UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Julio 2025, Vol. 18, No. 1, pp. 87–101

Spatial Process Models for Social Network Analysis

Modelos de Procesos Espaciales para el Análisis de Redes Sociales

Juan Sosa[a]
jcsosam@unal.edu.co

Jesús David Solano Velásquez[b]
jdsolanov@unal.edu.co

Abstract

Statistical modeling of networks enables us to fully characterize the entire system as well as make predictions regarding link formation. Latent models address these tasks by incorporating non-conditional dependencies through random effects. A notable example is the Bayesian spatial process-based model, which is particularly useful for avoiding overfitting issues that may arise in latent distance space models. In this paper, we provide the computational implementation of the model and evaluate its goodness of fit and predictive performance using synthetic networks. The model demonstrates strong capabilities in replicating network statistics and estimating the corresponding latent surface. We also propose an alternative fitting approach using a case-control algorithm. Based on the estimated log-likelihood, the model exhibits good performance in terms of prediction as well as goodness of fit.

***Keywords*:** Spatial Process Models, Latent Space Models, Social Network Analysis, Bayesian Estimation, Case-Control Algorithm..

Resumen

El modelado estadístico de redes nos permite caracterizar completamente todo el sistema, así como realizar predicciones sobre la formación de enlaces. Los modelos latentes abordan estas tareas incorporando dependencias no condicionales a través de efectos aleatorios. Un ejemplo notable es el modelo Bayesiano basado en procesos espaciales, que resulta particularmente útil para evitar problemas de sobreajuste que pueden surgir en modelos de espacio de distancia latente. En este trabajo, presentamos la implementación computacional del modelo y evaluamos su bondad de ajuste y desempeño predictivo utilizando redes sintéticas y reales. El modelo demuestra fuertes capacidades para replicar estadísticas de la red y estimar

[a]Departamento de Estadística, Universidad Nacional de Colombia
[b]Departamento de Estadística, Universidad Nacional de Colombia

la superficie latente correspondiente. También proponemos un enfoque alternativo de ajuste utilizando un algoritmo de caso-control. Con base en la verosimilitud estimada, el modelo muestra un buen desempeño tanto en términos de predicción como en bondad de ajuste.

***Palabras clave*:** Modelos de Procesos Espaciales, Modelos de Espacio Latente, Análisis de Redes Sociales, Estimación Bayesiana, Algoritmo de Caso-Control..

1. Introduction

The relational structure resulting from interactions between individuals is referred to as a network. Analyzing phenomena through the lens of network structures is essential in various fields of knowledge. For instance, in epidemiology, it facilitates preventive and corrective decision-making to address the spread of viruses; in marketing, it aids in designing sales strategies based on customer purchasing behavior; and in biology, it helps study interactions between birds and plants, shedding light on seed dispersal behaviors, among many other examples.

A statistical model is a collection of probability distributions indexed by an unknown parameter vector. Statistical modeling of networks enables the identification of the data-generating process mechanism and the prediction (imputation) of unobserved (missing) values. However, traditional methods, such as the Erdős–Rényi model Erdős and Rényi (1959, 1960), are unsuitable for network modeling because individuals' relationships are typically not independent. To address this issue, several alternatives have been proposed in recent years, with one of the most popular being latent space modeling (Hoff et al., 2002).

A statistical model is a collection of probability distributions indexed by an unknown parameter vector. Statistical modeling of networks enables the identification of the data-generating process mechanism and the prediction (imputation) of unobserved (missing) values. However, traditional methods, such as the Erdős–Rényi Model Erdős and Rényi (1959, 1960), are unsuitable for network modeling because individuals' relationships are typically not independent. To address this issue, several alternatives have been proposed in recent years, with one of the most popular being latent space modeling (Hoff et al., 2002).

Early network models, such as the Random Graph Models (Gilbert, 1959), laid the foundation for understanding network structure but were limited by their assumption of independence between edges. Extensions to these models, referred to as the Generalized Random Graph Models, aimed to capture more realistic features of networks. Examples include the Configuration Models (Bender and Canfield, 1978), Small-World Models (Watts and Strogatz, 1998), and Preferential Attachment Models (Barabási and Albert, 1999), which account for clustering, degree heterogeneity, and other network characteristics.

Exponential Random Graph Models (ERGMs) (Frank and Strauss, 1986) represent another important advancement, extending generalized linear models to networks.

ERGMs allow for the inclusion of endogenous variables, such as network structural features (e.g., transitivity), alongside exogenous covariates, such as node attributes. By incorporating Markov structures, these models enable the representation of dependencies between links. However, ERGMs often struggle with high-dimensional and complex dependencies, which can limit their applicability in certain contexts.

