Goodness of Fit of Logistic Regression Models for Random Graphs

Pierre Latouche\textsuperscript{a}, Stéphane Robin\textsuperscript{b,c}, and Sarah Ouadah\textsuperscript{b,c}

\textsuperscript{a}Laboratoire SAMM, EA 4543, Université Paris 1 Panthéon-Sorbonne, France; \textsuperscript{b}AgroParisTech, UMR 518, MIA, Paris, France; \textsuperscript{c}INRA, UMR 518, MIA, Paris, France

ABSTRACT

Logistic regression is a natural and simple tool to understand how covariates contribute to explain the topology of a binary network. Once the model is fitted, the practitioner is interested in the goodness of fit of the regression to check if the covariates are sufficient to explain the whole topology of the network and, if they are not, to analyze the residual structure. To address this problem, we introduce a generic model that combines logistic regression with a network-oriented residual term. This residual term takes the form of the graphon function of a $W$-graph. Using a variational Bayes framework, we infer the residual graphon by averaging over a series of blockwise constant functions. This approach allows us to define a generic goodness-of-fit criterion, which corresponds to the posterior probability for the residual graphon to be constant. Experiments on toy data are carried out to assess the accuracy of the procedure. Several networks from social sciences and ecology are studied to illustrate the proposed methodology. Supplementary material for this article is available online.

1. Introduction

Networks are now used in many scientific fields (Snijders and Nowicki 1997; Watts and Strogatz 1998; Nowicki and Snijders 2001; Hoff, Raftery, and Handcock 2002; Handcock, Raftery, and Tantrum 2007; Zhangi, Ambroise, and Miele 2008) from biology (Albert and Barabási 2002; Newman 2003; Barabási and Oltvai 2004; Lacroix, Fernandes, and Sagot 2006) to historical sciences (Villa, Rossi, and Truong 2008; Jernite et al. 2014) and geography (Ducruet 2013). Indeed, while being simple data structures, they are yet capable of describing complex interactions between entities of a system. A lot of effort has been put, especially in social sciences, in developing methods to characterize the heterogeneity of these networks using latent variables, covariates, or both (Hoff, Raftery, and Handcock 2002; Handcock, Raftery, and Tantrum 2007; Mariadassou, Robin, and Vacher 2010; Zhangi, Volant, and Ambroise 2010).

In this article, we are interested in the contribution of covariates to explain the topology of an observed network. To this aim, we consider standard logistic models that are a simple way to account for the possible effect of covariates, assuming edges to be independent conditionally on the covariates. Our goal is to provide the practitioners with tools to check the fit of the model and/or to analyze the residual structure. This goes along with the characterization of some residual structure present in the network that is not explained by the covariates. Our approach consists in combining logistic regression with the graphon function of a $W$-graph. This additional term plays the role of a very flexible, network-oriented residual term that can be visualized and on which a goodness-of-fit criterion can be based.

The $W$-graph can be casted among the latent-variable network models (Goldenberg et al. 2010; Matias and Robin 2014). It is characterized by a function $W$ called graphon where $W(u, v)$ is the probability for two nodes, with latent coordinates $u$ and $v$, each sampled from a uniform distribution over $[0, 1]$, to connect. As shown in Lovász and Szegedy (2006), it is the limiting adjacency matrix of the network. This result comes from graph limit theory for which Diaconis and Janson (2008) gave a proper definition using Aldous–Hoover theorem, which is an extension of de Finetti’s theorem to exchangeable arrays. Until recently, few inference techniques had been proposed to infer the graphon function of a network. The earliest reference is Kallenberg (1999). Since then, both parametric (Hoff 2008; Palla, Lovasz, and Vicsek 2010) and nonparametric (Chatterjee 2015) techniques have been developed. Graphon inference is a particularly challenging problem that has received strong attention in the last few years (Airoldi, Costa, and Chan 2013; Wolfe and Olhede 2013; Asta and Shalizi 2014; Chan and Airoldi 2014; Yang, Han, and Airoldi 2014; Chatterjee 2015). In the present article, we follow Latouche and Robin (2015) who took advantage of the fact that the well-known stochastic block model (SBM: Holland, Laskey, and Leinhardt 1983; Wang and Wong 1987; Nowicki and Snijders 2001) is a special case of $W$-graph corresponding to a blockwise constant graphon. This enables them to derive a variational Bayes EM (VBEM) procedure to estimate the graphon function as an average of SBM models with increasing number of blocks.

As mentioned above, the model we consider combines a logistic regression term with a residual graphon function.
Following the Bayesian framework of Latouche and Robin (2015), we estimate the residual graphon by averaging over a series of SBM including the one-block SBM, which corresponds to a constant residual graphon. We interpret a constant residual graphon as an absence of residual structure in the network. This approach enables us

(a) to assess the goodness of fit of the logistic regression through the posterior probability for the residual graphon to be constant and

(b) to display an estimate of the residual graphon that allows a visual inspection of the residual structure.

As the exact Bayesian inference of this new model for networks is not tractable, we make an intensive use of variational approximations to achieve the inference. Because of the combination of logistic regression and SBM, two different types of variational approximations are actually required.

In Section 2, we introduce the general model and we define the goodness-of-fit criterion. Technical issues and theoretical aspects are addressed in Section 3. Finally, toy and real datasets are analyzed in Section 4 and 5, respectively, to illustrate the proposed methodology. In the body of the article, only undirected networks are considered. The extension to directed networks (with proofs and update formulas) is derived in the supplementary materials. The proposed methodology is implemented in the R package GOFNetwork (github.com/platouche/gofNetwork) based on the R packages Rcpp (Eddelbuettel and Francois 2011; Eddelbuettel 2013), RcppArmadillo (Eddelbuettel and Sanderson 2014), doMC (Analytics and Weston 2015), lattice (Sarkar 2008), and mixer (Daudin, Picard, and Robin 2008; Zangi, Ambroise, and Miele 2008; Zangi, Volant, and Ambroise 2010; Latouche, Birmelé, and Ambroise 2012).

2. Assessing Goodness of Fit

We consider a set of individuals among which interactions are observed. The observed interaction network is encoded in the binary adjacency matrix \( Y = (Y_{ij})_{1 \leq i, j \leq n} \) where \( Y_{ij} \) is 1 if nodes \( i \) and \( j \) are connected, and 0 otherwise. We further assume that a \( d \)-dimensional vector, \( d \geq 1 \), of covariates \( x_{ij} \) is available for each pair of nodes. In the following, we denote as \( X = (x_{ij})_{1 \leq i, j \leq n} \) the set of all covariates.

