



HAL
open science

The Random Subgraph Model for the Analysis of an Ecclesiastical Network in Merovingian Gaul

Yacine Jernite, Pierre Latouche, Charles Bouveyron, Patrick Rivera, Laurent Jegou, Stéphane Lamassé

► **To cite this version:**

Yacine Jernite, Pierre Latouche, Charles Bouveyron, Patrick Rivera, Laurent Jegou, et al.. The Random Subgraph Model for the Analysis of an Ecclesiastical Network in Merovingian Gaul. 2012. hal-00764160v1

HAL Id: hal-00764160

<https://hal.science/hal-00764160v1>

Preprint submitted on 21 Dec 2012 (v1), last revised 5 Jul 2013 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THE RANDOM SUBGRAPH MODEL FOR THE ANALYSIS OF AN ECCLESIASTICAL NETWORK IN MEROVINGIAN GAUL

Laboratoire SAMM, EA 4543, Université Paris 1 Panthéon–Sorbonne,
Laboratoire LAMOP, UMR 8589, Université Paris 1 Panthéon–Sorbonne†
and Ecole Polytechnique‡*

BY YACINE JERNITE^{*,‡}, PIERRE LATOUCHE^{*}, CHARLES BOUVEYRON^{*},
PATRICK RIVERA[†], LAURENT JEGOU[†] AND STÉPHANE LAMASSÉ[†]

In the last two decades, many random graph models have been proposed to extract knowledge from networks. Most of them look for communities or more generally clusters of vertices with homogeneous connection profiles. While the first models focused on networks with binary edges only, extensions now allow to deal with valued networks. Recently, new models were also introduced in order to characterize connection patterns in networks through mixed memberships. This work was motivated by the need of analyzing a historical network where a partition of the vertices is given and where edges are typed. A known partition is seen as a decomposition of a network into subgraphs that we propose to model using a stochastic model with unknown latent clusters. Each subgraph has its own mixing vector and sees its vertices associated to the clusters. The vertices then connect with a probability depending on the subgraphs only, while the types of the edges are assumed to be sampled from the latent clusters. A variational Bayes expectation-maximization algorithm is proposed for inference as well as a model selection criterion for the estimation of the cluster number. Experiments are carried out on simulated data to assess the approach. The proposed methodology is then applied to an ecclesiastical network in merovingian Gaul. An R code implementing the inference algorithm is available from the authors upon request.

1. Introduction. Since the original work of Moreno (1934) on sociograms, network data has become ubiquitous in Biology (Albert and Barabási, 2002; Milo et al., 2002; Palla et al., 2005) and computational social sciences (Snijders and Nowicki, 1997). Applications range from the study of gene regulation processes to that of social interactions. Network analysis was also applied recently to a medieval social network in Villa, Rossi and Truong (2008), where the authors find a clustering of vertices through kernel methods. Both deterministic and probabilistic methods have been used to seek structure in these networks, depending on prior knowledge and assumptions on the form of the data. For example, Hofman and Wiggins (2008) looked for

a partition of the vertices where the clusters exhibit a transitivity property. The model of Handcock, Raftery and Tantrum (2007) on the other hand assumes the relations to be conditioned on the projection of the vertices in a latent social space. Notable among the community discovery methods, though asymptotically biased (Bickel and Chen, 2009), are those based on the modularity score given by Girvan and Newman (2002).

Many of the other currently used methods derive from the stochastic block model (SBM) (Nowicki and Snijders, 2001), which is a probabilistic generalization (Fienberg and Wasserman, 1981) of the method applied by White, Boorman and Breiger (1976) to Sampson’s famous monastery data. SBM assumes that each vertex belongs to a hidden cluster and that connection probabilities between a pair of vertices depend exclusively on their clusters, as in Frank and Harary (1982). The parameters and clusters are then inferred to optimize a criterion, usually a lower bound of an integrated log-likelihood. Thus, Latouche, Birmelé and Ambroise (2011) used an approximation of the marginal likelihood, while Daudin, Picard and Robin (2008) considered a Laplace approximation of the integrated classification likelihood. A non parametric Bayesian approach was also introduced by Kemp et al. (2006) to estimate the number of clusters while clustering the vertices. SBM was extended by the mixed membership stochastic block model (MMSBM) (Airoldi et al., 2008), which allows a vertex to belong to different clusters in its relations towards different vertices, and by the overlapping stochastic block model (Latouche, Birmelé and Ambroise, 2011), which allows a vertex to belong to no cluster or to several at the same time. More recent works focused on extending MMSBM to dynamic networks (Xing, Fu and Song, 2010), or dealing with non-binary networks, such as networks with weighted edges (Soufiani and Airoldi, 2012). Goldenberg, Zheng and Fienberg (2010) provide an extensive review of statistical network models.

However, one kind of network which, to the best of our knowledge, has not been explored yet is when a partition of the nodes is observed and bears some importance in their behaviour. For example, when analyzing a worldwide social network describing social interactions between individuals, it is reasonable to assume that different countries, or at a different scale, different regions of the world, will have different connectivity patterns. We might also observe the same kind of phenomenon between different scientific fields in a citation network.

Our aim in this paper is to provide insight into the relationships between ecclesiastics and notable people in the kingdoms that made up Merovingian Gaul, by analyzing a network characterizing their different kinds of relations. Specifically, the data set focuses on the relationships between individuals

built during the ecclesiastical councils which took place in Gaul during the 6th century. These councils were convened under the authority of a bishop to discuss specific questions relating to the Church. Though consisting mainly of clergymen, laics would also occasionally take part, as representatives of the secular power or experts in the questions discussed. These assemblies shaped a significant part of that period, and we are interested in discovering how they reflected the relationships between various groups of individuals. For this network, extra information on the vertices, namely a geographical partition, is available, associating each individual to a specific kingdom. This partition induces a decomposition of the network into subgraphs and we aim at modelling the connection pattern of each subgraph through latent clusters.

Thus, in this paper, we propose a new model, that we call the random subgraph model (RSM), for the analysis of directed networks with typed edges for which a partition of the vertices is available. On the one hand, we consider that the probability of observing an edge between two vertices depends solely on the subgraphs to which the vertices belong. On the other hand, we assume that each vertex belongs to a hidden cluster, with a probability depending on its subgraph. Then, if a relation is present, its type is drawn from a multinomial distribution whose parameters depend on the clusters to which the vertices belong. Let us emphasize that the latter property allows, once the inference is done, to compare the different subgraphs.

The article is organized as follows. The random subgraph model is presented along with its inference algorithm in Section 2, then tested on simulated data and compared to other models in Section 3. Our model is then applied to the ecclesiastical network and the results are analyzed from the historical point of view in Section 4. Concluding remarks and possible extensions are finally discussed in Section 5.

2. The random subgraph model. We consider a directed graph \mathcal{G} with N vertices represented by its $N \times N$ adjacency matrix \mathbf{X} along with a known partition \mathcal{P} of the vertices into S classes. Our goal is to cluster the network into K groups with homogeneous connection profiles, *i.e.* estimating a binary matrix \mathbf{Z} such that $Z_{ik} = 1$ if vertex i belongs to cluster k , 0 otherwise.

Let us now detail the notations. Each edge X_{ij} , describing the relation between the vertices i and j , is typed, *i.e.* takes its values in a finite set $\{0, \dots, C\}$. Note that $X_{ij} = 0$ corresponds to the absence of an edge. We assume that \mathcal{G} does not have any self loop and therefore the entries X_{ii} will not be taken into account. In order to simplify the notations when describing

Notations	Description
\mathbf{X}	Adjacency matrix. $X_{ij} \in \{0, \dots, C\}$ indicates the edge type
\mathbf{A}	Binary matrix. $A_{ij} = 1$ indicates the presence of an edge
\mathbf{Z}	Binary matrix. $Z_{ik} = 1$ indicates that i belongs to cluster k
N	Number of vertices in the network
K	Number of latent clusters
S	Number of subgraphs
C	Number of edge types
$\boldsymbol{\alpha}$	α_{sk} is the proportion of cluster k in subgraph s
$\boldsymbol{\Pi}$	Π_{klc} is the probability of having an edge of type c between vertices of clusters k and l
$\boldsymbol{\gamma}$	γ_{rs} probability of having an edge between vertices of subgraphs r and s

TABLE 1

Summary of the notations used in the paper.

the model, we also consider the binary matrix \mathbf{A} with entries A_{ij} such that $A_{i,j} = 1 \iff X_{i,j} \neq 0$.

We also emphasize that the observed partition \mathcal{P} induces a decomposition of the graph into subgraphs where each class of vertices corresponds to a specific subgraph. We introduce the variable s_i which takes its values in $\{1, \dots, S\}$ and is used to indicate in which of the subgraphs vertex i belongs, for $i \in \{1, \dots, N\}$.