To overcome these challenges, Latent Space Models (LSMs; (Hoff et al., 2002)) have emerged as a powerful alternative. These models adopt a marginal perspective, incorporating non-conditional dependencies through random effects. Spatial Process Models (SPMs; Linkletter 2007), a prominent subclass, assume that each individual occupies an unobserved (latent) position in a d-dimensional Euclidean space, known as the "Social Space." This assumption allows spatial latent models to uncover hidden structures in networks while mitigating overfitting issues that may arise with other approaches.

The importance of LSMs is highlighted by their extensions and applications across various fields. For instance, they have been used to analyze settings in social networks (Schweinberger and Snijders, 2003), multiview network data (Salter-Townshend and McCormick, 2017; Durante et al., 2018; Wang et al., 2019; Sosa and Betancourt, 2022), network perception data (Sewell, 2019; Sosa and Rodríguez, 2021), and dynamic networks (Sewell and Chen, 2015; Kim, 2018). These applications demonstrate the versatility of latent space models in addressing diverse research questions. For a more exhaustive review, readers are referred to Sosa and Buitrago (2021).

From a computational perspective, several tools have been developed to implement and fit these network models. In R, the `ergm` library provides a suite of functions for ERGM modeling, while the `latentnet` library (Handcock and Krivitsky, 2008) facilitates the fitting of LSMs. Practical applications and case studies using these tools can be found in Kolaczyk and Csárdi (2020), which serves as a comprehensive resource for applied researchers.

In this way, LSMs represent a significant advancement in network modeling, offering a robust framework for capturing hidden structures and addressing complex dependencies. Their extensions and computational implementations have made them an indispensable tool for analyzing networks across a wide range of applications, further cementing their importance in the field of network science. That is why, this work develops the theoretical and computational implementation of a LSM based on spatial processes, which addresses the overfitting issues inherent in the distance LSM. Here, we provide a comprehensive description of the SPM, a detailed explanation of the Markov chain Monte Carlo algorithm (e.g., Gamerman and López 2006), and an exhaustive evaluation of its goodness of fit and predictive performance, using both synthetic and real networks. Additionally, we propose and implement an alternative approach to fit the model using the case-control algorithm from Raftery et al. (2012).

The paper is structured into six sections. The Introduction discusses the importance of network modeling and introduces latent space models (LSMs). Section 2

presents the theoretical foundation of LSMs and Spatial Process Models (SPMs). Section 3 details the computational methodology using Bayesian estimation with MCMC. Section 4 introduces the case-control algorithm to improve computational efficiency. Section 5 illustrates the application of these methods using synthetic networks, evaluating model fit and predictive performance. Finally, the Discussion summarizes findings, highlights the efficiency of the proposed approaches, and suggests directions for future research.

2. Latent Space Modeling

In statistical network modeling, we associate a probability distribution with the adjacency matrix $\mathbf{Y} = [y_{i,j}]$ of a given network to capture essential features such as homophily and structural equivalence. Since network observations are inherently dependent, traditional modeling approaches, such as generalized linear models, are suboptimal. To address this challenge, we use latent space models (LSMs) as a means to incorporate the dependence structures between observations directly into the modeling framework.

2.1. Latent Spatial Models

Let n denote the number of actors in the network. Latent space models (LSMs) estimate the interaction probabilities $\pi_{i,j} = \mathsf{Pr}(y_{i,j} = 1 \mid \beta_0, \boldsymbol{\beta}, \zeta_{i,j})$, where $i, j = 1, \ldots, n$ and $i < j$ (utilizing only the upper triangular portion of the adjacency matrix for undirected networks). The sampling distribution, which represents the data-generating process, is given by:

$$y_{i,j} \overset{\text{ind}}{\sim} \mathsf{Bernoulli}(\pi_{i,j}) \qquad \Longleftrightarrow \qquad \mathsf{p}(\mathbf{Y} \mid \boldsymbol{\pi}) = \prod_{i<j} \pi_{i,j}^{y_{i,j}} (1 - \pi_{i,j})^{1-y_{i,j}},$$

where $\boldsymbol{\pi} = [\pi_{i,j}]$ is the matrix of interaction probabilities. The linear predictor is specified as:

$$\eta_{i,j} = \mathsf{logit}(\pi_{i,j}) = \log\left(\frac{\pi_{i,j}}{1 - \pi_{i,j}}\right) = \beta_0 + \boldsymbol{x}_{i,j}^{\top}\boldsymbol{\beta} + \zeta_{i,j},$$

where $\boldsymbol{x}_{i,j} = (\mid x_{i,1} - x_{j,1} \mid, \ldots, \mid x_{i,p} - x_{j,p} \mid)$ denotes the vector of observed covariates, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is the corresponding parameter vector, and $\zeta_{i,j}$ accounts for dependencies between $y_{i,j}$ that are not explained by the fixed effects $\beta_0 + \boldsymbol{x}_{i,j}^{\top}\boldsymbol{\beta}$. While the logit function serves as the default link function, alternative link functions, such as the probit function, can also be employed. The sampling distribution implies that $y_{i,j}$ are conditionally independent given the interaction probabilities $\pi_{i,j}$. However, this conditional independence does not extend to marginal independence of $y_{i,j}$, as is often assumed in generalized linear models.