2.1. Logistic Regression and Residual Structure

The influence of the covariates on the network topology can be easily accounted for using a logistic regression model. Such a model assumes that the edges \( (Y_{ij}) \) are independent (conditionally on the covariates) with respective distribution

\[
H_0 : \quad Y_{ij} \sim \mathcal{B} \left[ g \left( x_{ij}^T \beta + \alpha \right) \right],
\]

where \( \beta \in \mathbb{R}^d, \alpha \in \mathbb{R}, g \) stands for the logistic function \( g(t) = 1/(1 + \exp(-t)), t \in \mathbb{R} \). Our goal is to check if model \( H_0 \) is sufficient to explain the whole topology of the network. Note that the network structure does not explicitly appear in this model, as edges are considered as independent outcomes of a (generalized) linear model.

To assess the fit of Model \( H_0 \), we define a generic alternative network model. The alternative we consider is inspired from the \( W \)-graph model. More precisely, we consider the model

\[
H_1 : \quad Y_{ij} \sim \mathcal{B} \left[ g \left( x_{ij}^T \beta + \phi(U_i, U_j) \right) \right],
\]

where the \( (U_i)_{1 \leq i \leq n} \) are independent unobserved latent variables, with uniform distribution over the \((0, 1)\) interval. The nonconstant function \( \phi : [0, 1]^2 \to \mathbb{R} \) encodes a residual structure in the network that is not accounted for by Model \( H_0 \).

Note that, in absence of covariate, this model corresponds to a \( W \)-graph (Lovász and Szegedy 2006) with graphon function \( g \circ \phi \). Model \( H_0 \) corresponds to the case where the residual function \( \phi \) is constant.

The present article focuses on the goodness of fit of a regression model, therefore, the interpretation of the residual term \( \phi(U_i, U_j) \) is not critical but its visual inspection may help to better understand where the residual heterogeneity does come from. Note this generic form encompasses additive node effect, which, in absence of regression term, would result in a model close to the expected degree model (Chung and Lu 2002).

The inference of the function \( \phi \) in Model \( H_1 \) is not an easy task and, following Airoldi, Costa, and Chan (2013) and Latouche and Robin (2015), we consider a class of blockwise constant \( \phi \) function. More precisely, we define the Model

\[
M_K : \quad Y_{ij} \sim \mathcal{B} \left[ g \left( x_{ij}^T \beta + Z_i^T \alpha Z_j \right) \right],
\]

where \( \alpha \) is a \( K \times K \) real matrix \((K \geq 1)\) and where the \((Z_i)_{1 \leq i \leq n} \) are independent vectors with \( K \) coordinates, all zero except one. We denote \( \pi_k (1 \leq k \leq K) \) the probability that the \( k \)-th coordinate is nonzero. Briefly speaking, each vector \( Z_i \) has multinomial distribution \( \mathcal{M}(1, \pi) \) where \( \pi = (\pi_k)_{1 \leq k \leq K} \). The set of parameters of such a model is \( \theta = (\beta, \pi, \alpha) \). Note that in the absence of covariate, this model corresponds exactly to an SBM model. The ability of the stochastic block model to approximate the \( W \)-graph model was demonstrated by Airoldi, Costa, and Chan (2013) and Latouche and Robin (2015) and is not the purpose of this article.

Model \( H_0 \) is then equivalent to Model \( M_1 \) so the goodness-of-fit problem can be rephrased as the comparison between Model \( H_0 \) and \( H_1' \), where

\[
H_0 = M_1 \quad \text{and} \quad H_1' = \bigcup_{K \geq 2} M_K.
\]

2.2. Bayesian Model Comparison

Now, we are given a series of Models \( M_K (K \geq 1) \) indexed by \( K \), which characterize \( H_0 \) and \( H_1' \). In this article, we propose to compare \( H_0 \) and \( H_1' \) using a Bayesian model comparison framework.

Thus, each Model \( M_K \) is associated with a prior probability \( p(M_K) \). The parameter \( \theta \) is then drawn conditionally on \( M_K \) according to the prior distribution \( p(\theta|M_K) \). Given \( \theta, M_K \) and the given set \( X \) of covariates, the graph is finally assumed to be sampled according to Model (2). In this framework, the prior probability of Models \( H_0 \) and \( H_1' \) are

\[
p(H_0) = p(M_1) \quad \text{and} \quad p(H_1') = \sum_{K \geq 2} p(M_K).
\]
Moreover, the posterior probability of Model $M_K$ is
\[ p(M_K|Y) = \frac{p(Y|M_K)p(M_K)}{p(Y)} = \frac{p(Y|M_K)p(M_K)}{\sum_{K'\geq 1} p(Y|M_{K'})p(M_{K'})}. \] (3)

The goodness of fit of Model $H_0$ can then be assessed by computing the posterior probability of $H_0$:
\[ p(H_0|Y) = p(M_1|Y). \] (4)

The Bayes factor (Kass and Raftery 1995) between Models $H_0$ and $H_1$ can be computed in a similar way as
\[ B_{01} = \frac{p(Y|H_0)}{p(Y|H_1')}. \]
where
\[ p(Y|H_1') = \frac{1}{p(H_1')} \sum_{K>2} p(M_K)p(Y|M_K). \] (5)

### 3. Inference

The goodness-of-fit criteria introduced in the previous section all depend on marginal likelihood terms $p(Y|M_K)$, which have to be estimated from the data in practice. This is the object of this section. The prior distributions $p(M_K)$ and $p(\theta|M_K)$ are first introduced. A variational three-steps optimization scheme, based on global and local variational methods, is then derived for inference.

In the following, we focus on undirected networks and therefore both the adjacency matrix $Y$ and the matrix $X$ of covariates are symmetric: $Y_{ij} = Y_{ji}$ and $X_{ij} = X_{ji}, \forall i \neq j$. Moreover, we do not consider self-loops, that is, the connection of a node to itself and therefore the pairs $(i, i), \forall i$ are discarded from the sums and products involved. The complete derivation of the model and the inference procedure in the directed case are given as supplementary materials. The Appendix with all proofs in the undirected case is also provided as supplementary materials.

#### 3.1. Prior Distributions

With no prior information on which model should be preferred, we give equal weights $p(H_0) = p(H_1') = 1/2$ to $H_0$ and $H_1'$. Therefore, $p(M_1) = 1/2$. Alternative choices can be made by integrating expert knowledge at hand. Recall that $p(H_1') = \sum_{K>2} p(M_K)$.