2.1. The probabilistic model. The data is assumed to be generated in three steps. First, the presence of an edge from vertex i to vertex j is supposed to follow a Bernoulli distribution whose parameter depends on the subgraphs s_i and s_j only:

$$A_{i,j} \sim \mathcal{B}(\gamma_{s_i, s_j}).$$

Each vertex i is then associated to a latent cluster with a probability depending on s_i . In practice, if we assume for now that the number K of latent clusters is known, the variable \mathbf{Z}_i is drawn from a multinomial distribution:

$$\mathbf{Z}_i \sim \mathcal{M}(1; \boldsymbol{\alpha}_{s_i}),$$

where

$$\forall s \in 1, \dots, S, \sum_{k=1}^K \alpha_{sk} = 1.$$

A notable point of the model is that we allow each subgraph to have different mixing proportions $\boldsymbol{\alpha}_s$ for the latent clusters. We denote hereafter $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_S)$. Finally, if an edge between i and j is present, *i.e.* $A_{ij} = 1$, its type X_{ij} is sampled from a multinomial distribution with parameters

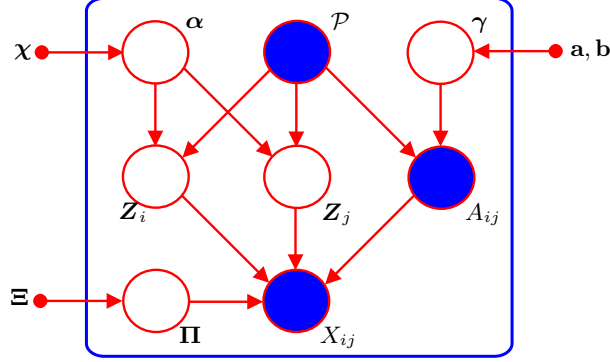


FIG 1. Graphical representation of the RSM model.

depending on the latent clusters. Thus, if i belongs to cluster k and j to cluster l :

$$X_{i,j}|Z_{ik}Z_{jl} = 1, A_{ij} = 1 \sim \mathcal{M}(1, \mathbf{\Pi}_{kl}),$$

where the sum over the C types of each vector $\mathbf{\Pi}_{kl} = (\Pi_{kl1}, \dots, \Pi_{klC})$ is:

$$\forall (k, l) \in \{1, \dots, K\}^2, \sum_{c=1}^C \Pi_{klc} = 1.$$

If there is no edge between the two vertices, the entry X_{ij} is simply set to $X_{ij} = A_{ij} = 0$. Figure 1 presents the graphical model associated with the RSM model and all notations are summarized in Table 1.

The model is therefore defined through the joint distribution:

$$\begin{aligned} p(\mathbf{X}, \mathbf{A}, \mathbf{Z} | \boldsymbol{\alpha}, \boldsymbol{\gamma}, \mathbf{\Pi}) &= p(\mathbf{X}, \mathbf{A} | \mathbf{Z}, \boldsymbol{\gamma}, \mathbf{\Pi})p(\mathbf{Z} | \boldsymbol{\alpha}) \\ &= p(\mathbf{X} | \mathbf{A}, \mathbf{Z}, \mathbf{\Pi})p(\mathbf{A} | \boldsymbol{\gamma})p(\mathbf{Z} | \boldsymbol{\alpha}), \end{aligned}$$

where

$$p(\mathbf{X} | \mathbf{A}, \mathbf{Z}, \mathbf{\Pi}) = \prod_{k,l} \prod_{c=1}^C (\Pi_{klc}^c)^{\sum_{i \neq j} \delta(X_{ij}=c) A_{ij} Z_{ik} Z_{jl}},$$

and

$$p(\mathbf{A} | \boldsymbol{\gamma}) = \prod_{i \neq j} \gamma_{r_i, r_j}^{A_{ij}} (1 - \gamma_{r_i, r_j})^{1 - A_{ij}}.$$

Finally,

$$p(\mathbf{Z} | \boldsymbol{\alpha}) = \prod_{i=1}^N \prod_{k=1}^K \alpha_{r_i, k}^{Z_{ik}}.$$

We refer to the appendix for the detailed calculation of the complete data log likelihood associated to the RSM model.

2.2. Bayesian framework. In such a context, a popular way to perform the inference of the model is to consider a Bayesian framework and to use a variational Bayes expectation-maximization (VBEM) algorithm. In order to obtain analytical optimization equations, we consider conjugate prior distributions. Thus, since \mathbf{Z}_i is sampled from a multinomial distribution, we rely on a Dirichlet prior to model the parameters $\boldsymbol{\alpha}_s$:

$$p(\boldsymbol{\alpha}_s) = \text{Dir}(\boldsymbol{\alpha}_s; \chi_{s1}^0, \dots, \chi_{sK}^0), \forall s \in \{1, \dots, S\}.$$

A similar distribution is used as a prior distribution for the parameters $\boldsymbol{\Pi}_{kl}$:

$$p(\boldsymbol{\Pi}_{kl}) = \text{Dir}(\boldsymbol{\Pi}_{kl}; \Xi_{kl1}^0, \dots, \Xi_{klC}^0), \forall (k, l) \in \{1, \dots, K\}^2.$$

If no prior information is available, a common choice in the literature consists in fixing the hyperparameters of the Dirichlet to $1/2$, *i.e.* $\chi_{sk}^0 = 1/2, \forall (s, k)$ and $\Xi_{klc} = 1/2, \forall (k, l, c)$. Such a distribution corresponds to a non informative Jeffreys prior distribution which is known to be proper (Jeffreys, 1946). A uniform distribution can also be obtained by setting the hyperparameters to 1.

Finally, since the presence or absence of an edge between a pair of vertices is drawn from a Bernoulli distribution, we rely on a beta prior for the parameters γ_{rs} :

$$p(\gamma_{rs}) = \text{Beta}(\gamma_{rs}; a_{rs}^0, b_{rs}^0), \forall (r, s) \in \{1, \dots, S\}^2.$$

Again, if no prior information is available, both hyperparameters a_{rs}^0 and b_{rs}^0 can be set to $1/2$ or 1 to obtain non informative prior distributions, respectively a Jeffreys or a uniform distribution.

2.3. Inference with the variational Bayes EM algorithm. Given the observed matrices \mathbf{X} and \mathbf{A} , we aim at estimating the posterior distribution $p(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\Pi} | \mathbf{X}, \mathbf{A})$, which in turn will allow us to compute a maximum a posteriori estimate of the clustering structure \mathbf{Z} as well as the parameters $(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\Pi})$. Because this distribution is not tractable, we rely on a variational approximation. Thus, given a distribution $q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\Pi})$, the marginal likelihood can be computed in two terms:

$$\log p(\mathbf{X}, \mathbf{A}) = \mathcal{L}(q) + KL(q(\cdot) || p(\cdot | \mathbf{X})),$$

Algorithm 1 VBEM algorithm for the RSM model

```

Initialize matrix  $\tau = \mathbf{Z}$  with kmeans
Initialize hyperparameters  $\theta^0 = (\chi^0, \gamma^0, \Xi^0)$ 
Compute  $\mathcal{L}(q)$ 
while  $|\theta^{new} - \theta^{old}| \geq \epsilon$  do
  E step: update  $\tau$ 
  M step: update  $\theta^{new} = (\chi, \gamma, \Xi)$ 
  Compute  $\mathcal{L}$ 
end while

```

where \mathcal{L} is defined as follows:

$$\mathcal{L}(q) = \sum_{\mathbf{Z}} \int_{\alpha, \gamma, \Pi} q(\mathbf{Z}, \alpha, \gamma, \Pi) \log\left(\frac{p(\mathbf{X}, \mathbf{A}, \mathbf{Z}, \alpha, \gamma, \Pi)}{q(\mathbf{Z}, \alpha, \gamma, \Pi)}\right) d\alpha d\gamma d\Pi,$$

and the KL divergence is given by:

$$KL(q(\cdot) || p(\cdot | \mathbf{X})) = - \sum_{\mathbf{Z}} \int_{\alpha, \gamma, \Pi} q(\mathbf{Z}, \alpha, \gamma, \Pi) \log\left(\frac{p(\mathbf{Z}, \alpha, \gamma, \Pi | \mathbf{X})}{q(\mathbf{Z}, \alpha, \gamma, \Pi)}\right) d\alpha d\gamma d\Pi.$$

Finding the best approximation of the posterior distribution $p(\mathbf{Z}, \alpha, \gamma, \Pi | \mathbf{X}, \mathbf{A})$ in the sense of the KL divergence becomes equivalent to finding $q(\cdot)$ that maximizes the lower bound $\mathcal{L}(q)$ of the integrated log-likelihood. To obtain a tractable algorithm, we assume that $q(\cdot)$ can be fully factorized, that is:

$$q(\mathbf{Z}, \alpha, \gamma, \Pi) = \left(\prod_{i=1}^N q(\mathbf{Z}_i) \right) \left(\prod_{s=1}^S q(\alpha_s) \prod_{t=1}^S q(\gamma_{s,t}) \right) \prod_{k,l} q(\Pi_{k,l}).$$

The functional optimization of the lower bound with respect to $q(\cdot)$ is performed using a VBEM algorithm (see Algorithm 1). All the optimization equations are given in the appendix.