Hoff et al. (2002) propose embedding actors in a d-dimensional Euclidean space, referred to as *social space*. Typically, $d = 2$ is chosen to facilitate visuali-

zation and interpretation of the relational system, although higher-dimensional spaces can also be used. In this framework, each actor is assigned a latent position $\boldsymbol{z}_i = (z_{i,1}, \ldots, z_{i,d}) \in \mathbb{R}^d$, representing unobserved (latent) characteristics of the individual. The random effects in the latent factor model are then reformulated as:

$$\zeta_{i,j} = -\|\boldsymbol{z}_i - \boldsymbol{z}_j\|, \tag{1}$$

where $\|\cdot\|$ denotes the Euclidean norm. Greater values of $\|\boldsymbol{z}_i - \boldsymbol{z}_j\|$ indicate larger distances between nodes in Social Space. The negative sign in Equation (1) ensures that as the distance increases, the probability of a link between nodes i and j decreases. This structure naturally incorporates homophily and transitivity, which are reflected in $\boldsymbol{x}_{i,j}$ and $\zeta_{i,j}$, respectively. Such a model is known as distance LSM. Lastly, the model without considering exogenous data is given by

$$\eta_{i,j} = \mathsf{logit}(\pi_{i,j}) = \beta_0 - \|\boldsymbol{z}_i - \boldsymbol{z}_j\|. \tag{2}$$

Bayesian methods are commonly employed for estimation as they provide a natural framework for parameter regularization through prior distributions. In this approach, prior distributions are assigned to each regression parameter β_h, for $h = 0, 1, \ldots, p$, as well as to the latent positions $\boldsymbol{z}_i$, for $i = 1, \ldots, n$. Commonly used priors that perform well in practice are:

$$\beta_h \sim \mathsf{N}(0, \psi_\beta) \qquad \text{and} \qquad \boldsymbol{z}_i \sim \mathsf{N}_d(\mathbf{0}, \psi_z \mathbf{I}),$$

where ψ_β and ψ_z are hyperparameters (fixed values chosen properly to ensure model performance), and $\mathbf{I}$ denotes the identity matrix. Since the full conditional distributions do not follow standard forms, estimation typically relies on Metropolis-Hastings within a Markov chain Monte Carlo (MCMC) framework (e.g., Gamerman and López 2006). Alternatively, variational inference methods (e.g., Blei et al. 2017) can also be applied for computational efficiency. An estimate for $\pi_{i,j}$ is:

$$\hat{\pi}_{i,j} = \mathsf{E}(\pi_{i,j} \mid \mathbf{Y}) = \frac{1}{B} \sum_{b=1}^{B} \frac{\exp \eta_{i,j}^{(b)}}{1 + \exp \eta_{i,j}^{(b)}},$$

where

$$\eta_{i,j}^{(b)} = \beta_0^{(b)} + \boldsymbol{x}_{i,j}^\top \boldsymbol{\beta}^{(b)} - \|\boldsymbol{z}_i^{(b)} - \boldsymbol{z}_j^{(b)}\|,$$

and $\beta_0^{(b)}$, $\boldsymbol{\beta}^{(b)}$, and $\boldsymbol{z}_1^{(b)}, \ldots, \boldsymbol{z}_n^{(b)}$ are samples from the posterior distribution of $\beta_0, \boldsymbol{\beta}, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, for $b = 1, \ldots, B$.

2.2. Spatial Process Models

Distance LSMs may suffer from overfitting problems (a modeling issue in which a function corresponds uniquely with the dataset used for estimation; Linkletter 2007, p. 38), given that they are not oriented to estimate the interaction probabilities for nodes outside the observed sample. Thus, Linkletter (2007) propose to flexibilize the model by letting

$$\eta_{i,j} = \mu - \mid z_i - z_j \mid,$$

where $z_i \equiv z(\boldsymbol{x}_i)$ is a real-valued latent function of the covariates vector $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,p})$. This formulation is known as the Spatial Process Model (SPM). Under this formulation, μ represents the average log-ratio of connection for two nodes with the same latent process score. Similar to distance LSMs, nodes with similar values for z_i are more likely to be connected. Additionally, the absolute value of $z_i - z_j$ induces transitivity, which is a fundamental aspect of social network modeling. Note that in contrast to distance LSMs, where the $\boldsymbol{z}_i$ are vectors, for SPMs the distance between the z_i is measured through the absolute value, as they are scalars.

