Universidade Federal de São Carlos
Centro de Ciências Exatas e de tecnologia
Departamento de Estatística

Inferência em Grafos Aleatórios Exponenciais através de métodos MCMC

Guilherme Antonio Alves de Lima

Trabalho de Conclusão de Curso

Guilherme Antonio Alves de Lima

Inferência em Grafos Aleatórios Exponenciais através de métodos MCMC

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por "Guilherme Antonio Alves de Lima" e aprovado pela banca examinadora.

São Carlos, 14 de Janeiro de 2021.

Banca Examinadora

- Andressa Cerqueira (Orientadora)
- Luis Ernesto Bueno Salasar
- Rafael Bassi Stern

Resumo

Neste Trabalho de Conclusão de Curso, estudamos Redes Aleatórias através de Grafos Aleatórios, explorando a literatura para introduzirmos o tópico de Redes e sua interpretação computacional e estatística. Este trabalho explora técnicas e modelos, destacando o modelo de Grafos Aleatórios Exponenciais e Métodos de Monte Carlo baseado em Cadeia de Markov (MCMC), não vistas na graduação, em um tópico interdisciplinar que apresenta vários usos em diversas áreas do conhecimento. Uma das suas aplicações, redes aéreas, é o foco de uma aplicação prática do modelo, onde modelamos e analisamos dados da rede aérea brasileira utilizando dados disponibilizados pela Agência Nacional de Aviação Civil.

Palavras-chave: *ANAC, Grafos Aleatórios Exponenciais, Inferência Bayesiana, MCMC.*

Sumário

Capítulo 1

Introdução

Uma rede é um grupo de pontos conectados por linhas, que podem representar vários objetos de diversas áreas e suas interações entre si, matematicamente, representamos redes através de grafos.

Em um mundo cada vez mais conectado, diversos aspectos da vida moderna se veem compostos por redes de diversos tipos, desde as amplamente reconhecidas como a Internet e redes de transporte, outras sentidas mas não tão notadas como redes de amizade e outras não tão óbvias para o público como redes de citação e redes ecológicas (Newman (2012)).

Independente do tipo de rede, existe um interesse crescente em estudar suas estruturas, para esta finalidade, representamos cada rede através de grafos, onde representamos cada objeto ou ponto da rede por um vértice, e cada conexão ou interação por uma aresta conectando dois vértices, formando a imagem representada na Figura 1.1.

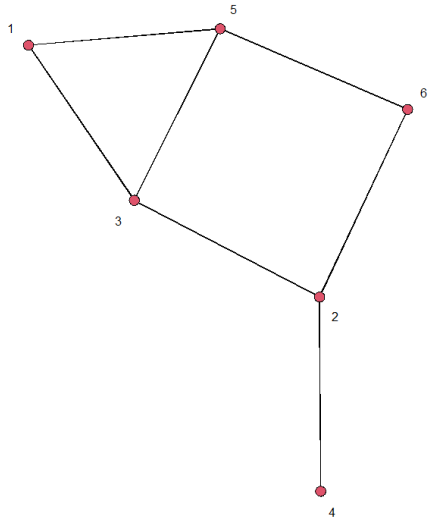

Figura 1.1: Exemplo de Grafo

Existem várias razões para estudarmos a estrutura e o comportamento de redes, representadas pela sua forma e pelos padrões que ela segue. Utilizando os exemplos já mencionados, o estudo da Internet, tendo como vértices seus servidores e arestas as transmissões de dados, estudar sua estrutura é fundamental para identificar o fluxo de dados, e a partir dessa informação, melhorar a estrutura e design da rede para otimizar o roteamento de dados e oferecer uma experiência mais estável e rápida para o usuário.

Para redes de transporte, como redes aéreas, ferroviárias, rodoviárias entre outras, a utilidade do seu estudo é evidente, sendo fundamental para o planejamento de rotas, medir demandas e custos para o transporte de passageiros ou carga.

Para redes de amizade, seu estudo pode identificar vários fenômenos sociais, como indivíduos de alta influência em um grupo, e até a maior chance de um individuo se conectar a outro quando há vários amigos em comum.

Para o exemplo de redes de citações, seu estudo permite identificar várias propriedades da criação acadêmica, identificando artigos altamente influentes e conexões dentro de uma área e entre áreas do conhecimento.

Por fim, um exemplo de rede ecológica é simplesmente a cadeia alimentar de um ecossistema, um conceito bem conhecido e que intuitivamente se usa grafos para representá-lo. Uma vantagem desta representação é a possibilidade de ter grafos direcionados e que apresentam pesos nas suas conexões, facilitando o estudo do fluxo de energia em um ecossistema.

Estes são alguns exemplos de aplicações do estudo de redes entre muitos outros, motivados por sua utilidade, temos interesse em estudar este conceito utilizando técnicas estatísticas, explorando em particular modelos não vistos no curso de graduação e principalmente o modelo de Grafos Aleatórios Exponenciais.

1.1 Objetivos

O objetivo principal do Trabalho de Conclusão de Curso é estudar o modelo de Grafos Aleatórios Exponencias e estudar sua aplicação em dados reais.

Nosso enfoque, primeiramente, é o estudo teórico de um método de inferência estatística aplicado à Grafos Aleatórios Exponenciais proposto em Caimo e Friel (2011). Nesse trabalho, os autores propõe um método de inferência com objetivo de estimar os parâmetros do modelo. Para isso é tomado um enfoque Bayesiano utilizando o Método

de Monte Carlo Baseado em Cadeias de Markov (MCMC).

Para tal, iremos necessitar um estudo teórico de métodos MCMC para simular amostras de grafos do Modelo de Grafos Exponencias e a análise da implementação desses métodos através de estudos de simulação.

Estudaremos a eficiência do método proposto em dados simulados de redes e também o seu uso para inferir os parâmetros do Modelo.

Por fim, iremos utilizar o método de inferência estudado para analisar uma rede real ajustando o Modelo de Grafos Exponenciais aos dados e inferindo os parâmetros do modelo. A rede real que será estudada corresponde a rede aérea brasileira, construída a partir de voos nacionais entre aeroportos.

1.2 Capítulos

No Capítulo 2, definiremos grafos e alguns conceitos importantes que serão utilizados nesse trabalho, apresentaremos sua representação matemática e definiremos os conceitos de Grafos Aleatórios e Grafos Aleatórios Exponenciais.

No Capítulo 3, exploraremos os conceitos do método de Monte Carlo Baseado em Cadeias de Markov e o principal algoritmo a ser utilizado no nosso trabalho.

No Capítulo 4, implementaremos a simulação por MCMC através de exemplos, utilizando os pacotes "ergm" (Hunter *et al.*, 2008) e "Bergm" (Caimo e Friel, 2014) em um banco de dados normalmente utilizado como exemplo e redes simuladas.

No Capítulo 5, processaremos dados reais da rede Aérea Brasileira de 2019 em um Grafo e aplicaremos o algoritmo estudado neste trabalho para determinar qual modelo melhor descreve a rede aérea.

Por fim, no Capítulo 6, apresentamos e comentamos os resultados obtidos na elaboração deste trabalho. Além disso, discutimos possíveis direções futuras do estudo de redes na estatística.

No fim deste trabalho, disponibilizamos figuras adicionais, que não foram incluídas no seu corpo, no Apêndice A, e apresentamos as Referencias Bibliográficas.

Capítulo 2

Grafos e Grafos aleatórios

Neste capítulo, exploramos a definição de Grafos, introduzimos alguns conceitos principais que serão utilizados ao longo do trabalho e um método de representá-los que será necessário para a implementação computacional. Por fim, introduzimos o modelo de Grafos aleatórios e o subsequente modelo de Grafos Aleatórios Exponenciais, o qual focaremos nesse trabalho.

2.1 Grafos

Como anteriormente afirmado, uma rede é um grupo de pontos conectados por linhas, que representam objetos e as conexões entre si, respectivamente. No contexto matemático, representamos essas redes visualmente e formalmente através de grafos, onde temos vértices representando os objetos e arestas representando as conexões, em que n é o número de vértices e m o número de arestas.

Para ilustrar certos aspectos e propriedades de grafos, retomamos a Figura 1.1. Nesse exemplo de grafo, temos 6 vértices, correspondendo a 6 objetos, e 7 arestas, correspondendo a 7 conexões dentre esses objetos. Nesse caso, temos no máximo uma aresta entre dois vértices e não temos vértices que conectem a si mesmos. Embora seja possível a representação de um grafo que não apresente esses dois fatos, um grafo com com essas restrições é chamado de Grafo Simples, e neste trabalho só iremos trabalhar com esse tipo de grafo.

Logo, iremos apresentar em seguida os principais conceitos e propriedades de Grafos, que serão importantes para a elaboração do restante do nosso trabalho.

Direção: Um grafo pode apresentar arestas direcionadas ou não direcionadas. O

exemplo na Figura 1.1 representa um grafo com arestas não direcionadas. O uso de arestas direcionadas é útil quando é interessante estudar relações específicas entre vértices, como nos exemplos mencionados de cadeia alimentar, redes de citação e redes aéreas. Nesses casos, as arestas entre vértices são representadas por flechas, para indicar a direção da influência.

Conexão: Como já afirmado, a presença de uma aresta entre dois vértices representa uma conexão no grafo. Desse modo, contar o número de conexões é uma estatística básica e importante de grafos. No caso direcionado, temos diferentes tipos de conexões possíveis dependendo da direção das arestas e de conexões mútuas, que podem indicar certos comportamentos da rede como reciprocidade.

Grau: O grau de um vértice do grafo representa o número de arestas conectadas a ele. Graus são de particular interesse, pois não só representam centralidade no grafo, como no exemplo de um indivíduo de alta influência como mencionado em redes de amizade, mas também a sua distribuição na rede pode revelar informações e comportamentos relevantes para a análise do grafo.

Pesos: Grafos podem apresentar conexões mais complexas, como um fluxo de passageiros entre aeroportos, que indicam algo além da simples presença de conexões. Em casos como esses, cabe utilizar pesos para as arestas, indicando conexões mais fortes ou mais fracas comparativamente. Normalmente representamos arestas entre vértices por 0, caso não exista, e 1, caso ela exista, porém se utilizarmos pesos, podemos ter valores como por exemplo 0.5 e 2, indicando conexões mais fracas ou mais fortes respectivamente.

Caminho: Como seu nome implica, esse conceito representa o caminho tomado através de arestas partindo de um vértice e resultando em outro, com seu comprimento mensurado pelo número de arestas utilizadas. Enquanto existem diversos caminhos possíveis entre dois vértices que apresentam pelo menos um caminho entre si, é interessante o uso de caminhos mais específicos com limites impostos. Um importante exemplo é o caminho geodésico, sendo o menor caminho possível entre dois vértices. O estudo dos caminhos em um grafo pode revelar varias informações úteis, visto que em redes como a internet por exemplo, deseja-se otimizar seus comprimentos.

Transitividade: Semelhante ao conceito matemático, esta propriedade implica conexões não diretas entre vértices, porém uma transitividade perfeita, onde para quaisquer pontos a, b e c de um grafo, onde a e b estão ambos conectados por arestas à c, implica uma conexão por aresta entre a e b, não é um conceito apropriado para grafos. Neste

caso, seria apropriado apenas criar uma nova conexão por aresta entre estes vértices, resultando em um grafo composto por subgrafos completamente conectados, sendo um caso extremamente raro e geralmente pouco vantajoso para a análise de redes . Para grafos, é interessante a transitividade parcial, onde por exemplo, em uma rede de amizade, uma pessoa com dois amigos não garante que estes dois sejam amigos entre si, porém é mais provável que este seja o caso comparado com outras duas pessoas quaisquer.

Triângulos e estrelas: Triângulos em um grafo não direcionado são definidos através de 3 vértices, cada um conectado com os dois outros por arestas, enquanto estrelas são definidas através de um vértice conectado por arestas a n outros vértices, formando uma n-estrela. No caso direcionado, estrelas se diferenciam pela direção da conexão, sendo de entrada ou saída, já triângulos são semelhantes a tríades cíclicas, que serão definidas a seguir. A quantidade da ocorrência destes conceitos são estatísticas bastante úteis para a modelagem de grafos.

Ciclos e Tríades: No caso direcionado, temos ciclos, definidos como caminhos em que o primeiro vértice é também o vértice final do caminho, com o resto de seus vértices sendo únicos. Um ciclo de comprimento 3, denominado tríade cíclica representa um conceito semelhante a triângulos do caso não direcionado. Existem também tríades transitivas, definidas como um par de vértices que apresentam alguma conexão entre si, ambas se conectando a um terceiro vértice em comum direcionado a este. A Figura 2.1 ilustra ambos tipos de tríade.

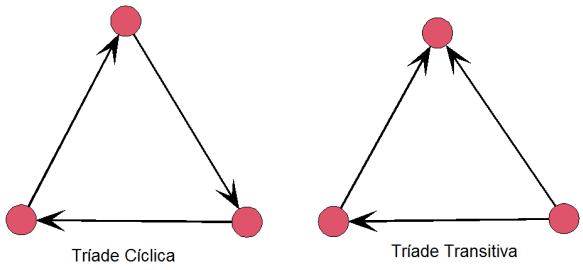

Figura 2.1: Exemplo de Tríades

Coeficiente de Agrupamento: Para medir o nível da transitividade parcial mencionada em um grafo, podemos utilizar esse coeficiente, que varia de 0 a 1, onde 0 indica nenhuma transitividade e 1 indica transitividade perfeita. Este coeficiente é calculado pela proporção de caminhos de comprimento 2 que formam um triângulo, ou seja, o último vértice deste caminho está conectado por uma aresta ao vértice inicial.

Embora uma imagem seja intuitiva para analisar um grafo, precisamos de um

método para representá-lo matematicamente, existem vários métodos propostos para este fim, para nosso trabalho, utilizaremos matrizes de adjacência para nossa representação.

Utilizando matriz de adjacência, representamos grafos por uma matriz $n \times n$ denotada por $\mathbf{X}$, onde cada entrada x_{ij} representa a existência da aresta do vértice i para o vértice j. Este valor pode depender caso o grafo apresente pesos, mas em geral, no caso sem pesos, utilizaremos o valor 0 caso não exista aresta e 1 caso exista aresta. Esta matriz é simétrica no caso não direcionado. No caso em que os vértices não conectam a si mesmos, a diagonal dessa matriz é formada por zeros. Portanto, no caso de grafos simples, a matriz de adjacência do grafo é uma matriz simétrica com zeros em sua diagonal. Abaixo temos o grafo da Figura 1.1 representado por sua matriz de adjacência.

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{2.1}$$

Agora, podemos definir matematicamente certos conceitos mencionados:

O número de arestas, no caso de grafos simples, é definido por:

$$\sum_{i<j} x_{ij} \, . \tag{2.2}$$

No caso direcionado, o número de arestas mútuas entre dois vértices é definido por:

$$\sum_{i<j} x_{ij} x_{ji} \, . \tag{2.3}$$

O grau de um vértice k é definido por:

$$\sum_{j} x_{kj} \, . \tag{2.4}$$

O número de triângulos, no caso de grafos simples, é definido por:

$$\sum_{i<j<k} x_{jk} x_{ik} x_{ij} \, . \tag{2.5}$$

No caso direcionado, o número de tríades cíclicas é definido por:

$$\sum_{i<j<k} x_{ij}x_{jk}x_{ki} \, . \tag{2.6}$$

O número de 2-estrela, no caso de grafos simples, é definido por:

$$\sum_{i<j<k} x_{ik}x_{jk} \, . \tag{2.7}$$

O número de 3-estrela, no caso de grafos simples, é definido por:

$$\sum_{i<j<k<l} x_{il}x_{jl}x_{kl} \, . \tag{2.8}$$

E similarmente para n-estrela maiores.

2.2 Grafos Aleatórios

Grafos aleatórios são modelos de redes onde fixamos certos propriedades de interesse e a partir destas, as conexões entre vértices são geradas aleatoriamente, ou alternativamente, é escolhido aleatoriamente um grafo dentro do conjunto de todos os grafos que apresentem certas propriedades de interesse. Em geral, essas propriedades são escolhidas utilizando as estatísticas anteriormente definidas.

De destaque na literatura, existe o modelo de Erdős–Rényi, onde fixamos o número de vértices e a probabilidade de existência de arestas entre os vértices, formando um modelo $G(n, p)$, sendo n e p o número de vértices e a probabilidade de existir uma conexão entre dois dados vértices, respectivamente.

O modelo resultante apresenta várias propriedades interessantes, como a distribuição de cada aresta sendo uma Bernoulli independente de cada uma, a de número de arestas e do grau sendo ambas distribuições Binomiais, entre outras bastante exploradas na literatura. Formalmente, o grafo Y é uma matriz cujas entradas são variáveis aleatórias assumindo valores 0 ou 1, e seja $\mathcal{G}_n$ o conjunto de todos os grafos simples com n arestas. Então utilizando o modelo de Erdős–Rényi a probabilidade de selecionar um grafo $y \in \mathcal{G}_n$ é dada por:

$$\pi(y) = p^m (1-p)^{\binom{n}{2}-m}. \tag{2.9}$$

Porém, embora bastante estudado e utilizado, existem problemas com a aplicação do modelo de Erdős–Rényi em redes reais. Em primeiro lugar, o modelo supõe independência das conexões entre os vértices, com a existência de cada aresta tendo a mesma probabilidade, não levando em consideração a estrutura geral do grafo.

Em segundo lugar, quando estudamos dados reais, estamos trabalhando com um número de vértices grande, cuja distribuição dos graus está convergindo, no limite, em uma Distribuição Poisson. Esta distribuição acaba não condizendo com a maioria das redes reais, onde a maioria dos vértices apresenta grau baixo, e poucos apresentam graus extremamente altos, retomamos o exemplo dado anteriormente sobre redes de amizade, onde poucos indivíduos apresentam grande número de conexões enquanto a maioria apresenta poucas.

Por essas razões e outras fora do escopo deste trabalho, decidimos utilizar outro modelo para esse trabalho, o Modelo de Grafos Aleatórios Exponenciais, que de destaque,

não supõe a independência das probabilidades e apresenta maior abrangência na sua modelagem. De fato, apresentaremos como o próprio modelo de Erdős–Rényi é um caso particular de Modelos de Grafos Aleatórios Exponenciais.

2.3 Grafos Aleatórios Exponenciais

A ideia central deste modelo, primeiramente proposto em Holland e Leinhardt (1981), é ao invés de fixarmos as propriedades de interesse, estatísticas do grafo, levamos em consideração diferentes possibilidades próximas a nossas estatísticas, dando maior liberdade para a modelagem ao incluir grafos que poderiam ter sido observados na rede em questão, mas que apresentam pequenas diferenças nas estatísticas utilizadas.

O desenvolvimento desse modelo é longo e envolve conceitos que fogem do escopo desse trabalho, caso interessado ler Grandy Jr (2012) para os conceitos e Newman (2012) para o desenvolvimento. Segundo esse modelo, a probabilidade de selecionar o grafo $y \in \mathcal{G}_n$ é dada por:

$$\pi(y|\theta) = \frac{\exp(\theta^T s(y))}{z(\theta)}\,, \tag{2.10}$$

onde θ é um vetor que contém os parâmetros do modelo, $s(y)$ é um vetor que contém as estatísticas do grafo, e $z(\theta)$ é uma constante normalizadora definida por

$$z(\theta) = \sum_{y \in \mathcal{G}_n} \exp(\theta^T s(y))\,, \tag{2.11}$$

onde $\mathcal{G}_n$ é o grupo de todos os grafos possíveis com n vértices.

Podemos notar que cada estatística do modelo está acompanhada por um parâmetro θ correspondente, o que torna a interpretação destes parâmetros simples e intuitiva, quanto maior o parâmetro positivo ou menor o parâmetro negativo, maior a influência desta estatística correspondente no modelo.

Como mencionamos, é possível definir o modelo de Erdős–Rényi, se utilizarmos apenas o número esperado de arestas como estatística de $s(y)$, chegamos no mesmo modelo G(n,p), com $p = E(m)/\binom{n}{2}$, onde $E(m)$ representa o número esperado de arestas. O desenvolvimento deste resultado pode ser conferido em Newman (2012).

Com o modelo definido, temos um problema, encontrar um valor para $z(\theta)$ é extremamente difícil para a maior parte de grafos, pois o número de grafos possíveis rapida-

mente explode junto ao número de vértices, que tende a ser grande para dados reais.Esse fato apresenta um obstaculo para realizar a inferência dos parâmetros θ desejados. Porém, é possível ultrapassar esta dificuldade através do uso de métodos de Monte Carlo Baseado em Cadeias de Markov (MCMC)(Caimo e Friel (2011)).

Capítulo 3

Monte Carlo Baseado em Cadeias de Markov

Neste Capítulo, introduzimos o conceito Monte Carlo e os algoritmos de Monte Carlo Baseado em Cadeias de Markov (MCMC) como uma solução para o problema introduzido no capítulo anterior, descrevendo em especial o Algoritmo de Metropolis-Hastings que será utilizado no restante do trabalho.

