An alternative class of models to position social network groups in latent spaces

1st Izabel Nolau
Dept. of Statistical Methods (DME)
Federal University of Rio de Janeiro (UFRJ)
Rio de Janeiro, Brazil
nolau@dme.ufrj.br

2nd Gustavo S Ferreira
National School of Statistical Sciences (ENCE)
Rio de Janeiro, Brazil
gustavo.ferreira@ibge.gov.br

***Abstract*—Identifying key nodes, estimating the probability of connection between them, and distinguishing latent groups are some of the main objectives of social network analysis. In this paper, we propose a class of blockmodels to model stochastic equivalence and visualize groups in an unobservable space. In this setting, the proposed method is based on two approaches: latent distances and latent dissimilarities at the group level. The projection proposed in the paper is performed without needing to project individuals, unlike the main approaches in the literature. Our approach can be used in undirected or directed graphs and is flexible enough to cluster and quantify between and within-group tie probabilities in social networks. The effectiveness of the methodology in representing groups in latent spaces was analyzed under artificial datasets and in a case study.**

***Index Terms*—blockmodel, social networks, multidimensional scaling, latent space, visualization**

I. Introduction

In social network analysis, it is usual to examine the association of n individuals through a matrix $\mathbf{Y}_{n\times n}$, whose elements y_{ij} describe the connection between the i-th and j-th components of the network. These elements can be represented as a graph, where each node (or vertex, or point) represents an individual, and the edges (or links, or ties) represent relationships among them. Identifying key nodes, estimating the probability of connection between them, and distinguishing latent groups are some of the main objectives of social network analysis. Several methods have been being proposed, including deterministic techniques for graph analysis and, more recently, sophisticated statistical models using latent effects.

In this paper, we propose a simple latent class model to represent groups in an unobservable space. In this setting, we propose an alternative class of models to position groups based on two approaches: latent distances and latent dissimilarities at the group level. The novelty of our approach is to propose a class of simple blockmodels that allows to position groups in latent spaces aiming:

- to model the relationship between groups as in [9] without needing to represent individuals in the latent space;
- to estimate within and between-groups probabilities of ties to properly represent stochastic equivalence, as in the traditional latent class models; and
- to use a random version of multidimensional scaling based on samples from the posterior distribution of a set of unknown dissimilarities provided by data.

The paper is organized as follows. In Section II, an alternative class of models to positioning groups in latent spaces is introduced, and aspects of inference are discussed. III evaluate the effectiveness of the proposed models according to real data. Finally, a discussion about the results and suggestions for further research are presented in Section IV.

A. Related models

Our general model, presented in Section II, brings new features about blockmodelling and has similarities with alternative models in the literature. The Latent Dissimilarities Model - LDM - and the Latent Positions Model - LPM - , respectively described in Sections II-A and II-B, extend the ideas from [14] and [1] to spatially represent network's groups in unobserved spaces.

These representations are performed assuming that the latent positions are parameters, in the LPM model, and via Multidimensional Scaling, in the LDM model. Both models allow the visualization of the latent structure between groups and also give an intuitive interpretation for between-group probabilities.

Instead of representing individuals in latent spaces, as the distance model proposed by [9], the LPM model places groups. On the other hand, the LDM model can be seen as a reparametrization of a blockmodel by decomposing the between-group tie probability into two components: an intercept and a distance between groups, which are estimated from data. This set of distances is then used to represent latent positions in a second stage via Multidimensional Scaling. Thus, the novelty of the LDM model is to provide a set of distances for the Multidimensional Scaling directly estimated from data by using a simple and intuitive model structure.

To compare our approach to other models that represent individuals in latent spaces, we will point out below the main differences and similarities between the proposed models, the Latent Position Cluster Model - LPCM -[8] and the Latent Space Stochastic Blockmodel - LSSBM [6].

The LPCM models within and between-group tie probabilities in the network, projecting all nodes in a latent space. In this setting, each latent position is modeled as Gaussian mixtures and the latent distances between all nodes are used to model tie probabilities, regardless of whether individuals are in the same group or not. Thus, while projecting all nodes in the latent space, the LPCM model makes no distinction between tie probabilities for nodes in the same or different groups and also does not provide the cluster's projection in the latent space. Despite the fact that the vector of centers from the Gaussian mixture in the LPCM model can be indirectly assumed as a set of latent group's positions — since the position of a node is affected by the weighted average of the positions of all centers and not only by the center of its group —, obtaining their estimates requires modeling the latent positions of all individuals in the network.

In turn, the LSSBM model decomposes network structure into two components: one that describes between-community relations, and another describing within-community relations. This approach also uses the concept of latent distances [9], but only to model the within-group tie probabilities. Thus, unlike our approach, the LSSBM model provides latent representations only for nodes inside each group — in a multiresolution perspective — but does not provide groups' latent positions. Furthermore, the LSSBM model also requires modeling the latent positions of all individuals in the network and focuses on undirected relations. On the other hand, LDM and LPM models do not require nodes' latent positions to obtain the groups' latent positions and are suitable to model undirected or directed relational data.