For Model $M_K$, the prior distribution over the model parameters in $\theta$ is defined as a product of conjugate prior distributions over the different sets of parameters: $p(\theta|M_K) = p(\beta|M_K)p(\pi|M_K)p(\alpha|M_K)$. Since $\pi$ is involved in a multinomial distribution to sample the vectors $Z_i$, a Dirichlet prior distribution is chosen
\[ p(\pi|M_K) = \text{Dir}(\pi; \epsilon), \]
where $\epsilon$ is a vector with $K$ components such that $\epsilon_k = \epsilon_0 > 0, \forall k \in \{1, \ldots, K\}$. Note that fixing $\epsilon_0 = 1/2$ induces a non-informative Jeffreys’ prior distribution, which is known to be proper (Jeffreys 1946). It is also possible to obtain a uniform distribution over the $K - 1$ dimensional simplex by setting $\epsilon_0 = 1$.

To characterize the $d$-dimensional regression vector $\beta$, a Gaussian distribution is considered
\[ p(\beta|\eta, M_K) = \mathcal{N}(\beta; 0, \frac{I_d}{\eta}) = \prod_{j=1}^{d} \mathcal{N}(\beta_j; 0, \frac{1}{\eta}), \]
with $I_d$ the $d \times d$ identity matrix and $\eta > 0$ a parameter controlling the inverse variance. Similarly, the matrix $\alpha$ is modeled using a product of Gaussian distributions with $\gamma > 0$ controlling the variance
\[ p(\alpha|\gamma, M_K) = \prod_{k \leq l} \mathcal{N}(\alpha_{kl}; 0, \frac{1}{\gamma}). \]

Since we focus on undirected networks, $\alpha$ has to be symmetric and therefore the product involves the $k \leq l$ terms of $\alpha$. In the directed case (see supplementary materials), the product is over all terms $k, l$, and the vec operator, which stacks the columns of a matrix into a vector, is used to simplify the calculations.

Finally, Gamma distributions are considered for $\gamma$
\[ p(\gamma|M_K) = \text{Gam}(\gamma'; a_0, b_0), \quad a_0, b_0 > 0, \]
and $\eta$
\[ p(\eta|M_K) = \text{Gam}(\eta; c_0, d_0), \quad c_0, d_0 > 0. \]

By construction, Gamma distributions are informative. To limit the influence on the posterior distributions, the hyperparameters controlling the scale ($a_0, c_0$) and rate ($b_0, d_0$) are usually set to low values in the literature.

The choice of modeling the prior information on the parameters $\alpha$ and $\beta$ from such Gaussian-Gamma distributions has been widely considered both in standard Bayesian linear regression and Bayesian logistic regression (see, e.g., Bishop and Svensén 2003; Bishop 2006). The prior distributions $p(\beta|M_K)$ and $p(\alpha|M_K)$ are then obtained by marginalizing over $p(\eta|M_K)$ and $p(\gamma|M_K)$, respectively. This results in prior distributions from the class of generalized hyperbolic distributions. For more details, we refer to Caron and Doucet (2008).

In the following, and to simplify the notations, the dependency on $M_K$ is omitted in the prior and posterior distributions.

#### 3.2. Variational Approximations

Denoting $Z$ the set of all latent vectors $(Z_i)$, the marginal log-likelihood of Model $M_K$, also called the integrated observed data log-likelihood, is given by
\[
\log p(Y|M_K) = \log \left\{ \sum_Z \int p(Y|Z, \alpha, \beta)p(Z|\pi)p(\alpha|\gamma) \times p(\beta|\eta)p(\pi)p(\gamma)\,d\pi\,d\alpha\,d\beta\,d\gamma\,d\eta \right\}. \] (6)

It requires a marginalization over the prior distributions of all parameters. In particular, it involves testing all the $K^n$ configurations of $Z$. Unfortunately, (6) is not tractable and therefore we propose to rely on variational approximations for inference purposes. Let us first consider the global
To maximize the lower bound, we assume that the distribution \( q \) is equivalent to minimizing the Kullback–Leibler divergence between \( \mathbb{K} \) and the unknown posterior distribution \( p(\cdot|Y) \). \( \mathcal{L}_K(\cdot) \) is given by

\[
\mathcal{L}_K(q) = \sum_Z \int q(Z, \pi, \alpha, \beta, \gamma, \eta) \times \log \frac{p(Y, Z, \pi, \alpha, \beta, \gamma, \eta)}{q(Z, \pi, \alpha, \beta, \gamma, \eta)} d\pi d\alpha d\beta d\gamma d\eta.
\]

To maximize the lower bound, we assume that the distribution can be factorized as follows:

\[
q(Z, \pi, \alpha, \beta, \gamma, \eta) = q(\pi)q(\alpha)q(\beta)q(\gamma)q(\eta) \prod_{i=1}^{n} q(Z_i).
\]

Unfortunately, \( \mathcal{L}_K(\cdot) \) is still intractable due to the logistic function in \( p(Y|Z, \alpha, \beta) \). Following the work of Jaakkola and Jordan (2000), a tractable lower bound is derived.

**Proposition 1.** Given any \( n \times n \) positive real matrix \( \xi = (\xi_{ij})_{1 \leq i, j \leq n} \), a lower bound of the first lower bound is given by

\[
\log p(Y|M_K) \geq \mathcal{L}_K(q) \geq \mathcal{L}_K(q; \xi),
\]

where

\[
\mathcal{L}_K(q; \xi) = \sum_Z \int q(Z, \pi, \alpha, \beta, \gamma, \eta) \times \log \mathcal{L}_K(q) \times d\pi d\alpha d\beta d\gamma d\eta,
\]

and

\[
\log h(Z, \alpha, \beta, \xi) = \sum_{i \neq j} \left\{ \left( Y_{ij} - \frac{1}{2} \right) \left( Z_i^\top \alpha Z_j + x_{ij}^\top \beta \right) + \log g(\xi_{ij}) - \frac{\xi_{ij}}{2} - \lambda(\xi_{ij}) \left( \left( Z_i^\top \alpha Z_j + x_{ij}^\top \beta \right)^2 - \xi_{ij}^2 \right) \right\},
\]

with \( \xi_{ij} \in \mathbb{R}^+ \), \( \xi_{ij} = \xi_{ji} \). Moreover, \( \lambda(\xi_{ij}) = (g(\xi_{ij}) - 1/2)/(2\xi_{ij}) \), \( g \) being the logistic function.