3.1 Monte Carlo e MCMC

Métodos de Monte Carlo tem como principal objetivo aproximar um valor esperado, que geralmente não é não é possível ser obtido via cálculo analítico, através de uma estimativa utilizando amostragem de variáveis aleatórias simuladas.

Em outros termos, é a aplicação de conceitos como Lei dos Grandes Números e o Teorema do Limite Central para um cálculo de um valor esperado. Embora pareça uma aplicação restringida, vale notar que é fácil expressar outros conceitos através de valores esperados, como probabilidades e integrais.

Existem várias aplicações de Monte Carlo, uma classe destas aplicações são os métodos de Monte Carlo Baseado em Cadeias de Markov (MCMC), onde o valor esperado desejado, que em geral representa uma distribuição de probabilidade, é definido como medida invariante de uma cadeia de Markov, deste modo, só precisamos simular a cadeia para obtermos uma amostra da distribuição de probabilidade desejada.

Para simular a cadeia mencionada, utilizaremos o Algoritmo de Metropolis-Hastings (Gamerman e Lopes (2006)).

3.2 Algoritmo de Metropolis-Hastings

Nosso objetivo é inferir os parâmetros θ do Modelo de Grafos Aleatórios Exponenciais, descrito em (2.10), a partir de um grafo observado y. Em Caimo e Friel (2011), os autores propõe um método Bayesiano para inferir θ. Esse é o método que estudaremos e aplicaremos nesse trabalho.

Utilizando conceitos de inferência Bayesiana, podemos definir uma distribuição a priori para os valores de θ do Modelo de Grafos Exponenciais, dada por $\pi(\theta)$, e temos que a distribuição a posteriori é dada por:

$$\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta)\,, \tag{3.1}$$

Notamos que $z(\theta)$ ainda está presente em $\pi(\theta|y)$. Desse modo, a distribuição a posteriori não pode ser obtida facilmente, já que não podemos obter analiticamente $z(\theta)$, como discutido na Seção 2.2. Assim, a média, a moda e a mediana a posteriori não podem ser obtidas analiticamente. Então, a ideia é simular uma amostra $\theta_1, \theta_2, \ldots, \theta_k$ da distribuição a posteriori para estimar a média, a mediana e a moda a posteriori utilizando o método de Monte Carlo. Note que simular diretamente de (3.1) não é possível por causa da constante normalizadora $z(\theta)$. Porém, podemos contornar esse problema através de um algoritmo MCMC utilizando um *Update* de Gibbs.

3.2.1 *Update* de Gibbs

Primeiramente, definimos uma distribuição arbitraria $h(\theta^*|\theta)$, que pode ou não depender de θ, os valores de θ^* são novos valores propostos para os parâmetros do modelo. A partir disto, definimos:

$$\pi(\theta^*, y^*, \theta|y) \propto \pi(y|\theta)\pi(\theta)h(\theta^*|\theta)\pi(y^*|\theta^*)\,. \tag{3.2}$$

com y^* sendo um novo grafo proposto, e $\pi(y^*|\theta^*)$ pertencendo a mesma distribuição de $\pi(y|\theta)$.

Desta forma, geramos uma proposição de valores novos para θ e um novo grafo baseado nestes valores. Essa proposta ainda não é incorporada pelo algoritmo, então teremos uma probabilidade de aceitar essa proposta como novas amostras do modelo,

definida por

$$r = \min\left(1, \frac{\pi(y^*|\theta)\pi(\theta^*)h(\theta|\theta^*)\pi(y|\theta^*)}{\pi(y|\theta)\pi(\theta)h(\theta^*|\theta)\pi(y^*|\theta^*)}\right). \tag{3.3}$$

Abrindo parte desta expressão, temos

$$\frac{\exp(\theta^T s(y^*))\pi(\theta^*)h(\theta|\theta^*)\exp(\theta^{*T}s(y))}{\exp(\theta^T s(y))\pi(\theta)h(\theta^*|\theta)\exp(\theta^{*T}s(y^*))} \times \frac{z(\theta)z(\theta^*)}{z(\theta)z(\theta^*)}, \tag{3.4}$$

Logo, conseguimos eliminar a necessidade de se calcular $z(\theta)$. Podemos reduzir essa expressão (3.4) ainda mais se utilizarmos uma distribuição simétrica para $h(\theta^*|\theta)$. Como essa escolha é arbitrária, é útil fazer a escolha de uma distribuição que apresenta essa propriedade.

Por fim, simplificamos a probabilidade de aceitação para:

$$r = \min\left(1, \frac{\exp(\theta^T s(y^*))\pi(\theta^*)\exp(\theta^{*T}s(y))}{\exp(\theta^T s(y))\pi(\theta)\exp(\theta^{*T}s(y^*))}\right). \tag{3.5}$$

Existem vários fatores que influenciam a eficácia deste algoritmo, em particular a escolha de distribuição priori dos parâmetros. Em geral, a escolha destes depende do contexto dos grafos sendo trabalhados, logo é útil se familiarizar com o tipo de grafo para evitar cadeias com taxas de aceitação extremamente baixas. Em geral, os autores de Caimo e Friel (2011) utilizam normais como distribuições priori dos parâmetros e para $h(\theta^*|\theta)$, sendo uma distribuição simétrica e bem conhecida, sendo implementada no pacote "Bergm"Caimo e Friel (2014).

O algoritmo em Etapas

Explicados os conceitos, a implementação do algoritmo é dada pelo seguinte pseudo-código:

1. Determinar o Modelo e distribuições:
   (a) Escolher as estatísticas associadas aos parâmetros θ de interesse do Modelo
   (b) Escolher a distribuição a priori dos parâmetros de interesse
   (c) Escolher uma distribuição simétrica para $h(\theta^*|\theta)$
   (d) Determinar valores iniciais de θ
2. Update de Gibbs:

(a) Gerar novo valor θ^* utilizando $h(\theta^*|\theta)$

(b) Gerar novo grafo y^* através de um MCMC auxiliar, denominado cadeia auxiliar, utilizando o modelo e os novos valores θ^*

3. Proposta:

(a) Calcular a probabilidade de aceitação r utilizando os valores atuais e os valores gerados na última etapa

(b) Gerar um número U de uma distribuição uniforme $[0, 1]$

(c) Se U for inferior a r, aceitar a proposta e definir θ^* como novos valores atuais de θ

(d) Caso U seja superior a r, manter θ como valor atual

(e) Repetir etapas 2 e 3 até convergência dos valores de θ

Na cadeia auxiliar, iremos utilizar um método MCMC com amostrador "tie no tie" (TNT) para gerar um grafo do Modelo Exponencial, implementado no pacote "ergm" (Hunter *et al.*, 2008). De destaque, este amostrador decide primeiro entre criar ou deletar uma aresta, com probabilidade uniforme, para depois selecionar aleatoriamente um par de vértices que não tenha ou tenha aresta, respectivo a decisão anterior, e obter a proposta do novo grafo para este MCMC auxiliar. Em geral, este amostrador é vantajoso para grafos esparsos, que são extremamente comuns devido ao número de possíveis pares de vértices explodir para quantidades razoáveis de vértices em um grafo (Byshkin *et al.*, 2016).

O algoritmo referente ao pseudo-código descrito acima está implemento no R no pacote "Bergm" (Caimo e Friel, 2014), utilizando normais multivariadas como priori dos parâmetros e para $h(\theta^*|\theta)$, como anteriormente mencionado.

Capítulo 4

Implementação para Exemplos

Neste capítulo, utilizamos o pacote "bergm" no software R para aplicar o modelo de Grafos Aleatórios Exponenciais em um banco de dados vastamente utilizado na literatura como exemplo e em um grafo simulado, observando os resultados para entender o comportamento do algoritmo e como melhorar sua performance.

4.1 Monastério Sampson

O banco de dados, denominado SAMPSON devido ao autor do estudo original Sampson (1969), representa interações sociais de um grupo de monges, sendo um exemplo de uma rede de amizade direcionada. Para o grafo, estamos apenas interessados na afeição positiva entre monges como conexão. O grafo é um grafo direcionado e apresenta 18 vértices e 88 arestas.

Para entender o comportamento do algoritmo, implementado pelo pacote "bergm", modelamos o grafo utilizando as estatísticas de número de arestas, conexões mútuas e tríades cíclicas utilizando os argumentos padrões do pacote, e depois modificamos estes para observar o comportamento.

Estamos interessados em observar o tempo levado pelo algoritmo em segundos, o desvio padrão das estimativas dos parâmetros resultantes, a taxa de aceitação média obtida pelo algoritmo e por fim, a performance em testes de Bondade De Ajuste Bayesiano, também implementados pelo pacote.

Testamos além dos argumentos padrões, uma diminuição das iterações auxiliares para 100 de 1000, um aumento destas para 2000, uma diminuição das iterações principais para 100 de 1000, um aumento destas para 5000, uma diminuição do número de cadeias

para 3 de 6, e um aumento para 8 destas. Os valores de interesse de cada uma destas mudanças, realizadas separadamente, estão disponíveis na Tabela 4.1.

Modelo	Tempo de execução	Taxa de aceitação	Desvio Padrão Aresta	Desvio Padrão Mútua	Desvio Padrão Tríade
Padrão	18.26	0.39	0.30	0.40	0.16
- Aux	15.26	0.43	0.40	0.62	0.26
+ Aux	25.48	0.40	0.29	0.41	0.14
- Prin	3.45	0.4	0.29	0.47	0.12
+ Prin	91.00	0.38	0.30	0.43	0.17
- Cadeia	9.81	0.36	0.28	0.48	0.15
+ Cadeia	26.92	0.38	0.30	0.40	0.17

Tabela 4.1: Métricas dos Modelos Sampson

De destaque, observamos que a quantidade de iterações auxiliares influência principalmente o desvio padrão das estimativas obtidas até certo ponto, com este sendo pior caso tenhamos muito poucas interações auxiliares, mas um aumento mas um aumento desse valor não aparenta causar uma grande variação nos desvios padrões.

As iterações principais apresentam grande efeito no tempo que o algoritmo leva para rodar mas não aparenta ter muito efeito no valor da taxa de aceitação e nos desvios padrões.

Por fim, observando o número de cadeias, seu principal efeito aparenta ser uma redução na taxa de aceitação com poucas cadeias e além disso, fora das métricas utilizadas, foi obtido um valor para o parâmetro das tríades significantemente diferente dos outros modelos, podendo indicar problemas na convergência.

A implementação dos testes de Bondade de Ajuste Bayesiano do pacote "bergm" nos fornecem uma comparação do grafo real observado com uma amostra de 100 realizações do modelo estimado, utilizando estatísticas do grafo não modeladas explicitamente, sendo estas a distribuição do grau (separado em grau de entrada e grau de saída para o caso direcionado), da distância geodésica miníma e de "edge-wise shared partners", que é o número de arestas do grafo que apresentam X caminhos de comprimento 2 entre as vértices ligadas pela aresta, para $X = 1, 2....$

Observando a performance dos testes de Bondade de Ajuste Bayesiano, incluídos no Apêndice A deste trabalho, não observamos problemas particulares a qualquer uma das mudanças feitas.

Em geral, o algoritmo se apresentou robusto em termo de seus argumentos neste

caso, apresentando resultados razoáveis mesmo em seus piores casos.

4.2 Grafo Simulado

Utilizando o pacote “ergm”, simulamos um grafo simples não direcionado com 100 vértices, utilizando arestas e 2-estrelas como estatísticas com coeficientes -1.8 e 0.03 respectivamente para o modelo.

Obtemos o seguinte grafo:

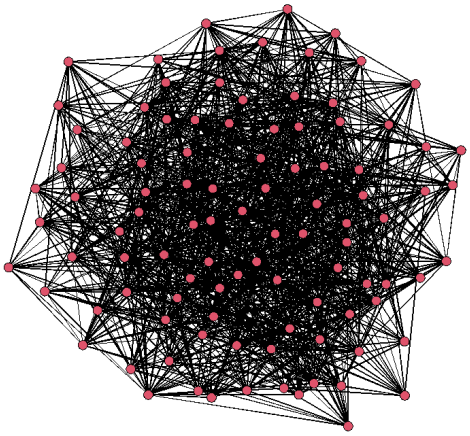

Figura 4.1: Grafo Simulado

Como é possível observar, o grafo obtido pela simulação apresentou uma grande quantidade de conexões, formando 1248 arestas e 31473 2-estrelas, por comparação, um grafo cheio de 100 vértices apresenta 4950 arestas total, logo temos presente em torno de um quarto de todas as arestas possíveis.

Utilizando o grafo simulado, aplicamos o algoritmo através do pacote “Bergm”, utilizando as mesmas estatísticas utilizadas para a simulação na modelagem e os argumentos padrões do algoritmo,os valores obtidos estão na Tabela 4.2.

	Média	Desvio	1^o Quartil	Mediana	3^o Quartil
Arestas	-1.934	0.424	-2.302	-1.969	-1.521
2-estrela	0.018	0.008	0.011	0.177	0.024

Tabela 4.2: Estatísticas da Amostra da Posteriori obtida pelo Algoritmo para os Dados Simulados.

Utilizando a média a posteriori como estimativa para os parâmetros, obtemos valores

relativamente próximos ao real para o parâmetro relacionado as arestas, enquanto para o parâmetro relacionado a 2-estrelas, temos um certo viés, mas ainda permanece próximo.

Porém, obtemos deste modelo um valor de 0.02 para a taxa de aceitação, e na Figura 4.2, observamos que o algoritmo apresentou problemas em convergindo a cadeia, mesmo que tenha chegado a resultados razoáveis, com uma densidade tri-modal e auto-correlação da cadeia que se mantém alta e quase constante, onde seu comportamento esperado é de queda.

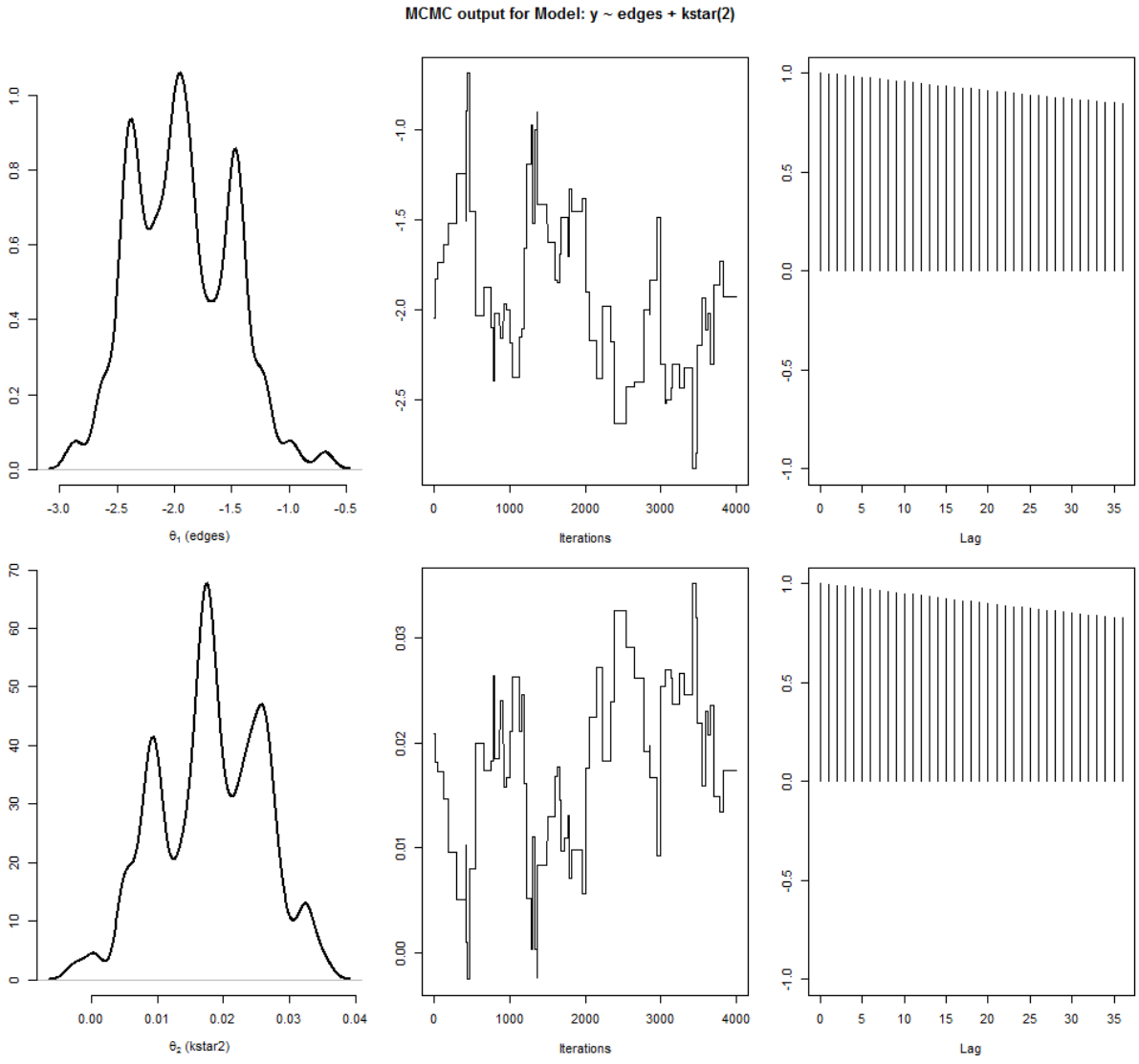

Figura 4.2: Diagnósticos da Cadeia Simulada.

Para verificar se seria apenas necessário um tempo maior para entrar em convergência, aplicamos o algoritmo novamente, desta vez utilizando 8 cadeias e 100000 iterações, os valores obtidos estão na Tabela 4.3.

Novamente, utilizando a média a posteriori como estimativa dos parâmetros, ob-

	Média	Desvio	1^o Quartil	Mediana	3^o Quartil
Arestas	-2.214	0.421	-2.498	-2.209	-1.940
2-estrela	0.023	0.008	0.017	0.022	0.028

Tabela 4.3: Estatísticas da Amostra da Posteriori obtida pelo Algoritmo para os Dados Simulados utilizando mais iterações

temos valores próximos aos valores reais, e também novamente obtemos uma taxa de aceitação pequena de 0.02.

Observando a Figura 4.3, não temos mais uma densidade tri-modal, mas o comportamento da auto-correlação da cadeia se mantém.

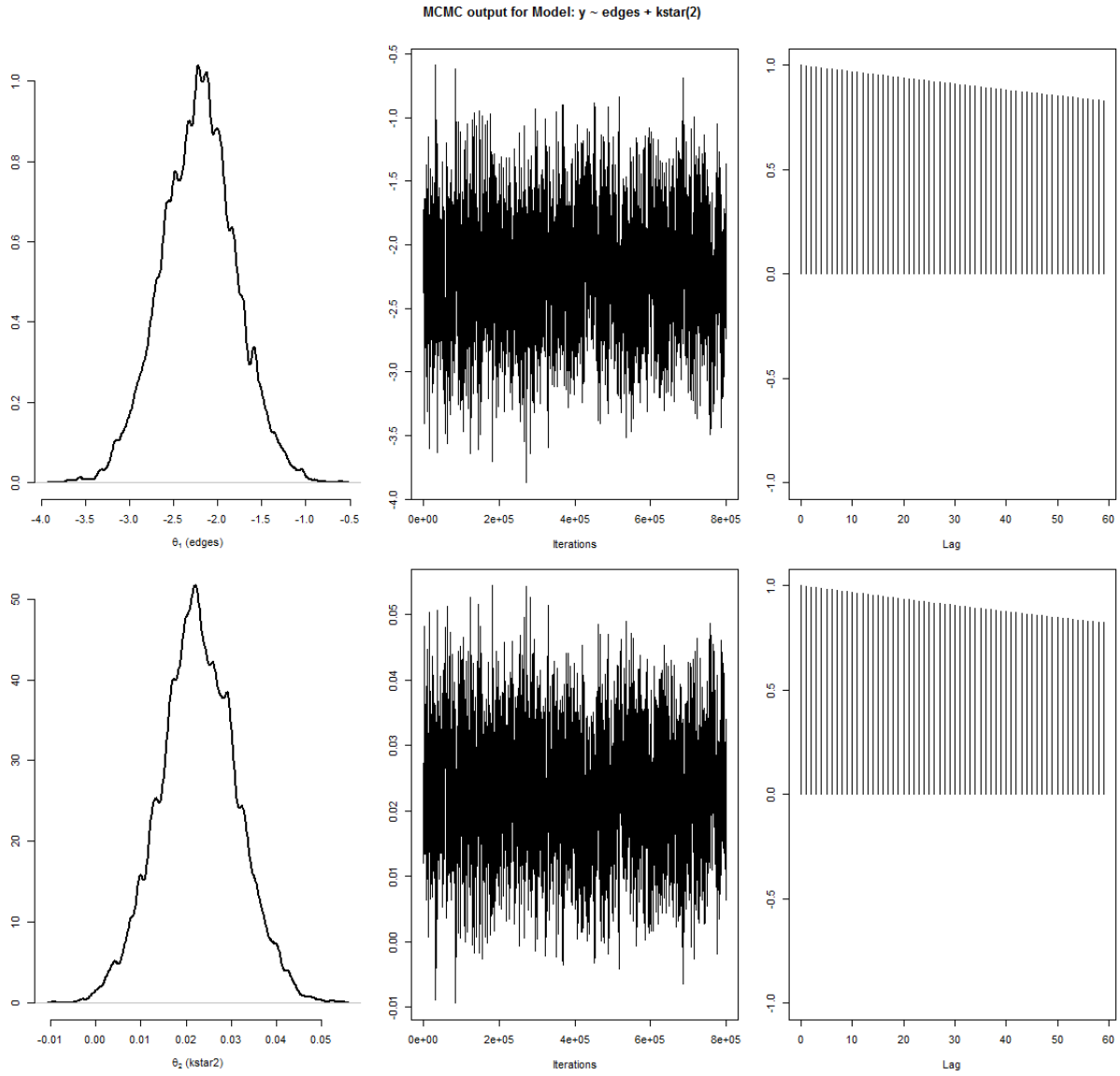

Figura 4.3: Diagnósticos da Cadeia Simulada com mais iterações

Como não aparenta ser um problema de convergência da cadeia, aplicamos o algoritmo novamente, desta vez voltando aos valores padrões de iterações, mas modificando

	Média	Desvio	1^o Quartil	Mediana	3^o Quartil
Arestas	-2.207	0.389	-2.484	-2.175	-1.935
2-estrela	0.022	0.008	0.017	0.022	0.028

Tabela 4.4: Estatísticas da Amostra da Posteriori obtida pelo Algoritmo para os Dados Simulados utilizando proposta alternativa

a proposta para ser muito mais especifica, utilizando uma normal com variância 0.00001 ao invés do 0.0025 padrão do pacote "bergm". Os valores obtidos estão na Tabela 4.4.