II. A NEW CLASS OF MODELS TO POSITION LATENT GROUPS

In general, social network data with n individuals produces a matrix $\mathbf{Y}_{n\times n}$, whose elements y_{ij} indicate the existence or non-existence of a connection — or simply the number of connections — between the i-th and j-th elements of the network. These matrices can be symmetric (undirected networks) or not (directed networks). In this paper, we consider the asymmetric case, where $y_{ij} \neq y_{ji}$, but all results can be easily extended to undirected graphs by adjusting indexes (from $i \neq j$ to $i < j$). Let μ_{ij} be the probability that two individuals i and j share a connection — or the expected number of connections between them. Now, we propose a novel class of models to position groups in a latent space, which general formulation is presented below:

$$\begin{aligned}
Y_{ij} &\sim F(\mu_{ij}), & i \neq j\\
g(\mu_{ij}) &= \alpha_0 + \beta(c_i, c_j), & \\
\mathbf{C}_i &\sim \text{Multinom}(1, \mathbf{p}_i), & i = 1, \ldots, n\\
\alpha_0 &\sim N(0, \sigma^2_\alpha), & \qquad (1)
\end{aligned}$$

where

- F represents a probability distribution with mean μ_{ij};
- $g(\cdot)$ is a link function;
- α_0 is a common level for all individuals in the social network;
- $\mathbf{C}_i = (C_{i_1}, \ldots, C_{i_K})'$ is a vector with $K-1$ zeros and one number one, that indicates the group of the $i-$th individual and $\mathbf{C} = (\mathbf{C}_1, \ldots, \mathbf{C}_n)'$;
- c_i is the class of the $i-$th individual, i.e., $c_i = \{k \mid C_{i_k} = 1\}$, for $i = 1, \ldots, n$;
- $\mathbf{p}_i = (p_{i_1}, \ldots, p_{i_K})'$ is a vector indicating the prior probabilities of belonging to each group for the $i-$th individual;
- K is the number of latent groups.

Here, the $\beta(c_i, c_j)$ parameter can be defined in many ways to accommodate, or not, between-groups' and within-groups' effects. The prior distribution of this quantity will be initially denoted by $p(\beta(c_i, c_j))$, for all $i \neq j$. On the other hand, the α_0 parameter assumes the role of a sparsity parameter, controlling the average number of connections between individuals observed in the sociomatrix $\mathbf{Y}$. Finally, the vector $\mathbf{p}_i = (p_{i_1}, \ldots, p_{i_K})'$ can be fixed — e.g., assigning prior probabilities equal to $1/K$ — or modeled according to a Dirichlet distribution.

Under this formulation, the posterior distribution is given as follows:

$$\begin{aligned}
p(\alpha_0, \boldsymbol{\beta}, \mathbf{C} \mid \mathbf{y}) \quad &\propto \quad \left[\prod_{j=1}^{n}\prod_{i\neq j} \mu_{ij}^{y_{ij}}(1-\mu_{ij})^{1-y_{ij}} \times p(\beta(c_i, c_j))\right]\\
&\times \left[\prod_{i=1}^{n}\prod_{k=1}^{K} p_{i_k}^{C_{i_k}}\right] \times \exp\left(-\frac{\alpha_0^2}{2\sigma_\alpha^2}\right).
\end{aligned}$$

The complexity level of the general model depends on the structure specified for $\beta(c_i, c_j)$. Henceforward, without loss of generality, it will be assumed that each observation y_{ij} assumes only 0 or 1. In this way, set F as the Bernoulli distribution, μ_{ij} as the probability that two individuals i and j share a connection, and g as the logistic link function.

In Subsections II-A and II-B, four possible formulations of the model (1) are presented. These different approaches are subdivided into two types: models based on latent dissimilarities and models based on latent positions.

A. Latent dissimilarities models

In this subsection, the proposed models aim to place groups in space based on the concept of latent dissimilarities. In this formulation, it is considered that the tie probability depends on a set of dissimilarities between groups. This approach allows a posterior spatial representation of groups through multidimensional scaling, without the need to previously specify the set of dissimilarities, denoted by $\boldsymbol{\delta}$. Then, one can use samples drawn from the posterior distribution of the dissimilarities to place groups in the latent space, through Multidimensional Scaling.

1) Latent dissimilarities model with no within-group variation: The latent dissimilarities model with no within-group variation (LDM) is the simplest proposed formulation of (1). The structure of this model depends on a set of $\binom{K}{2}$ latent dissimilarities (for each group pair, there is an associated dissimilarity) that account for estimating the between-group tie probabilities. This model's construction considers that individuals belonging to the same group have a constant connection probability, regardless of the group that they are part of, denying variation in within-group tie probabilities. In this case, the structure specified for $\beta(c_i, c_j)$ is given by:

$$\beta(c_i, c_j) = \begin{cases} 0, & \text{if } c_i = c_j \\ -\delta_{c_i c_j}, & \text{if } c_i \neq c_j \end{cases}, \qquad (2)$$

for $i, j = 1, \ldots, n, \ i \neq j$, where δ_{kl} represents a symmetric dissimilarity measure between groups k and l. Note that, since the dissimilarity among groups k and l and between groups l and k are the same, then $\delta_{kl} = \delta_{lk}$. Thus, to unify the notation, the groups that compose the subindex kl will be displayed in the ascending order. Moreover, if $c_i = c_j$, the probability $\mu_{ij} = \text{logit}^{-1}(\alpha_0 + \beta(c_i, c_j))$ will only depend on α_0. Then, $\text{logit}^{-1}(\alpha_0)$ can be interpreted as the probability of a tie between two elements belonging to the same group. Moreover, the higher the dissimilarity between two groups, the smaller the probability of a tie between two elements from different groups.

Following the Bayesian paradigm, it is assumed that $\delta_{kl} \sim \text{Gamma}(a_\delta, b_\delta)$, for $l = 1, \ldots, K$ and $k < l$, with $a_\delta = 0.01$ and $b_\delta = 0.01$ to ensure a non-informative prior.

2) Latent dissimilarities model with within-group variation: The latent dissimilarities model with within-group variation (GLDM) is an extension of the LDM, to represent variation among within-group tie probabilities, accommodating different levels of homogeneity inside each group. The structure of this model depends on a set of $\binom{K}{2}$ latent dissimilarities that account for estimating the between-group tie probabilities, and on a set of K parameters, given in the vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, that account for estimating the within-group tie probabilities. It is required to include a sum-to-zero constraint in the vector $\boldsymbol{\alpha}$ to make its components identifiable, as described in Subsection II-C. The effect of belonging to group k is represented by α_k, for $k = 1, \ldots, K$. In this case, the structure specified for $\beta(c_i, c_j)$ is given by:

$$\beta(c_i, c_j) = \begin{cases} \alpha_{c_i}, & \text{if } c_i = c_j \\ -\delta_{c_i c_j}, & \text{if } c_i \neq c_j \end{cases}, \tag{3}$$

for $i, j = 1, \ldots, n, \ i \neq j$, where δ_{kl} represents a symmetric dissimilarity measure between groups k and l.