The proof is given in Appendix A.1. The quality of the lower bound \( \mathcal{L}_K(q; \xi) \), which was obtained through a series of Taylor expansions, clearly depends on the choice of the matrix \( \xi \). As we shall see in Section 3.2.2, \( \xi \) can be estimated from the data to obtain tight bounds.

### 3.2.1. Variational Bayes EM

For now, we assume that the matrix \( \xi \) is fixed and we rely on \( \mathcal{L}_K(q; \xi) \) as a lower bound of \( \log p(Y|M_K) \). To maximize the lower bound, a VBEM algorithm (Beal and Ghahramani 2002) is applied on \( \mathcal{L}_K(q; \xi) \). This optimization scheme is iterative and is related to the EM algorithm (Dempster, Laird, and Rubin 1977). Keeping all distributions fixed except one, the bound is maximized with respect to the remaining distribution. This procedure is repeated in turn until convergence of the bound. The optimization of the distribution \( q(Z) \) over the latent variables usually refers to the variational E step. The updates of \( q(\pi), q(\alpha), q(\beta), q(\gamma), \) and \( q(\eta) \) refer here to the variational M step. Proposition 2 provides the update formula of the E step and Propositions 3 to 7 provide those of the M step. The corresponding proofs are given in Appendix A.2 to A.7.

**Proposition 2.** The variational E update step for each distribution \( q(Z) \) is given by

\[
q(Z_i) = \mathcal{M}(Z_i; 1, \tau_i),
\]

where \( \sum_{k=1}^K \tau_{ik} = 1 \) and

\[
\tau_{ik} \propto \exp \left\{ \sum_{l=1}^K (m_{dkl}) \sum_{j \neq i} \left( Y_{ij} - \frac{1}{2} \right) - 2\lambda(\xi_{ij})x_{ij}^\top m_{i} \right\} \tau_{ij} - \sum_{k=1}^K E_{ik} \left[ \alpha_i^\top \right] \sum_{j \neq i} \lambda(\xi_{ij})\tau_{ij} + \psi\left( \epsilon_i^\alpha \right) - \psi\left( \sum_{k=1}^K \epsilon_k^\alpha \right),
\]

\( \psi(\cdot) \) denotes the digamma function, which is the logarithmic derivative of the gamma function.

**Proposition 3.** The variational M update step for the distribution \( q(\pi) \) is given by

\[
q(\pi) = \text{Dir}(\pi; \epsilon^\pi),
\]

where, \( \forall k \in \{1, \ldots, K\}, \epsilon_k^\pi = \epsilon_0 + \sum_{l=1}^n \tau_{ik}, \tau_{ik} \) being given by Proposition 2.

**Proposition 4.** The variational M update step for the distribution \( q(\beta) \) is given by

\[
q(\beta) = \mathcal{N}(\beta; m_{\beta}, S_{\beta}),
\]

where

\[
S_{\beta} = S_{\beta} - \frac{1}{2} \sum_{i \neq j} \left( Y_{ij} - \frac{1}{2} - 2\lambda(\xi_{ij})\tau_i^\top m_{i} \right) x_{ij}.
\]

**Proposition 5.** The variational M update step for the distribution \( q(\gamma) \) is given by

\[
q(\gamma) = \text{Gam}(\gamma; a_n, b_n),
\]

where \( a_n = a_0 + \frac{K(K+1)}{4} \) and \( b_n = b_0 + \frac{1}{2} \sum_{k \leq j} E_{ik} \left[ \alpha_i^\top \right] \).

**Proposition 6.** The variational M update step for the distribution \( q(\eta) \) is given by

\[
q(\eta) = \text{Gam}(\eta; c_n, d_n),
\]

where \( c_n = c_0 + \frac{d}{2} \) and \( d_n = d_0 + \frac{1}{2} \text{Tr}(S_{\beta}) + \frac{1}{2} m_{\beta}^\top m_{\beta}, S_{\beta} \) and \( m_{\beta} \) being given by Proposition 4.
Proposition 7. The variational M update step for the distribution $q(\alpha)$ is given by

$$q(\alpha) = \prod_{k=1}^{K} \mathcal{N}(a_{ik}; (m_{\alpha})_{kl}, (\sigma_{\alpha}^2)_{kl}),$$

where

$$(\sigma_{\alpha}^{2})^{-1}_{kk} = a_{kn} + \sum_{i \neq j} \lambda(\xi_{ij}) \tau_{ik} \tau_{jk}, \quad \forall k,$$

$$(\sigma_{\alpha}^{2})^{-1}_{kl} = a_{kn} + \frac{2}{n} \sum_{i \neq j} \lambda(\xi_{ij}) \tau_{ik} \tau_{jl}, \quad \forall k \neq l,$$

$$(m_{\alpha})_{kk} = (\sigma_{\alpha}^2)_{kk} \sum_{i \neq j} \left( \frac{1}{2} \left( Y_{ij} - \frac{1}{2} \right) - \lambda(\xi_{ij}) \alpha_{ij} m_{\beta} \right) \tau_{ik} \tau_{jk}, \quad \forall k,$$

$$(m_{\alpha})_{kl} = (\sigma_{\alpha}^2)_{kl} \sum_{i \neq j} \left( \frac{1}{2} \left( Y_{ij} - \frac{1}{2} \right) - 2 \lambda(\xi_{ij}) \alpha_{ij} m_{\beta} \right) \tau_{ik} \tau_{jl}, \quad \forall k \neq l.$$

### 3.2.2. Optimization of $\xi$

So far, we have seen how the lower bound $\mathcal{L}_{K}(\xi)$ of $\log p(Y|M_{K})$ could be maximized with respect to the distribution $q(\xi)$. However, we have not addressed yet how $\xi$ could be estimated from the data. Given a distribution $q(\xi)$, we propose to maximize $\mathcal{L}_{K}(q; \xi)$ with respect to each variable $\xi_{ij}$ to obtain the tightest bound $\mathcal{L}_{K}(q; \xi)$ of $\log p(Y|M_{K})$. This follows the work of Bishop and Svensén (2003) on Bayesian hierarchical mixture of experts and Latouche, Birmelé, and Ambroise (2011, 2014) on the overlapping stochastic block model. As shown in the following proposition, this leads to new estimates $\hat{\xi}_{ij}$ of $\xi_{ij}$.