Novamente obtemos valores adequados, mas levemente afastados. Porém de destaque, temos uma taxa de aceitação de 0.24 agora. Observando a Figura 4.4, finalmente obtemos um comportamento mais adequado de queda da autocorrelação da cadeia simulada.

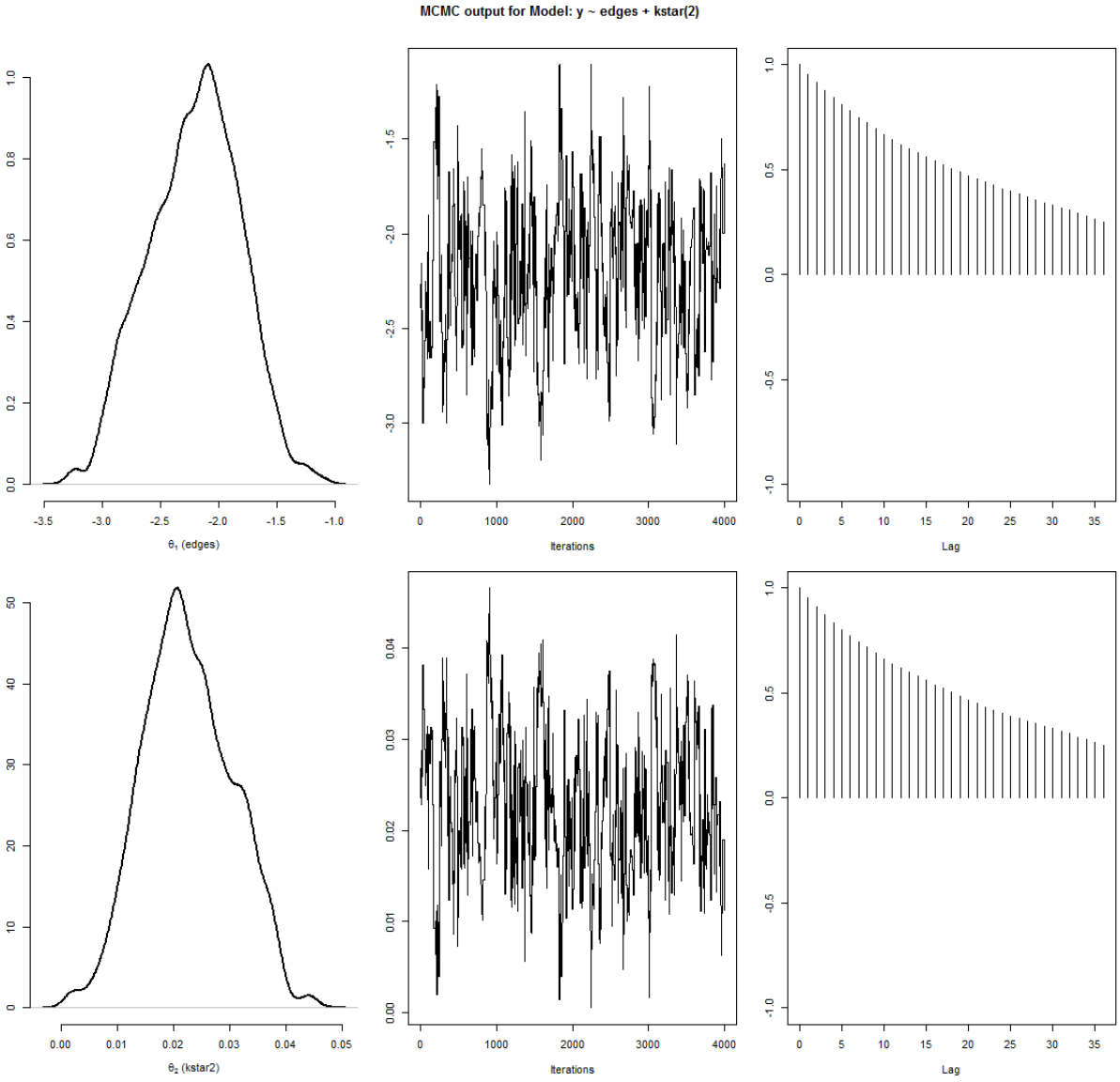

Figura 4.4: Diagnósticos da Cadeia Simulada com proposta Alternativa

Nossas expectativas eram de obter um resultado mais exato para estes modelos utilizando o grafo simulado, porém obtemos modelos adequados para os nossos parâmetros da simulação, mas levemente viesados. Suspeitamos este comportamento ser fruto do fato de nossas estatísticas utilizadas serem intrinsecamente dependentes uma da outra, o que recorda a própria teoria do modelo, onde damos maior liberdade aos parâmetros para aceitar valores próximos e obter um modelo mais abrangente.

Também observamos uma relação entre a taxa de aceitação e uma proposta não adequada, revisando a Figura 4.2, podemos atribuir o comportamento tri-modal da densidade para os longos intervalos observados nos traços da cadeia onde não há proposta aceita, o que contribui também para a alta autocorrelação.

Capítulo 5

Aplicação em Dados De Aeroportos Brasileiros

Neste capítulo, aplicamos o modelo de Grafos Aleatórios Exponenciais para os dados da rede de aeroportos brasileiros disponibilizado pela Agência Nacional De Aviação Civil (ANAC) referente ao ano de 2019, realizando um pré-processamento e análise dos dados para criar um grafo e utilizando o pacote "Bergm" para modelagem e validação do modelo.

5.1 Banco de dados

O banco de dados, disponibilizado pela Agência Nacional De Aviação Civil (ANAC), contém várias variáveis relacionadas a cada voo realizado envolvendo aeroportos brasileiros e aeroportos internacionais. Para nossos fins, só estamos interessados em voos nacionais e em parte dessas variáveis, sendo estas e suas descrições, disponibilizadas em ANAC (2016), citadas abaixo.

Natureza do Voo: Se refere a forma do voo, podendo ser "Doméstico" caso ocorra exclusivamente no Brasil ou "Internacional" caso inclua outros países.

Tipo de Voo: Existem três possíveis tipos de voo:

- **Improdutivas (Non-revenue flights):** Voos que não são feitos comercialmente, geralmente relacionados a treinamentos e manutenção;

- **Regulares (Scheduled revenue flights):** Voos realizados conforme planejamentos de horários e itinerários, ocorrem regularmente;

- **Não Regulares (Non-scheduled revenue flights):** Voos realizados comercialmente, porém sem obedecer planejamentos e realizados sem continuidade, como por exemplo voos Charter que ocorrem em função da demanda.

Aeroporto de Origem e Destino: Nome da cidade dos Aeroportos envolvidos com o voo.

Para representar os dados, utilizaremos um grafo direcionado sem pesos. Como estamos interessados apenas na rede aérea brasileira, selecionamos apenas voos de natureza doméstica e não levamos em consideração voos de tipo Improdutivo, cuja natureza técnica não é relevante para a rede aérea.

Em seguida, representamos os aeroportos como vértices do grafo e a ocorrência de um voo como uma conexão direcionada entre dois aeroportos, dependendo da origem e do destino e consideramos apenas a existência de pelo menos um voo em uma direção específica. A partir disto, formamos a matriz de adjacência que utilizaremos para esta aplicação.

5.2 Análise Descritiva e Exploratória

No grafo obtido, temos 162 vértices, representando 162 aeroportos brasileiros. Para entendermos o comportamento do grafo obtido, realizamos uma análise exploratória utilizando sua representação gráfica.

Primeiramente, representamos o Grafo pela Figura 5.1, sem afixar coordenadas na sua representação para poder observar o espaçamento relativo entre os vértices. Podemos notar que existe um grande aglomerado de vértices que formam uma grande "teia" de conexões no centro, enquanto os vértices mais afastados dependem de um caminho que passa por vértices intermediários para se conectar ao centro, com alguns destes vértices intermediários formando um grande número de conexões separadas do centro, criando outros "polos" para caminhos. Vale observar que em geral, são poucas as conexões não mútuas, e que quando não temos estas, observamos caminhos cíclicos de conexões, o que faz sentido dada a natureza do grafo, aviões que decolam precisam retornar ao seu ponto de origem eventualmente.

Em seguida, como podemos observar na Figura 5.2, adequamos a posição dos vértices do grafo para as coordenadas das localizações físicas do respectivo aeroporto de cada

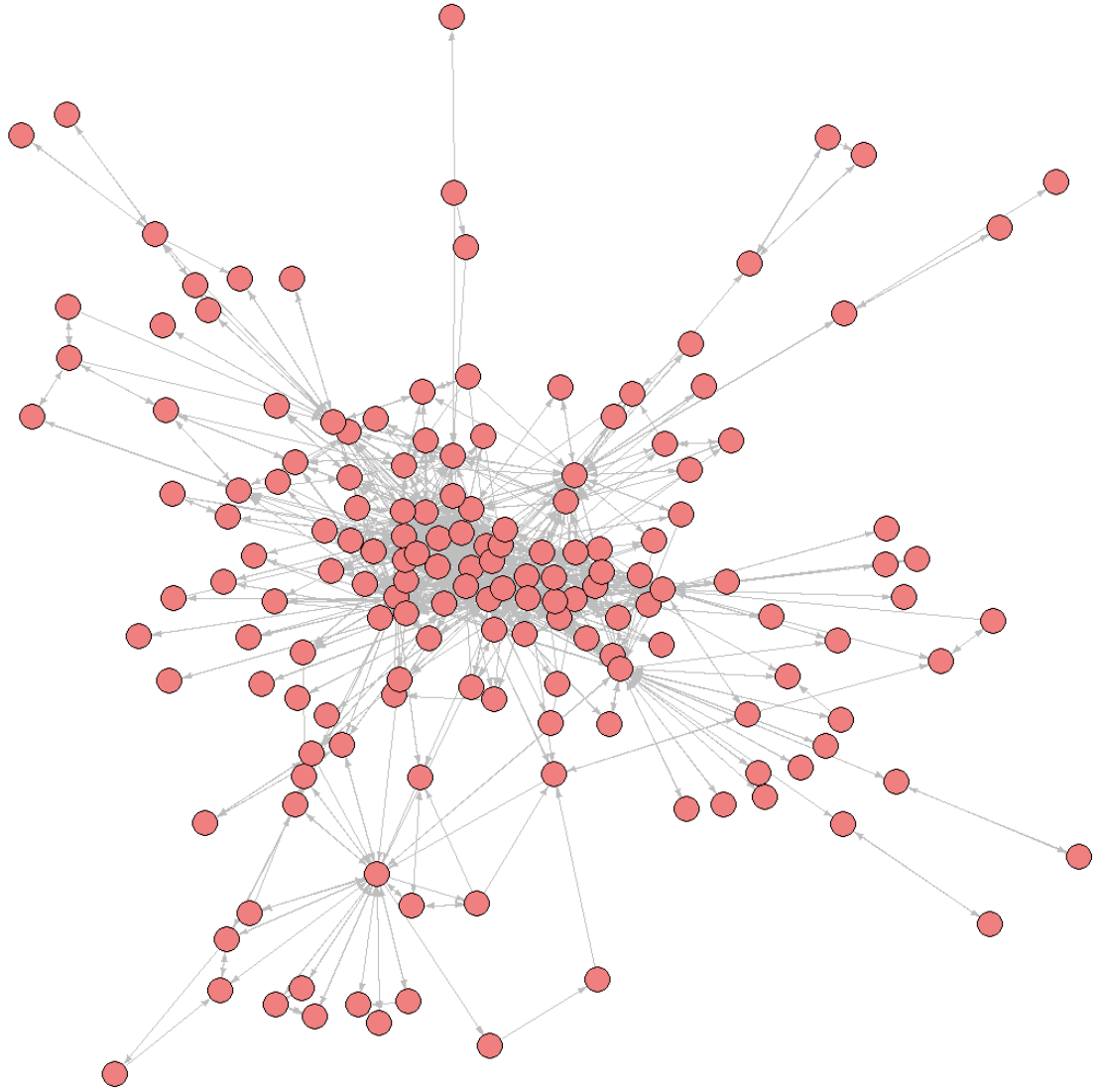

Figura 5.1: Grafo não organizado

vértice, obtendo não só uma boa representação do território, mas justificando as centralidades observadas previamente, que estão conectadas aos grandes aeroportos.

Para facilitar a observação da rede de aeroportos e obtermos uma ideia mais clara do comportamento dos aeroportos de cada porte, separamos os vértices que possuem 10 ou mais conexões daqueles que possuem menos que 10 conexões. A partir deste critério, classificamos 34 aeroportos como de Porte Grande e 128 como de Porte Pequeno.

No grafo com os aeroportos de porte pequeno, Figura 5.3, observamos que muitos aeroportos só apresentam conexões com os aeroportos de porte grande, mas existem várias formações de tríades entre eles, especialmente no norte. Já no grafo com os aeroportos de porte grande, Figura 5.4, como é esperado, é extremamente denso, com cada vértice se

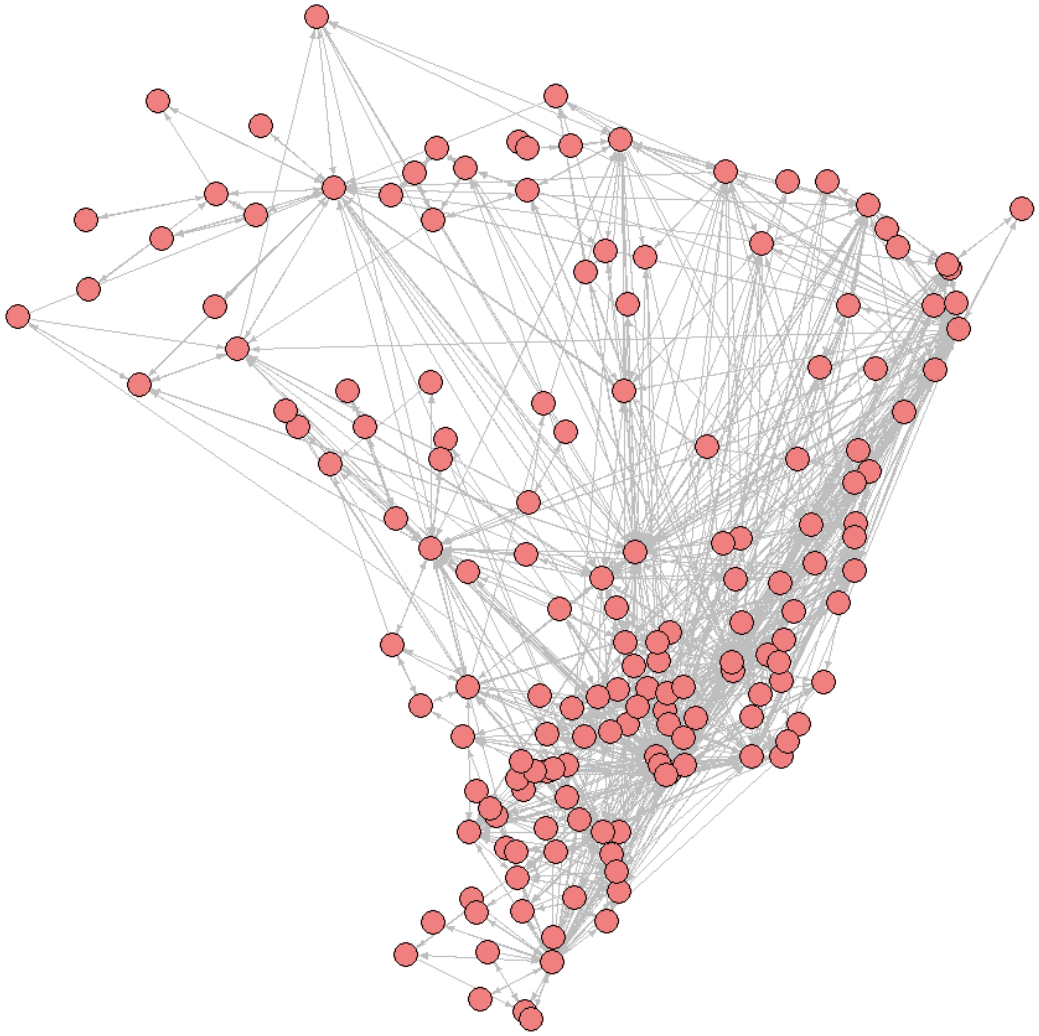

Figura 5.2: Grafo Inteiro Organizado

conectando com a maioria dos outros vértices. Este comportamento merece uma atenção para tentarmos replicar no modelo utilizado.

Com base nas observações realizadas, calculamos as seguintes estatísticas do grafo e suas proporções comparadas a um grafo completo de 162 vértices:

	Arestas	Conexões Mútuas	Tríades Transitivas	Tríades Cíclicas
Valor	1140	428	6125	1999
Proporção	0.044	0.033	0.001	0.001

Tabela 5.1: Estatísticas do Grafo da Rede Aérea

Observamos que estamos lidando com uma rede pouco “densa”, ou seja, o seu número de conexões está bem distante do número total de conexões possíveis para um

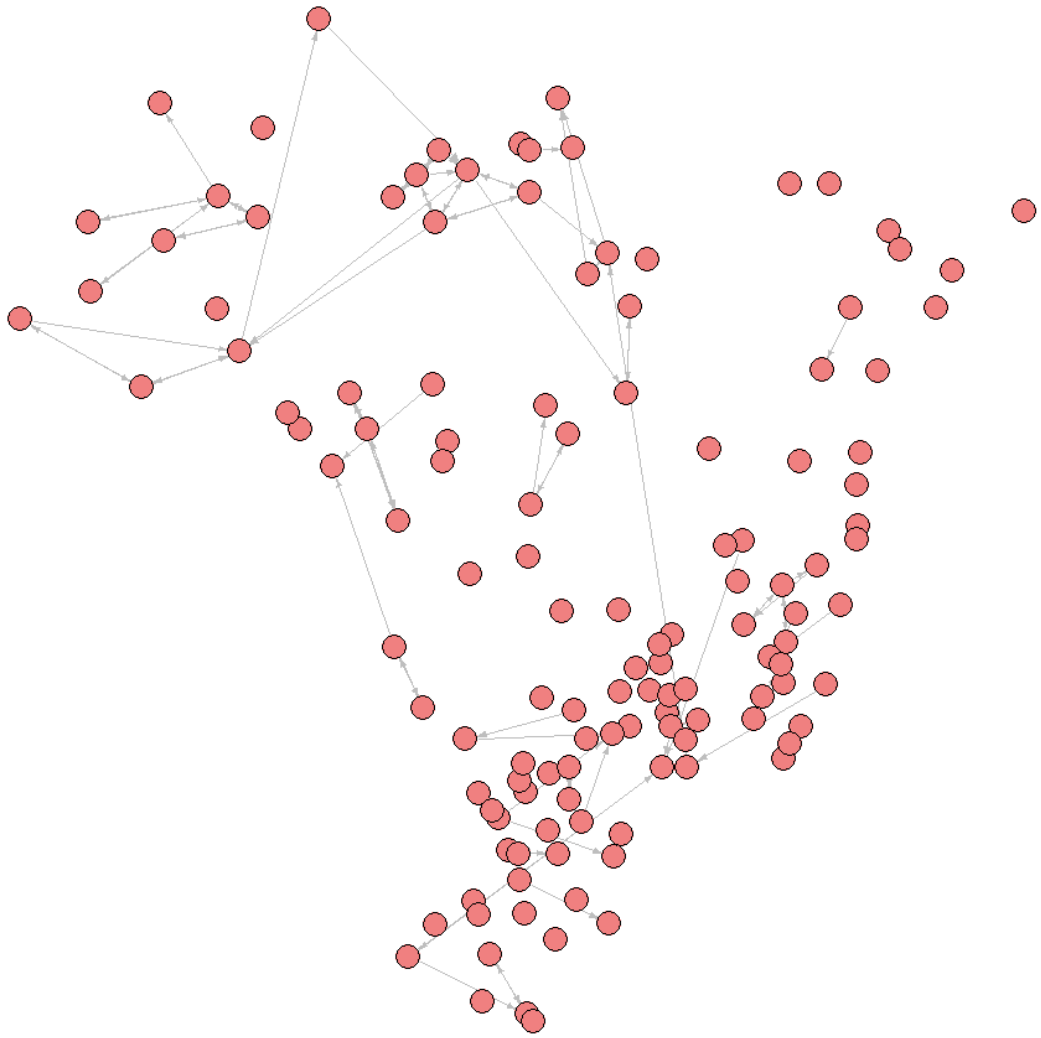

Figura 5.3: Grafo representando os aeroportos de Porte Pequeno.

grafo direcionado de 162 vértices, o que é lógico devido a presença de aeroportos de baixo porte.