The main difference between LDM and GLDM occurs when two individuals belong to the same group. Now, when i and j belong to the same group k, the probability $\mu_{ij} = \text{logit}^{-1}(\alpha_0 + \beta(c_i, c_j))$ will depend on $\alpha_0 + \alpha_k$. Then, $\text{logit}^{-1}(\alpha_0 + \alpha_k)$ represent the tie probability between two elements belonging to the same group k, for $k = 1, \ldots, K$. Thus, α_k is responsible for increase, or decrease, the probability that two individuals i and j belonging to the same group share a connection.

Following the Bayesian paradigm, it is assumed that $\delta_{kl} \sim \text{Gamma}(a_\delta, b_\delta)$, for $l = 1, \ldots, K$ and $k < l$, with $a_\delta = 0.01$ and $b_\delta = 0.01$ to ensure a non-informative prior. For α_k it is assumed a $N(0, \sigma^2_\alpha)$ distribution, for $k = 1, \ldots, K$, with $\sigma^2_\alpha = 9$, that is a low-informative prior due to the logit link function structure.

B. Latent positions models

In this subsection, the proposed models aim to place groups in space based on the concept of latent distances [9]. In this formulation, it is considered that the tie probability depends on the latent distance between groups' positions. This approach provides a spatial representation of groups through the set of positions in the latent space, denoted by $\mathbf{a}$. The LDM and GLDM depend on dissimilarities δ_{kl}, which can be interpreted as the distance between groups k and l since dissimilarities are positive numbers, for $l = 1, \ldots, K$ and $k < l$. Thus, in this context, LDM and GLDM can be viewed as simplified versions of the models based on latent positions, that will be presented in Subsections II-B1 and II-B2.

1) Latent positions model with no within-group variation: The latent positions model with no within-group variation (LPM) can be considered as an extension of the LDM to spatially place groups through a set of latent positions. The structure of this model depends on a set of $K \times d$ latent positions (the position of each group is a point in the $d-$dimensional latent space) that account for estimating the between-group tie probabilities. This model's construction considers that individuals belonging to the same group have a constant connection probability, regardless of the group that they are part of, denying variation in within-group tie probabilities. In this case, the structure specified for $\beta(c_i, c_j)$ is given by:

$$\beta(c_i, c_j) = -|\mathbf{a}_{c_i} - \mathbf{a}_{c_j}|, \tag{4}$$

for $i, j = 1, \ldots, n, \ i \neq j$, where $|\cdot|$ is a distance measure satisfying the triangular inequality, $\mathbf{a}_k = (a_{k1}, \ldots, a_{kd})'$ is the vector containing the position of class k, for $k = 1, \ldots, K$, in the latent space $D \subset \mathbb{R}^d$ and $\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_K)'$. Note that, if $c_i = c_j$, the probability $\mu_{ij} = \text{logit}^{-1}(\alpha_0 + \beta(c_i, c_j))$ will only depend on α_0 since $|\mathbf{a}_{c_i} - \mathbf{a}_{c_i}| = 0$. Then, $\text{logit}^{-1}(\alpha_0)$ can be interpreted as the probability of a tie between two elements belonging to the same group. Moreover, the greater the distance between two groups' position, the smaller the probability of a tie between two elements from different groups.

Following the Bayesian paradigm, it is assumed that $a_{k_l} \sim N(0, \sigma^2_a)$, for $k = 1, \ldots, K$ and $l = 1, \ldots, d$, with $\sigma^2_a = 25$, that is a low-informative prior due to the logit link function structure.

2) Latent positions model with within-group variation: The latent positions model with within-group variation (GLPM) is an extension of the LPM, to represent variation among within-group tie probabilities, accommodating different levels of homogeneity inside each group; and can be considered as an extension of the GLDM to spatially place groups through a set of latent positions. The structure of this model depends on a set of $K \times d$ latent positions that account for estimating the between-group tie probabilities, and on a set of K parameters, given in the vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, that account for estimating the within-group tie probabilities. It is required to include a sum-to-zero constraint in the vector $\boldsymbol{\alpha}$ to make its components identifiable, as described in Subsection II-C. The effect of belonging to group k is represented by α_k, for $k = 1, \ldots, K$. In this case, the structure specified for $\beta(c_i, c_j)$ is given by:

$$\beta(c_i, c_j) = \begin{cases} \alpha_{c_i}, & \text{if } c_i = c_j \\ -|\mathbf{a}_{c_i} - \mathbf{a}_{c_j}|, & \text{if } c_i \neq c_j \end{cases}, \tag{5}$$

for $i, j = 1, \ldots, n, \ i \neq j$, where $|\cdot|$ is a distance measure satisfying the triangular inequality, $\mathbf{a}_k = (a_{k1}, \ldots, a_{kd})'$ is the vector containing the position of class k, for $k = 1, \ldots, K$, in the latent space $D \subset \mathbb{R}^d$ and $\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_K)'$.

Similarly to LDM and GLDM models, the main difference between LPM and GLPM occurs when two individuals belong to the same group. Thus, the interpretation of the parameters is similar to that of models based on latent dissimilarities.

Following the Bayesian paradigm, it is assumed that $a_{k_l} \sim N(0, \sigma_a^2)$, for $k = 1, \ldots, K$ and $l = 1, \ldots, d$, with $\sigma_a^2 = 25$. For α_k it is assumed a $N(0, \sigma_\alpha^2)$ distribution, for $k = 1, \ldots, K$, with $\sigma_\alpha^2 = 9$. Both distributions can be considered as low-informative priors due to the logit link function structure.