For carrying out Bayesian estimation, prior distributions must be assigned to model parameters. In this way, μ is assigned a Normal distribution $\mu \sim \mathsf{N}(0, \psi_\mu)$, and each z_i is assigned a Gaussian Process (GP; e.g., Rasmussen and Williams 2006). Specifically, we let

$$z(\boldsymbol{x}) = \sum_{r=1}^{m} \alpha_r k(\boldsymbol{x} - \boldsymbol{w}_r), \tag{3}$$

where $k(\cdot)$ is a p-dimensional Gaussian kernel, $\alpha_1, \ldots, \alpha_m$ is a sequence of model parameters, and $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m$ represents a grid obtained from a Latin hypercube design (a statistical method to draw a pseudo-random sample of the values of the parameters of a given joint distribution).

Assuming that the kernel bandwidth varies in the direction of each component, we consider the parameter vector $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_p)$ so that $k(\cdot)$ can be expressed as

$$k(\boldsymbol{x}_i - \boldsymbol{w}_r) = \prod_{h=1}^{p} \rho_h^{(w_{r,h} - x_{i,h})^2},$$

where $w_{r,h}$ and $x_{i,h}$ are the h-th elements of $\boldsymbol{w}_r$ and $\boldsymbol{x}_i$, respectively, $\rho_h = e^{-\frac{1}{2\sigma_h^2}}$, and σ_h is the standard deviation of the kernel in the h-th direction. Finally, to complete the prior specification, the prior distributions for the expansion coefficients and the bandwidths are $\boldsymbol{\alpha} \sim \mathsf{N}_m(\mathbf{0}, \mathbf{I})$ and $\rho_h \sim \mathsf{U}(0, 1)$, for $h = 1, \ldots, p$.

Under this formulation, an estimate of the process at a given point $\boldsymbol{x}_0$ can be obtained using posterior samples as follows:

$$\hat{z}(\boldsymbol{x}_0) = \mathsf{E}(z(\boldsymbol{x}_0) \mid \boldsymbol{y}) = \frac{1}{B} \sum_{b=1}^{B} \sum_{r=1}^{m} \alpha_r^{(b)} k^{(b)}(\boldsymbol{x}_0 - \boldsymbol{w}_r),$$

where $\mu^{(b)}, \boldsymbol{\alpha}^{(b)}$, and $\boldsymbol{\rho}^{(b)}$ are samples from the posterior distribution of $\mu, \boldsymbol{\alpha}, \boldsymbol{\rho}$, for $b = 1, \ldots, B$, and

$$k^{(b)}(\boldsymbol{x}_0 - \boldsymbol{w}_r) = \prod_{h=1}^{p} \rho_h^{(b)^{(w_{r,h} - x_{i,h})^2}}.$$

Finally, the interaction probabilities $\pi_{i,j}$, given vectors of covariates $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, can be estimated as:

$$\hat{\pi}_{i,j} = \frac{1}{B} \sum_{b=1}^{B} \frac{\exp\{\eta_{i,j}^{(b)}\}}{1 + \exp\{\eta_{i,j}^{(b)}\}},$$

where

$$\eta_{i,j}^{(b)} = \mu^{(b)} - \mid z^{(b)}(\boldsymbol{x}_i) - z^{(b)}(\boldsymbol{x}_j) \mid .$$

All the necessary code to reproduce the model fitting and replicate our findings is available in the repository: `https://github.com/DavidSolan0/bayesian_spatial_process_models_social_network_analysis/tree/repo_refactor`.

3. Illustration: Latent Process Model using MCMC

Here, we illustrate the model's properties by analyzing a synthetic dataset. The network is generated by setting $n = 40$, $\mu = -0.5$, and a covariate process defined as $g(x, y) = 1.5\, x^2 \exp\{x^2\}$. Covariates are randomly generated within $[0, 1] \times [0, 1]$ for each actor in the network. The model is then fitted using a MCMC algorithm with a burn-in period of 50,000 iterations, followed by 10,000 effective iterations with systematic thinning every 10 iterations, resulting in a total of 150,000 iterations. From now on, specifications for burn-in, sample size, and thinning are as described here unless stated otherwise.

Figure 1 presents the estimated surface, which closely resembles $g(x, y)$. This result provides confidence that parameter estimation is not an issue and that the parameters are being accurately recovered.

To assess the goodness-of-fit of the model, we follow the approach outlined in Gelman et al. (2013), where the posterior predictive distribution is used to evaluate