Uma observação de particular interesse nestas estatísticas é a proporção de conexões mútuas com o número total de arestas. Por sua própria definição, o número de conexões mútuas é no máximo a metade do número de arestas do grafo, e neste caso temos cerca de 75% deste máximo presente no grafo, reafirmando nossas observações feitas na análise exploratória.

5.3 Modelagem

Utilizando as estatísticas calculadas, formulamos e utilizamos o algoritmo do pacote "bergm" para os 4 modelos seguintes

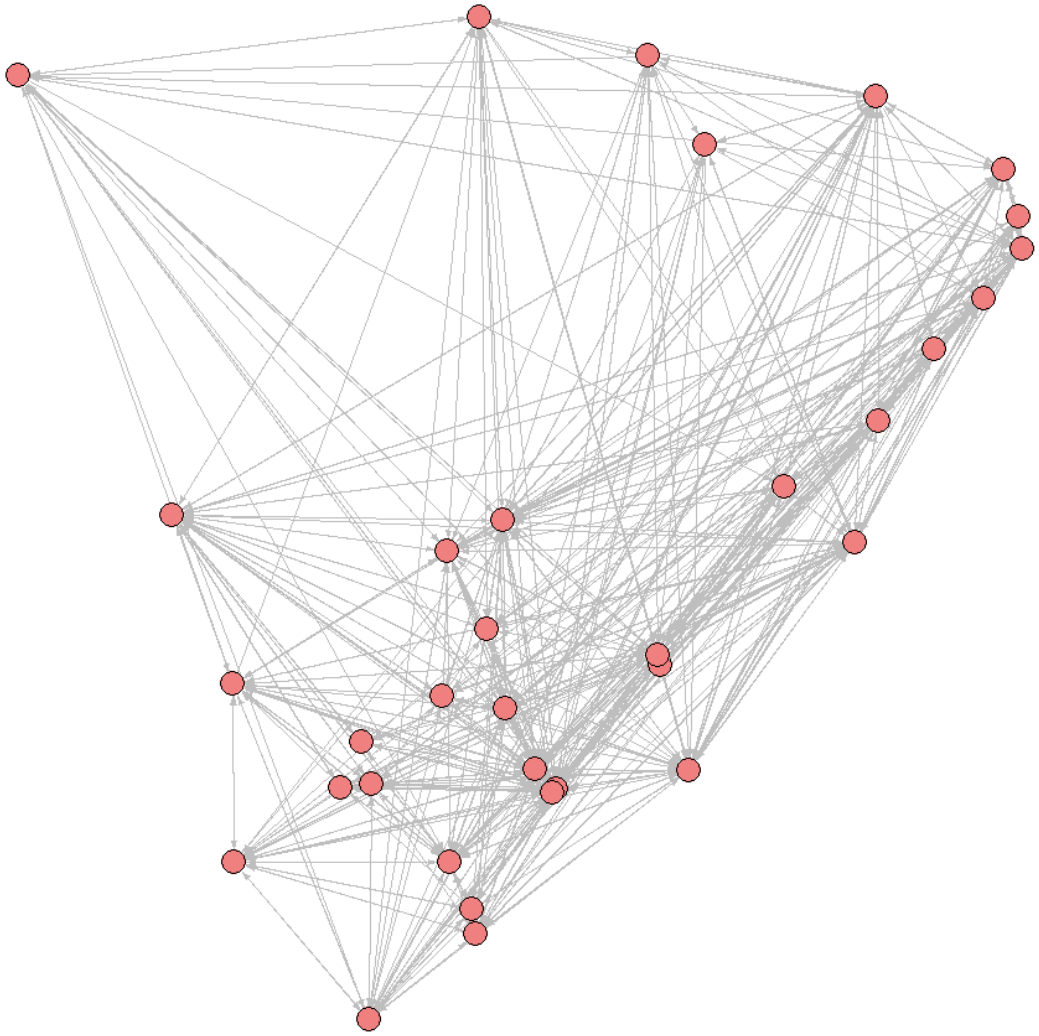

Figura 5.4: Grafo representando os aeroportos de Porte Grande.

1. Arestas e Tríades Transitivas;

2. Arestas, Conexões Mútuas e Tríades Transitivas;

3. Arestas, Conexões Mútuas, Tríades Transitivas e Tríades Cíclicas;

4. Arestas, Conexões Mútuas e Tríades Cíclicas;

Como argumentos, utilizamos 4 cadeias com 1000 iterações auxiliares e 5000 iterações principais. Os resultados obtidos para cada modelo estão descritos na Tabela 5.3.

Modelo: Arestas e Tríades Transitivas					
	Média	Desvio Padrão	1^o Quartil	Mediana	3^o Quartil
Arestas	-4.4048	0.1119	-4.4807	-4.4027	-4.3266
Tríades T	0.2474	0.0226	0.2326	0.2460	0.2615
Modelo: Arestas, Conexões Mútuas e Tríades Transitivas					
	Média	Desvio Padrão	1^o Quartil	Mediana	3^o Quartil
Arestas	-5.5413	0.2067	-5.6864	-5.5284	-5.3864
Mútuas	5.3509	0.6342	4.9038	5.3192	5.7750
Tríades T	0.2236	0.0266	0.2054	0.2221	0.2402
Modelo: Arestas, Conexões Mútuas, Tríades Transitivas e Tríades Cíclicas					
	Média	Desvio Padrão	1^o Quartil	Mediana	3^o Quartil
Arestas	-5.5543	0.1957	-5.6802	-5.5382	-5.4255
Mútuas	5.4306	0.6632	4.9315	5.3558	5.8547
Tríades T	0.2423	0.0632	0.2031	0.2382	0.2840
Tríades C	-0.0603	0.1854	-0.1894	-0.0593	0.0578
Modelo: Arestas, Conexões Mútuas e Tríades Cíclicas					
	Média	Desvio Padrão	1^o Quartil	Mediana	3^o Quartil
Arestas	-5.4220	0.2033	-5.5490	-5.4148	-5.2759
Mútuas	5.3164	0.6370	4.8677	5.2632	5.6915
Tríades C	0.6414	0.0823	0.5851	0.6350	0.6928

Tabela 5.2: Estatísticas da distribuição à Posteriori de cada Modelo

Adicionalmente, na Tabela 5.3, obtemos as seguintes métricas para taxas de aceitação e tempo de execução em minutos do algoritmo para avaliar sua performance.

Os parâmetros estimados para cada modelo reforçam nossas observações anteriores, com valores absolutos elevados para conexões mútuas, percebemos também que tríades cíclicas aparentam ter efeito mais relevante que tríades transitivas, aparentando haver redundância entre estas duas estatísticas para a modelagem. Antes de interpretarmos mais a fundo os resultados obtidos, realizaremos testes para a validação de cada modelo.

Modelo: Arestas e Tríades Transitivas	
Taxa de Aceitação	Tempo de Execução
0.14	2.243
Modelo:Arestas, Conexões Mútuas e Tríades Transitivas	
Taxa de Aceitação	Tempo de Execução
0.09	2.374
Modelo:Arestas, Conexões Mútuas, Tríades Transitivas e Tríades Cíclicas	
Taxa de Aceitação	Tempo de Execução
0.06	2.544
Modelo:Arestas, Conexões Mútuas e Tríades Cíclicas	
Taxa de Aceitação	Tempo de Execução
0.14	1.160

Tabela 5.3: Métricas da Performance Algoritmo de cada Modelo

5.4 Validação

Para validarmos os modelos ajustados na Seção 5.3, novamente utilizaremos testes de Bondade de Ajuste Bayesiano, implementados pelo pacote "bergm"e analisaremos se cada modelo conseguiu capturar propriedades da rede verdadeira. Trabalharemos com gráficos recortados para melhor visibilidade, com as versões completas disponíveis no Apêndice A.

Primeiramente, para o modelo de Arestas e Tríades Transitivas, observamos na Figura 5.5 que o Modelo 1 apresenta grande falha para capturar o comportamento da distribuição de "edge-wise shared partners" da rede real, que é uma estatística que nos interessa a capturar por estar relacionada ao comportamento de ida e retorno dos aviões observado anteriormente. Para as outras estatísticas, os seus formatos aparentam estar capturados, mas existem certos vieses na distância e uma tendência a limitar altos graus cedo demais.

Observamos na Figura 5.6 que no Modelo 2, utilizando Arestas,Conexões Mútuas e Tríades Transitivas como estatísticas , a incorporação de conexões mútuas auxiliou bastante para melhor capturar os comportamentos em que o Modelo 1 falhava em capturar.

Observamos na Figura 5.7 que no Modelo 3, utilizando Arestas,Conexões Mútuas, Tríades Transitivas e Tríades Cíclicas, a inclusão de tríades cíclicas não aparenta ter efeito significante não só no seu parâmetro apresentando valor pequeno, mas também na captação de comportamentos do grafo não sofrendo mudanças notáveis.

Observamos na Figura 5.8 que no Modelo 4, utilizando Arestas,Conexões Mútuas, e Tríades Cíclicas, com a inclusão de apenas tríades cíclicas obtemos resultados semelhantes ao Modelo 2 e 3 novamente.

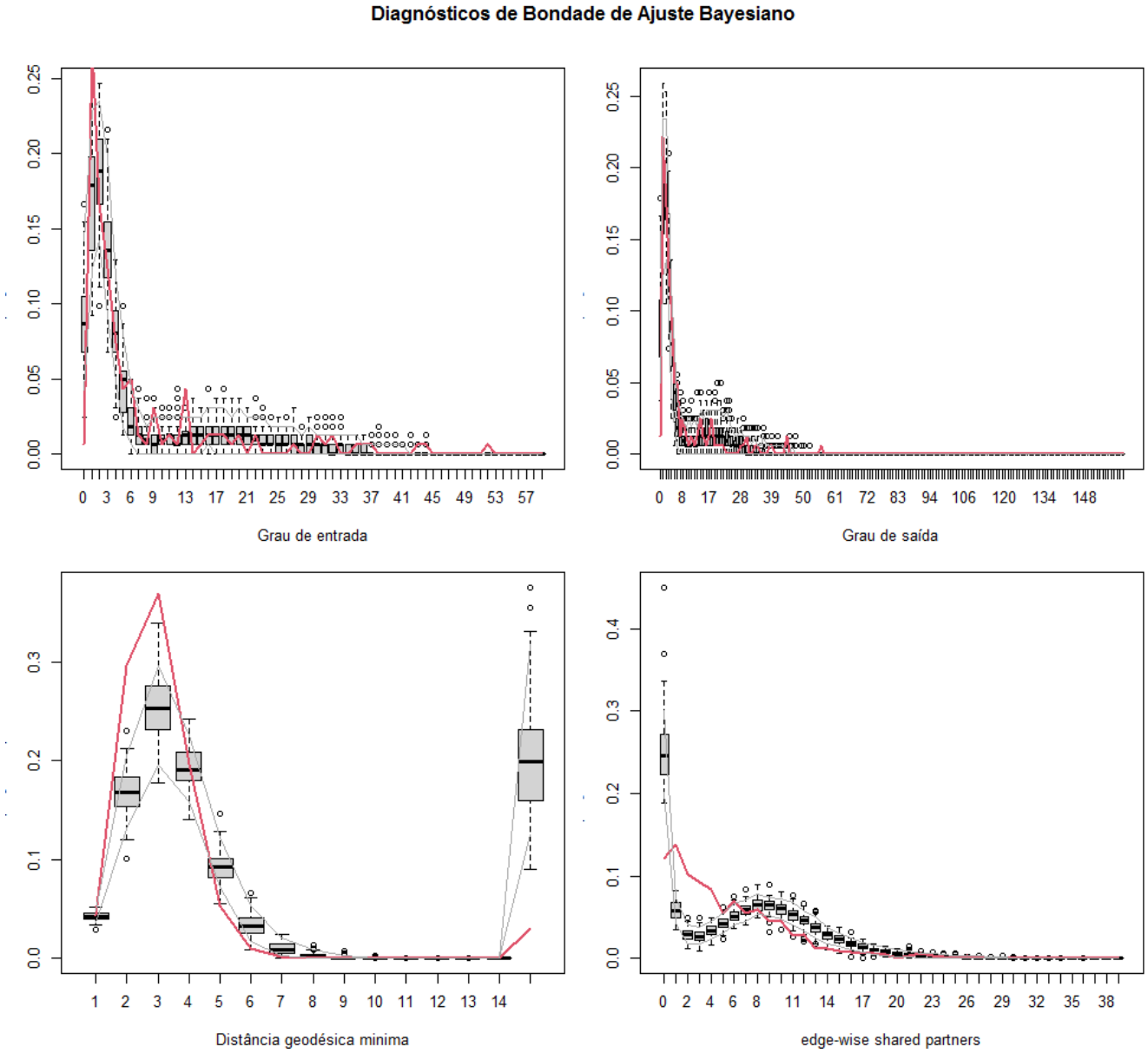

Figura 5.5: Gráficos de Bondade de Ajuste Bayesiano para o Modelo 1 de Arestas e Tríades Transitivas

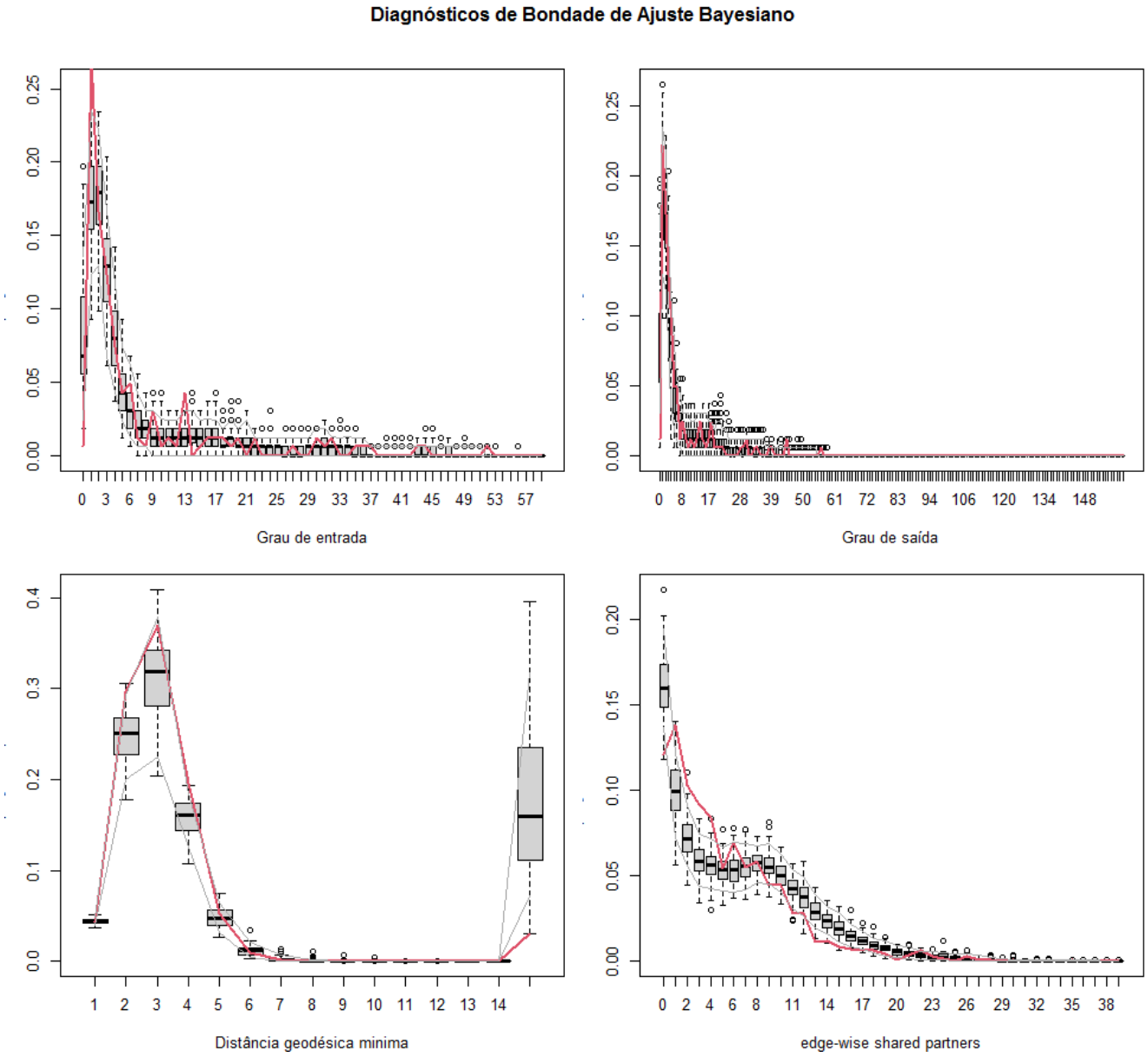

Figura 5.6: Gráficos de Bondade de Ajuste Bayesiano para o Modelo 2 de Arestas,Conexões Mútuas e Tríades Transitivas

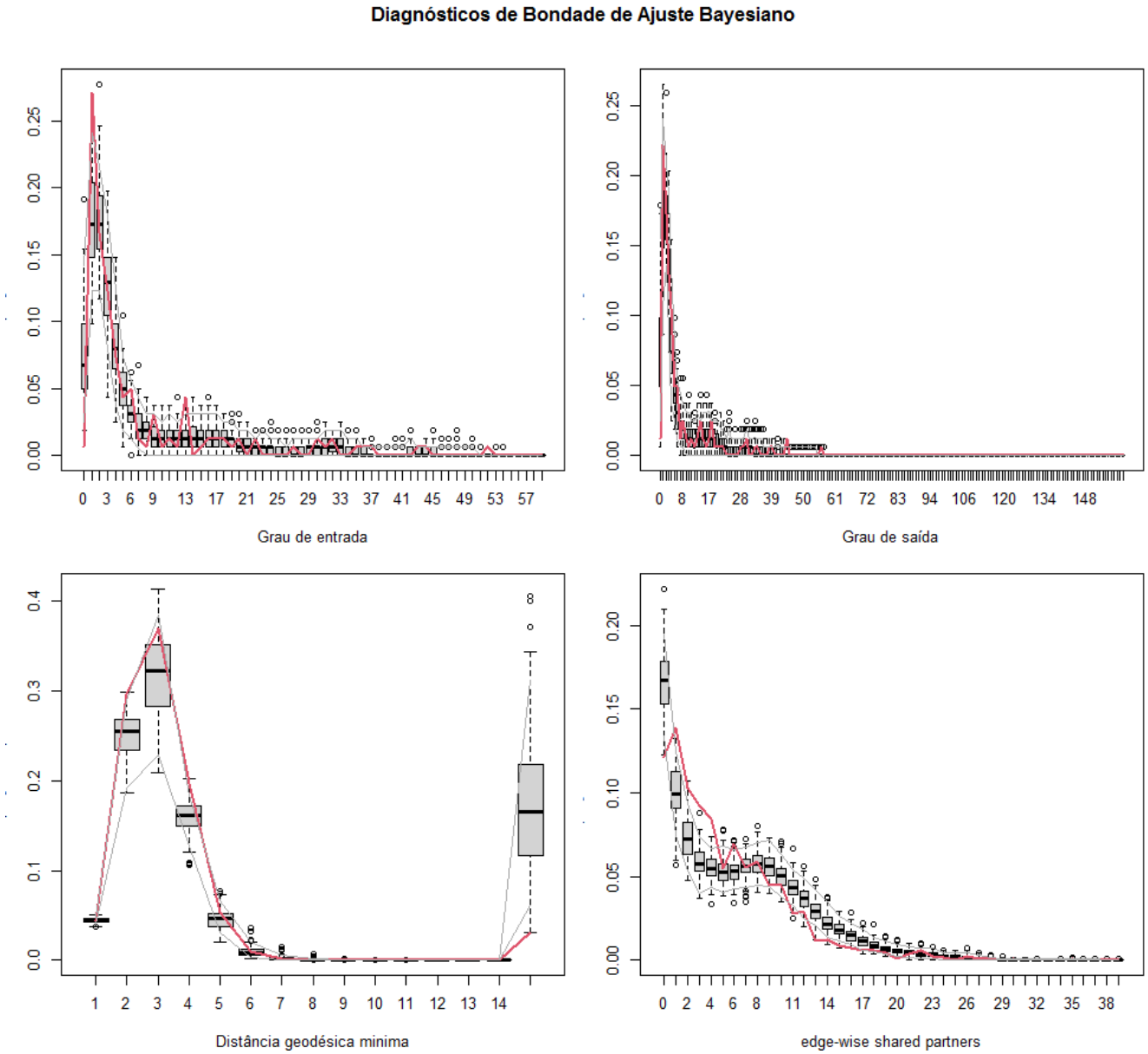

Figura 5.7: Gráficos de Bondade de Ajuste Bayesiano para o Modelo 3 de Arestas,Conexões Mútuas, Tríades Transitivas e Tríades Cíclicas