C. Model inference

We perform inference via MCMC to obtain samples from the resulting posterior distribution of each proposed model. To describe the inference procedure for the latent dissimilarity models presented in Section II-A, we show the estimation procedure for the GLDM model, which consists of the following steps:

(1) Initialize the counter $j = 2$ and set initial values for the parameters of the model: α_0, $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$ and $\mathbf{C}$;

(2) Update the model parameters $\alpha_0, \alpha_2, \ldots, \alpha_K$ from their conditional distributions;

(3) Set $\alpha_1 = -\sum_{k=2}^{K} \alpha_k$, according to the identification procedure described ahead;

(4) Update the model parameters $\boldsymbol{\delta}$ and $\mathbf{C}$ from their conditional distributions;

(5) Increment the counter j to $j+1$ and iterate from (2).

For the LDM model, the estimation procedure's step (2) does not update α_k parameters, for $k = 1, \ldots, K$, since this model does not consider these parameters. The inference procedure for the latent positions models, presented in Section II-B, follows the same steps of the GLDM estimation procedure, updating $\mathbf{a}$ instead of $\boldsymbol{\delta}$.

Some parameters of the model demand specific inference strategies to become identifiable. To identify the K parameters responsible for estimating the within-group tie probabilities, $\alpha_1, \ldots, \alpha_K$, it is required to include a sum-to-zero constraint in the vector $\boldsymbol{\alpha}$. This restriction is added to the GLDM and GLPM models. To estimate the set of positions in the latent space $\mathbf{a}$, which provide a spatial representation of groups in the LPM and GLPM models, it is necessary to eliminate translation, rotation, and reflection effects in the configuration of the latent positions $\mathbf{a}$ via procrustes transformation as in [9]. Henceforward, without loss of generality, it will be assumed that the distance measure is the Euclidean distance.

A usual problem in social network blockmodels is identifying the group labels for each partition of the n individuals obtained while performing inference. More specifically, given a partition obtained, one can not directly determine which label is assigned to each group, since all $K!$ label permutations produce the same likelihood function value. To deal with this identification problem, known as label switching, we use a deterministic online procedure of classification, performed during the MCMC method, based on a classification method described in [5].

More specifically, the MCMC method is divided into two stages. In the first stage, we obtain reference centers and dispersion measures for a set of parameters ξ from the first m_1 posterior distribution samples. From $j = m_1 + 1$ onwards, the permutation p of labels of $\mathbf{C}^{(j)}$ that produces a set of parameters $\xi^j_{(p)}$ closest to ξ is chosen, and $\mathbf{C}^{(j)}$ is switched. After identifying the optimal label configuration, the reference measures associated with ξ are then updated, and this procedure is repeated in each of the next m_2 iterations.

This method is sensitive to the choice of m_1 since high values can be affected by label switching, and low values may not be enough to ensure a good estimate of ξ. To improve the label switching detection, we used the median and the median absolute deviation as robust measures of centrality and dispersion, respectively, and we also set $m_1 = 50$ after discarding the first 50 samples in MCMC.

Note that, since we have only $K!$ label permutations, this procedure has a low computational cost for a small number of groups, and is considered a reasonable and simple solution to the label switching problem in the simulations performed. Other methods based in order constraints or more sophisticated methods of classification are also available to deal with the switching label problem (see [20, 4, 17, 15, 16], among others), although they were not considered in this paper.

III. Case study

In this section, we analyze the performance of each of the four proposed models using a real dataset. The data contains information about the relationship of trust among eighteen monks in an American monastery. [18] suggested the existence of four factions in the monastery: Loyal Opposition (LO), Young Turks (YT), Outcasts (O) and Waverers (W). The Loyal Opposition represents the oldest members of the monastery, while the Young Turks are the newest ones. The Outcasts are the members of the monastery that were not accepted in any of the previous groups, and the Waverers are the members who did not take sides. These monks are usually labeled according to their names, or according to a sequence of numbers from 1 to 18.

The sociomatrix $\mathbf{Y}$ with the relationship of trust between the eighteen monks was obtained from `latentnet` [10], an `R` package utilized to fit and evaluate the statistical latent position and cluster models for networks. [18] performed three studies over time, which resulted in a dataset of a time-aggregated network. In this dataset, a tie from monk A to monk B exists if A nominated B as one of his three (or four, in case of a draw) best friends at any of the three-time points. Based on Figure 1, which presents the sociomatrix of monks data (a black pixel represents a link between two monks), it is possible to distinguish block structures of the social relationship between each monk faction. Note that,

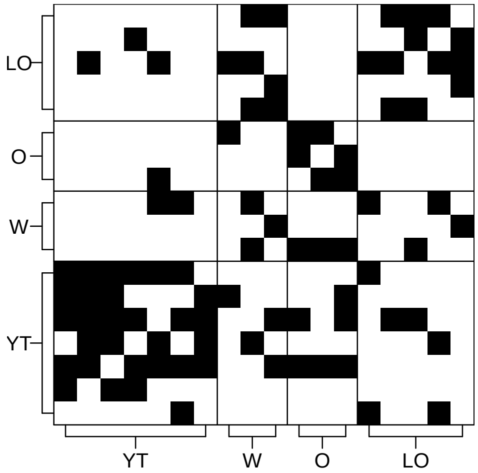

Fig. 1. Monks sociomatrix $\mathbf{Y}$ blocked according to the different factions (black pixels indicate the ties between monks).

analyzing the communication within-blocks (black pixels), the Loyal Opposition group presents fewer connections between its members than other groups.

Several authors that analyzed this dataset point that there were three prominent latent groups in the monastery (see [3, 21, 1, 8, 9, 19], among others). Since the main objective of this analysis is to determine the latent social structure within the monastery, the LDM, LPM, GLDM and GLPM models were fitted to this data considering the existence of $K = 2$ and $K = 3$ latent classes. For each scenario and number of latent groups, we let each chain run for 60,000 iterations, discarded the first 10,000 as burn-in, and stored every 5th iteration to obtain 10,000 independent samples.