2.4. Initialization. The VBEM algorithm, though useful in approximating posterior distributions of graphical models, is only guaranteed to converge to a local optimum (Bilmes, 1998). Strategies to tackle this issue include simulated annealing and the use of multiple initializations (Biernacki, Celeux and Govaert, 2003). In this work, we choose the latter option. In order to have a better chance of reaching a global optimum, VBEM is run for several initializations of a kmeans like algorithm with the following distance $d(i, j)$ between the vertices i and j :

$$(1) \quad d(i, j) = \sum_{h=1}^N \delta(X_{ih} \neq X_{jh}) A_{ih} A_{jh} + \sum_{h=1}^N \delta(X_{hi} \neq X_{hj}) A_{hi} A_{hj}.$$

The first term looks at all possible edges from i and j towards a third vertex h . If both i and j are connected to h , *i.e.* $A_{ih}A_{jh} = 1$, the edge types X_{ih} and X_{jh} are compared. By symmetry, the second term looks at all possible edges from a vertex h to both i as well as j , and compare their types. Thus, the distance computes the number of discordances in the way both i and j connect to other vertices or vertices connect to them. The algorithm starts by sampling the cluster centers among all the vertices of the network. It then iterates a two-step procedure until convergence of the cluster centers. In the first step, the vertices are classified into the cluster with the closest center. Each cluster center is then associated to a vertex minimizing its distance with all the vertices of the corresponding cluster. In Section 3, we illustrate that the distance defined in (1) and the associated kmeans like algorithm leads to a good starting point for the VBEM algorithm.

2.5. Choice of K . So far, the number K of latent clusters has been assumed to be known. Given K , we showed in Section 2.3 how an approximation of the posterior distribution over the latent structure and model parameters could be obtained. We now address the problem of estimating the number of clusters directly from the data. Given a set of values of K , we aim at selecting K^* for which the marginal log likelihood $\log p(\mathbf{X} | K)$ is maximized. However, because this integrated likelihood involves a marginalization over all the model parameters and latent variables, it is not tractable. Therefore, we propose to replace the marginal likelihood with its variational approximation, as in Bishop (2006); Latouche, Birmelé and Ambroise (2009, 2012). Thus, for each value of K considered, the VBEM algorithm is applied. We recall that the maximization of the lower bound induces a minimization of the KL divergence. After convergence of the algorithm, the lower bound is used as an approximation of $\log p(\mathbf{X} | K)$ and K is chosen such that the lower bound is maximized. We prove in the appendix that, if computed right after the M step of the variational Bayes EM, the lower bound has the following expression:

$$\mathcal{L}(q) = \sum_{r,s} \log\left(\frac{B(a_{rs}, b_{rs})}{B(a_{rs}^0, b_{rs}^0)}\right) + \sum_{s=1}^S \log\left(\frac{C(\boldsymbol{\chi}_s)}{C(\boldsymbol{\chi}_s^0)}\right) + \sum_{k,l} \log\left(\frac{C(\boldsymbol{\Xi}_{kl})}{C(\boldsymbol{\Xi}_{kl}^0)}\right) - \sum_{i=1}^N \sum_{k=1}^K \tau_{ik} \log(\tau_{ik}),$$

where $C(x) = \frac{\prod_{d=1}^D \Gamma(x_d)}{\Gamma(\sum_{d=1}^D x_d)}$ if $x \in \mathbb{R}^D$ and $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, $\forall (a, b) \in \mathbb{R}^2$.

3. Numerical experiments and comparisons. In this section, we first run experiments aiming at proving the validity of our model, focusing on the ability of its inference procedure to find the right clustering. We

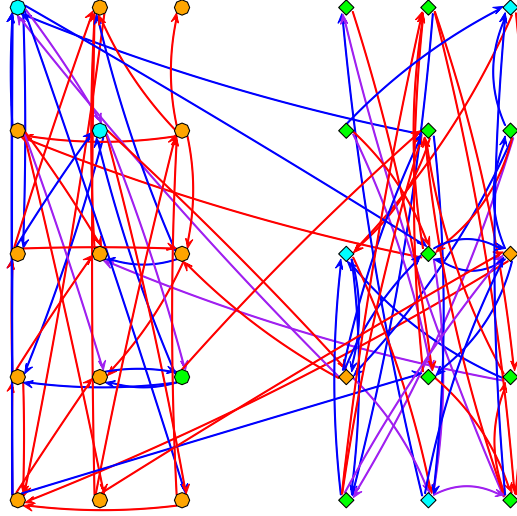


FIG 2. Example of a RSM network for $S = 2$ subgraphs (indicated by the node forms), $C = 3$ types of edges (indicated by the edge colors) and $K = 3$ clusters to identify (indicated by the node colors).

then compare its performance to that of other stochastic models for graph clustering.

3.1. *Experimental setup.* In order to evaluate the performance of our approach, we applied it on data generated according to the RSM model. To simplify the parameterization and facilitate the reproducibility of the experiments, we constrained the parameters $\mathbf{\Pi}$ and γ to have the following forms:

$$\mathbf{\Pi} = \begin{bmatrix} \mathbf{u} & \mathbf{v} & \cdots & \mathbf{v} \\ \mathbf{v} & \mathbf{u} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{v} \\ \mathbf{v} & \cdots & \mathbf{v} & \mathbf{u} \end{bmatrix}, \gamma = \begin{bmatrix} \lambda & \epsilon & \cdots & \epsilon \\ \epsilon & \lambda & \ddots & \vdots \\ \vdots & \ddots & \ddots & \epsilon \\ \epsilon & \cdots & \epsilon & \lambda \end{bmatrix},$$

where $\lambda, \epsilon \in [0, 1]$ and $\mathbf{u}, \mathbf{v} \in [0, 1]^K$. With such a parameterization, the probability λ of an edge within a subgraph is assumed to be common be-

Parameters	Scenario 1	Scenario 2	Scenario 3									
N	100	100	100									
S	1	1	3									
C	3	3	3									
K	3	3	3									
α	(0.3,0.3,0.4)	(0.3,0.3,0.4)	<table border="1"> <tr> <td>0</td> <td>0.5</td> <td>0.5</td> </tr> <tr> <td>0.5</td> <td>0</td> <td>0.5</td> </tr> <tr> <td>0.5</td> <td>0.5</td> <td>0</td> </tr> </table>	0	0.5	0.5	0.5	0	0.5	0.5	0.5	0
0	0.5	0.5										
0.5	0	0.5										
0.5	0.5	0										
\mathbf{u}	(0.8,0.1,0.1)	(0.5,0.45,0.05)	(0.5,0.45,0.05)									
\mathbf{v}	(0.1,0.1,0.8)	(0.1,0.45,0.45)	(0.1,0.45,0.45)									
λ	0.2	0.2	0.2									
ϵ	0.06	0.06	0.1									

TABLE 2

Parameter values for the three types of graphs used in the experiments.

tween subgraphs and the probability ϵ of a connection between different subgraphs is also assumed to be the same for all couples of subgraphs. Similarly, the vector \mathbf{u} controls the probability of each edge type between nodes of a same cluster whereas \mathbf{v} defines the edge type probabilities between nodes of different clusters. We recall that the prior probabilities of each group within each subgraph are given by the parameter $\alpha = (\alpha_1, \dots, \alpha_S)$.

Figure 2 presents an example of a network generated this way with parameters $S = 2$, $C = 3$, $K = 3$, $\alpha = \begin{bmatrix} 0.1 & 0.3 & 0.6 \\ 0.6 & 0.3 & 0.1 \end{bmatrix}$, $\lambda = 0.6$, $\epsilon = 0.06$, $\mathbf{u} = (0.8, 0.1, 0.1)$ and $\mathbf{v} = (0.1, 0.3, 0.6)$. This RSM network is made of 30 nodes with $S = 2$ subgraphs (indicated by the node forms), $C = 3$ types of edges (indicated by the edge colors) and $K = 3$ clusters that have to be identified in practice (indicated by the node colors).

In order to illustrate, on various situations, that RSM is a relevant model and that its corresponding inference procedure provides an accurate estimation of the true clustering structure, we rely in the following paragraphs on three types of graphs, described in Table 3.1. The three scenarii considered correspond to different situations ranging from a almost classical setup to a more specific one. The first scenario considers networks with no subgraphs ($S = 1$) and with a preponderant proportion of edges of type 1 ($U_1 = 0.8$) and 3 ($U_3 = 0.8$). The second scenario still considers networks with no subgraphs ($S = 1$) but with balanced proportions of edge types. Finally, the third scenario considers networks with several subgraphs ($S = 3$) and balanced proportions for edge types. Therefore, the latter case should be the more complex situation to fit.