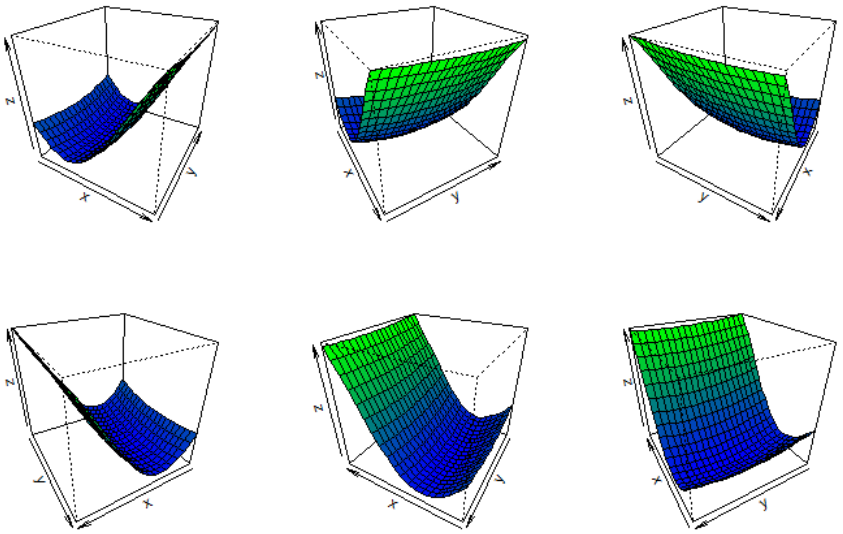
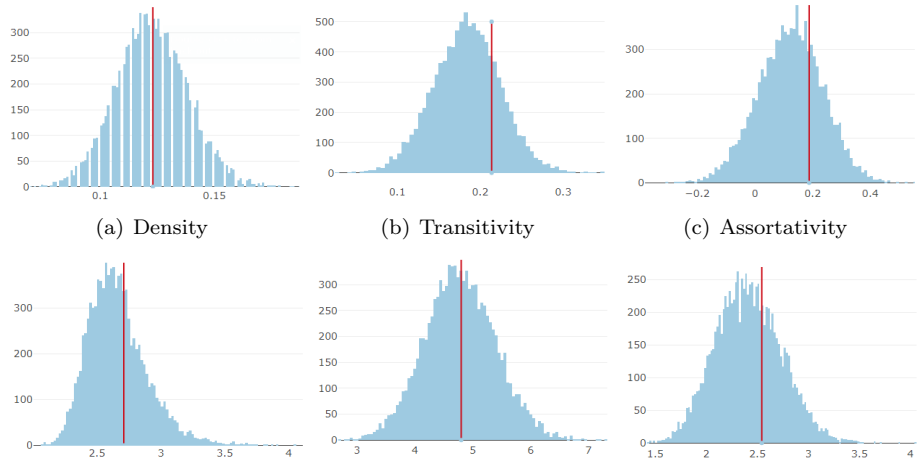


Figure 1: Estimated surface for the synthetic dataset.

a set of predetermined test statistics. If the observed values of these statistics are typical within their corresponding posterior predictive distributions, we have sufficient evidence to conclude that the model demonstrates good internal performance. In this analysis, we consider key structural network characteristics, including density, transitivity, assortativity, average geodesic distance, and node degree (mean and standard deviation). Figure 2 illustrates that the model successfully replicates these statistics, indicating that it performs well in fitting the data.



(a) Density (b) Transitivity (c) Assortativity

(d) Mean distance (e) Mean degree (f) SD degree

Figure 2: Goodness-of-fit evaluation for the synthetic dataset. The red line corresponds to the observed value.

To evaluate the predictive performance of the model, we conduct a 5-fold cross-validation experiment (e.g., James et al. 2023), where dyads are randomly assigned to five folds. For each fold, all dyads in that fold are excluded from the training data and predicted using a model fitted on the remaining folds. After obtaining the predictions, we calculate a confusion matrix, the corresponding Receiver Operating Characteristic (ROC) curve, and the Area Under the Curve (AUC; e.g., Fawcett 2006). The AUC quantifies a classifier's ability to distinguish between classes, with values closer to 1 indicating excellent performance, 0.5 representing random guessing, and values below 0.5 indicating performance worse than random. Figure 3 shows the ROC curves for each fold in the cross-validation experiment. The model demonstrates strong predictive performance, with AUC values ranging from 0.7 to 0.9. The corresponding AUC values for the folds are $0.73, 0.73, 0.84, 0.73, 0.76$, resulting in a mean AUC of 0.76.

Another simulation study, involving a synthetic network generated with $\mu = -0.7$, $n = 50$, and $g(x, y) = 1.5 \exp\{x^2\} \sin((x + y)^2)$ as part of the data-generating process, demonstrates very promising results. The test statistics closely match their corresponding posterior predictive distributions (not shown here for brevity), and the mean AUC is approximately 0.7. These results highlight the model's strong