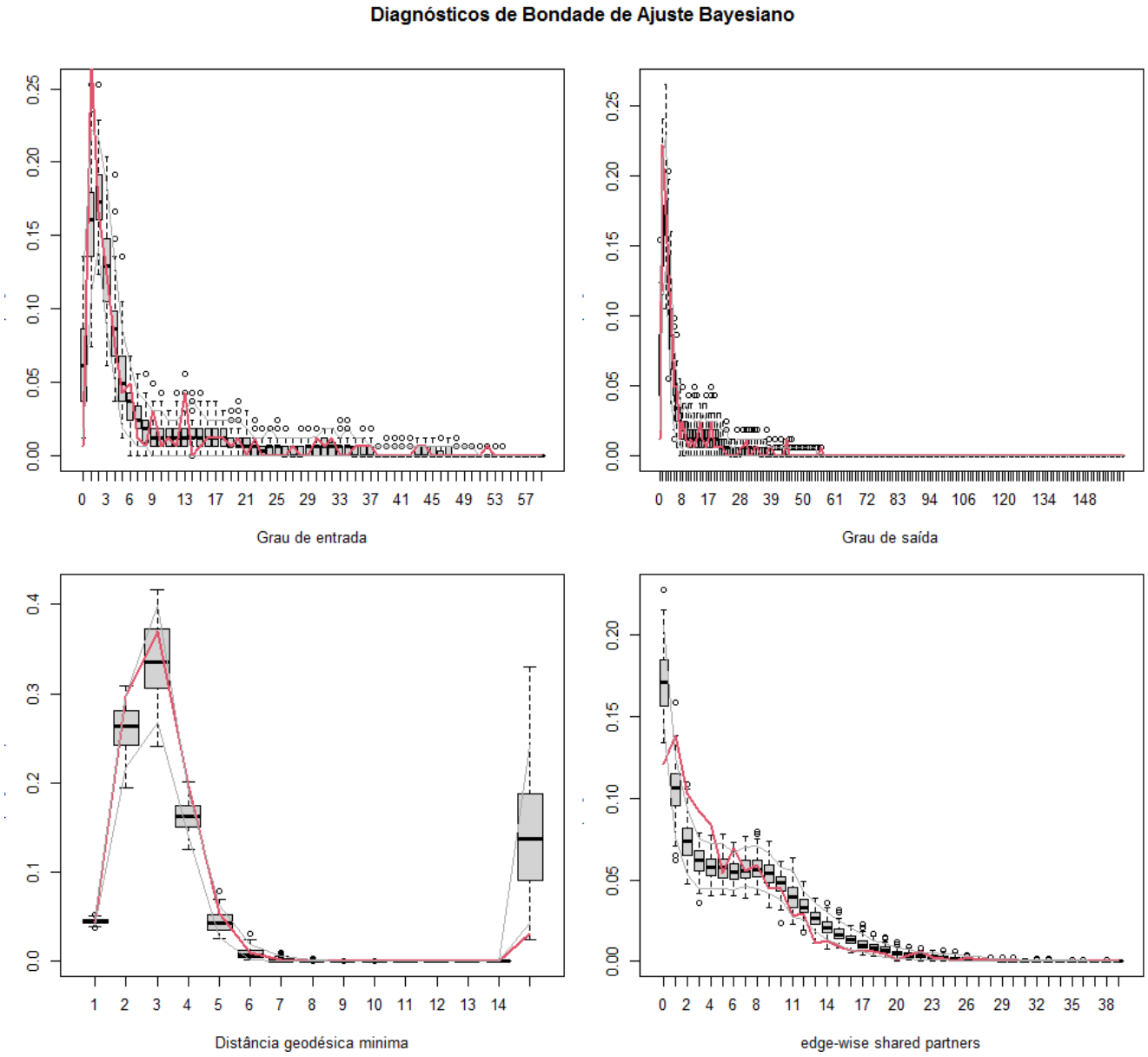

Figura 5.8: Gráficos de Bondade de Ajuste Bayesiano para o Modelo 4 de Arestas,Conexões Mútuas, e Tríades Cíclicas

Dentre os quatro modelos, descartamos o Modelo 1 por não capturar o grafo original adequadamente e julgamos os outros 3 modelos adequados após os testes de Bondade de Ajuste Bayesiano. Dentre estes 3 modelos, selecionamos o Modelo 4 como melhor modelo para explicar a rede e interpretá-la, baseando-se em um menor número de variáveis do que o Modelo 3, que contribui para sua simplicidade sem sacrificar informações importantes. Além disso, o Modelo 4 envolve Tríades Cíclicas ao invés de Tríades Transitivas, que melhor correspondem a um comportamento de "ida e volta" relacionada a voos aéreos.

5.5 Interpretação

Do modelo escolhido, que utiliza parâmetros para arestas, conexões mútuas e tríades cíclicas, obtemos estimativas, utilizado a média a posteriori dos valores simulados na cadeia, de -5.422, 5.3164 e 0.6414 respectivamente.

Devido a fórmula conveniente para o modelo de Grafos Aleatórios Exponenciais, podemos inferir que a rede de Aeroportos Brasileiros em 2019 apresenta uma estrutura com relativamente poucas conexões entre si, representada pelo alto parâmetro negativo associado a arestas, mas uma grande quantidade de conexões mútuas entre aeroportos, representada pelo alto parâmetro positivo associado a conexões mútuas, reforçando nossa primeira análise exploratória.

Tríades cíclicas também estão presentes significantemente na estrutura da rede, apresentando um valor positivo não nulo no parâmetro relacionado, embora não seja tão importante quanto conexões mútuas para a "ida e volta" de aviões entre aeroportos.

Capítulo 6

Considerações Finais

Grafos trazem uma representação interessante para vários tipos de dados, proporcionando propriedades e estatísticas que são intuitivas e fáceis de serem observadas visualmente.

O Modelo de Grafos Aleatórios estudado neste trabalho utiliza bem os conceitos de Grafos, com uma estrutura intuitiva e resultados de fácil interpretação que o torna bem útil para análises e inferência.

Computacionalmente, devido a métodos MCMC, é fácil de se trabalhar e processar dados para seu formato, e devido a sua natureza intuitiva, é fácil perceber quando algo não está sendo propriamente realizado. O pacote "bergm" fornece ferramentas ótimas para sua implementação, que depois de um estudo utilizando exemplos, representou uma clara ligação dos conceitos teóricos do Modelo de Grafos Aleatórios com a prática.

Em geral, realizamos nosso objetivo do estudo deste tópico novo ao aluno, cujo tal possibilitou a aplicação prática realizada neste trabalho.

Para análise destes dados, aplicamos com sucesso o conhecimento obtido durante este trabalho, elaborando um modelo que aparentemente bem representa a Rede Aérea Brasileira e captura os principais aspectos que estávamos interessados em estudar.

Em particular, identificamos Arestas, Conexões Mútuas e Tríades Cíclicas como melhores estatísticas para representar a rede aérea brasileira e utilizando a interpretabilidade intuitiva do modelo, observamos que com base nas médias a posteriori dos parâmetros simulados, obtemos uma tendência de haver poucas arestas na rede, ou seja, poucas conexões entre aeroportos brasileiros em geral, mas com uma grande tendência das conexões que existem formarem uma conexão mútua. Também observamos uma tendência a presença significante das tríades cíclicas, que junto as conexões mútuas, são facilmente

associadas ao comportamento de “ida e volta” que é esperado de aviões em transito.

Esses dados ainda apresentam outras propriedades interessantes de serem estudadas em outros trabalhos com métodos mais sofisticados, como por exemplo uma maneira de diferenciar na modelagem os aeroportos de porte grande dos de porte pequeno, uma distinção que ficou evidente durante a análise exploratória. Também é possível a utilização de outras informações do banco de dados na formação do Grafo, como por exemplo o fluxo de passageiros para a criação de um Grafo com pesos.

Apêndice A

Apêndice

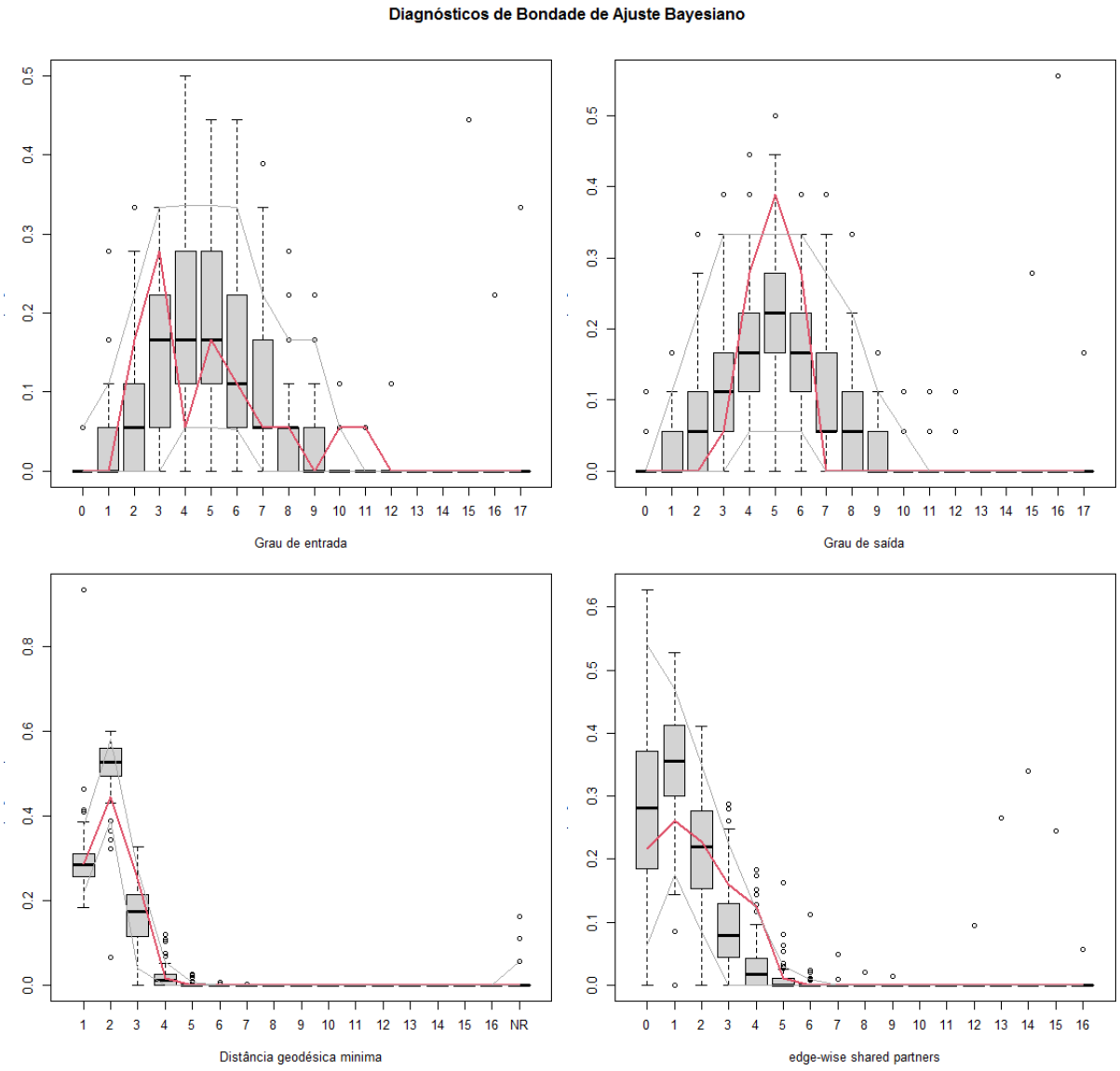

Figura A.1: Gráficos de Bondade do Ajuste Bayesiano do Modelo para o Monastério Sampson Padrão

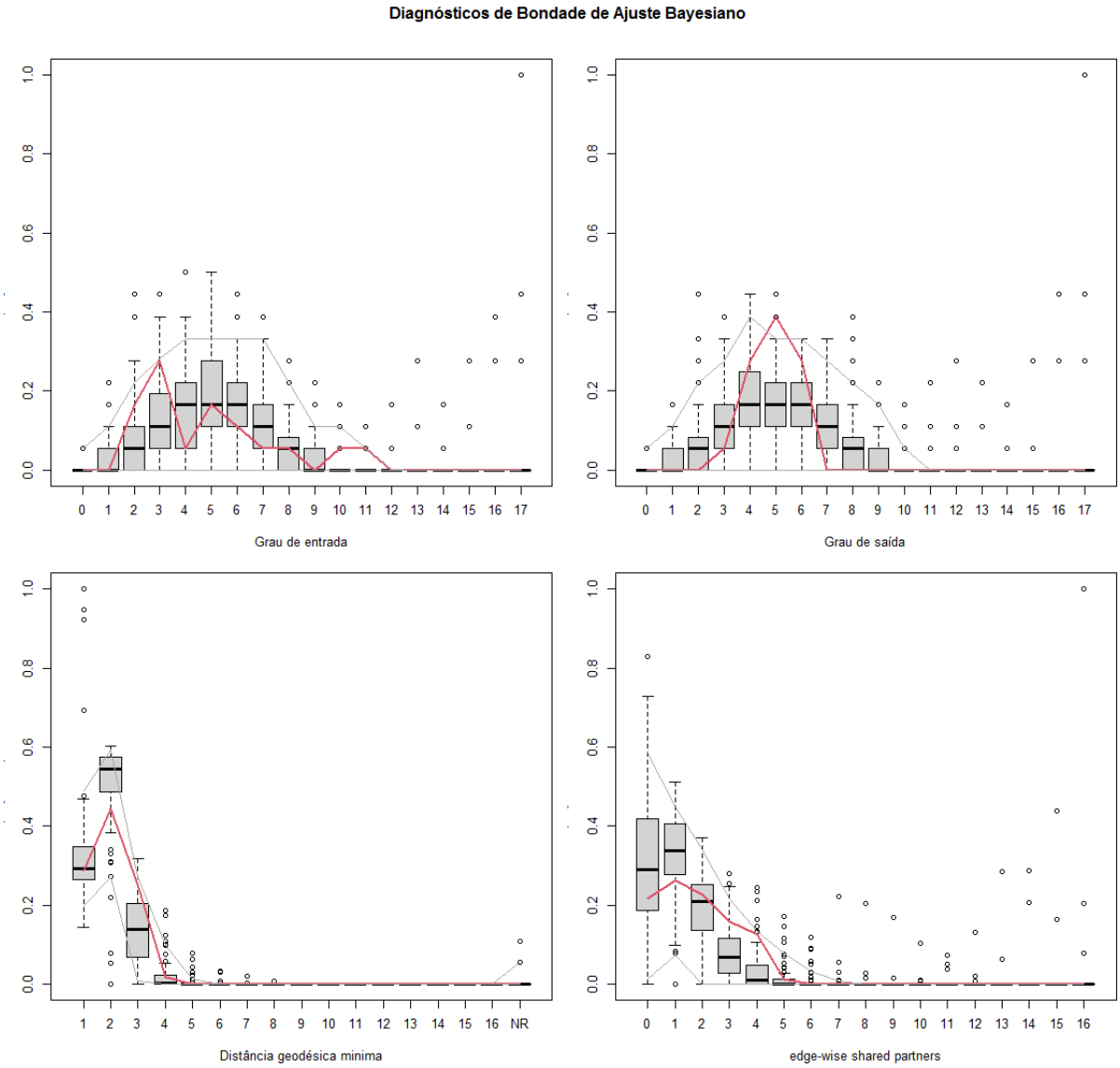

Figura A.2: Gráficos de Bondade do Ajuste Bayesiano do Modelo para o Monastério Sampson com 100 iterações Auxiliares

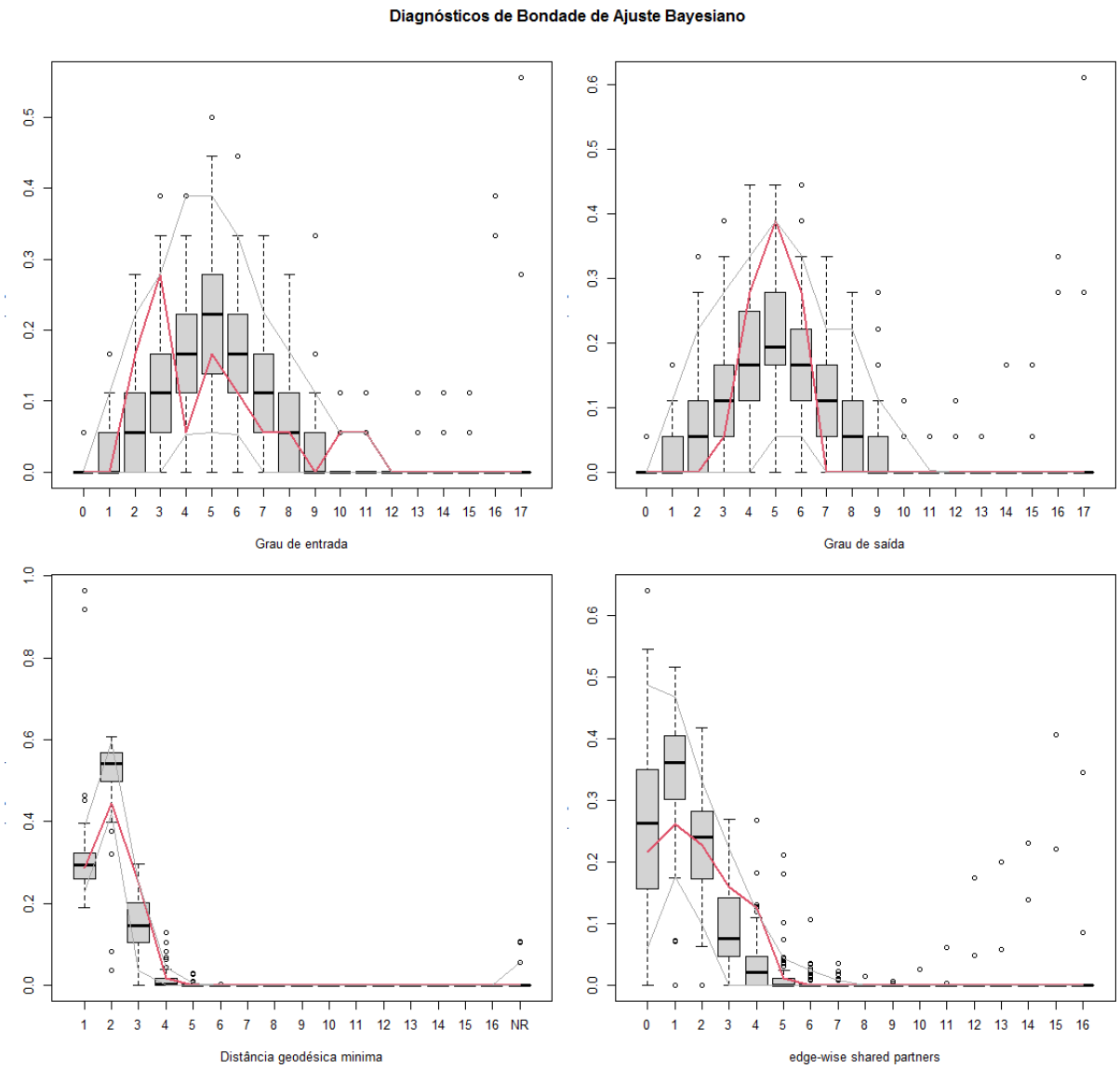

Figura A.3: Gráficos de Bondade do Ajuste Bayesiano do Modelo para o Monastério Sampson com 2000 iterações Auxiliares