The VBEM algorithm with multiple initializations, presented in Section 2,

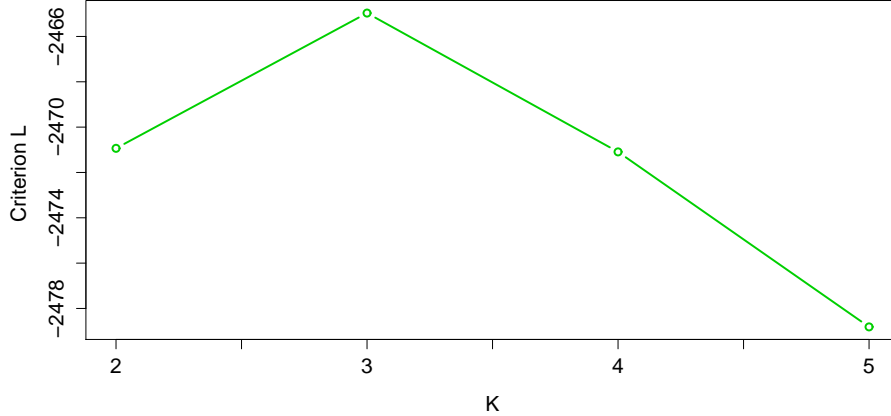


FIG 3. Criterion values $\mathcal{L}(q)$ vs. number K of groups for a graph simulated according to scenario 1.

is used in the following experiments. For a given value of K , the result with the best value for $\mathcal{L}(q)$ is chosen among the multiple initializations. Then, a clustering partition is deduced from the posterior probabilities τ_{ik} using the maximum *a posteriori* (MAP) rule, *i.e.* a node is assigned to the group with the highest posterior probability. The resulting partition is finally compared with the actual partition using the Adjusted Rand Index (ARI) (Rand, 1971), which serves as a widely accepted criterion for the difficult task of clustering evaluation.

3.2. Choice of K and inference results. In this first simulation study, we aim at evaluating the ability of the lower bound $\mathcal{L}(q)$ to serve as a criterion for selecting the appropriate number K of clusters. To this end, the VBEM algorithm for the RSM model was first run on a graph simulated according to scenario 1 for several values of K . The highest criterion value among the different initializations obtained for each value of K are presented in Figure 3. The figure indicates that $K = 3$ seems to be the appropriate number of groups for the studied network, which is the actual number of group.

We then replicated this experiment over 50 networks, still simulated according to scenario 1, for both verifying the consistency of $\mathcal{L}(q)$ and studying the clustering ability of our approach. Figure 4 shows the repartition of the criterion values (left panel) as well as the associated ARI values (right panel).

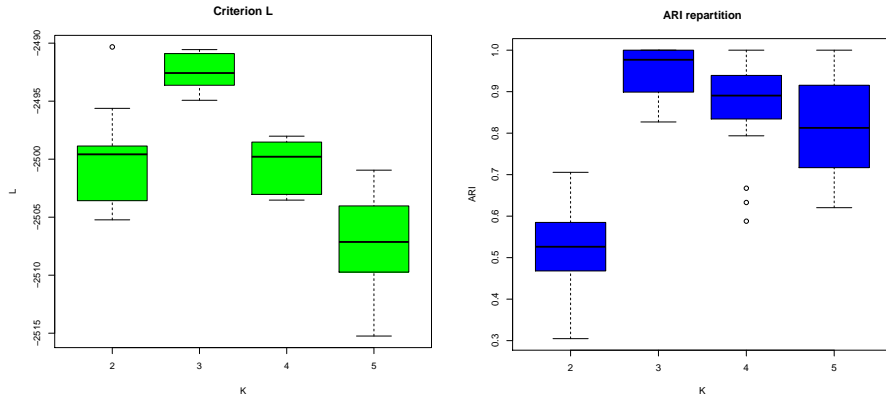


FIG 4. *Repartition of the criterion (left panel) and ARI (right panel) over 50 networks generated with the parameters of the first scenario.*

These results confirm that the lower bound $\mathcal{L}(q)$ is a valid criterion for selecting the number of groups. One can also observe that the partition resulting from our VBEM algorithm has, for the selected number of groups, a good adequation with the actual partition of the data.

3.3. Comparison with the stochastic block model. Our second set of experiments compares the performance of RSM to that of other models on data drawn according to its generative process. We were interested in the comparison with the following models:

- *binary SBM (presence)*: We fit a binary SBM using the R package mixer (Ambroise et al., 2010) on the data by considering only the presence of the edges and not the type of the edges.
- *binary SBM (type 1, 2 or 3)*: We fit a binary SBM, still using the Mixer package, on the graphs defined by taking only the edges of one type.
- *typed SBM*: We consider here a SBM with discrete edges. Although SBM was originally proposed in Nowicki and Snijders (2001) with discrete edges, existing softwares only propose to fit a SBM on binary networks. We therefore had to implement a version of the SBM which supports typed edges. Note that, in this case, the types of edges are in $\{0, \dots, C\}$, where 0 corresponds to the absence of a relation.
- *RSM*: We run the VBEM algorithm, that we proposed in Section 2 for the inference of the RSM model, with the available subgraph partition and with 5 random initializations for each run.

Method	Scenario 1	Scenario 2	Scenario 3
binary SBM (presence)	0.001 \pm 0.012	0.001 \pm 0.013	0.239 \pm 0.061
binary SBM (type 1)	0.976 \pm 0.071	0.494 \pm 0.233	-0.372 \pm 0.262
binary SBM (type 2)	0.001 \pm 0.006	-0.003 \pm 0.006	0.179 \pm 0.097
binary SBM (type 3)	0.959 \pm 0.121	0.519 \pm 0.219	0.367 \pm 0.244
Typed SBM	0.694 \pm 0.232	0.472 \pm 0.339	0.360 \pm 0.162
RSM	1.000 \pm 0.000	0.918 \pm 0.154	0.820 \pm 0.155

TABLE 3

Average ARI values and standard deviations for binary SBM, typed SBM and RSM according to the three simulation scenarii. The results are averaged on 25 simulated graphs for each scenario.

Table 3 presents the average ARI values and standard deviations on 50 simulated graphs for each scenario and with binary SBM, typed SBM and RSM. We point out that the inference is done with the actual number of clusters and this for each method. One can observe that, for the first scenario, the binary SBM based on the link presences and the type 2 SBM always fail whereas type 1, type 3 and typed SBM work pretty well. Those behaviors can be explained by the nature of scenario 1 which is a rather easy situation with no subgraphs and a predominant presence of type 1 and type 2 links. However, we can remark that it seems easier in this case to fit a binary SBM on type 1 or type 2 edges than to fit a typed SBM. This is due to the high discriminative power of type 1 and type 2 edges in this specific scenario. Let us also remark that RSM works perfectly here even though the network does not contain any subgraphs.

Regarding scenario 2, which considers a situation where there is still no subgraphs but with more balanced proportions of the different edge types, one can first notice that binary SBM and type 2 SBM fail once again. The type 1 and type 3 SBM have now a behavior closer to the one of typed SBM whereas RSM gives very accurate results once again. Finally, scenario 3 considers a RSM-type network, *i.e.* with several subgraphs, and all SBM-based algorithms are significantly outperformed by RSM which succeeds in exploiting both the information carried by different edge types and by the different subgraphs. To summarize, the RSM model and its associated VBEM algorithm turn out to be effective on situations ranging from classical setups without subgraphs to complex scenarii with subgraphs and typed edges.

4. Ecclesiastical network. This section now focuses on applying the RSM model to the ecclesiastical network, that we briefly described in the introduction and that initially motivated this work, and on analyzing its

results from the historical point of view.

4.1. *Description of the data.* Having shown that our variational Bayes EM algorithm can be used to infer the hidden clusters of data generated under the assumptions of the RSM model, we now apply it to analyze relational data built around the ecclesiastical councils that took place in Merovingian Gaul during the 6th century. Since 511, Kings as well as bishops could call for a council, which means that some political, judiciary or legal issues were addressed at these occasions, and that laics (kings, dukes or counts for instance) would attend. During the 6th century, 46 councils took place in Gaul, which was divided into the kingdoms of Neustria, Austrasia and Burgundy, and the provinces of Aquitaine and Provence (the latter changed hands several times during this period). Although there were mostly local or regional councils, attended by individuals from a specific ecclesiastical province, there were some national councils convened under the authority of a king.

We know the composition of some of these councils thanks to the acts written at the end of the meeting, and which were signed by all attending members. For others, our sources only mention the presence of “bishops, abbots and clergymen”. For this reason, we had to resort to other types of sources to build our database. In addition to the council acts, we used narrative texts (among which the famous *Ten History Books* by Gregory of Tours), hagiographies or letters. Let us notice that a “source effect” is expected due to the possible overrepresentation of a place (Neustria by Gregory of Tours or Austrasia by Fredegar) or one individual (in letters or hagiographies). The database, that took over 18 months to build, contains 1331 individuals who held one or several offices in Gaul between the years 480 and 614, and who we know to have been related or to have met during their lifetime. However, the poverty of the sources only allowed for an approximate characterization of the relations between individuals: they are either positive, negative, variable or neutral (when the type was unknown).

We expect the statistical analysis of this network to help us understand how the behaviour of an individual can be modeled through their belonging to a group. In History, this kind of approach is more common to modernists or contemporarists than to medievalists who rarely have access to this kind of data.

4.2. *Results.* The RSM model was run on the network defined by these relations, where the subgraphs are the provinces in which the individuals lived (Aquitaine, Austrasia, Burgundy, Neustria, Provence or Unknown), and the use of the criterion $\mathcal{L}(q)$ allowed us to find 6 clusters. Figure 5 presents the repartition of the different functions in the found clusters.

