Trabalho de Grafos

Pedro Loes

14/03/2021

Introdução

O tema escolhido foi a portabilidade de clientes da plataforma de cursos on-line DataCamp. O banco de dados utilizado recupera as informações sobre portabilidade de alunos da plataforma. Este banco de dados faz parte do curso *Predictive Analytics Using Networked Data in R* oferecido pela mesma plataforma.

A primeira parte do trabalho consistiu na análise exploratória e descritiva do grafo por meio da ilustração da rede e cálculo de estatísticas descritivas. A segunda parte consistiu na modelagem da cooperação entre clientes da rede utilizando o modelo de Grafos Aleatórios Exponenciais.

```
# Carrega Bibliotecas
library(igraph)
library(tidyverse)
library(knitr)
library(ergm)
library(broom)

# Carrega dados
load("dados/StudentEdgelist.RData")
load("dados/StudentCustomers.RData")

# Renomeia categorias de portabilidade
clientes <- customers %>% mutate(portabilidade = ifelse(churn == 0, "Não", "Sim"))
```

- Os pacotes **tidyverse** e **knitr** foram utilizados para manipular e imprimir os dados. O pacote **igraph** foi utilizado para construção, visualização e o cálculo das estatísticas da rede. O pacote **ergm** e o pacote **broom** foram utilizados para modelar e extrair os resultados do modelo.
- O data frame de **arestas** ou ligações entre clientes possui **12491** observações.
- O data frame de **vértices** com a variável indicadora de portabilidade de clientes possui **4964** observações de clientes.

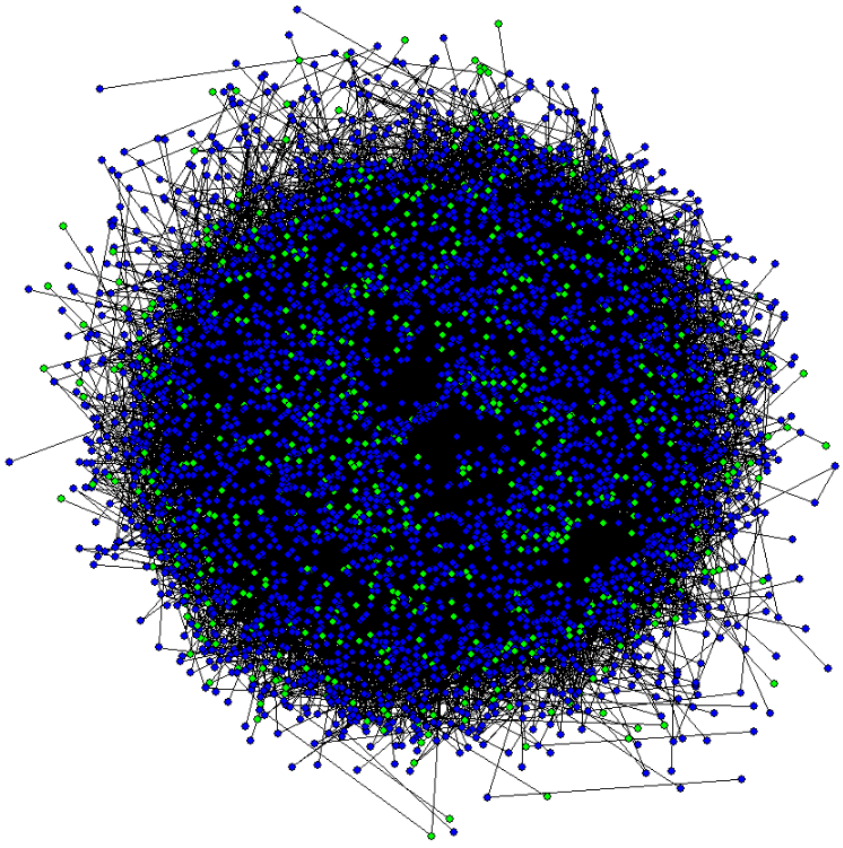
Análise Exploratória e Descritiva

- A análise exploratória e descritiva da rede consistiu em:
  - Ilustrar o grafo e um subgrafo da rede.
  - Calcular estatísticas utilizadas no universo de grafos para compreender a relações entre os vértices.
  - Produzir um gráfico do tipo histograma para visualizar a distribuição dos graus da rede.
  - Produzir um gráfico de barras para visualizar a quantidade de clientes que optaram pela portabilidade.