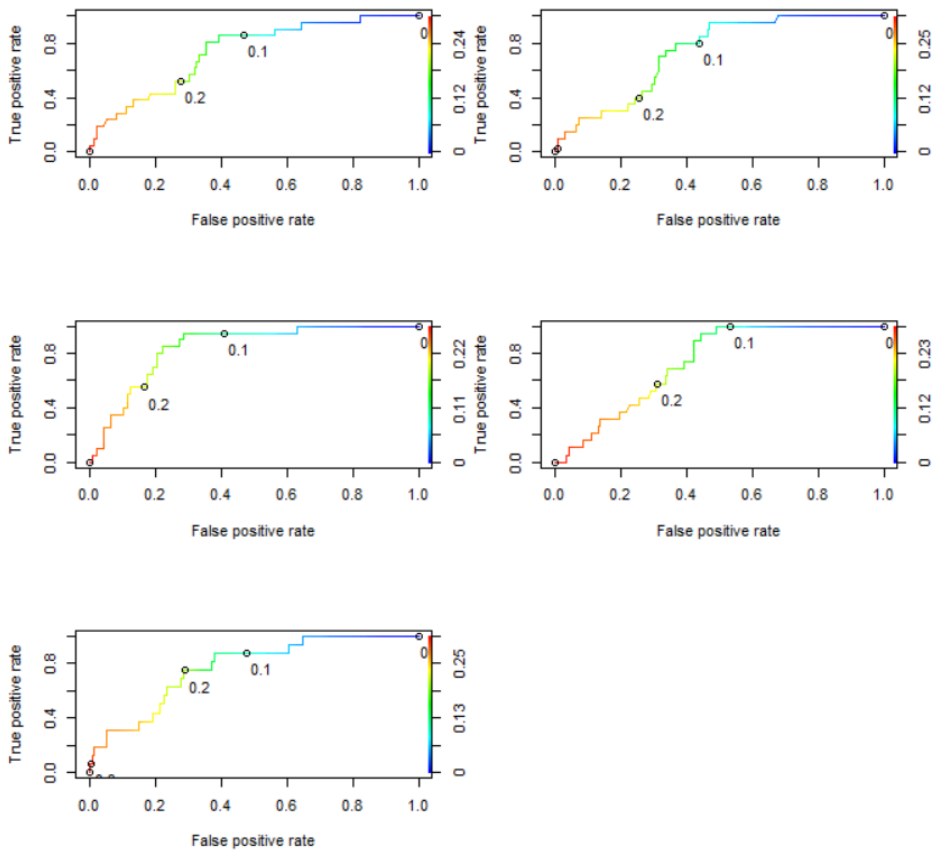


Figure 3: ROC curves of the cross-validation experiment for the synthetic dataset.

performance in terms of both goodness-of-fit and predictive accuracy.

4. Case-Control Algorithm

The computational complexity of the latent space model presented in Section 2 for a network with n nodes is $O(n^2)$, which makes its implementation infeasible for large networks. To address this issue, Raftery et al. (2012) propose replacing the full likelihood function in the MCMC procedure with an unbiased estimate based on the epidemiological approach of cases and controls, reducing the computational cost to $O(n)$. In the network context, cases correspond to the formation of a link, while absences are treated as controls.

To determine which factors are essential for link creation, the log-likelihood can

be rewritten as:

$$l = \log \mathsf{p}(\mathbf{Y} \mid \boldsymbol{\pi}) = \sum_{i=1}^{N} l_i, \tag{4}$$

where

$$\begin{aligned} l_i &= \sum_{i \neq j} (\eta_{i,j} y_{i,j} - \log(1 + \exp\, \eta_{i,j})) \\ &= \sum_{j \neq i, y_{i,j}=1} (\eta_{i,j} - \log(1 + \exp\, \eta_{i,j})) + \sum_{j \neq i, y_{i,j}=0} (-\log(1 + \exp\, \eta_{i,j})) \\ &= l_{i,1} + l_{i,0}. \end{aligned} \tag{5}$$

Thus, a point estimate of $l_{i,0}$ can be obtained using an unbiased estimator for the total under simple random sampling. Let $N_{i,0}$ and $n_{i,0}$ represent the total number of nodes such that $y_{i,j} = 0$ and the corresponding sample size, respectively. By choosing $n_{i,0}$ small enough to reduce computational time, the point estimate $\tilde{l}_{i,0}$ is given by:

$$\tilde{l}_{i,0} = \frac{N_{i,0}}{n_{i,0}} \sum_{k=1}^{n_{i,0}} \log(1 + \exp\, \eta_{i,j}). \tag{6}$$

The right-hand term of Equation (5) can be approximated using stratified random sampling across M strata, which are defined based on the concept of node closeness. This leads to the following decomposition of Equation (4):

$$\begin{aligned} l_i = &\sum_{j \neq i, y_{i,j}=1} (\eta_{i,j} - \log(1 + \exp\, \eta_{i,j})) + \sum_{j: D_{i,j}=2} \log(1 + \exp\, \eta_{i,j}) \\ &+ \cdots + \sum_{j: D_{i,j}=M} \log(1 + \exp\, \eta_{i,j}), \end{aligned}$$

where $D_{i,j}$ represents the geodesic distance between nodes i and j. Therefore, an unbiased estimator for l_i based on stratified random sampling is then defined as:

$$\hat{l}_i = \sum_{j \neq i, y_{i,j}=1} (\eta_{i,j} - \log(1 + \exp\, \eta_{i,j})) + \sum_{h=2}^{M} \frac{N_{i,h}}{n_{i,h}} \sum_{j: D_{i,j}=h} \log(1 + \exp\, \eta_{i,j}), \tag{7}$$

where $N_{i,h}$ is the number of nodes j such that $D_{i,j} = h$, $n_{i,h}$ is the number of nodes j sampled with $D_{i,j} = h$, and $D_{i,j}$ represents the geodesic distance between nodes i and j.