Proposition 8. The estimate $\hat{\xi}_{ij}$ of $\xi_{ij}$ maximizing $\mathcal{L}_{K}(q; \xi)$ is given by

$$\hat{\xi}_{ij} = \frac{1}{2} \sum_{i \neq j} \left( \sum_{k=1}^{K} \tau_{ik} \tau_{jk} \left[ \frac{a_{kn}}{b_{n}} + 2 \sum_{l \neq k} \tau_{il} \tau_{jl} (m_{\alpha})_{kl} \alpha_{ij} m_{\beta} \right] + \text{Tr} \left( \frac{X_{ij}}{\beta} \beta \right) \frac{a_{kn}}{b_{n}} + m_{\beta} m_{\beta} \right).$$

Note that $\hat{\xi}_{ij} = \hat{\xi}_{ji}$, $\forall i \neq j$ since the networks considered are undirected.

This gives rise to a three-step optimization scheme. Given a matrix $\xi$, the variational E and M steps of the VBEM algorithm are used to maximize $\mathcal{L}_{K}(q; \xi)$ with respect to $q(\xi)$. This distribution is then held fixed and the bound is maximized with respect to $\xi$. These three steps are repeated until convergence of the lower bound. The proof is given in Appendix A.8.

### 3.3. Estimation

**Goodness of fit.** For any $K$, we have seen how variational techniques could be used to approximate the marginal log-likelihood $\log p(Y|M_{K})$ using a lower bound $\tilde{L}_{K} := \max_{q, \xi} \mathcal{L}_{K}(q, \xi)$. As exposed in Section 2.1, our goodness-of-fit procedure relies on the posterior probability of $K$, that is, $p(M_{K}|Y)$. Indeed, this posterior distribution cannot be derived in an exact manner but, as shown by Volant, Magniette, and Robin (2012), the distribution $\tilde{p}(M_{K}|Y)$ that minimizes the Kullback–Leibler divergence with $p(M_{K}|Y)$ satisfies

$$\tilde{p}(M_{K}|Y) \propto p(M_{K}) \exp \{ \tilde{L}_{K} \}.$$

The approximate posterior probability of $H_{0}$ is then $\tilde{p}(H_{0}|Y) = \tilde{p}(M_{1}|Y)$ and the corresponding approximate posterior Bayes factor $\tilde{B}_{01}$, defined in (5), can be computed in the same manner.

The following proposition, which is proved in Appendix A.9, shows that many terms of $\mathcal{L}_{K}(q; \xi)$ vanish, when computed after a specific optimization step, so that the lower bound takes a simpler form.

Proposition 9. If computed right after the variational M step, the lower bound is given by

$$\mathcal{L}_{K}(q; \xi) = \frac{1}{2} \sum_{i \neq j} \left( \log g(\xi_{ij}) - \frac{\xi_{ij}}{2} + \lambda(\xi_{ij}) \xi_{ij} \right) + \log \frac{C(e)}{C(e)} + \log \frac{\Gamma(a_{e})}{\Gamma(a_{e})} + \log \frac{\Gamma(c_{e})}{\Gamma(c_{e})} + a_{0} \log b_{0} + a_{n} \left( 1 - \frac{b_{0}}{b_{n}} - \log b_{n} \right) + c_{0} \log d_{0} + c_{n} \left( 1 - \frac{d_{0}}{d_{n}} - \log d_{n} \right) + \frac{1}{2} \sum_{k \neq l} \log (\sigma_{\alpha}^{2})_{kl} + \frac{1}{2} \sum_{i=1}^{n} \tau_{ik} \log \tau_{ik} + \frac{1}{2} \sum_{k \neq l} (\sigma_{\alpha}^{2})^{-1}_{kl} (m_{\alpha}^{2})_{kl} - \frac{1}{2} m_{\beta}^{2} \beta \right) + m_{\beta} \sum_{i \neq j} \left( Y_{ij} - \frac{1}{2} \right),$$

where $C(x) = \prod_{k=1}^{K} \Gamma(x_{ik}) / \Gamma(\sum_{k=1}^{K} x_{ik})$ and $\Gamma(\cdot)$ is the gamma function.

**Residual structures.** While the main object of this work is to provide tools to assess the goodness of fit of a logistic regression model for networks, the considered variational algorithm also provides a natural way to estimate the residual structure $\phi$. We recall that, under Model $H_{0}$, that is, the network is completely explained by the covariates, the function $\phi$ is constant.

Still, under the alternative Model $H_{1}$, a residual structure remains, which is encoded in $\phi$. As a consequence, an estimate of this function can be useful to investigate the residual structure, similarly to the residual plot classically used in a regression context. Removing the covariate effect, recall that $M_{K}$ is an SBM model. Therefore, an approximate posterior mean can be derived, relying on the VBEM model averaging approach considered in Latouche and Robin (2015) for SBM. Proposition 10 provides the approximate posterior mean of the function $\phi$, which we propose as the network counterpart of the residual plot in regression. Note that it results from an integration over all model parameters and Models $M_{K}$.

Proposition 10. From Proposition 1 in Latouche and Robin (2015), for $(u, v) \in [0, 1]^{2}$, $u \leq v$, the approximate posterior
mean of the residual structure $\phi$ is
\[
\mathbb{E}[\phi(u,v)|Y] = \sum_{K \geq 1} \hat{p}(M_K|Y) \mathbb{E}[\phi(u,v)|Y, M_K],
\]
where
\[
\mathbb{E}[\phi(u,v)|Y, M_K] = \sum_{k \leq 1} (m_u)_{ik} \left[ F_{k-1,l-1}(u,v;e) - F_{k-1,l-1}(u,v;e) - F_{k-1,l-1}(u,v;e) + F_{k,l-1}(u,v;e) \right].
\]
$F_{k,l}(u,v;e)$ denotes the joint cdf of the Dirichlet variables $(\sigma_k, \sigma_l)$ such that $\sigma_k = \sum_{l=1}^{d} \pi_l$ and $\pi$ has a Dirichlet distribution $\text{Dir}(e)$.

As mentioned in Section 2.1, the residual structure $\phi$ is related to the graphon function of $W$-graph models, which suffer from identifiability issues. Indeed, for any measure preserving transformation $\sigma$ of $[0,1]$ to $[0,1]$, the function $\phi_\sigma(u,v) = \phi(\sigma(u), \sigma(v))$ leads to the same model as with the function $\phi(u,v)$. To tackle this issue, the common approach is to assume that the mean function $\int \phi(u,v) dv$ is increasing in $u$. This identifiability constraint was applied when producing the residual structure plots presented in the following section.