Grafo da Rede de Clientes

O layout **kamada.kawai** foi utilizado para desenhar o grafo. Os clientes que não optaram pela portabilidade receberam a cor **azul** e os clientes que optaram pela portabilidade receberam a cor **verde**. O parâmetro tamanho dos vértices recebeu o valor **2** para evitar a sobreposição de clientes.

```
# Declara a rede apartir do data frame de arestas
rede <- graph_from_data_frame(edgeList, directed = FALSE)

# Atribui cor verde para clientes que portaram
V(rede)$color <- gsub("1", "green", clientes$churn)

# Atribui a cor azul para os clientes que não portaram
V(rede)$color <- gsub("0", "blue", clientes$churn)

# Desenha o grafo
plot(rede,
     vertex.label = NA,
     edge.label = NA,
     edge.color = "black",
     vertex.size = 2,
     layout = layout_with_kk)
```



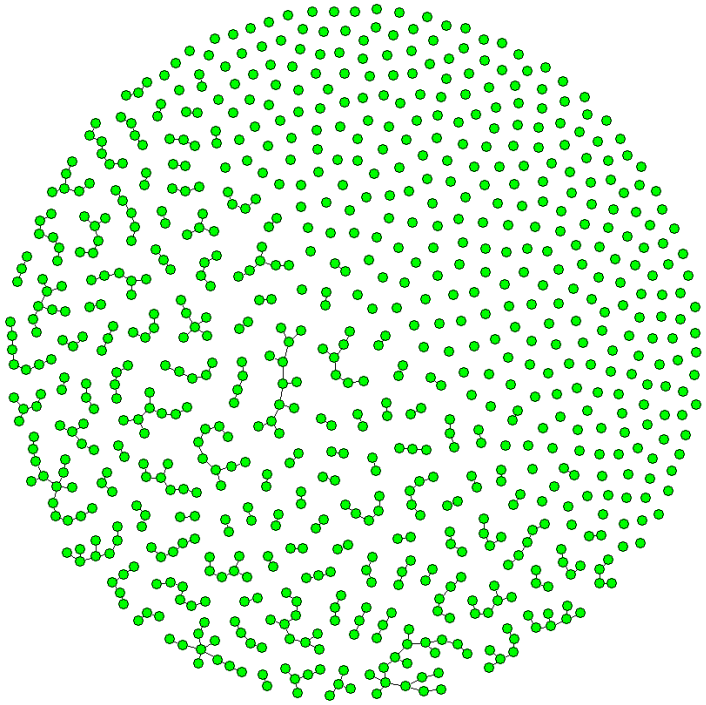
A ilustração gráfica não possibilita a percepção visual de algum padrão na estrutura da rede de clientes. Porém, é possível verificar que o número de clientes que optaram pela portabilidade é visivelmente inferior aos que permaneceram na plataforma.

Subgrafo da Rede de Clientes

O segundo grafo ilustra um subgrafo da rede completa considerando apenas com os clientes que optaram pela portabilidade. O layout utilizado para desenhar a subrede de clientes que optaram pela portabilidade foi **fruchterman reingold** porque facilitou a separação dos clientes isolados em relação aos conectados.

Todos os clientes deste subgrafo optaram pela portabilidade e portanto foram coloridos com a cor verde. As arestas receberam a cor preta. O tamanho dos vértices foi indicado com o valor **3** e o das arestas recebeu o valor **4** para facilitar a visualização das relações e evitar sobreposições de clientes tendo em vista que esta subrede possui menos vértices e menos arestas que a rede completa.

```
# Declara subgrafo de clientes que optaram pela portabilidade
rede_portabilidade <- induced_subgraph(rede, v = V(rede)[which(V(rede)$churn == 1)])

# Desenha grafico
plot(rede_portabilidade,
     vertex.label = NA,
     vertex.size = 2,
     edge.size = 4,
     edge.color = "black",
     vertex.color = "green",
     layout = layout_with_fr)
```



Pode-se observar que dos **774** vértices deste subgrafo de clientes que optaram pela portabilidade, **438**, ou **57 %** desses clientes não se comunicam com nenhum outro cliente dentre os que optaram pela portabilidade.

Dentre os clientes que optaram pela portabilidade e se conectam a outros clientes que optaram pela portabilidade, pode-se observar que o grafo **não é conectado**. As relações entre estes clientes apresentaram diversas estruturas de **árvores** e **estrelas**, mas **nenhum triângulo**.