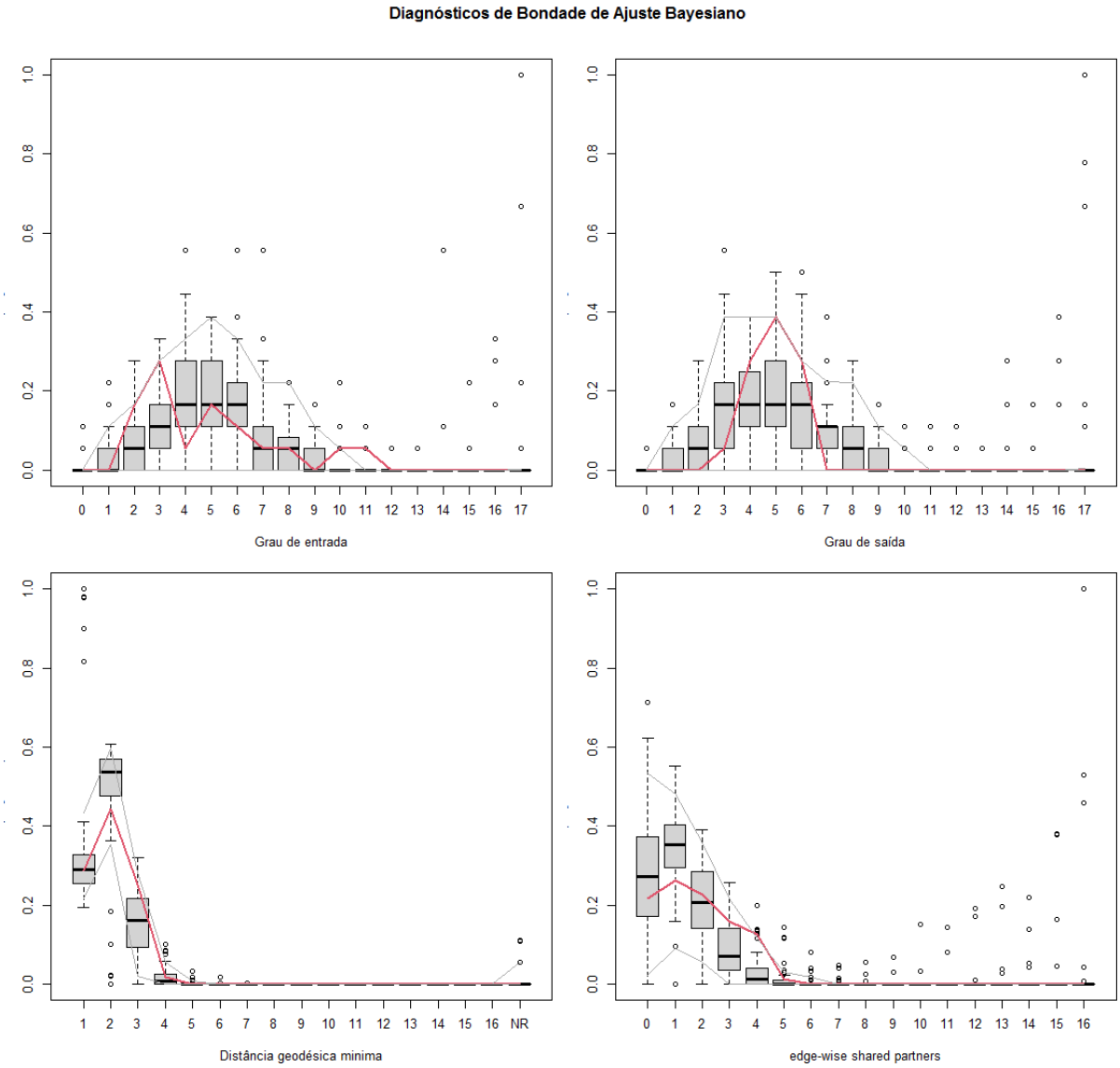

Figura A.4: Gráficos de Bondade do Ajuste Bayesiano do Modelo para o Monastério Sampson com 100 iterações Principais

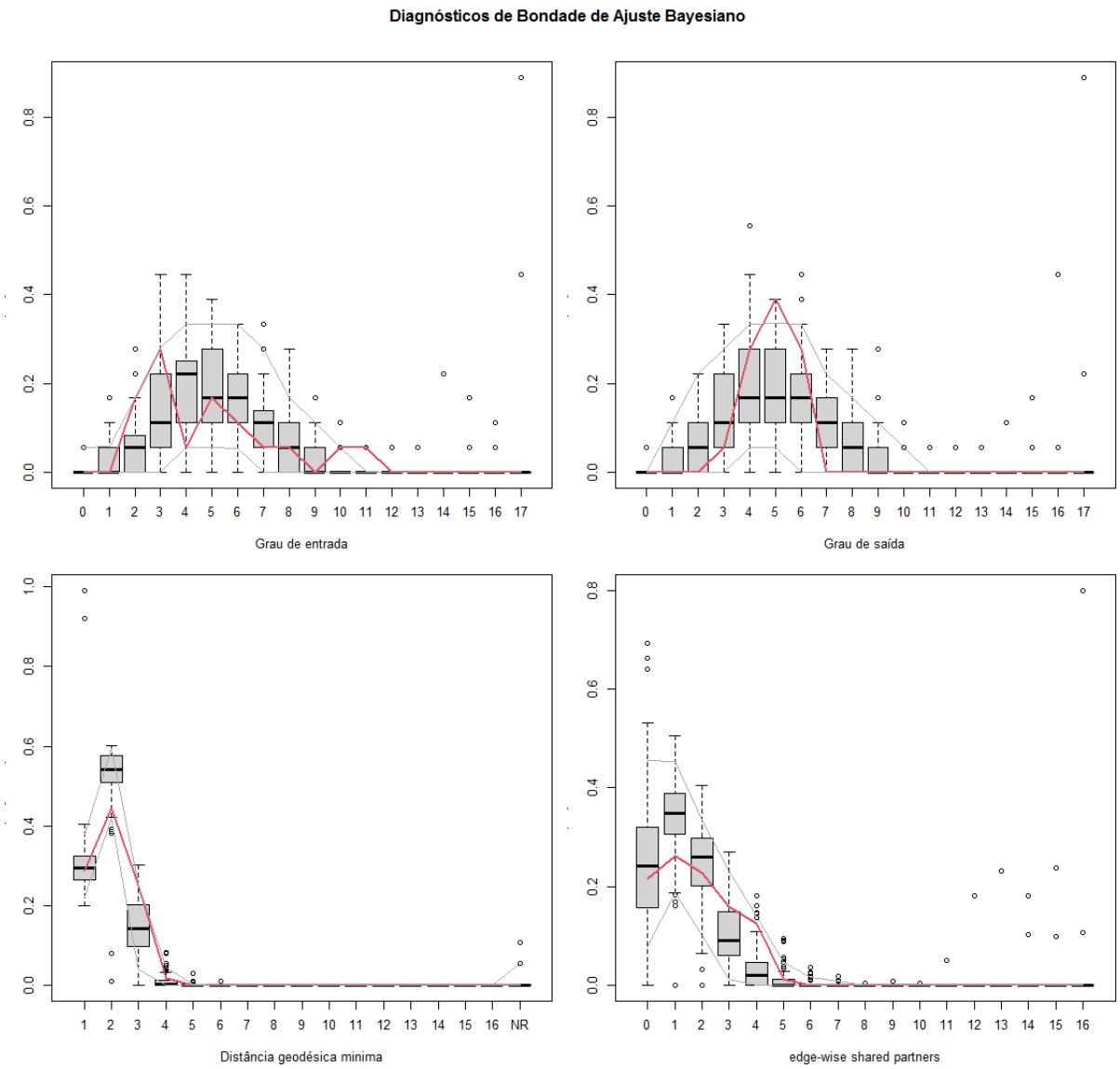

Figura A.5: Gráficos de Bondade do Ajuste Bayesiano do Modelo para o Monastério Sampson com 5000 iterações Principais

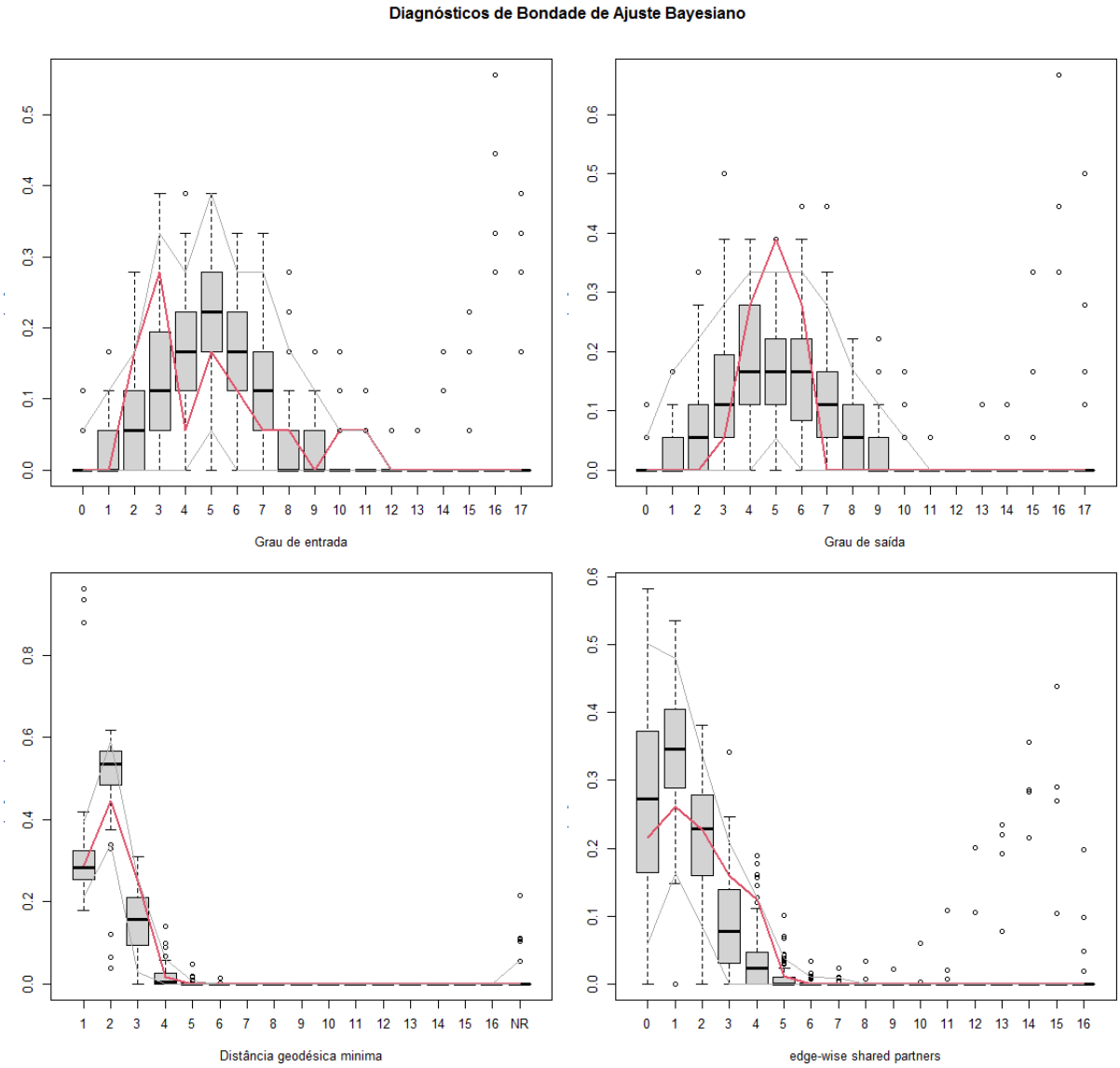

Figura A.6: Gráficos de Bondade do Ajuste Bayesiano do Modelo para o Monastério Sampson com 3 cadeias

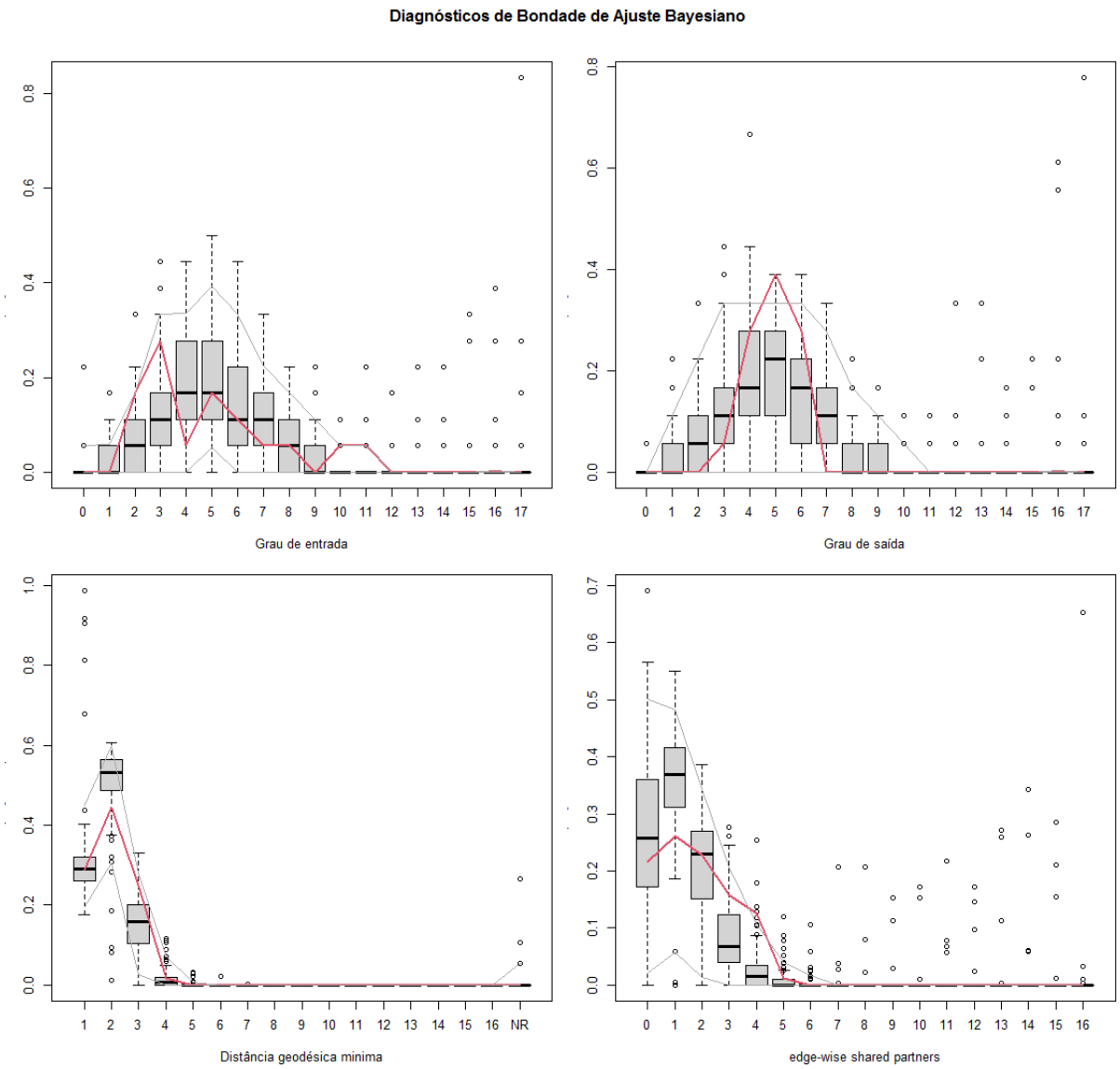

Figura A.7: Gráficos de Bondade do Ajuste Bayesiano do Modelo para o Monastério Sampson com 8 cadeias

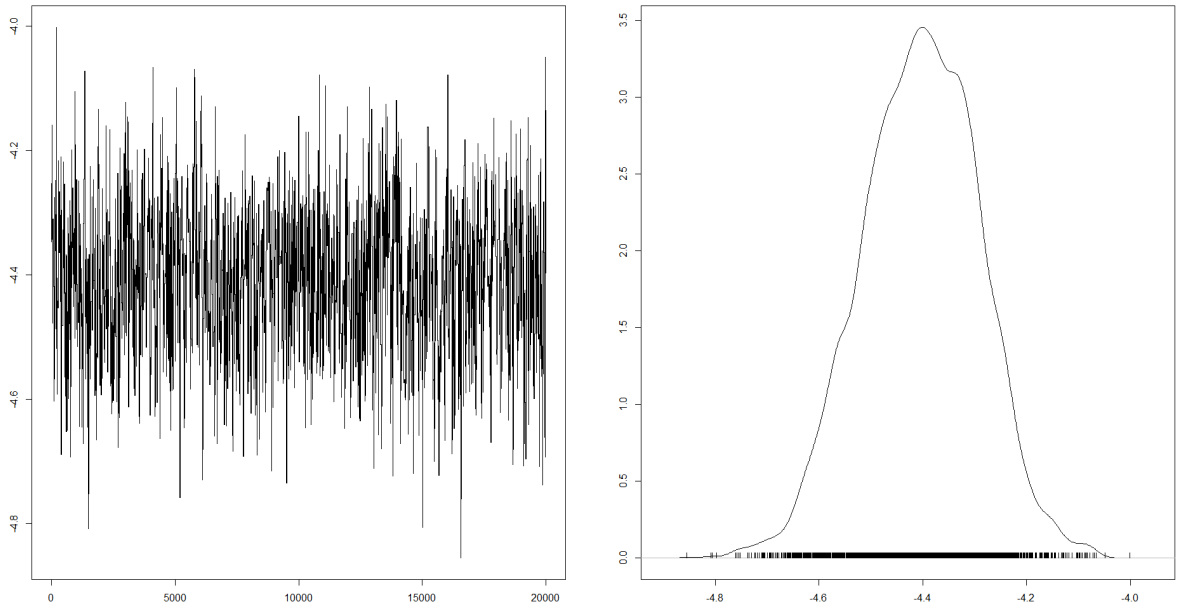

Figura A.8: Densidade e Traço do parâmetro relacionado a Aresta do Modelo de Aresta e Tríade Transitiva

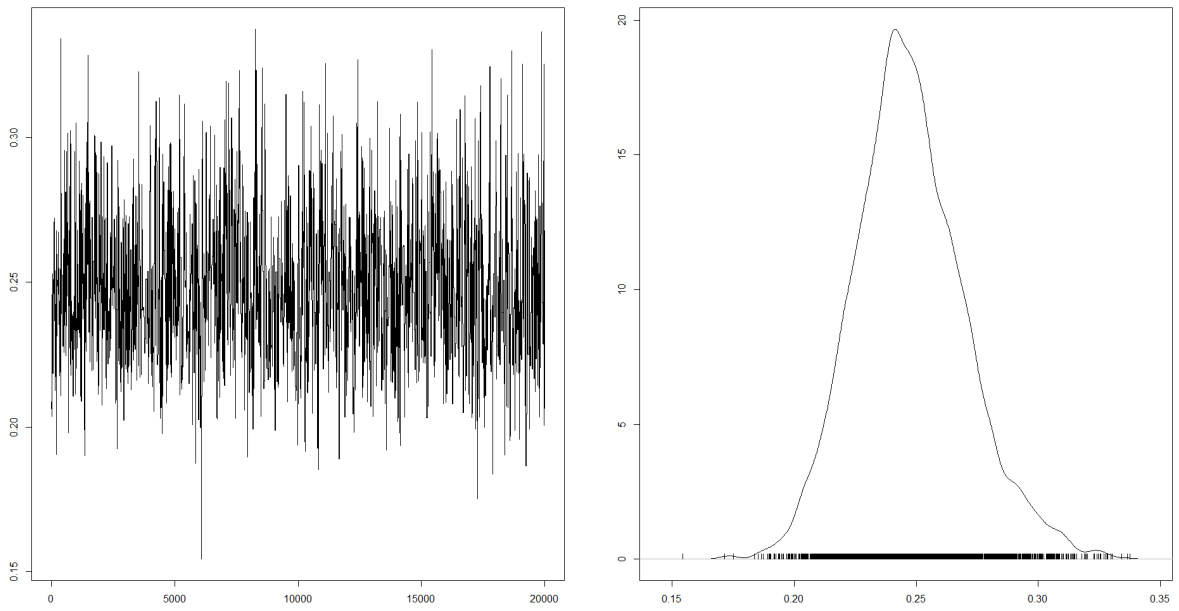

Figura A.9: Densidade e Traço do parâmetro relacionado a Tríade Transitiva do Modelo de Aresta e Tríade Transitiva