4. Simulation Study

To assess the proposed methodology, we carried out a series of experiments on simulated data first and then on real data. In this section, we focus on the estimation of the posterior probability $\hat{p}(H_0|Y)$. We aim at evaluating the capacity of the approach to detect $H_0$ using toy data. Similar results were obtained for the estimated Bayes factors $\hat{B}_{01}$ and identical conclusions were drawn.

4.1. Simulation Design

We simulated networks using Model $H_1$. Thus, each node is first associated with a latent position $U_i$ sampled from a uniform distribution over the $(0,1)$ interval. Then, a vector of covariates $X_i \in \mathbb{R}^d$ is drawn for each node, using a standardized Gaussian distribution, that is, with zero mean and covariance matrix set to the identity matrix, with $d=2$. To construct the covariate vector $X_{ij} \in \mathbb{R}^d$ for each edge $(i,j)$ with $(i < j)$, we fixed $X_{ij} = X_i - X_j$. For the function $\phi(\cdot, \cdot)$, we considered a design inspired by the one proposed in Latouche and Robin (2015). In this work, the graphon function is $W(u,v) = \rho \lambda^2 (uv)^{\lambda-1}$ where the parameter $\rho > 0$ controls the graph density and $\lambda > 0$ the degree concentration. For more details, we refer to Latouche and Robin (2015). Note that the maximum of the graphon function is $\rho \lambda^2$ so $\lambda < 1/\sqrt{\rho}$ must hold since $W(\cdot, \cdot)$ is a probability. In our case, the probabilities for nodes to connect are given through a logistic function $g(\cdot)$ and therefore we set $\phi(u,v) = g^{-1}(\rho \lambda^2 (uv)^{\lambda-1})$. For $\lambda = 1$, the function $\phi(\cdot, \cdot)$ is constant and so the networks are actually sampled from Model $H_0$. Conversely, for all $\lambda > 1$, datasets come from Model $H_1$. As $\lambda$ increases, the residual structure, not accounted for by Model $H_0$, becomes sharper and thus easier to detect.

We considered networks of size $n = 100$ and $n = 150$ as well as three values for the parameter $\rho \in \{10^{-2}, 10^{-1.5}, 10^{-1}\}$ helping controlling the sparsity. Finally, we tested 20 different values of $\lambda$ in $[1,5]$. For each of the triplets $(n, \rho, \lambda)$, we simulated 100 networks and we applied the methodology we propose for values of $K$ between 1 and 10. Because the variational algorithm depends on the initialization, as any EM like procedure, for each $K$ it was run twice and the best run was selected, such that the lower bound was maximized. Note that equal prior probabilities were given for the Models $M_K (K \geq 2)$ such that $p(H_1') = 1/2$. Moreover, we set $a_0 = b_0 = a_1 = d_0 = e_0 = 1$.

4.2. Results

Estimation of $p(H_0|Y)$. The results are presented in Figure 1. It appears that for low values of $\lambda$, the median (indicated in bold on the boxplots) of the estimated values of $p(H_0|Y)$ is 1 and goes to 0, when $\lambda$ increases, as expected. The results for the scenario with the highest sparsity ($\rho = 10^{-2}$) and $n = 100$ are unstable although the median values share this global property. More stable results were obtained for larger networks. Interestingly, experiments can be distinguished in the way Model $H_1$ is detected. As soon as $\lambda > 1$, then the true model responsible for generating the data is $H_1$ and so the probability of Model $H_0$ should be lower than 1/2. In practice, the estimated probability $\hat{p}(H_0|Y)$ is lower than 1/2 for slightly larger values of $\lambda$. For instance, for $\rho = 10^{-1.5}$ and $n = 150$, $\hat{p}(H_0|Y) \approx 0$ for $\lambda = 1.8$. For $\rho = 10^{-1}$ and $n = 100$ the detection threshold appears sooner, for $\lambda = 1.6$. The experiments illustrate that $H_1$ is detected more easily, as the network size $n$ and (density) parameter $\rho$ increase. Overall the results are encouraging with particularly low detection threshold. For $\rho = 10^{-1}$ and $n = 150$, Model $H_1$ is always detected when present as soon as $\lambda \geq 1.2$.

Computational cost. To give some insight into the computational cost of the proposed methodology, we recorded the running time for the estimation of $p(H_0|Y)$, in various conditions. Note that the inference strategy can easily be parallelized. Therefore, to give a fair evaluation, we applied the methodology once for each network generated, on a unique core. The results presented in Table 1 were obtained on an Intel Xeon CPU 3.07 GHz, for $\lambda = 2$ and $\rho = 10^{-1}$. As expected, the running time increases as the network size $n$ becomes higher. Similarly, increasing the number $d$ of covariates induces an additional computational effort. Again, the methodology proposed involves testing various values of $K$ (from 1 to 10 in these experiments), which can be done in parallel to reduce significantly the running times. If a core is used for each value of $K$, then the running time is given essentially by the slowest run, usually for the largest value of $K$. For information, the corresponding running times are also indicated in parenthesis in Table 1.

| $n$ | $d=2$ | $d=5$ | $d=10$ |
|-----|-------|-------|--------|
| 100 | 0.47 (0.1) | 0.6 (0.12) | 0.72 (0.14) |
| 250 | 3.42 (0.73) | 4.74 (0.88) | 5.97 (1.26) |
| 500 | 18.03 (3.73) | 20.28 (4.17) | 24.43 (4.91) |

Note: In parenthesis, the average running times (in minutes) for $K = 10$. Networks with as much as 500 nodes can be analyzed in less than half an hour.
Figure 1. Boxplots of the estimated values $\hat{p}(H_0|Y)$ of the posterior probability $p(H_0|Y)$, obtained with the variational approximations, for values of $\lambda$ ranging from 1 to 5. Six scenarios considered with the number $n$ of nodes in $\{100, 150\}$ and the sparsity parameter $\rho$ in $\{10^{-2}, 10^{-1.5}, 10^{-1}\}$. Model $H_0$ is true for $\lambda = 1$ and false for $\lambda > 1$. As long as $\rho$ exceeds $10^{-1.5}$ or $n \geq 150$, a sharp detection boundary is observed.

5. Illustrations

We applied our approach to analyze a series of networks (see supplementary materials) of various sizes and densities, from social sciences and ecology. For all studies, equal prior probabilities were given for the Models $M_K$ ($K \geq 2$) such that $p(H_1') = 1/2$. Moreover, we set $a_0 = b_0 = c_0 = d_0 = e_0 = 1$. The variational algorithm was run on each network for $K$ between 1 and 16. For each $K$, the procedure was repeated 20 times and the run maximizing the lower bound was selected.