Table I displays the log-likelihood function evaluated in the posterior mean of $\boldsymbol{\mu}$, i.e., $\log p(\mathbf{y}|\boldsymbol{\mu})$, and the Expected Akaike Information Criterion (EAIC) as a measure of goodness-of-fit. Note that the greater the number of latent positions considered, the higher the log-likelihood value, indicating that the models considering $K = 3$ performed better than those with $K = 2$ latent groups. These results corroborate with the EAIC values since they are notoriously lower for $K = 3$. Among the models based on latent positions, the ones with within-group variation have a higher log-likelihood value. Furthermore, among the models based on dissimilarities, those with variation within the group also have a higher log-likelihood value. This conclusion is analogous to the models based on latent positions. Finally, there is evidence of a better fitting of the GLPM model with $K = 3$ latent classes.

TABLE I
LOG-LIKELIHOOD FUNCTION EVALUATED IN THE POSTERIOR MEAN OF $\boldsymbol{\mu}$ AND EAIC, OBTAINED FOR THE FITS OF THE MODELS TO MONKS' DATASET.

	$\log L(\hat{\mu})$				EAIC			
K	LDM	LPM	GLDM	GLPM	LDM	LPM	GLDM	GLPM
2	-156.58	-156.38	-154.19	-156.08	323.94	325.63	327.71	328.96
3	-136.75	-136.64	-134.02	-133.33	286.33	291.28	297.61	304.94

Figure 2 presents the posterior mean of tie probabilities $\boldsymbol{\mu}$ between monks obtained from each model. Note that models assuming the existence of $K = 3$ latent groups performed better than models assuming $K = 2$ since they were able to distinguish Outcasts and Waverers from Young Turks and Loyal Opposition groups. Moreover, based on the range of values (represented by the shades of gray), it is possible to point out that models considering variation in the within-group tie probabilities produce more extreme values, i. e., values closer to 0 and 1. These results highlight the ability of GLDM and GLPM models to properly predict ties in comparison to LDM and LPM models, considering this dataset. Finally, the results corroborate the previous evidence, given in Table I, of a better fitting of the models with different within-group tie probabilities and $K = 3$ latent classes.

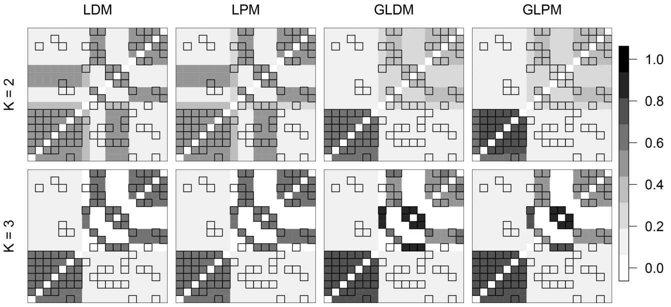

Fig. 2. Posterior mean of the tie probabilities between monks obtained from LDM, LPM, GLDM, and GLPM models (bordered pixels represent true ties and the shades of gray represent the estimated tie probabilities).

In the present study, the latent groups to which the monks belong to are not previously known, except for [18]'s factions suggestion. Thus, it is possible to analyze how the proposed models classify the monastery's monks into 2 and 3 latent groups in comparison with [18]' classification. Table II presents the clustering obtained from the posterior mode of $\mathbf{C}$, considering each number of latent positions, for all proposed models. Note that all models were able to distinguish the Young Turks and the Loyal Opposition groups satisfactorily. Moreover, for $K = 2$ latent groups, the LDM and LPM models grouped the monks in the same way, as well as the GLDM and LPM models, and for $K = 3$, all four proposed models equally grouped monks.

TABLE II
MONKS GROUPING OBTAINED FROM THE POSTERIOR MODE OF $\mathbf{C}$ FOR THE FITS OF THE MODELS TO MONKS' DATASET, CONSIDERING 2 AND 3 LATENT GROUPS.

		$K = 2$				$K = 3$			
Monks	[18]	LDM	LPM	GLDM	GLPM	LDM	LPM	GLDM	GLPM
Albert	YT	1	1	1	1	1	1	1	1
Boniface	YT	1	1	1	1	1	1	1	1
Gregory	YT	1	1	1	1	1	1	1	1
Hugh	YT	1	1	1	1	1	1	1	1
John Bosco	YT	1	1	1	1	1	1	1	1
Mark	YT	1	1	1	1	1	1	1	1
Winfrid	YT	1	1	1	1	1	1	1	1
Amand	W	1	1	2	2	2	2	2	2
Romauld	W	2	2	2	2	3	3	3	3
Victor	W	2	2	2	2	3	3	3	3
Basil	O	1	1	2	2	2	2	2	2
Elias	O	1	1	2	2	2	2	2	2
Simplicius	O	1	1	2	2	2	2	2	2
Ambrose	LO	2	2	2	2	3	3	3	3
Berthold	LO	2	2	2	2	3	3	3	3
Bonaventure	LO	2	2	2	2	3	3	3	3
Louis	LO	2	2	2	2	3	3	3	3
Peter	LO	2	2	2	2	3	3	3	3

According to Table II, considering $K = 2$, LDM and LPM grouped the Young Turks, the Outcasts, and Amand (from Waverers) in the same class, and grouped the Loyal Opposition and the remaining Waverers in another class. On the other hand, GLDM and GLPM grouped all Outcasts, Waverers, and Loyal Opposition groups in the same class. In the $K = 3$ latent groups' configuration, all models led to the same partition composed of Young Turks in a class, the Outcasts, and Amand (from Waverers) in another class, and the Loyal Opposition and the remaining Waverers in a third class. These results are in accordance with [21], [8], [11] and [19] as showed in Table III, which also presents the posterior probabilities of belonging to each class for each monk, for the GLPM model.