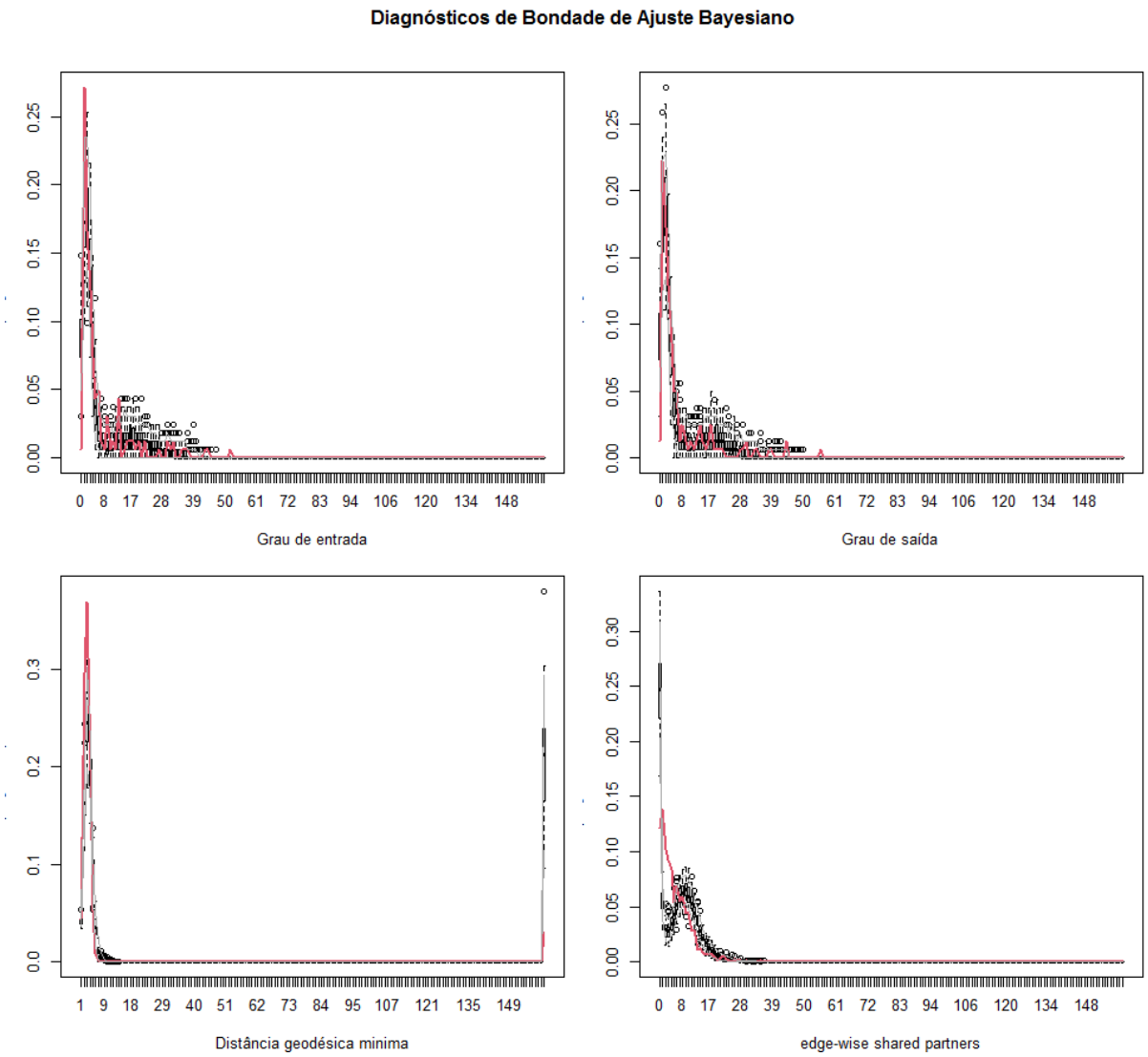

Figura A.10: Gráficos não recortados de Bondade do Ajuste Bayesiano do Modelo de Aresta e Tríade Transitiva

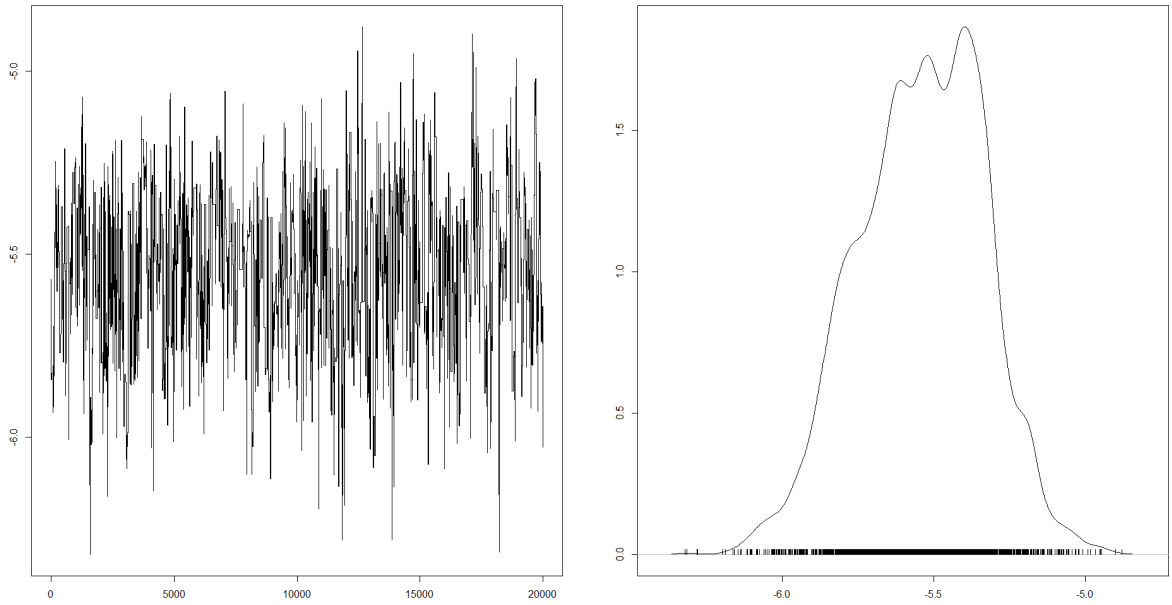

Figura A.11: Densidade e Traço do parâmetro relacionado a Aresta do Modelo de Aresta, Conexões Mútuas e Tríade Transitiva

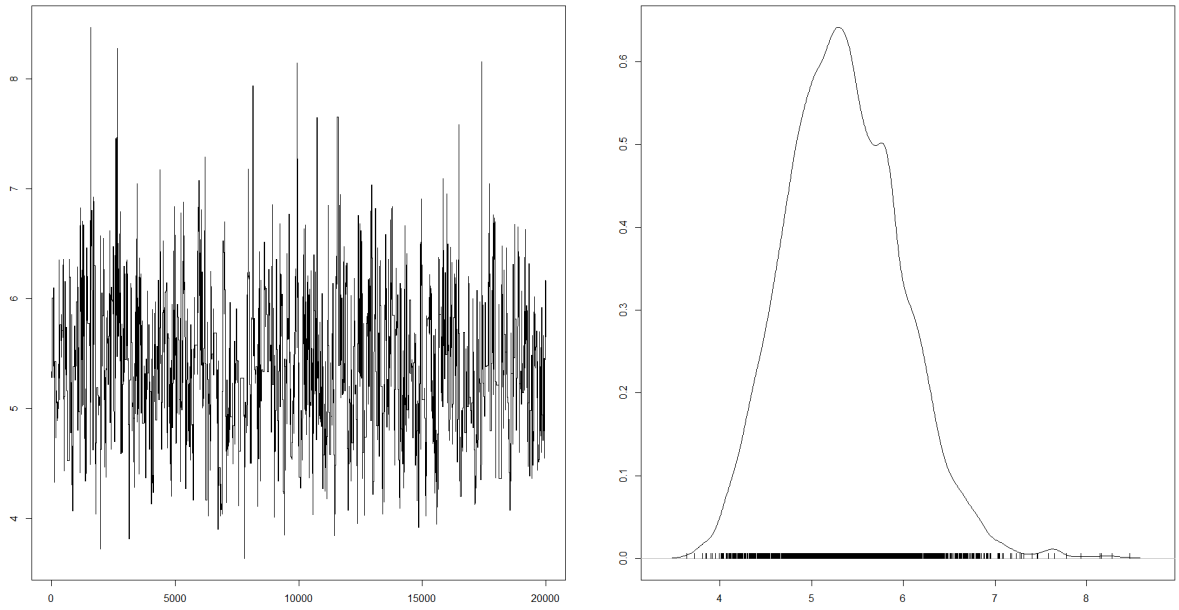

Figura A.12: Densidade e Traço do parâmetro relacionado a Conexões Mútuas do Modelo de Aresta, Conexões Mútuas e Tríade Transitiva

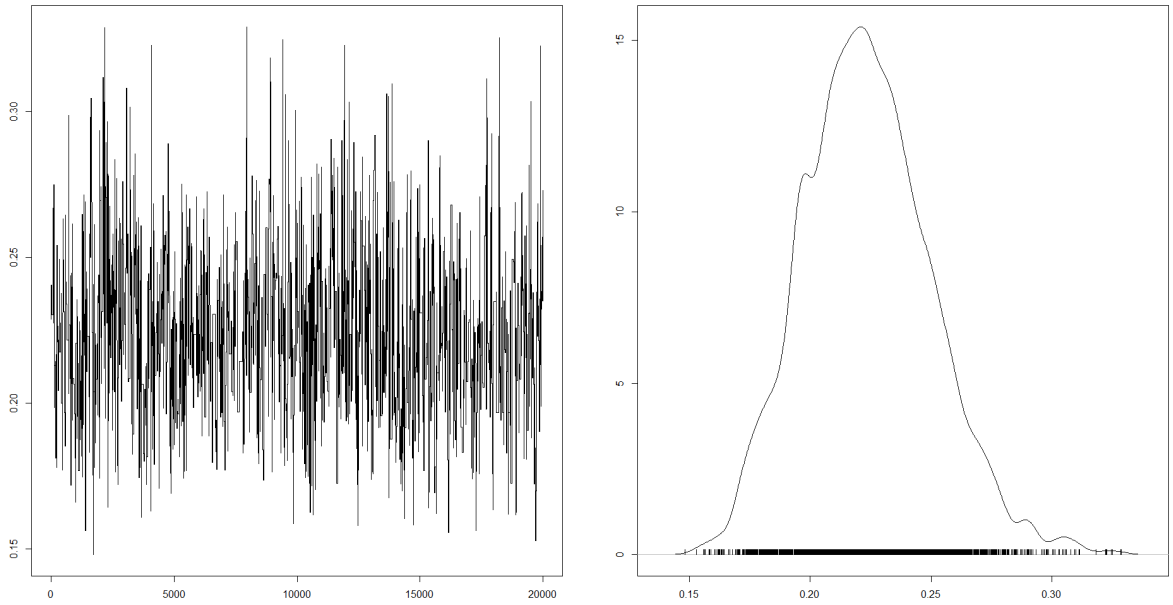

Figura A.13: Densidade e Traço do parâmetro relacionado a Tríade Transitiva do Modelo de Aresta, Conexões Mútuas e Tríade Transitiva

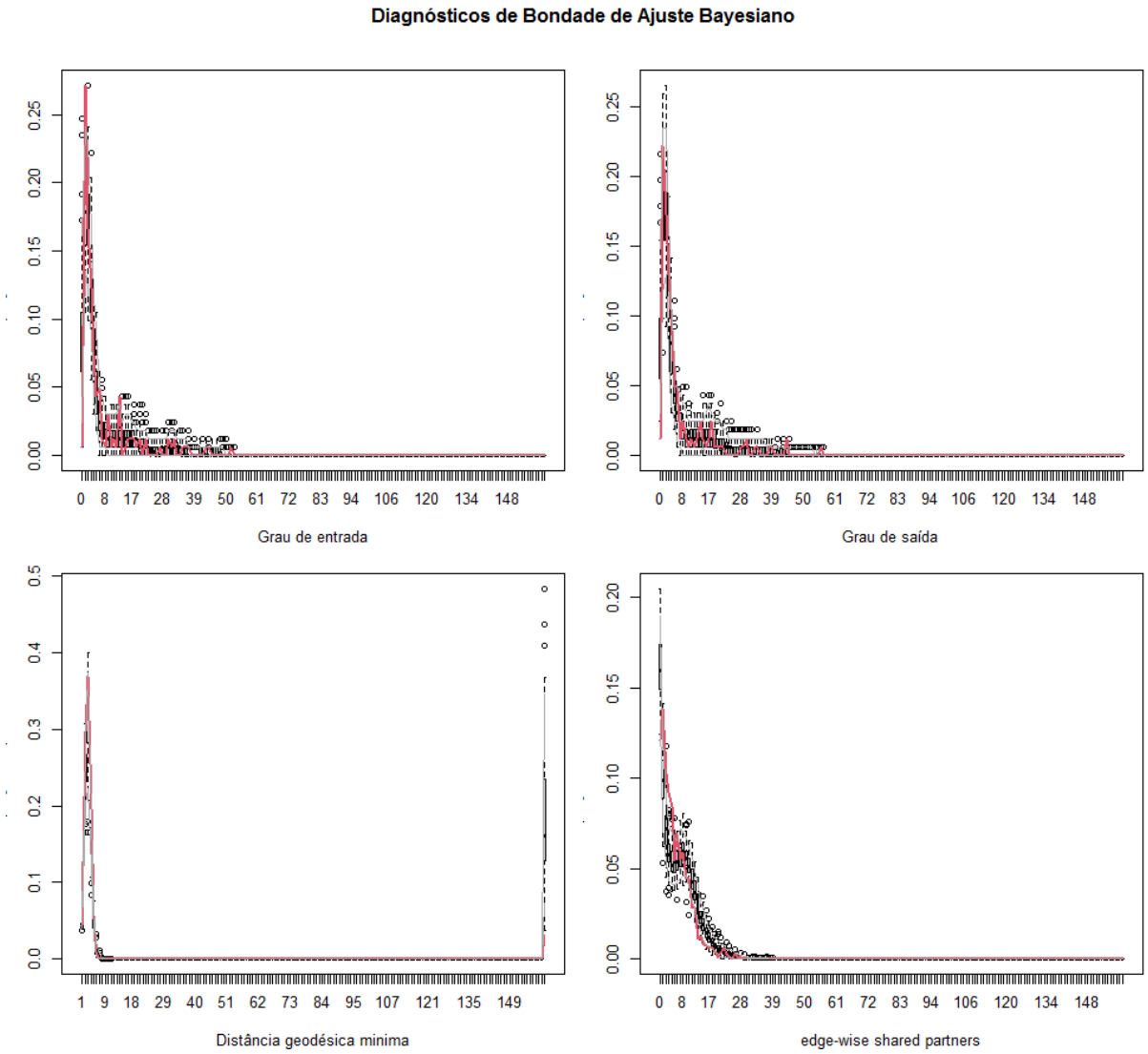

Figura A.14: Gráficos não recortados de Bondade do Ajuste Bayesiano do Modelo de Aresta, Conexões Mútuas e Tríade Transitiva

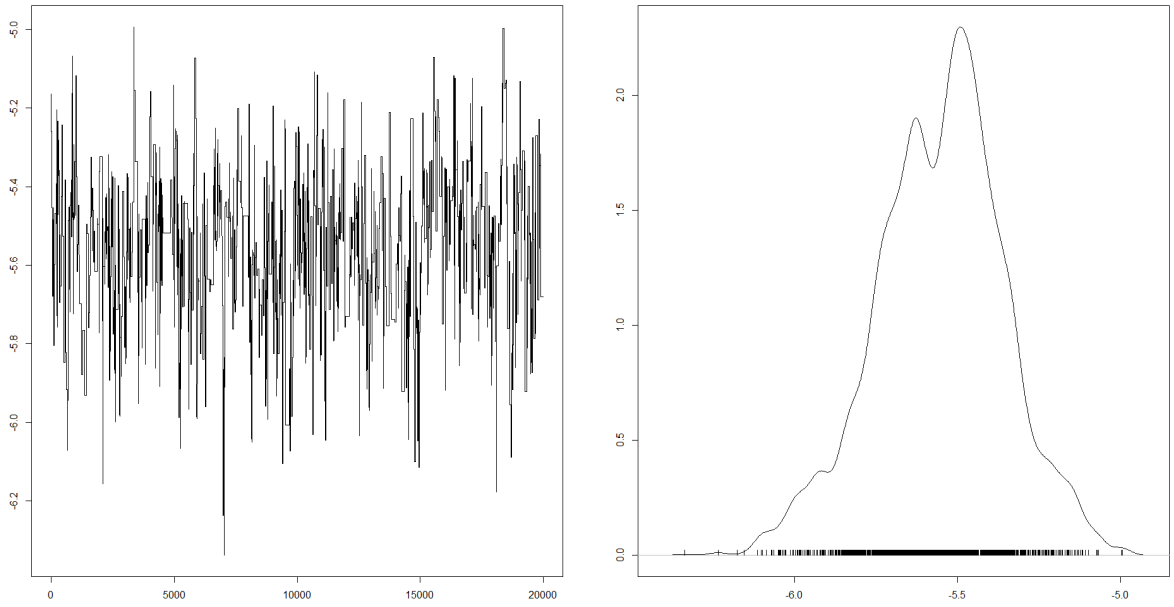

Figura A.15: Densidade e Traço do parâmetro relacionado a Aresta do Modelo de Aresta, Conexões Mútuas, Tríade Transitiva e Tríade Cíclica

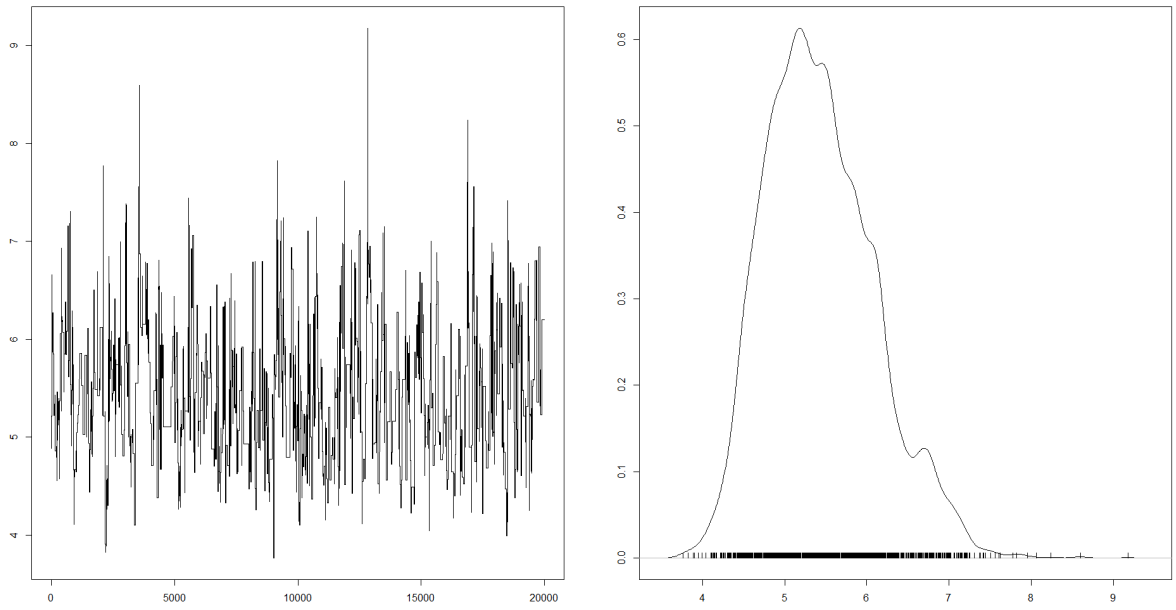

Figura A.16: Densidade e Traço do parâmetro relacionado a Conexões Mútuas do Modelo de Aresta, Conexões Mútuas Tríade Transitiva e Tríade Cíclica