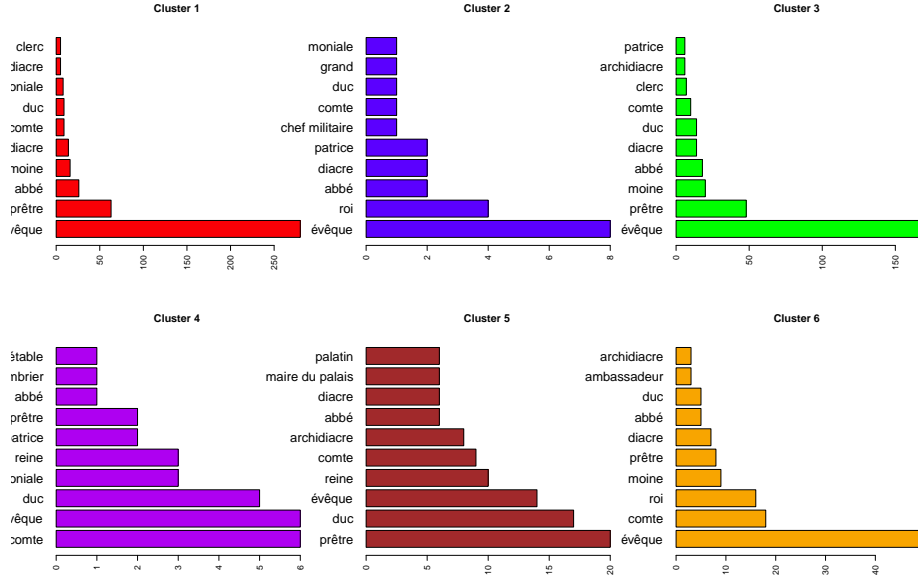


FIG 5. *Repartition of the different functions in the found clusters (restricted to the 10 most frequent functions per cluster).*

In view of these results, we can first analyze the composition of the different clusters found by the algorithm. First, clusters 1 and 3 appear to be made of the people who would attend local assemblies, provincial or diocesan councils. The council of Arles which took place in 554, would have had the same kind of composition as cluster 3, while that of Auxerre, in 585 could well represent cluster 1. Second, clusters 4 and 5 are more characteristic of aristocratic assemblies, such as the council of Orange in 529. Third, clusters 2 and 6 have the same compositions as councils concerned with more political issues (those usually convened by a king). Such a council took place in Orleans in 511. Let us however notice that cluster 2 is composed of very few individuals, which might hurt the relevance of its interpretation. Also, we might be able to further our understanding of the composition of these clusters by taking into account the similarity of certain functions (such as “comte” and “grand”).

The relations between the different clusters, shown in Figure 6, inform us further. Though the limitations expressed above about the roughness of the types still apply, they nevertheless provide us with interesting elements to confirm the coherence of the proposed model. First, it is natural that we should find “neutral” relations at the local level, between clusters 1, 3 and 6.

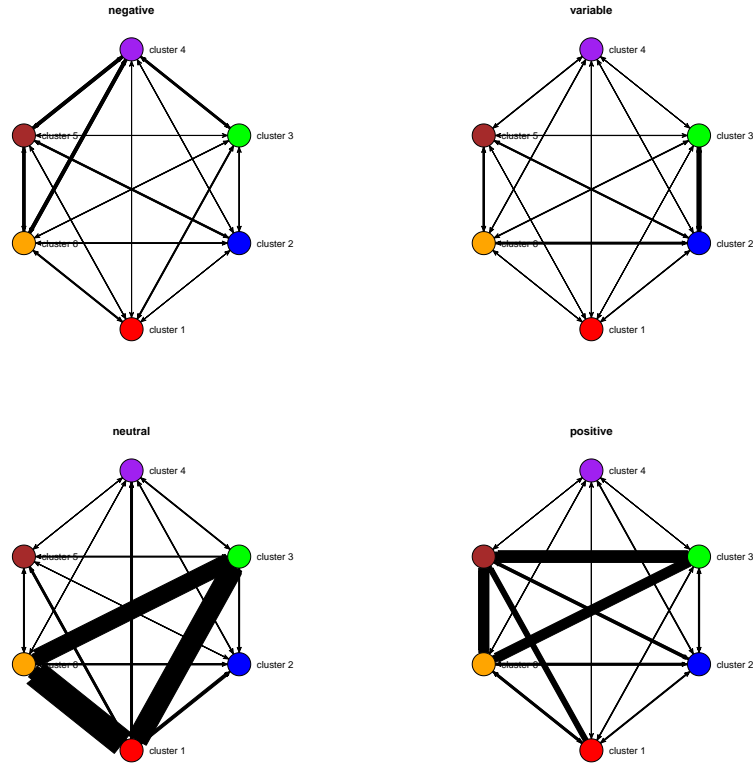


FIG 6. *Relations between clusters, for each relation type: negative, variable, neutral and positive.*

Indeed, local assemblies were the most documented ones in our sources. On the other hand, the links between high level individuals are better known, because councils used to settle conflicts between aristocrats, which explains the presence of “negative” and “variable” relations. Finally, the positive relations between cluster 3, 5 and 6 could represent the personal friendships documented by collection of letters between bishops.

After having described the political background represented by each of the clusters, we can compare the organization of the different regions. Figure 7 presents the cluster repartition in the different provinces. One can observe that the clergy and noblemen of the different regions were concerned with very different issues: Provence and Burgundy were more concerned with local questions (clusters 1 and 3), and less with political ones (clusters 2 and 6). The clusters concerned both with local (clusters 1 and 3) and high level (clus-

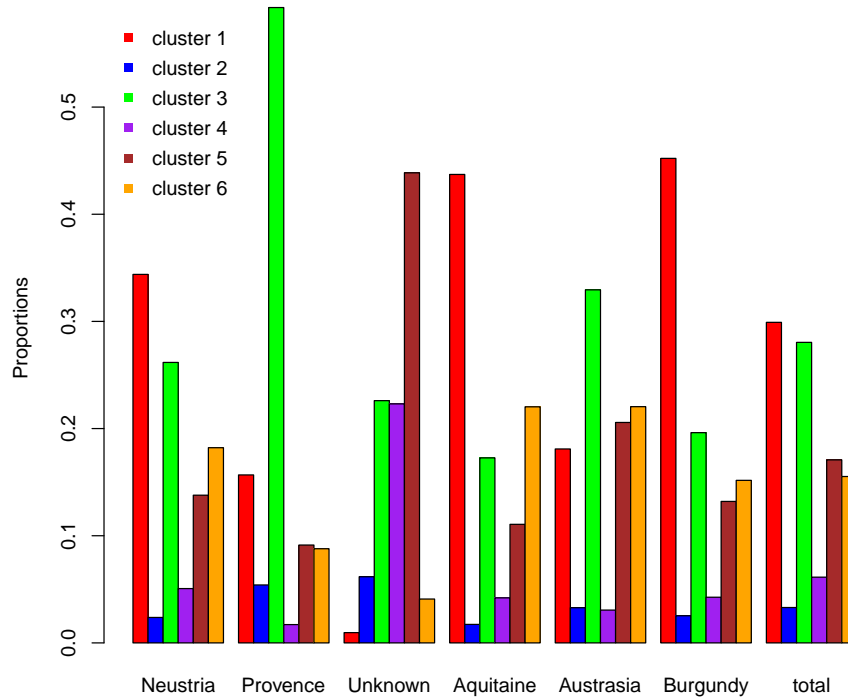


FIG 7. *Repartition of the 6 found clusters in the different Provinces.*

ter 6) questions are represented in Aquitaine. Conversely, all levels of power are represented in Neustria. This could be the result of a “source effect”, as mentioned above. Let us also notice that council structures seem similar in Austrasia and Aquitaine: sovereigns (kings and queens) are involved in Church, and convene councils in order to discuss political questions.

Some of these observations are confirmed by the estimate of parameter γ which is given in a log scale by Figure 8. First, it shows a greater frequency of relations between Aquitaine and Neustria, which comes both from a geographical and political proximity (Aquitaine is absorbed into Clovis’ kingdom in 507, then divided and absorbed by Neustria in 511). One can also see there another example of “source” effect, as our main source, Gregory of Tours, was bishop in Neustria and raised in Aquitaine (next to his uncle, the bishop of Clermont), which gave him a good knowledge of both provinces.

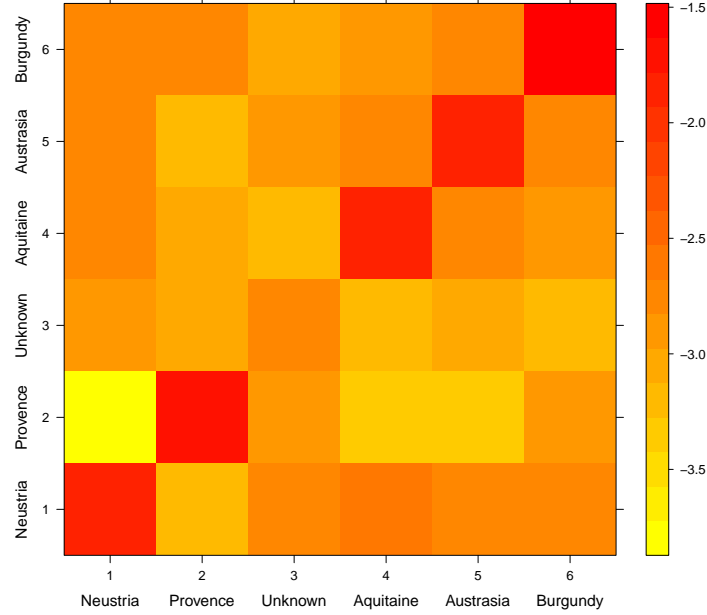


FIG 8. *Estimated values for the parameter γ (in log scale).*

More enlightening, is the relative disconnection of Burgundy and Provence, especially in regard to the provinces of Austrasia and Neustria, both heavily connected.