4.1. Sample Size Determination

The sample size $n_{i,h}$ can be determined as follows:

1. Determine r: Compute r using $n_{i,0} = r\bar{d} = n_0$, where $\bar{d}$ is the mean degree of the network.

2. Run a pilot MCMC:

   (a) For each iteration b, compute:

$$\begin{aligned}\Delta\tilde{l}_i^{(b)} &= \tilde{l}_i(z_i^{(b)\,*}) - \tilde{l}_i(z_i^{(b)}) \\ &= l_{i,1}(z_i^{(b)\,*}) - l_{i,1}(z_i^{(b)}) + \sum_h \left(\tilde{l}_{i,h}(z_i^{(b)\,*}) - \tilde{l}_{i,h}(z_i^{(b)})\right) \\ &= \Delta l_{i,1}^{(b)} + \sum_h \Delta\tilde{l}_{i,h}^{(b)},\end{aligned}$$

   where $z_i^{(b)\,*}$ is the updated value for z_i in iteration b.

   (b) Define:

$$w_{i,h}^{(b)} = \left| \frac{\Delta\tilde{l}_{i,h}^{(b)}}{\sum_{g=2}^{M} \Delta\tilde{l}_{i,g}} \right|.$$

   (c) Compute:

$$w_{i,h} = \frac{1}{B-1} \sum_{b=1}^{B-1} w_{i,h}^{(b)},$$

   where B is the number of iterations in the burn-in period.

3. Set the size of stratum h for actor i:

$$n_{i,h} = \frac{n_{i,0}\, w_{i,h}}{\sum_{g=2}^{M} w_{i,g}},$$

   for $h = 2, \ldots, M$ and $i = 1, \ldots, n$.

4.2. Algorithm

Given the point estimates for $l_{i,0}$ and l_i, as defined in Equations (6) and (7), the estimation algorithm proceeds as follows:

1. Execute the pilot MCMC:

   (a) Define $n_{i,0}$, for each actor $i = 1, \ldots, N$.

   (b) Using the estimate $\tilde{l}_{i,0}$ for $l_{i,0}$ from Equation (6), execute the MCMC algorithm, replacing l with:

$$\tilde{l} = \sum_{i=1}^{N} \left(l_{i,1} + \tilde{l}_{i,0}\right).$$

2. For each geodesic distance $h = 2, \ldots, M$ and each node $i = 1, \ldots, n$:

   (a) Determine $n_{i,h}$.

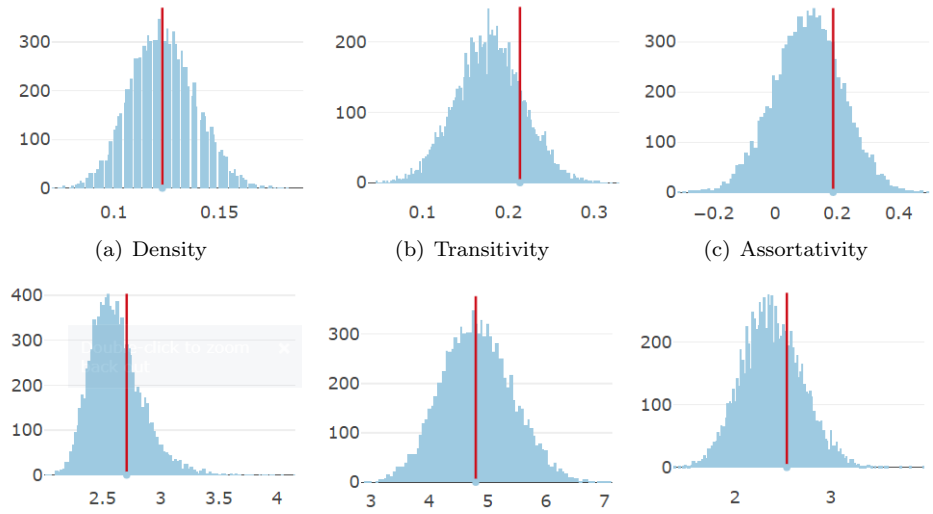
(b) Sample $n_{i,h}$ nodes such that $y_{i,j} = 0$ and geodesic distance $D_{i,j} = h$.