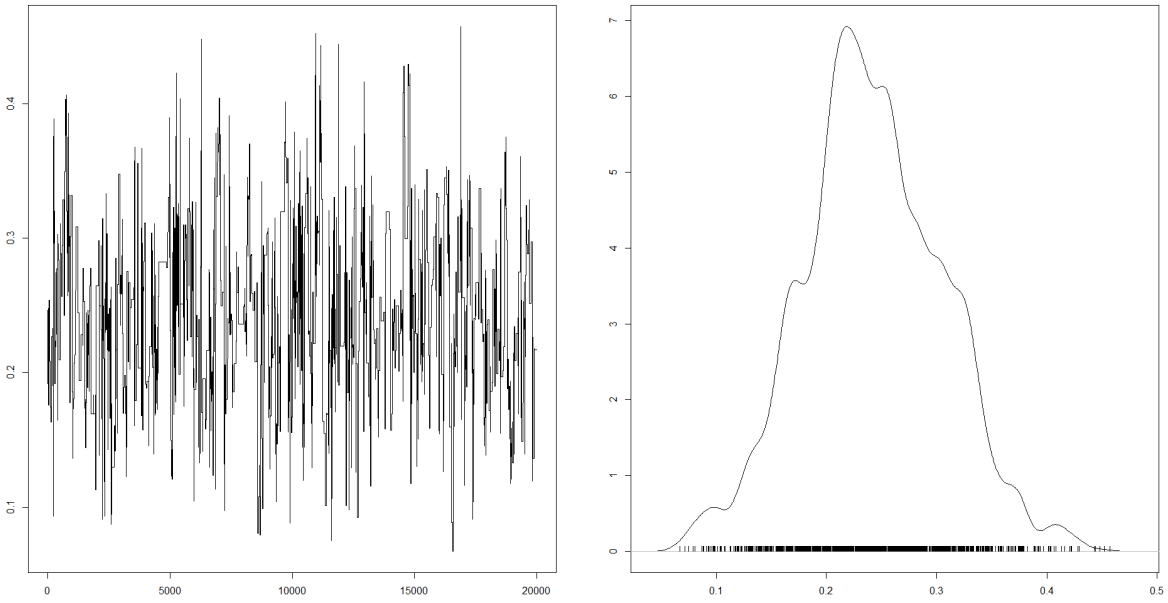

Figura A.17: Densidade e Traço do parâmetro relacionado a Tríade Transitiva do Modelo de Aresta, Conexões Mútuas Tríade Transitiva e Tríade Cíclica

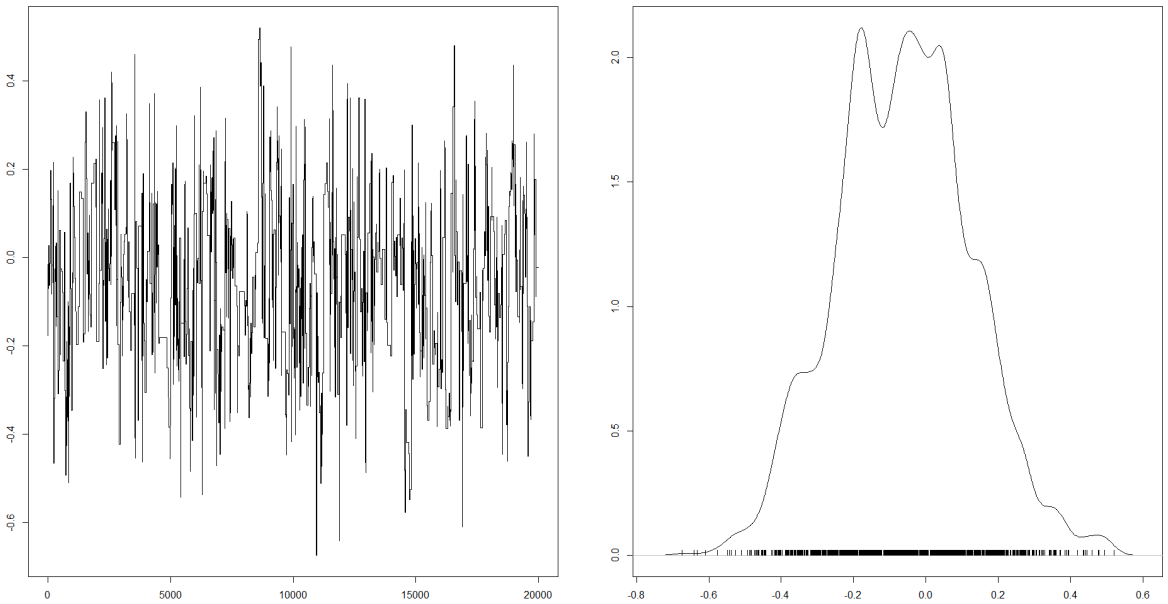

Figura A.18: Densidade e Traço do parâmetro relacionado a Tríade Cíclica do Modelo de Aresta, Conexões Mútuas Tríade Transitiva e Tríade Cíclica

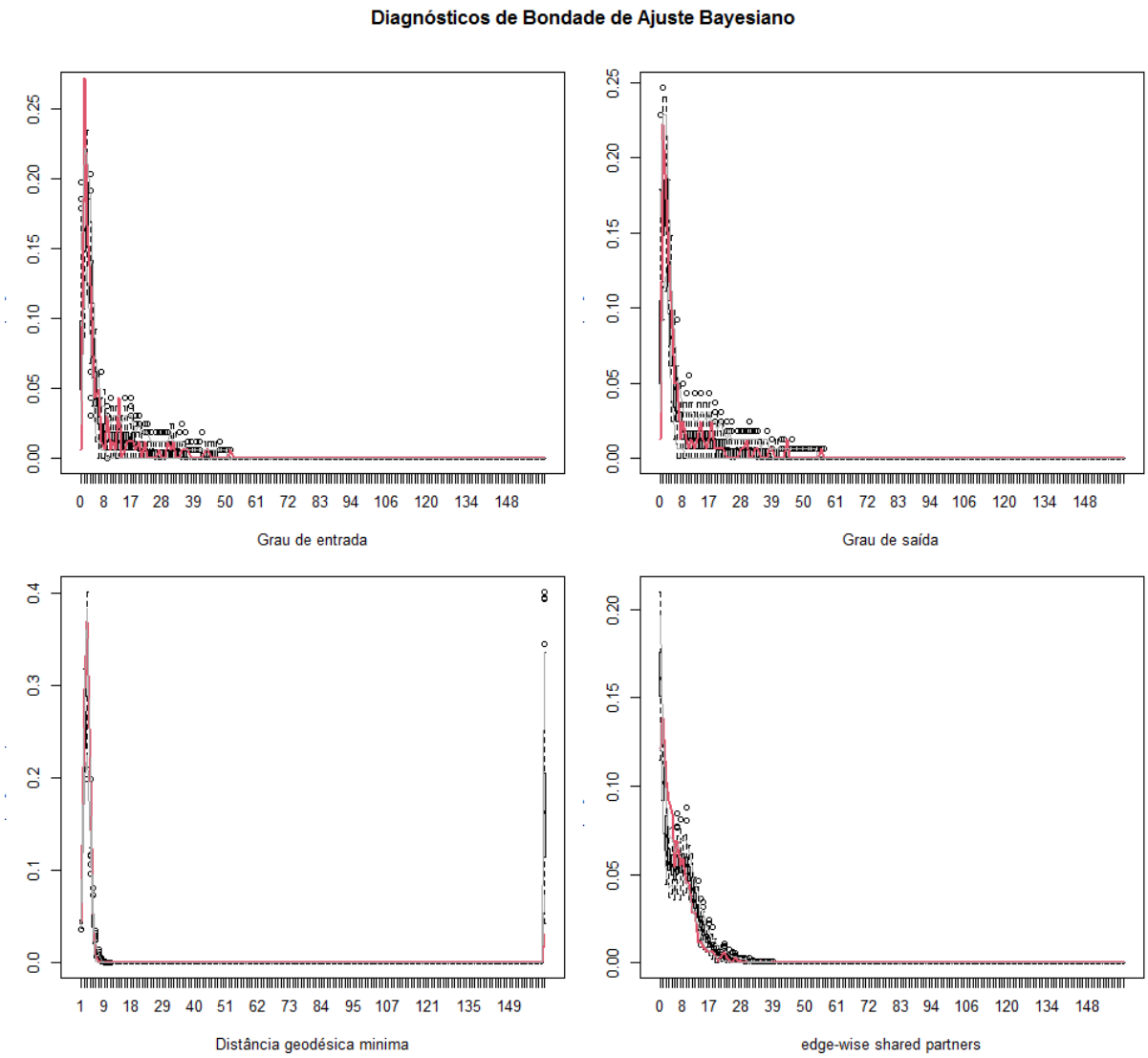

Figura A.19: Gráficos não recortados de Bondade do Ajuste Bayesiano do Modelo de Aresta, Conexões Mútuas Tríade Transitiva e Tríade Cíclica

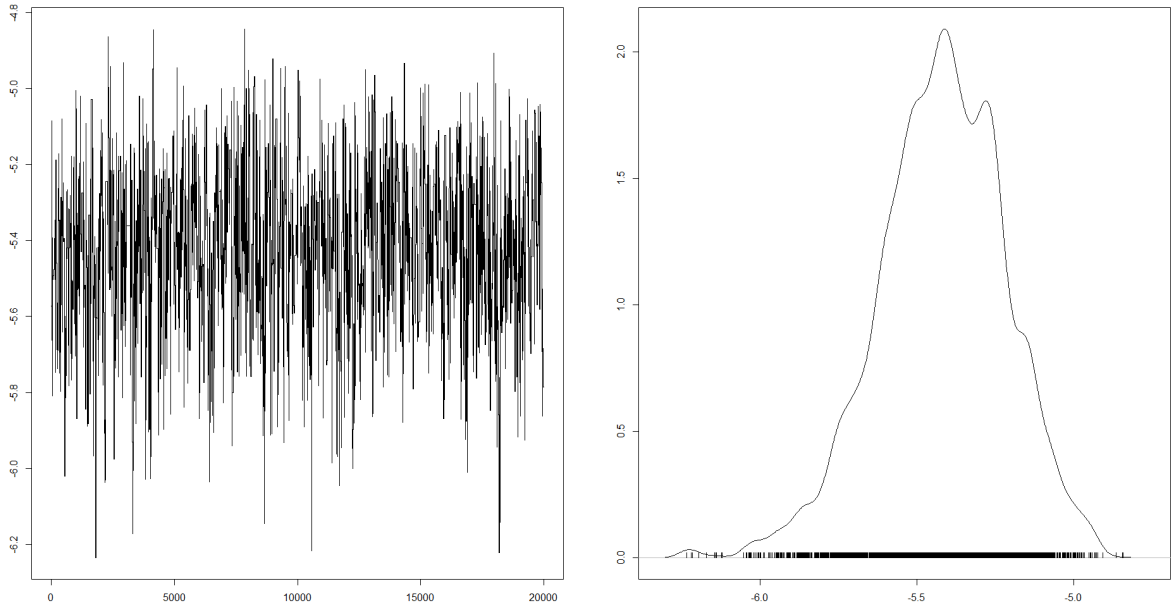

Figura A.20: Densidade e Traço do parâmetro relacionado a Aresta do Modelo de Aresta, Conexões Mútuas e Tríade Cíclica

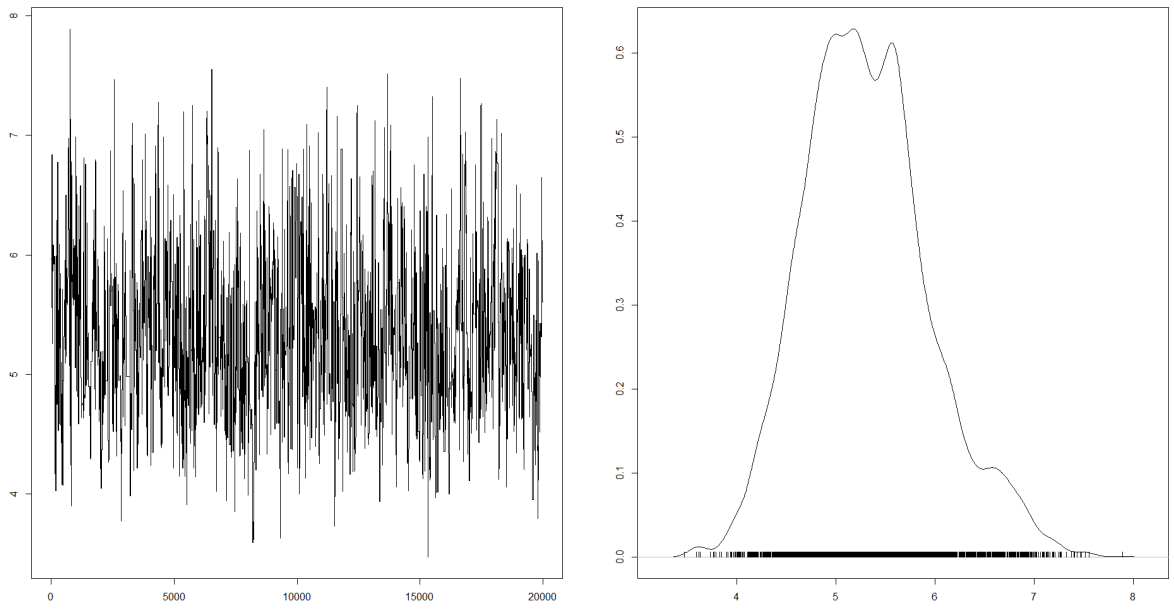

Figura A.21: Densidade e Traço do parâmetro relacionado a Conexões Mútuas do Modelo de Aresta, Conexões Mútuas e Tríade Cíclica

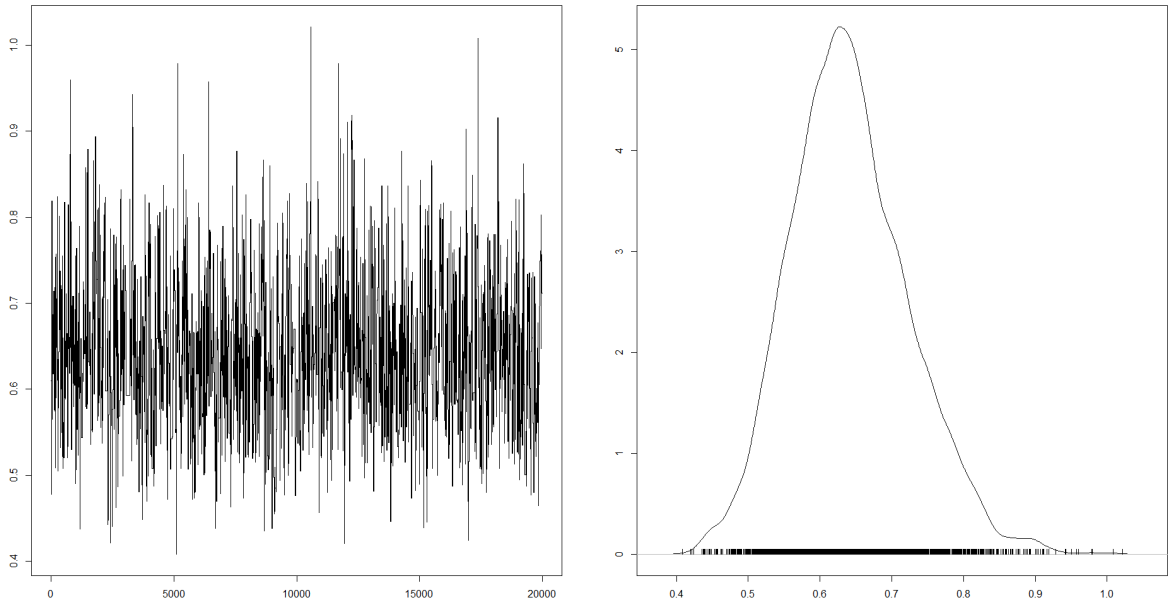

Figura A.22: Densidade e Traço do parâmetro relacionado a Tríade Cíclica do Modelo de Aresta, Conexões Mútuas e Tríade Cíclica

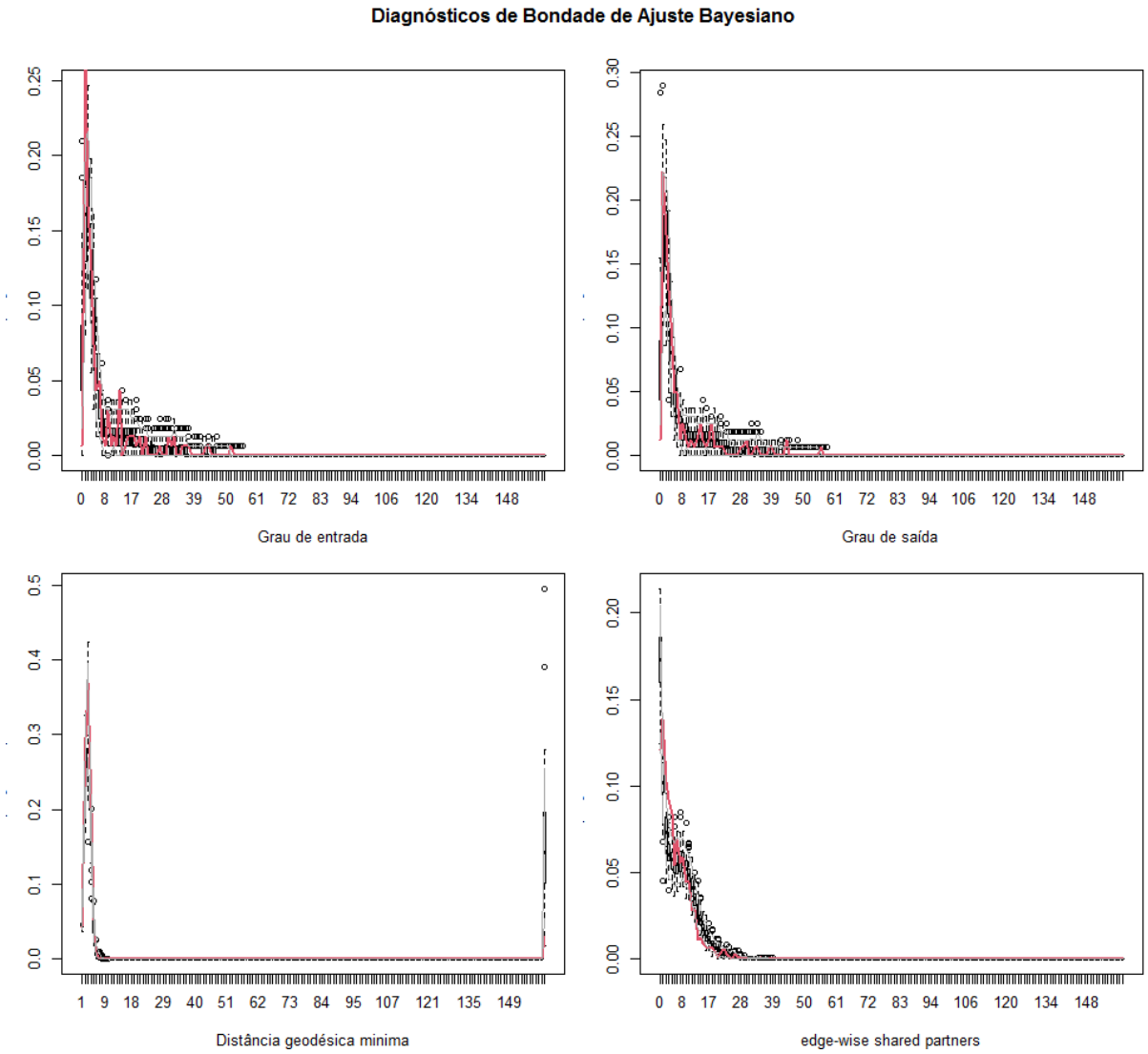

Figura A.23: Gráficos não recortados de Bondade do Ajuste Bayesiano do Modelo de Aresta, Conexões Mútuas Tríade Transitiva e Tríade Cíclica

Referências Bibliográficas

ANAC (2016). Descrição de variáveis. `https://www.anac.gov.br/assuntos/dados-e-estatisticas/descricao-de-variaveis`. Acessado em 27/11/2020.

Byshkin, M., Stivala, A., Mira, A., Krause, R., Robins, G. e Lomi, A. (2016). Auxiliary parameter mcmc for exponential random graph models. *Journal of Statistical Physics*, **165**(4), 740–754.

Caimo, A. e Friel, N. (2011). Bayesian inference for exponential random graph models. *Social Networks*, **33**(1), 41–55.

Caimo, A. e Friel, N. (2014). Bergm: Bayesian exponential random graphs in R. *Journal of Statistical Software*, **61**(2), 1–25.

Gamerman, D. e Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press.

Grandy Jr, W. T. (2012). *Foundations of statistical mechanics: Equilibrium theory*, volume 19. Springer Science & Business Media.

Holland, P. W. e Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the american Statistical association*, **76**(373), 33–50.

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M. e Morris, M. (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, **24**(3), 1–29.

Newman, M. (2012). Networks: An introduction. 2010: Oxford university press. *Artificial Life*, **18**, 241–242.

Sampson, S. F. (1969). A novitiate in a period of change: An experimental and case study of social relationships.