4.3. *Conclusion from the historical point of view.* A first analysis of the results of the RSM model confirms two well known general facts. Indeed, our results confirm the preponderance of local assemblies in 6th century Gaul and the “source effect”. Nevertheless, further analysis of the found clusters and their relations yields a better understanding of the period. In particular, the composition of the found clusters reflect different archetypes of councils, and different levels of political concerns. Our results have also highlighted that the type of concerns of each province are closely related to the frequency of their communications with others. However, two limitations to these results remain. First, we are limited by the scarcity of the historical documentation. It would be interesting to see whether the use of more precise types of relations (ecclesiastical or secular, through which media, ...) could improve the results. Second, it would also be interesting for the model to take into account temporal evolutions of the relations and clusters.

5. Conclusion and further work. In this work, we proposed a new stochastic graph model, the random subgraph model, to deal with networks where a vertex behavior is influenced by an observed partition variable. We derived a variational Bayes EM algorithm to infer the model parameters from data and applied it to an ecclesiastical network from Merovingian Gaul. The results of the fitted RSM enlightened us on the different levels of power present at this time in Gaul, and on the different power structures of different regions. Let us highlight that the RSM model allows in addition the comparison of subgraphs through the model parameters, in particular the cluster proportions. One aspect, however, that RSM does not currently address is the temporality of the data. Since this aspect can be found in many of the data sets we wish to apply the RSM model to, we believe that a natural continuation of this work would be a dynamic extension of the RSM model. Moreover, we plan to introduce a Chinese restaurant process on the latent cluster structure in order to automatically estimate the number of clusters while clustering the vertices. Finally, we would like to consider the problem of visualizing such networks with typed edges and known subgraphs.

APPENDIX A: VARIATIONAL BAYES

In this final section, we detail the computations that lead to the update rules given in Section 2, and provide an explicit expression of the criterion $\mathcal{L}(q)$.

PROPOSITION A.1. *The complete data log likelihood the RSM model is given by:*

$$\begin{aligned} \log p(\mathbf{X}, \mathbf{A}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\Pi}) &= \sum_{i \neq j}^N \sum_{c=1}^C \sum_{k,l}^K \{ \delta(X_{ij} = c) Z_{ik} Z_{jl} \log(\Pi_{klc}) \} \\ &+ \sum_{i=1}^N \sum_{k=1}^K Z_{ik} \log(\alpha_{r_i,k}) \\ &+ \sum_{i \neq j}^N \{ A_{ij} \log(\gamma_{r_i,r_j}) + (1 - A_{ij}) \log(1 - \gamma_{r_i,r_j}) \} \\ &+ \sum_{s=1}^S \log p(\boldsymbol{\alpha}_s) + \sum_{r,s}^S \log p(\gamma_{rs}) + \sum_{k,l}^K \log p(\boldsymbol{\Pi}_{kl}). \end{aligned}$$

PROOF.

$$\begin{aligned}
\log p(\mathbf{X}, \mathbf{A}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\Pi}) &= \log p(\mathbf{X} | \mathbf{A}, \mathbf{Z}, \boldsymbol{\Pi}) + \log p(\mathbf{A} | \boldsymbol{\gamma}) + \log p(\mathbf{Z} | \boldsymbol{\alpha}) + \log p(\boldsymbol{\alpha}) + \log p(\boldsymbol{\gamma}) \\
&\quad + \log p(\boldsymbol{\Pi}) \\
&= \sum_{i \neq j}^N \sum_{c=1}^C \sum_{k,l}^K \{\delta(X_{ij} = c) Z_{ik} Z_{jl} \log(\Pi_{klc})\} \\
&\quad + \sum_{i=1}^N \sum_{k=1}^K Z_{ik} \log(\alpha_{r_i, k}) \\
&\quad + \sum_{i \neq j}^N \{A_{ij} \log(\gamma_{r_i, r_j}) + (1 - A_{ij}) \log(1 - \gamma_{r_i, r_j})\} \\
&\quad + \sum_{s=1}^S \log p(\boldsymbol{\alpha}_s) + \sum_{r,s}^S \log p(\gamma_{rs}) + \sum_{k,l}^K \log p(\Pi_{kl}).
\end{aligned}$$

PROPOSITION A.2. *The VBEM update step for the distribution $q(\gamma_{rs})$ is given by:*

$$q(\gamma_{rs}) = \text{Beta}(\gamma_{rs}; a_{rs}, b_{rs}), \forall (r, s) \in \{1, \dots, S\}^2,$$

where

$$a_{rs} = a_{rs}^0 + \sum_{r_i=r, r_j=s} (A_{ij}),$$

and

$$b_{rs} = b_{rs}^0 + \sum_{r_i=r, r_j=s} (1 - A_{ij}).$$

PROOF.

$$\begin{aligned}
\log q(\gamma_{rs}) &= \mathbb{E}_{\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\gamma} \setminus r_s, \boldsymbol{\Pi}} [\log p(\mathbf{X}, \mathbf{A}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\Pi})] + \kappa \\
&= \sum_{r_i=r, r_j=s} \{A_{ij} \log(\gamma_{rs}) + (1 - A_{ij}) \log(1 - \gamma_{rs})\} \\
&\quad + \log p(\gamma_{rs}) + \kappa \\
&= \sum_{r_i=r, r_j=s} \{A_{ij} \log(\gamma_{rs}) + (1 - A_{ij}) \log(1 - \gamma_{rs})\} \\
&\quad + (a_{rs}^0 - 1) \log(\gamma_{rs}) + (b_{rs}^0 - 1) \log(1 - \gamma_{rs}) + \kappa \\
&= (a_{rs}^0 - 1 + \sum_{r_i=r, r_j=s} A_{ij}) \log(\gamma_{rs}) \\
&\quad + (b_{rs}^0 - 1 + \sum_{r_i=r, r_j=s} (1 - A_{ij})) \log(1 - \gamma_{rs}) + \kappa,
\end{aligned}$$

where κ is a constant term. Hence, the functional form of the variational approximation $q(\gamma_{rs})$ corresponds to a Beta distribution with updated hyperparameters:

$$a_{rs} = a_{rs}^0 + \sum_{r_i=r, r_j=s} (A_{ij}),$$

and

$$b_{rs} = b_{rs}^0 + \sum_{r_i=r, r_j=s} (1 - A_{ij}).$$

PROPOSITION A.3. *The VBEM update step for the distribution $q(\mathbf{Z}_i)$ is given by:*

$$q(\mathbf{Z}_i) = \mathcal{M}(\mathbf{Z}_i; \mathbf{1}, \boldsymbol{\tau}_i), \forall i,$$

where

$$\begin{aligned} \tau_{ik} &\propto \exp \left(\psi(\chi_{r_i, k}) - \psi \left(\sum_{l=1}^K \chi_{r_i, l} \right) \right) \\ &+ \exp \left\{ \sum_{j \neq i}^N \sum_{c=1}^C \sum_{l=1}^K \delta(X_{ij} = c) \tau_{jl} \left(\psi(\Xi_{klc}) - \psi \left(\sum_{u=1}^C \Xi_{klu} \right) \right) \right\} \\ &+ \exp \left\{ \sum_{j \neq i}^N \sum_{c=1}^C \sum_{l=1}^K \delta(X_{ji} = c) \tau_{jl} \left(\psi(\Xi_{lkc}) - \psi \left(\sum_{u=1}^C \Xi_{lku} \right) \right) \right\}. \end{aligned}$$

PROOF.