TABLE III

MONKS' FACTIONS SUGGESTED BY [0] [18] AND LATENT GROUPS OBTAINED ACCORDING TO [1] [11], [21], [8] AND [19], [2] [1] AND [3] [3] IN COMPARISON WITH THE RESULTS OBTAINED FROM THE GLPM MODEL, AND POSTERIOR PROBABILITIES OF BELONGING TO EACH CLASS.

i	Monks	[0]	[1]	[2]	[3]	GLPM	$P(c_i = 1)$	$P(c_i = 2)$	$P(c_i = 3)$
1	Albert	YT	1	1	1	1	0.723	0.146	0.131
2	Boniface	YT	1	1	1	1	0.730	0.155	0.115
3	Gregory	YT	1	1	1	1	0.732	0.155	0.113
4	Hugh	YT	1	1	1	1	0.729	0.154	0.117
5	John Bosco	YT	1	1	1	1	0.731	0.153	0.117
6	Mark	YT	1	1	1	1	0.731	0.152	0.117
7	Winfrid	YT	1	1	1	1	0.731	0.156	0.113
8	Amand	W	2	2	3	2	0.125	0.740	0.135
9	Romauld	W	3	-	3	3	0.146	0.083	0.771
10	Victor	W	3	-	3	3	0.145	0.079	0.776
11	Basil	O	2	2	2	2	0.126	0.756	0.118
12	Elias	O	2	2	2	2	0.125	0.759	0.116
13	Simplicius	O	2	2	2	2	0.125	0.759	0.116
14	Ambrose	LO	3	3	3	3	0.141	0.077	0.782
15	Berthold	LO	3	3	3	3	0.138	0.079	0.783
16	Bonaventure	LO	3	3	3	3	0.144	0.079	0.776
17	Louis	LO	3	3	3	3	0.143	0.076	0.781
18	Peter	LO	3	3	3	3	0.145	0.084	0.771

According to Table III, the main difference between the partitions found in the literature involves the classification of monk Amand, probably due to his ambiguous positioning in the monastery (see [3, 1], among others). Moreover, for all monks, the modal class presents the highest probability, indicating a small probability of belonging to any other cluster class. These high probabilities are also following several authors in literature (see [8, 11], among others).

The visual display of clustering in the latent space for the GLPM model can be seen in Figure 3. This display presents both the different group cohesion levels — represented by distinct circles with a radius proportional to $1-$within-group tie probability — and a satisfactory notion of the distance between groups — represented by the Euclidean distance between estimated latent positions.

The estimated within and between-group tie probabilities for all models are presented in Table IV. As expected, the between-groups probabilities are considerably lower than within-groups probabilities, and the three groups presented different cohesion levels (see [21, 8, 1], among others). As observed in the sociomatrix $\mathbf{Y}$ (see Figure 1), the within-group tie probabilities of the class containing the Loyal Opposition members is lower than the one that holds the Young Turks members, for all four proposed models. In the models with $K = 3$ latent classes, all models presented higher within-group tie probabilities for the group containing the Outcasts members, as already pointed out by several authors in the literature (see [18, 3, 8], among others).

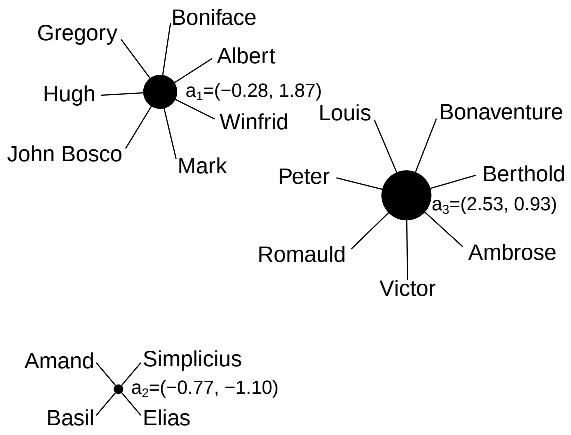

Fig. 3. Visual display of monks clustering in the latent space obtained from the GLPM model (circles radius are proportional to $1 -$ within-group tie probability).

TABLE IV

POSTERIOR MEAN OF AVERAGE LEVEL (AP), BETWEEN-GROUP (BP) AND WITHIN-GROUP PROBABILITIES (WP) OBTAINED FROM THE FITS OF THE LDM, LPM, GLDM, AND GLPM MODELS FOR $K = 2$ AND $K = 3$. THE WITHIN-GROUP PROBABILITIES ARE ORDERED ACCORDING TO LABELS 1, 2, AND 3, AND THE BETWEEN-GROUP PROBABILITIES ARE ORDERED ACCORDING TO PAIRS (1,2), (1,3) AND (2,3).

	K	LDM	LPM	GLDM	GLPM
AP	2	0.465 (0.386, 0.548)	0.468 (0.388, 0.547)	0.514 (0.398, 0.655)	0.532 (0.414, 0.672)
	3	0.646 (0.545, 0.744)	0.658 (0.561, 0.748)	0.733 (0.561, 0.887)	0.751 (0.548, 0.939)
BP	2	0.118 (0.068, 0.179)	0.114 (0.068, 0.169)	0.144 (0.083, 0.220)	0.144 (0.089, 0.212)
	3	0.175 (0.081, 0.311) 0.127 (0.054, 0.235) 0.063 (0.015, 0.146)	0.145 (0.056, 0.286) 0.147 (0.066, 0.265) 0.052 (0.011, 0.141)	0.153 (0.025, 0.408) 0.146 (0.026, 0.398) 0.104 (0.010, 0.379)	0.167 (0.019, 0.542) 0.168 (0.022, 0.545) 0.106 (0.006, 0.457)
WP	2			0.678 (0.484, 0.859) 0.346 (0.163, 0.552)	0.718 (0.529, 0.880) 0.335 (0.159, 0.541)
	3			0.675 (0.383, 0.885) 0.885 (0.708, 0.988) 0.516 (0.212, 0.798)	0.666 (0.272, 0.929) 0.906 (0.711, 0.997) 0.510 (0.124, 0.877)

To obtain the latent group positions for the LDM and GLDM models, the multidimensional scaling method was performed using the dissimilarity samples $\boldsymbol{\delta}$ as input. The chosen approach follows the analysis of [13] and was performed via `cmdscale` function in `R` software. The posterior distributions of the latent group positions for all four proposed models are presented in Figure 4.