3. Execute the main MCMC algorithm: Replace the usual log-likelihood with its unbiased estimator $\hat{l}$ from Equation (7).

The estimator in Equation (7) is unbiased irrespective of the value of r. However, the choice of r affects the efficiency of the estimator, enabling adjustments for more (or less) computational efficiency based on the researcher's needs.

5. Illustration: Latent Process Model using Case-Control Algorithm

Figure 4 shows the posterior predictive distributions obtained using the Case-Control algorithm described in Section 4 for the same set of test statistics used when fitting the model with MCMC. The resulting distributions closely resemble those generated via MCMC in Figure 2. This finding demonstrates that the approximation performs well while significantly reducing computational time.



(a) Density (b) Transitivity (c) Assortativity

(d) Mean distance (e) Mean degree (f) SD degree

Figure 4: Goodness-of-fit evaluation for the synthetic dataset using the Case-Control algorithm. The red line corresponds to the observed value.

6. Discussion

This article implements, for the first time, the sampling strategy proposed by Raftery et al. (2012) in the context of a Latent Process Model Linkletter (2007). Our findings demonstrate that the methodology performs well, as we obtain comparable results using both MCMC and the Case-Control algorithm. Therefore, this methodology is recommended for large datasets due to its equivalent performance relative to the original method, while significantly reducing computational time.

Finally, it would be valuable to investigate the goodness-of-fit and predictive power of the spatial model proposed by Ciminelli et al. (2019) using the Case-Control algorithm. This approach captures the spatial correlation inherent in social networks, simultaneously models the actors' features, and, based on these features and the network links, estimates their latent locations in the social space along with the underlying Spatial Process.

Statements and Declarations

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this part.

Recibido: Diciembre de 2024
Aceptado: Julio de 2025

Referencias

A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

E. A. Bender and E. R. Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307, 1978.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773.

J. T. Ciminelli, T. Love, and T. T. Wu. Social network spatial model. *Spatial statistics*, 29:129–144, 2019.

D. Durante, D. B. Dunson, et al. Bayesian inference and testing of group differences in brain networks. *Bayesian Analysis*, 13(1):29–58, 2018.

P. Erdős and A. Rényi. On random graphs i. *Publicationes Mathematicae*, 6: 290–297, 1959.

P. Erdős and A. Rényi. On the evolution of random graphs. *Magyar Tudományos Akadémia Matematikai Kutató Intézetének Kőzleményei*, 5:17–61, 1960.

T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8): 861–874, 2006.

O. Frank and D. Strauss. Markov graphs. *Journal of the american Statistical association*, 81(395):832–842, 1986.

D. Gamerman and J. M. López. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. CRC Press, Boca Raton, FL, 2006.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL, 3rd edition, 2013. ISBN 978-1-4398-4095-5.

E. N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4): 1141–1144, 1959.

M. S. Handcock and P. N. Krivitsky. Fitting latent cluster models for networks with latentnet. *Journal of Statistical Software*, 24(05), 2008.

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.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning with Applications in Python*. Springer, 2023. ISBN 978-1-0716-2803-8. doi: 10.1007/978-1-0716-2804-5.

B. Kim. *Latent Modeling of Dynamic Social Networks*. The Pennsylvania State University, 2018.

E. D. Kolaczyk and G. Csárdi. *Statistical analysis of network data with R,2nd edn,*. Springer, 2020.

C. D. Linkletter. *Spatial process models for social network analysis*. PhD thesis, Citeseer, 2007.

A. E. Raftery, X. Niu, P. D. Hoff, and K. Y. Yeung. Fast inference for the latent space network model using a case-control approximate likelihood. *Journal of Computational and Graphical Statistics*, 21(4):901–919, 2012.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006. doi: 10.7551/mitpress/3206.001.0001.

M. Salter-Townshend and T. H. McCormick. Latent space models for multiview network data. *The annals of applied statistics*, 11(3):1217, 2017.

M. Schweinberger and T. A. Snijders. Settings in social networks: A measurement model. *Sociological Methodology*, 33(1):307–341, 2003.

D. K. Sewell. Latent space models for network perception data. *Network Science*, 7(2):160–179, 2019.

D. K. Sewell and Y. Chen. Latent space models for dynamic networks. *Journal of the American Statistical Association*, 110(512):1646–1657, 2015.

J. Sosa and B. Betancourt. A latent space model for multilayer network data. *Computational Statistics & Data Analysis*, page 107432, 2022.

J. Sosa and L. Buitrago. A review of latent space models for social networks. *Revista Colombiana de Estadística*, 44(1):171–200, 2021.

J. Sosa and A. Rodríguez. A latent space model for cognitive social structures data. *Social Networks*, 65:85–97, 2021.

L. Wang, Z. Zhang, D. Dunson, et al. Common and individual structure of brain networks. *The Annals of Applied Statistics*, 13(1):85–112, 2019.

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.