$$\begin{aligned} \log q(\mathbf{Z}_i) &= \mathbb{E}_{\mathbf{Z} \setminus i, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\Pi}} [\log p(\mathbf{X}, \mathbf{A}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\Pi})] + \kappa \\ &= \mathbb{E}_{\mathbf{Z} \setminus i, \boldsymbol{\Pi}} \left[\sum_{j=1}^N \sum_{c=1}^C \left\{ \delta(X_{ij} = c) \sum_{k,l}^K Z_{ik} Z_{jl} \log(\Pi_{klc}) \right\} \right] \\ &+ \mathbb{E}_{\mathbf{Z} \setminus i, \boldsymbol{\Pi}} \left[\sum_{j=1}^N \sum_{c=1}^C \left\{ \delta(X_{ji} = c) \sum_{k,l}^K Z_{jk} Z_{il} \log(\Pi_{klc}) \right\} \right] \\ &+ \mathbb{E}_{\mathbf{Z} \setminus i, \boldsymbol{\alpha}} \left[\sum_{k=1}^K Z_{ik} \log(\alpha_{r_i, k}) \right] + \kappa \\ &= \sum_{k=1}^K Z_{ik} \mathbb{E}_{\boldsymbol{\alpha}} [\log(\alpha_{r_i, k})] \end{aligned}$$

$$\begin{aligned}
& + \sum_{j=1}^N \sum_{c=1}^C \sum_{l,k}^K Z_{ik} \delta(X_{ij} = c) \mathbb{E}_{Z \setminus i, \mathbf{\Pi}} [Z_{jl} \log(\Pi_{klc})] \\
& + \sum_{j=1}^N \sum_{c=1}^C \sum_{l,k}^K Z_{ik} \delta(X_{ji} = c) \mathbb{E}_{Z \setminus i, \mathbf{\Pi}} [Z_{jl} \log(\Pi_{lkc})] + \kappa \\
& = \sum_{k=1}^K Z_{ik} \left(\psi(\chi_{r_i,k}) - \psi\left(\sum_{l=1}^K \chi_{r_i,l}\right) \right) \\
& + \sum_{k=1}^K Z_{ik} \left\{ \sum_{j \neq i}^N \sum_{c=1}^C \sum_{l=1}^K \delta(X_{ij} = c) \tau_{jl} \left(\psi(\Xi_{klc}) - \psi\left(\sum_{u=1}^C \Xi_{klu}\right) \right) \right\} \\
& + \sum_{k=1}^K Z_{ik} \left\{ \sum_{j \neq i}^N \sum_{c=1}^C \sum_{l=1}^K \delta(X_{ji} = c) \tau_{jl} \left(\psi(\Xi_{lkc}) - \psi\left(\sum_{u=1}^C \Xi_{lku}\right) \right) \right\} + \kappa,
\end{aligned}$$

where κ is a constant term. Hence, the functional form of the variational approximation $q(\mathbf{Z}_i)$ corresponds to a multinomial distribution, with updated parameters:

$$\begin{aligned}
\tau_{ik} & \propto \exp \left(\psi(\chi_{r_i,k}) - \psi\left(\sum_{l=1}^K \chi_{r_i,l}\right) \right) \\
& + \exp \left\{ \sum_{j \neq i}^N \sum_{c=1}^C \sum_{l=1}^K \delta(X_{ij} = c) \tau_{jl} \left(\psi(\Xi_{klc}) - \psi\left(\sum_{u=1}^C \Xi_{klu}\right) \right) \right\} \\
& + \exp \left\{ \sum_{j \neq i}^N \sum_{c=1}^C \sum_{l=1}^K \delta(X_{ji} = c) \tau_{jl} \left(\psi(\Xi_{lkc}) - \psi\left(\sum_{u=1}^C \Xi_{lku}\right) \right) \right\}.
\end{aligned}$$

PROPOSITION A.4. *The VBEM update step for the distribution $q(\boldsymbol{\alpha}_s)$ is given by:*

$$q(\boldsymbol{\alpha}_s) = \text{Dir}(\boldsymbol{\alpha}_s; \chi_s), \forall s \in \{1, \dots, S\},$$

where

$$\chi_{sk} = \chi_{sk}^0 + \sum_{i=1}^N \delta(r_i = s) \tau_{ik}, \forall k \in \{1, \dots, K\}.$$

PROOF.

$$\begin{aligned}
\log q(\boldsymbol{\alpha}_s) &= \mathbb{E}_{\mathbf{Z}, \boldsymbol{\alpha} \setminus s, \boldsymbol{\gamma}, \boldsymbol{\Pi}}[\log p(\mathbf{X}, \mathbf{A}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\Pi})] + \kappa \\
&= \mathbb{E}_{\mathbf{Z}, \boldsymbol{\alpha} \setminus s} \left[\sum_{i=1}^N \sum_{k=1}^K Z_{ik} \log(\alpha_{r_i, k}) \right] + \log p(\boldsymbol{\alpha}_s) + \kappa \\
&= \sum_{k=1}^K \sum_{i=1}^N \delta(r_i = s) \log(\alpha_{sk}) \mathbb{E}_{\mathbf{Z}}[Z_{ik}] + \sum_{k=1}^K \log(\alpha_{sk})(\chi_{sk}^0 - 1) + \kappa \\
&= \sum_{k=1}^K \log(\alpha_{sk}) \left\{ \chi_{sk}^0 - 1 + \sum_{i=1}^n \delta(r_i = s) \tau_{ik} \right\} + \kappa,
\end{aligned}$$

where κ is a constant term. Hence, the functional form of the variational approximation $q(\boldsymbol{\alpha}_s)$ corresponds to a Dirichlet distribution with updated hyperparameters:

$$\chi_{sk} = \chi_{sk}^0 + \sum_{i=1}^N \delta(r_i = s) \tau_{ik}, \forall k \in \{1, \dots, K\}.$$

PROPOSITION A.5. *The VBEM update step for the distribution $q(\boldsymbol{\Pi}_{kl})$ is given by:*

$$q(\boldsymbol{\Pi}_{kl}) = \text{Dir}(\boldsymbol{\Pi}_{kl}; \Xi_{kl}), \forall (k, l) \in \{1, \dots, K\}^2,$$

where

$$\Xi_{klc} = \Xi_{klc}^0 + \sum_{i \neq j}^N \delta(X_{ij} = c) \tau_{ik} \tau_{jl}, \forall c \in \{1, \dots, C\}.$$

PROOF.

$$\begin{aligned}
\log q(\boldsymbol{\Pi}_{k,l}) &= \mathbb{E}_{\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\Pi} \setminus kl}[\log p(\mathbf{X}, \mathbf{A}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\Pi})] + \kappa \\
&= \mathbb{E}_{\mathbf{Z}, \boldsymbol{\Pi} \setminus kl} \left[\sum_{i \neq j}^N \sum_{c=1}^C \delta(X_{ij} = c) Z_{ik} Z_{jl} \log(\Pi_{klc}) \right] + \log p(\boldsymbol{\Pi}_{kl}) + \kappa \\
&= \sum_{c=1}^C \log(\Pi_{klc}) \left\{ \sum_{i \neq j}^N \delta(X_{ij} = c) \tau_{ik} \tau_{jl} \right\} + \sum_{c=1}^C \log(\boldsymbol{\Pi}_{klc})(\Xi_{klc}^0 - 1) + \kappa \\
&= \sum_{c=1}^C \log(\Pi_{klc}) \left\{ \Xi_{klc}^0 - 1 + \sum_{i \neq j}^N \delta(X_{ij} = c) \tau_{ik} \tau_{jl} \right\} + \kappa,
\end{aligned}$$

where κ is a constant term. Hence, the functional form of the variational approximation $q(\mathbf{\Pi}_{kl})$ corresponds to a Dirichlet distribution with updated hyperparameters:

$$\Xi_{klc} = \Xi_{klc}^0 + \sum_{i \neq j}^N \delta(X_{ij} = c) \tau_{ik} \tau_{jl}, \forall c \in \{1, \dots, C\}.$$

PROPOSITION A.6. *When computed right after the M step, the lower bound of the marginal likelihood is given by:*

$$\mathcal{L}(q) = \sum_{r,s}^S \log\left(\frac{B(a_{rs}, b_{rs})}{B(a_{rs}^0, b_{rs}^0)}\right) + \sum_{s=1}^S \log\left(\frac{C(\boldsymbol{\chi}_s)}{C(\boldsymbol{\chi}_s^0)}\right) + \sum_{k,l}^K \log\left(\frac{C(\boldsymbol{\Xi}_{kl})}{C(\boldsymbol{\Xi}_{kl}^0)}\right) - \sum_{i=1}^N \sum_{k=1}^K \tau_{ik} \log(\tau_{ik}),$$

where $C(x) = \frac{\prod_{d=1}^D \Gamma(x_d)}{\Gamma(\sum_{d=1}^D x_d)}$ if $x \in \mathbb{R}^D$ and $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, $\forall (a, b) \in \mathbb{R}^2$.