The latent group positions seem similar for all four proposed models, despite the models based on latent positions present more variability than the models based on latent dissimilarities. The resulting latent positions from models that allow variation in the within-group tie probabilities are quite similar to those obtained from models that do not allow this variation. Note that all models were successful in distinguishing the latent positions from all groups.

IV. DISCUSSION

In this paper, we proposed an alternative class of models for social networks to represent groups in an unobservable space. This class of models encompasses approaches based on latent dissimilarities — the LDM and GLDM models — and latent positions — the LPM and GLPM models — , allowing the researcher to visualize the latent groups of the social network; quantify tie probabilities for individuals

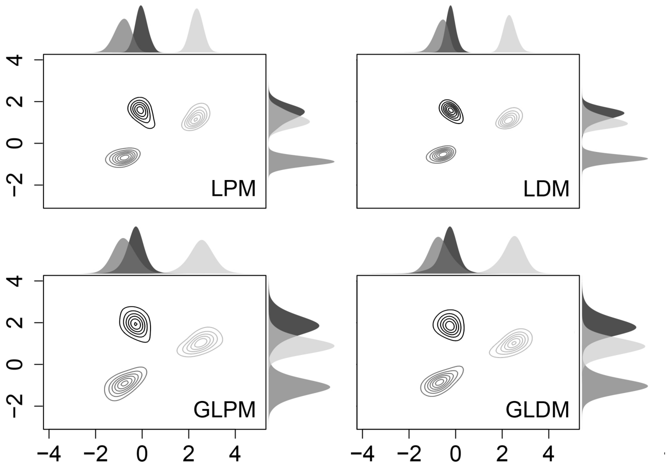

Fig. 4. Posterior distribution of the latent group positions for all four proposed models.

belonging to the same or different groups, without representing individuals in the latent space; and use a random version of multidimensional scaling based on samples from the posterior distribution of a set of unknown dissimilarities provided by data. Both approaches can be used in undirected or directed graphs, i.e., in cases where the sociomatrix $\mathbf{Y}$ is symmetric or non-symmetric, respectively. Remark that, if the sociomatrix is symmetrical, there will be less available data information to estimate the model parameters since $\mathbf{Y}$ will be an upper triangular matrix.

All four models proposed in this paper are related and suitable for classification problems in social networks. More specifically, the models where $\boldsymbol{\beta}$ are functions of dissimilarities $\boldsymbol{\delta}$ — the LDM and GLDM models —, can be viewed as simplified versions of the models based on latent positions $\mathbf{a}$ — the LPM and GLPM models. Regarding a scenario where the positions are represented in a two-dimensional latent space, the LDM and GLDM models can be more suitable when the number of groups is small. However, if $K > 5$, these models will contain more parameters associated with between-group probabilities than its associated models based on latent positions $\mathbf{a}$ and, consequently, it may not be the most suitable choice.

The LDM and GLDM models aim to represent groups in space based on the concept of latent dissimilarities. In the two-stage proposed methodology, the set of dissimilarities between groups is estimated in the first step, and the samples drawn from the posterior distribution of the dissimilarities are used as input in a multidimensional scaling technique, in the second stage. Once there is a sample of the posterior distribution of the set of dissimilarities available, it is possible to take into account the uncertainty associated with the multidimensional scaling result. Although there are several methods for modeling data through Multidimensional Scaling, the main differences between existing approaches have not been addressed in this paper.

Under the GLDM and GLPM models, individuals belonging to different groups have the same probabilities of tie to each other. Despite that, all four proposed models presented in this work can not properly represent transitivity and homophily on attributes. To represent these features, it would be required to model relational data at the individual level, or to include individuals' information. The individual-level modeling was not an aim of this paper, but rather to properly represent the relationships between groups in latent spaces.

The Multinomial distribution assigned for $\mathbf{C}_i$ — responsible for indicating the group of the $i-$th individual — depends on the hyperparameters $\mathbf{p}_i$, which represent prior probabilities of belonging to each group for the $i-$th individual, for $i = 1, \ldots, n$. For all cases analysed in this paper, we set $\mathbf{p}_i = K^{-1}\mathbf{1}_K$, for $i = 1, \ldots, n$, where $\mathbf{1}_K$ represents the K-dimensional vector of ones. Alternative approaches include assigning a prior distribution to these quantities, e.g., a Dirichlet distribution. However, simulated examples modeling $\mathbf{p}_i$ through the Dirichlet distribution, for $i = 1, \ldots, n$, have shown that inference about these parameters is quite sensitive to the choice of its hyperparameters.

Regarding the inference procedure, MCMC methods were used to obtain samples from the resulting posterior distribution from the proposed models. In this context, different distributions can be used to generate proposals in MCMC. Normal proposal distributions were assigned for α_0, $\boldsymbol{\alpha}$ and $\mathbf{a}$ parameters. For the set of dissimilarities $\boldsymbol{\delta}$, both Gamma and Truncated Normal distributions were examined, and the posterior results were quite similar. The posterior full conditional of $\mathbf{C}$ has an analytical closed-form. Despite that, to obtain a more efficient sampling scheme, the Metropolis-Hastings step was used to update $\mathbf{C}$'s chain instead of using Gibbs Sampling in MCMC [2, 7]. Two approaches were considered to generate proposal values for $\mathbf{C}$: Multinomial proposals based on prior probabilities $\mathbf{p}$; and proposals based on mutations of some elements of the configuration $\mathbf{C}$ drawn in the previous iteration. The main advantage of the latter approach is that it allows better control of the MCMC's acceptance rate.