Coding of the covariates. The model we propose involves a regression term $x_{ij}^T \beta$ where $x_{ij}$ is a vector of covariates for edge $(i,j)$. In some situations, edge descriptors $x_{ij}$, such as (phylogenetic, geographic) distances, are actually available. But in many situations, only node descriptors $x_i$ and $x_j$ are available and building an edge descriptor $x_{ij}$ from node descriptors is not a straightforward task (see, e.g., Hunter, Goodreau, and Handcock 2008). For all networks (except the blog network to be consistent with Latouche and Robin 2015), we adopted the following coding rules. Quantitative edge descriptors were treated as quantitative regressors. For quantitative node descriptors, the absolute difference $x_{ij} = |x_i - x_j|$ was used as a quantitative covariate. For ordinal node descriptors with $L$ levels were transformed into qualitative edge descriptors with $2L$ levels, each node level $\ell$ giving rise to two edge levels: one indicating if both $i$ and $j$ have level $\ell$ and one indicating if either $i$ or $j$ (but not both) has level $\ell$.

5.1. Description of the Datasets

Florentine marriage network. We considered the dataset analyzed by Breiger and Pattison (1981) in their study of local role analysis in social networks. It characterizes the social relations...
among 16 Renaissance Florentine families and was built by John Padgett from historical documents. Two nodes are linked if the two families share marriage alliances. Three quantitative node covariates are provided for each family, namely, the family’s net wealth in 1472 in thousands of lira, the family’s number of seats on the civic councils held between 1282 and 1344, and the family’s total number of business and marriage ties in the entire dataset.

Florentine business network. This dataset is similar to the Florentine marriage network described previously except that edges now describe business ties between families. We considered exactly the same covariates.

Karate network. The karate dataset describes the friendships between a subset of 34 members of a karate club at a university in the U.S., observed from 1970 to 1972. It was originally studied by Zachary (1977). When the study started, an incident occurred between the club president and a karate instructor, over the price of the karate lessons. The entire club then became divided over this issue, as time passed. The network is made of four known groups characterized by a node qualitative descriptor, taking four possible values, for each node in the network.

Tree network. This dataset was first introduced by Vacher, Piou, and Desprez-Loustau (2008) and further studied in Mariadassou, Robin, and Vacher (2010). We considered the tree network, which describes the interactions between 51 trees where two trees interact if they share at least one common fungal parasite. Three quantitative edge descriptors are available characterizing the genetic, geographic, and taxonomic distances between the tree species.

Blog network. The network is made of 196 vertices and was built from a single day snapshot of political blogs extracted on 14th October 2006 (Zanghi, Ambroise, and Miele 2008). Nodes correspond to blogs and an edge connect two nodes if there is a hyperlink from one blog to the other. They were annotated manually by the “Observatoire Présidentiel” project such that, for each node, labels are available. Thus, each node is associated with a political party from the left wing to the right wing and the status of the writer is also given (political analyst or not). This dataset has been studied in a series of works (Zanghi, Ambroise, and Miele 2008; Latouche, Birmelé, and Ambroise 2011, 2014) where all the authors pointed out the crucial role of the labels in the construction of the network. We considered a set of three covariates \( x_{ij} = (x_{ij}^1, x_{ij}^2, x_{ij}^3) \in \mathbb{R}^3 \) artificially constructed to analyze the influence of both the political parties and the writer status. We set \( x_{ij}^1 = 1 \) if blogs \( i \) and \( j \) have the same labels, 0 otherwise. Moreover, \( x_{ij}^2 = 1 \) if one of the two blogs \( i \) and \( j \) is written by political analysts, 0 otherwise. Finally, \( x_{ij}^3 = 1 \) if both are written by political analysts, 0 otherwise.

CKM. This dataset was created by Burt (1987) from the data originally collected by Coleman, Katz, and Menzel (1966). The network we considered characterizes the friendship relationships among physicians, each physician being asked to name three friends. The physicians were also asked to answer to a series of questions regarding their profession. We focused here on 13 questions corresponding to node covariates among which four are qualitative descriptors: city of practices (4 values), discussion with other doctors (3 values), specialty in a field of medicine (4 values), proximity with other physicians (4 values). All other node covariates were treated as quantitative variables. Note that we imputed the missing values in the dataset using the missMDA R package (Josse and Husson 2016).

Faux Dixon high network. Contrary to all networks presented in this work, this dataset is directed and therefore we employed the inference algorithm for the directed case, as presented in the supplementary materials. This network characterizes the (directed) friendship between 248 students. It results from a simulation based upon an exponential random graph model fit (Handcock et al. 2008) to data from one school community from the AddHealth Study, Wave I (Resnick et al. 1997). Node covariates are provided, namely, the grade, sex, and race of each student. The grade ordinal attribute has values 7 to 12, indicating each student’s grade in school. Moreover, the race qualitative attributes can take four values.

AddHealth 67. This dataset is related to the Faux Dixon network described previously. However, it was constructed from the original data of the AddHealth study, and not simulated from any random graph model. The AddHealth study was conducted using in-school questionnaires, from 1994 to 1995. Students were asked to designate their friends and to answer to a series of questions. Results were collected in schools from 84 communities. In our study, we considered a network associated with school community 67, which characterizes the undirected friendship relationships between 530 students. As for the Faux Dixon network, three node covariates are available. The sex qualitative covariate takes two values. Moreover, the grade ordinal attribute has values from 7 to 12. However, contrary to the Faux Dixon network, five values are present in the data for the race qualitative attribute.

5.2. Results

Goodness of fit. The estimated values of \( p(H_0|Y) \) for all networks are presented in Table 2. For illustration purposes, the estimations of the residual structures \( g \circ \phi \) are also provided in Figures 2–4. In practice, we used Proposition 10 to estimate \( \phi \) and then applied \( g(\cdot) \) to obtain graphon-like surfaces. There is no standard definition of \( W \)-graph models in the directed case and therefore, for the Faux Dixon high network, only the estimation of \( p(H_0|Y) \) is given.

As shown in Table 2, Model \( H_2 \) was rejected for the blog, tree, karate and AddHealth networks. Indeed, we obtained values of \( \hat{p}(H_0|Y) \) close to zero for the four datasets, indicating that the corresponding covariates cannot explain entirely the construction of these networks. For the blog network, we can observe in
Figure 2. Estimation of the blog (top) and tree (bottom) networks residual structure without (left) and with (right) covariates. $x$-axis: $u$, $y$-axis: $v$, $z$-axis: $g \circ \hat{\phi}(u,v)$. The residual graph is flatter than the one without covariates in both cases, but not enough to validate the goodness of fit of the regression model.