PROOF. The lower bound is given by:

$$\mathcal{L}(q) = \mathbb{E}_{\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \mathbf{\Pi}} \left[\log\left(\frac{p(\mathbf{X}, \mathbf{A}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \mathbf{\Pi})}{q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \mathbf{\Pi})}\right) \right],$$

where

$$\begin{aligned} \log\left(\frac{p(\mathbf{X}, \mathbf{A}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \mathbf{\Pi})}{q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \mathbf{\Pi})}\right) &= \sum_{i \neq j}^N \{A_{ij} \log(\gamma_{r_i, r_j}) + (1 - A_{ij}) \log(1 - \gamma_{r_i, r_j})\} \\ &+ \sum_{i \neq j}^N \sum_{c=1}^C \sum_{k,l}^K \{\delta(X_{ij} = c) Z_{ik} Z_{jl} \log(\Pi_{klc})\} \\ &+ \log\left(\frac{p(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \mathbf{\Pi})}{q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \mathbf{\Pi})}\right), \end{aligned}$$

and

$$\begin{aligned} \log\left(\frac{p(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \mathbf{\Pi})}{q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \mathbf{\Pi})}\right) &= \log\left(\frac{p(\mathbf{Z} | \boldsymbol{\alpha})}{q(\mathbf{Z})}\right) + \log\left(\frac{p(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \mathbf{\Pi})}{q(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \mathbf{\Pi})}\right) \\ &= \sum_{i=1}^N \sum_{k=1}^K Z_{ik} \log\left(\frac{\alpha_{r_i, k}}{\tau_{ik}}\right) + \sum_{r,s}^S \log\left(\frac{\text{Beta}(\gamma_{rs}; a_{rs}^0, b_{rs}^0)}{\text{Beta}(\gamma_{rs}; a_{rs}, b_{rs})}\right) \\ &+ \sum_{s=1}^S \log\left(\frac{\text{Dir}(\boldsymbol{\alpha}_s; \boldsymbol{\chi}_s^0)}{\text{Dir}(\boldsymbol{\alpha}_s; \boldsymbol{\chi}_s)}\right) + \sum_{k,l}^K \log\left(\frac{\text{Dir}(\mathbf{\Pi}_{kl}; \boldsymbol{\Xi}_{kl}^0)}{\text{Dir}(\mathbf{\Pi}_{kl}; \boldsymbol{\Xi}_{kl})}\right) \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^N \sum_{k=1}^K Z_{ik} \log\left(\frac{\alpha_{r_i,k}}{\tau_{ik}}\right) + \sum_{r,s}^S \log\left(\frac{B(a_{rs}, b_{rs})}{B(a_{rs}^0, b_{rs}^0) \gamma_{rs}^{a_{rs}-a_{rs}^0} (1-\gamma_{rs})^{b_{rs}-b_{rs}^0}}\right) \\
&+ \sum_{s=1}^S \log\left(\frac{C(\boldsymbol{\chi}_s)}{C(\boldsymbol{\chi}_s^0) \prod_{k=1}^K \alpha_k^{\chi_{sk}-\chi_{sk}^0}}\right) + \sum_{k,l}^K \log\left(\frac{C(\boldsymbol{\Xi}_{kl})}{C(\boldsymbol{\Xi}_{kl}^0) \prod_{c=1}^C \Pi_{klc}^{\Xi_{klc}-\Xi_{klc}^0}}\right).
\end{aligned}$$

If $x \in \mathbb{R}^D$ then $C(x) = \frac{\prod_{d=1}^D \Gamma(x_d)}{\Gamma(\sum_{d=1}^D x_d)}$ where $\Gamma(\cdot)$ is the gamma function. Moreover, if $(a, b) \in \mathbb{R}^2$ then $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. Finally

$$\begin{aligned}
\log\left(\frac{p(\mathbf{X}, \mathbf{A}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\Pi})}{q(\mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\Pi})}\right) &= \sum_{r,s}^S \left\{ \left(a_{rs}^0 - a_{rs} + \sum_{r_i=r, r_j=s} A_{ij} \right) \log(\gamma_{rs}) \right\} \\
&+ \sum_{r,s}^S \left\{ \left(b_{rs}^0 - b_{rs} + \sum_{r_i=r, r_j=s} (1 - A_{ij}) \right) \log(1 - \gamma_{rs}) \right\} \\
&+ \sum_{s=1}^S \sum_{k=1}^K \left\{ \left(\chi_{sk}^0 - \chi_{sk} + \sum_{i=1}^N \delta(r_i = s) Z_{ik} \right) \log(\alpha_{sk}) \right\} \\
&+ \sum_{k,l}^K \sum_{c=1}^C \left\{ \left(\Xi_{klc}^0 - \Xi_{klc} + \sum_{i \neq j} \delta(X_{ij} = c) Z_{ik} Z_{jl} \right) \log(\Pi_{klc}) \right\} \\
&- \sum_{i=1}^N \sum_{k=1}^K Z_{ik} \log(\tau_{ik}) + \sum_{r,s}^S \log\left(\frac{B(a_{rs}, b_{rs})}{B(a_{rs}^0, b_{rs}^0)}\right) \\
&+ \sum_{s=1}^S \log\left(\frac{C(\boldsymbol{\chi}_s)}{C(\boldsymbol{\chi}_s^0)}\right) + \sum_{k,l}^K \log\left(\frac{C(\boldsymbol{\Xi}_{kl})}{C(\boldsymbol{\Xi}_{kl}^0)}\right).
\end{aligned}$$

Therefore, that if $\mathcal{L}(q)$ is computed right after the M step:

$$\mathcal{L}(q) = \sum_{r,s}^S \log\left(\frac{B(a_{rs}, b_{rs})}{B(a_{rs}^0, b_{rs}^0)}\right) + \sum_{s=1}^S \log\left(\frac{C(\boldsymbol{\chi}_s)}{C(\boldsymbol{\chi}_s^0)}\right) + \sum_{k,l}^K \log\left(\frac{C(\boldsymbol{\Xi}_{kl})}{C(\boldsymbol{\Xi}_{kl}^0)}\right) - \sum_{i=1}^N \sum_{k=1}^K \tau_{ik} \log(\tau_{ik}).$$

REFERENCES

- AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. *The Journal of Machine Learning Research* **9** 1981–2014.
- ALBERT, R. and BARABÁSI, A. L. (2002). Statistical mechanics of complex networks. *Modern Physics* **74** 47–97.
- AMBROISE, C., GRASSEAU, G., HOEBEKE, M., LATOUCHE, P., MIELE, V. and PICARD, F. (2010). The mixer R package. <http://cran.r-project.org/web/packages/mixer/>.
- BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences* **106** 21068–21073.
- BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics and Data Analysis* **41** 561–575.
- BILMES, J. A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute* **4** 126.
- BISHOP, C. M. (2006). *Pattern recognition and machine learning*. Springer-Verlag.
- DAUDIN, J. J., PICARD, F. and ROBIN, S. (2008). A mixture model for random graphs. *Statistics and Computing* **18** 173–183.
- FIENBERG, S. E. and WASSERMAN, S. S. (1981). Categorical data analysis of single sociometric relations. *Sociological Methodology* **12** 156–192.
- FRANK, O. and HARARY, F. (1982). Cluster inference by using transitivity indices in empirical graphs. *Journal of the American Statistical Association* 835–840.
- GIRVAN, M. and NEWMAN, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* **99** 7821.
- GOLDENBERG, A., ZHENG, A. X. and FIENBERG, S. E. (2010). *A survey of statistical network models*. Now Publishers.
- HANDCOCK, M. S., RAFTERY, A. E. and TANTRUM, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **170** 301–354.
- HOFMAN, J. M. and WIGGINS, C. H. (2008). Bayesian approach to network modularity. *Physical review letters* **100** 258701.
- JEFFREYS, H. (1946). An invariant form for the prior probability in estimations problems. In *Proceedings of the Royal Society of London. Series A* **186** 453–461.
- KEMP, C., TENENBAUM, J. B., GRIFFITHS, T. L., YAMADA, T. and UEDA, N. (2006). Learning systems of concepts with an infinite relational model. In *Proceedings of the National Conference on Artificial Intelligence* **21** 381.
- LATOUCHE, P., BIRMELÉ, E. and AMBROISE, C. (2009). *Advances in Data Analysis Data Handling and Business Intelligence* Bayesian methods for graph clustering 229–239. Springer.
- LATOUCHE, P., BIRMELÉ, E. and AMBROISE, C. (2011). Overlapping stochastic block models with application to the French political blogosphere. *Annals of Applied Statistics* **5** 309–336.
- LATOUCHE, P., BIRMELÉ, E. and AMBROISE, C. (2012). Variational Bayesian inference and complexity control for stochastic block models. *Statistical Modelling* **12** 93–115.
- MILÓ, R., SHEN-ORR, S., ITZKOVITZ, S., KASHTAN, D., CHKLOVSKII, D. and ALON, U. (2002). Network motifs: simple building blocks of complex networks. *Science* **298** 824–827.
- MORENO, J. L. (1934). *Who shall survive?: A new approach to the problem of human*

- interrelations*. Nervous and Mental Disease Publishing Co.
- NOWICKI, K. and SNIJDERS, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association* **96** 1077–1087.
- PALLA, G., DERENYI, I., FARKAS, I. and VICSEK, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435** 814–818.
- RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association* 846–850.
- SNIJDERS, T. A. B. and NOWICKI, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification* **14** 75–100.
- SOUFIANI, H. A. and AIROLDI, E. M. (2012). Graphlet decomposition of a weighted network. *Arxiv preprint arXiv:1203.2821*.
- VILLA, N., ROSSI, F. and TRUONG, Q. D. (2008). Mining a medieval social network by kernel SOM and related methods. *Arxiv preprint arXiv:0805.1374*.
- WHITE, H. C., BOORMAN, S. A. and BREIGER, R. L. (1976). Social structure from multiple networks. I. Blockmodels of roles and positions. *American Journal of Sociology* 730–780.
- XING, E. P., FU, W. and SONG, L. (2010). A state-space mixed membership blockmodel for dynamic network tomography. *The Annals of Applied Statistics* **4** 535–566.