Some parameters of the model required specific inference strategies to become identifiable, as well as the latent groups' labels. In all proposed models, a deterministic classification procedure was performed during the MCMC method to overcome the label switching. Despite the satisfactory results obtained through this approach, other procedures can be successfully performed. As an alternative approach, one could previously estimate $\mathbf{C}$ based on a traditional cluster method, e.g., the K-means algorithm [12] — or using a crude estimate of $\mathbf{C}$ —, and then, use this estimate as a reference configuration to $\mathbf{C}$. To perform this, in each MCMC iteration, the configuration of labels $\mathbf{C}^{(j)}$ drawn in the $j-$th iteration would be switched to the configuration closest to the reference configuration of $\mathbf{C}$. The main limitation of this approach relies on the quality of the solution obtained to build the reference configuration, which can lead to unsatisfactory partitions of the n network individuals.

Since the labels associated with each latent group are arbitrary, there are $K!$ ways to represent identical groupings of the n individuals in the network. Thus, in all performed studies, the labels associated with the latent groups were relabeled to become comparable to each other. In particular, since the groups' labels are previously known when we are dealing with artificial data, the latent groups were relabeled according to the true partition.

In case study I, the performance of each of the four proposed models was analyzed using the Monks' dataset. The models considering $K = 3$ performed better than the ones with $K = 2$ latent groups. In particular, there is evidence of a better fitting of the model based on latent positions with different within-group tie probabilities, the GLPM model. Considering $K = 3$ latent groups, all models led to the same partition, which follows several authors in literature.

Still, in case study I, the performance of each of the four proposed models was analyzed using the Monks' dataset. The models considering $K = 3$ performed better than the ones with $K = 2$ latent groups. In particular, there is evidence of a better fitting of the model based on latent positions with different within-group tie probabilities, the GLPM model. Considering $K = 3$ latent groups, all models led to the same partition, well distinguishing Young Turks, Loyal Opposition, and Outcasts groups. All four proposed models' fit to this dataset showed that this class of models is flexible enough to properly clustering monks, and also to quantify between and within-group tie probabilities according to several authors in the literature.

In practical situations, the most common options for the dimension of the latent space are $d = 1$ or $d = 2$. The projection of networks or groups in latent spaces intends to allow better visualization of the network. Except in cases where interactive 3-D graphics are available, it is hard to achieve this aim considering $d > 2$. However, in cases where $K = 2$, the GLDM and LDM models do not provide projections of the groups in the latent space due to the limitations of the multidimensional scaling technique.

The main findings of this work encourage an extension of the proposed class of models to consider the number of groups as a random variable, i.e., assigning a zero truncated binomial distribution for K. In this case, it will also be necessary to specify a limit L for the number of classes K or to control its variation by choosing suitable hyperparameters in its prior distribution.

Finally, obtaining both the latent positions of nodes and groups is a challenging and promising future work. Since the groups have random latent positions, it would be necessary to propose a latent set of nodes for each group at each MCMC iteration. In addition, the switching label problem in this situation would need advanced treatment. Sequential approaches could be performed in this case, e.g., methods based on latent configurations obtained from [9]' model followed by a post-processing technique to estimate latent group positions.

References

[1] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *Journal of machine learning research*, 9(Sep):1981–2014, 2008.

[2] J. Besag, P. Green, D. Higdon, and K. Mengersen. Bayesian computation and stochastic systems. *Statistical science*, pages 3–41, 1995.

[3] R. L. Breiger, S. A. Boorman, and P. Arabie. An algorithm for blocking relational data, with applications to social network analysis and comparison with multidimensional scaling. *Journal of mathematical psychology*, 12:328–383, 1975.

[4] G. Celeux. Bayesian inference for mixture: The label switching problem. In *Compstat*, pages 227–232. Springer, 1998.

[5] G. Celeux, M. Hurn, and C. P. Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970, 2000.

[6] B. K. Fosdick, T. H. McCormick, T. B. Murphy, T. L. J. Ng, and T. Westling. Multiresolution network models. *Journal of Computational and Graphical Statistics*, 28(1):185–196, 2019.

[7] D. Gamerman and H. F. Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press, 2006.

[8] M. S. Handcock, A. E. Raftery, and J. M. Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007.

[9] P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

[10] P. N. Krivitsky and M. S. Handcock. Fitting position latent cluster models for social networks with latentnet. *Journal of Statistical Software*, 24, 2008.

[11] P. N. Krivitsky, M. S. Handcock, A. E. Raftery, and P. D. Hoff. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social networks*, 31(3):204–213, 2009.

[12] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif., 1967. University of California Press.

[13] K. V. Mardia. Some properties of clasical multi-dimesional scaling. *Communications in Statistics-Theory and Methods*, 7(13):1233–1241, 1978.

[14] K. Nowicki and T. A. B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American statistical association*, 96(455):1077–1087, 2001.

[15] P. Papastamoulis and G. Iliopoulos. An artificial allocations based solution to the label switching problem in bayesian analysis of mixtures of distributions. *Journal of Computational and Graphical Statistics*, 19(2):313–331, 2010.

[16] P. Papastamoulis and G. Iliopoulos. On the convergence rate of random permutation sampler and ecr algorithm in missing data models. *Methodology and Computing in Applied Probability*, 15(2):293–304, 2013.

[17] C. E. Rodriguez and S. G. Walker. Label switching in bayesian mixture models: Deterministic relabeling strategies. *Journal of Computational and Graphical Statistics*, 23(1):25–45, 2014.

[18] S. F. Sampson. *A novitiate in a period of change: An experimental and case study of social relationships.* PhD Thesis, Cornell University, 1968.

[19] M. Schweinberger and M. S. Handcock. Local dependence in random graph models: characterization, properties and statistical inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(3):647–676, 2015.

[20] T. A. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of classification*, 14(1):75–100, 1997.

[21] H. C. White, S. A. Boorman, and R. L. Breiger. Social structure from multiple networks. i. blockmodels of roles and positions. *American journal of sociology*, 81(4):730–780, 1976.