Figure 2 (top right) that $g \circ \hat{\phi}$ is not constant, which is coherent with Model $H_0$ being rejected. We also give in this figure (top left) the estimated residual structure without taking the covariates into account ($d = 0$). Clearly, the shape of $g \circ \hat{\phi}$ is simpler when $d = 3$. In particular, many of the hills on the diagonal vanish when adding the covariates. Thus, the covariates help in studying and explaining parts of the network. However, they are not sufficient and some of the heterogeneity observed in the network cannot be explained by political parties and writer status. Similar conclusions can be drawn from the tree, karate, and AddHealth networks (Figures 2 and 3). Indeed, the terms $g \circ \hat{\phi}$ simplify when adding the covariates but remain nonconstant.

In particular, for the tree network considered, this means that the interactions between trees through common fungal parasite cannot be entirely explained by the distances available, which is consistent with these from the valued version of this network, after taking the covariates into account.

For all other networks considered, model $H_0$ was chosen. Figure 4 provides the estimation of the respective $g \circ \hat{\phi}$ without covariates. Indeed, for the Florentine marriage and business networks, we found $\hat{p}(H_0|Y) = 0.995$ and $\hat{p}(H_0|Y) = 0.984$, respectively. As a consequence, the residual structures $g \circ \hat{\phi}$ were found almost constant when adding the covariates. They are therefore not given in Figure 4. Moreover, the variational

Figure 3. Estimation of the karate (top) and AddHealth (bottom) networks residual structure without (left) and with (right) covariates. $x$-axis: $u$, $y$-axis: $v$, $z$-axis: $g \circ \hat{\phi}(u,v)$. The residual graphons are not flat but emphasize a network structure not explained by the covariates.
approach led to $\hat{p}(H_0|Y) = 1$, for the Faux Dixon high and CKM networks. Thus, the statistical framework we propose shows that no other effect than these of the covariates contributes significantly to explain the structure of these networks. In other words, once corrected for the covariates, no residual heterogeneity is observed among the interactions.

A detailed example: the tree network. We now focus on the tree network to illustrate the use of the proposed procedure. Three distances were available to be used as edge covariates to explain the network topology. The residual graphons obtained with no covariate and the three covariates are given in Figure 2. Table 3 provides the (variational) posterior moments of the regression coefficients. To better understand the contribution of each covariate, we considered the four following regression models of type (1): $M_0$ with no covariates, $M_1$ involving only taxonomy, $M_2$ with both taxonomy and genetic, $M_3$ with both taxonomy and geography, and $M_4$ with all covariates. Using Volant, Magniette, and Robin (2012), we computed the (variational) posterior probability of each model and we got $\hat{Pr}(M_0) \simeq \hat{Pr}(M_2) \simeq 0$, $\hat{Pr}(M_1) = 72.7\%$, and $\hat{Pr}(M_4) = 27.3\%$. A first conclusion is that, among the three available distances, only taxonomy and, to a less extent, geography has an effect on the network topology.

Still Table 2 showed that a significant residual structure exists, which is not due to any of these distances. For each edge, we computed the Pearson residual $\hat{E}_{ij}$ for logistic regression models $M_1$ and, for each node $i$, we computed the sum of these residuals $\sum_j \hat{E}_{ij}$. We found a correlation of 0.96 between the cumulated residual and the degree of the node. Similar results were observed with models $M_2$, $M_3$, and $M_4$. This can be interpreted as strong individual effect, which could be related to a node covariate rather than to an edge covariate. Further investigations would require more ecological knowledge.

6. Conclusion

In this article, we proposed a framework to assess the goodness of fit of logistic models for binary networks. Thus, we added a generic term, related to the graphon function of $W$-graph models, to the logistic regression model. The corresponding new model was approximated with a series of models with block-wise constant residual structure. A Bayesian procedure was then considered to derive goodness-of-fit criteria. All these criteria depend on marginal likelihood terms for which we did provide estimates relying on variational approximations. The first approximation was obtained using a variational decomposition while the second involves a series of Taylor expansions. The approach was tested on toy datasets and encouraging results were obtained. Finally, it was used to analyze eight networks from social sciences and ecology. We believe the methodology has a large spectrum of applications since covariates are often given when analyzing binary networks.

### Supplementary Materials

Appendix: Give all proofs of the article. (appendix.pdf)

Directed case: Describe the inference procedure for the directed case. (directed.pdf)

Graphs: Plot of the networks studied in the Illustration section. (graphs.pdf)

Codes: Zip of all the codes required to reproduced our work, including a README file. (code.zip)
Villa, N., Rossi, F., and Truong, Q. (2008), “Mining a Medieval Social Network by Kernel SOM and Related Methods,” Technical Report. [98]

Volant, S., Magniette, M.-L. M., and Robin, S. (2012), “Variational Bayes Approach for Model Aggregation in Unsupervised Classification With Markovian Dependency,” *Computational Statistics & Data Analysis*, 56, 2375–2387. [102, 107]

Wang, Y., and Wong, G. (1987), “Stochastic Blockmodels for Directed Graphs,” *Journal of the American Statistical Association*, 82, 8–19. [98]

Watts, D., and Strogatz, S. (1998), “Collective Dynamics of Small-World Networks,” *Nature*, 393, 440–442. [98]

Wolfe, P. J., and Olhede, S. C. (2013), “Nonparametric Graphon Estimation,” Technical Report. arXiv:1309.5936. [98]

Yang, J. J., Han, Q., and Airoldi, E. M. (2014), “Nonparametric Estimation and Testing of Exchangeable Graph Models,” in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pp. 1060–1067. [98]

Zachary, W. (1977), “An Information Flow Model for Conflict and Fission in Small Groups,” *Journal of Anthropological Research*, 33, 452–473. [105]

Zanghi, H., Ambroise, C., and Miele, V. (2008), “Fast Online Graph Clustering via Erdős-Rényi Mixture,” *Pattern Recognition*, 41, 3592–3599. [98, 99, 105]

Zanghi, H., Volant, S., and Ambroise, C. (2010), “Clustering Based on Random Graph Model Embedding Vertex Features,” *Pattern Recognition Letters*, 31, 830–836. [98, 99]