{\beta_{kv}} = D_{vk}$, $\mathbb{E}_{\boldsymbol{\delta}}[\boldsymbol{\delta}_q] = \psi(\delta_q) - \psi\left(\sum\left(\sum_{l=1}^K \delta_l\right)\right) = G_q$ and $\psi(\cdot)$ is the Digamma function.

The parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta})$ of the conjugate prior distributions are updated in the Maximization Step (M-Step), according to:

Proposition 9 ([7])**.** *The optimisation of the lower variational bound (Free Energy) for $q(\lambda_{kl})$ for all $k,l = 1,\ldots,K$ shows, that $q(\lambda_{kl})$ has the functional form of a $\Gamma(\lambda_{kl};\alpha_{kl},\beta_{kl})$ distribution. It has the same functional form as the prior distribution $p(\lambda_{kl}^0) = \Gamma(\lambda_{kl};\alpha_{kl}^0,\beta_{kl}^0)$. The hyperparameters α_{kl} and β_{kl} for all $k,l = 1,\ldots,K$ for the partition matrix $\boldsymbol{Q}$ are:*

$$\alpha_{kl} = \sum_{i\neq j}^{N} Q_{ik}Q_{jl}A_{ij} + \alpha_{kl}^0, \tag{111}$$

$$\beta_{kl} = \sum_{i\neq j}^{N} Q_{ik}Q_{jl} + \beta_{kl}^0. \tag{112}$$

Proposition 10 ([13])**.** *The optimization of the lower bound (Free Energy) with respect to $q(\boldsymbol{\pi})$ produces a distribution with the same functional form as the prior $p(\boldsymbol{\pi})$*

$$q(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi};\boldsymbol{\delta}) \tag{113}$$

where

$$\delta_k = \sum_{i=1}^{K} Q_{ik} + \delta_k^0. \tag{114}$$

References

[1] J-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statist. Comput.*, 18:173–183, 2008.

[2] T. A. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14:75–100, 1997.

[3] K. Nowicki and T.A.B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 94:1077–1087, 2001.

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

[5] Y. Guan, J.G. Dy, and M.I. Jordan. A unified probabilistic model for global and local unsupervised feature selection. *Proceedings of the 28^{th} International Conference on Machine Learning, Bellevue, WA, USA, 2011*, 2011.

[6] K. Ishiguro, N. Ueda, and H. Sawada. Subset infinite relational models. *Proceedings of the* 15^{th} *International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands*, 2012.

[7] C.T. Willenbockel and C. Schütte. A variational bayesian algorithm for clustering of large and complex networks. *ZIB-Report 15–25 (April2015), Konrad–Zuse–Zentrum für Informationstechnik Berlin*, 2015.

[8] C.T. Willenbockel and C. Schütte. A variational bayesian expand and refine clustering algorithm for large and complex networks. *(in preparation)*, 2015.

[9] C. Kemp, T.L. Griffiths, and J.B. Tenenbaum. Discovering latent classes in relational data. *(Technical report). MIT, Massachusetts,USA*, 2004.

[10] P.D. Hoff. Model-based subspace clustering. *Bayesian Analysis*, 1, Number 2:321–344, 2006.

[11] P.D. Hoff. Subset clustering of binary sequences, with an application to genomic abnormality data. *Technical Report no. 456, Department of Statistics, University of Washington*, 2004.

[12] D. Blackwell and J.B. MacQueen. Ferguson distributions via polya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

[13] P. Latouche, E. Birmelé, and C. Ambroise. Variational bayesian inference and complexity control for stochastic block models. *Statistical Modelling*, 12:93, 2012.

[14] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on PAMI*, 22:888–905, 8 2000.

[15] A. McDaid, T. Murphy, and F.N.N. Hurley. Improved bayesian inference for the stochastic block model with application to large networks. *Computational Statistics and Data Analysis*, 60:12–31, 2013.

[16] E. Côme and P. Latouche. Model selection and clustering in stochastic block models with the exact integrated complete data likelihood. *arXiv:1303.2962*, 2013.

[17] J. M. Hofman and C. H. Wiggins. Bayesian approach to network modularity. *Phys. Rev. Lett.*, 100(258701), 2008.

[18] S. Abe and N. Suzuki. Scale-free network of earthquakes. *Europhys. Lett.*, 65(4):581, 2004.

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

[20] H. Attias. Inferring parameters and structure of latent variable models by variational bayes. In Laskey KB and Prade H, editors, *Uncertainty in artificial intelligence: proceedings of the fifth conference*, pages 21–30. Morgan Kaufmann, San Francisco, CA, 1999.

[21] A. Corduneanu and C. Bishop. Variational bayesian model selection for mixture distributions. In T. Richardson and T. Jaakkola, editors, *Artificial intelligence and statistics: proceedings of the eighth conference*, pages 27–34. Morgan Kaufmann, San Francisco, CA, 2001.

[22] M. Svensén and C. Bishop. Robust bayesian mixture modelling. *Neurocomputing*, 64:235–52, 2004.

[23] R. P. Feynman. *Statistical Mechanics, A Set of Lectures*. W.A. Benjamin, Reading, MA, 1972.

[24] C. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, New York, 2006.

[25] H. Zanghi, C. Ambroise, and V. Miele. Fast online graph clustering via erdös-rényi mixture. *Pattern Recognition*, 41:3592–3599, 2008.

[26] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 7:719–25, 2000.

[27] P. Gopalan and D. Blei. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, 110 (36):14534–14539, 2013.

[28] J.-B. Leger. Wmixnet: Software for clustering the nodes of binary and valued graphs using the stochastic block model. *arXiv:1402.3410v1*, 2014.

[29] P. Erdös and A. Rényi. On random graphs. i. *Publicationes Mathemticae*, 6:290–297, 1959.

[30] S. Abe and N. Suzuki. Complex-network description of seismicity. *Nonlin. Processes in Geophys.*, 13:145–150, 2006.

[31] S. Abe and N. Suzuki. Complex earthquake networks: Hierarchical organisation and assortative mixing. *Phys. Rev. E*, 74(026113), 2006.

[32] S. Abe and N. Suzuki. Aftershocks in modern perspectives: Complex earthquake network, aging, and non-markovianity. *arXiv:1202.4394v1 [physics.geo-ph]*, 2012.

[33] SCDEC. Southern california earthquake center caltech dataset, 2013. doi:10.7909/C3WD3xH1.

[34] K. Hutton, J. Woessner, and E. Handson. Earthquake monitoring in southern california for seventy years. *Bulletin of the Seismological Society of America*, 100:423–446, 2010.

[35] L. Danon, J. Duch, A. Diaz-Guilera, and A. Arenas. Comparing community structure identification. *J. Stat. Mech.*, P09008, 2005.

[36] P. Latouche, E. Birmelé, and C. Ambroise. Model selection in overlapping stochastic block models. *Electon. J. Statist.*, 8(1):762–794, 2014.

[37] J. Kerman. Neutral non informative and informative conjugate beta and gamma prior distributions. *Electronic Journal of Statistics*, 5:1450–1470, 2011.

[38] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):453–461, September 1946.