Estatísticas Descritivas

Para compreender as características da rede de portabilidade diversas estatísticas foram calculadas para verificar **simplicidade**, **conectibilidade**, **coesão**, **comunidades**, **subgrafos** e **semelhanças**.

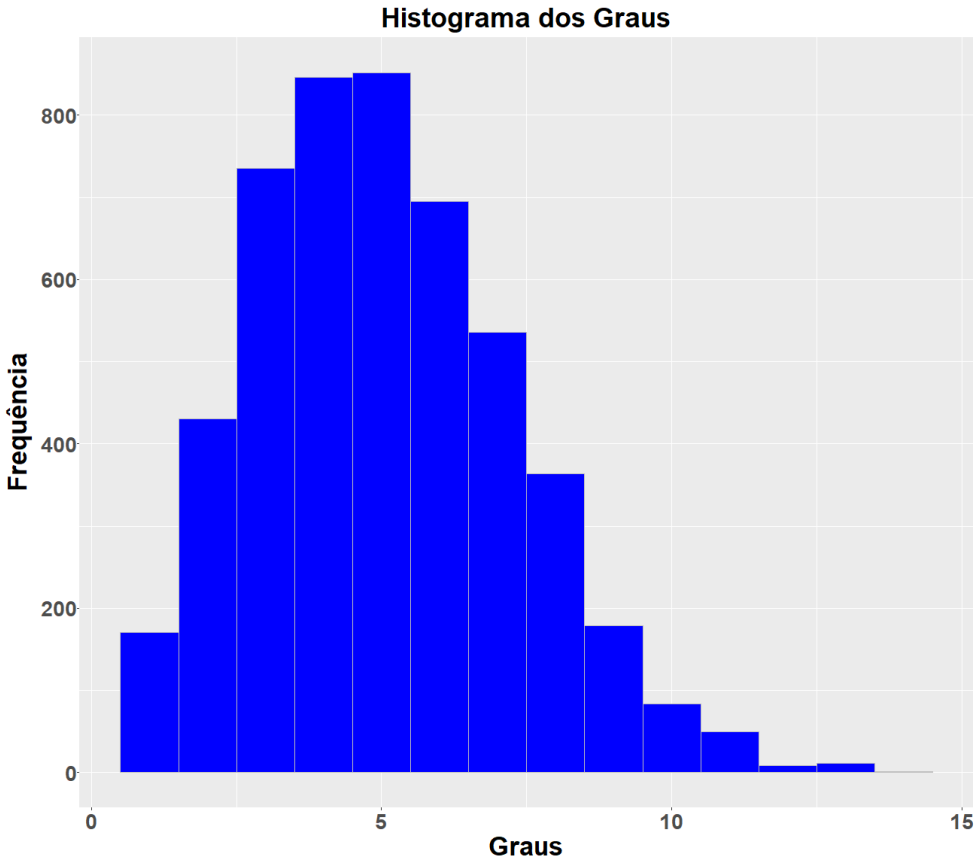
```
# Verifica se o grafo é simples
simples <- is.simple(rede)

# Verifica se a rede e conectada
conectada <- is.connected(rede)

# Calcula diâmetro
diametro <- diameter(rede)

# Calcula densidade
densidade <- graph.density(rede)

# Calcula transitividade
transitividade <- transitivity(rede)

# Calcula assortatividade nominal
assort_nominal <- assortativity_nominal(rede, (V(rede)$color == "1") + 1, directed = F)

# Calcula assortatividade de grau
assort_graus <- assortativity.degree(rede)

# Calcula número e tamanho de comunidades
comunidades <- sizes(fastgreedy.community(rede))
```

- **Estatísticas**:
  - **Simplicidade** e **Conectividade**:
    - A rede apresentou as características de ser **simples** pelo fato de não possuir loops ou múltiplas arestas e ser **conectada** pois cada vértice estava ao alcance de qualquer outro vértice.
  - **Diâmetro**:
    - A maior **distância geodésica** dos menores caminhos entre os vértices foi **11**. Esta estatística indicou que não é preciso passar por muitos vértices para atravessar a rede.
  - **Densidade**:
    - O **nº de arestas realizadas** dividido pelo **nº de arestas em potencial** foi **0.001**. A estatística indicou que a rede apresentava pouca densidade dado que $\boldsymbol{D \in [0, 1]}$.
  - **Transitividade**:
    - O percentual de **trincas** que se conectavam e formavam **triângulos** foi de **0.105%**. Esta estatística indicou uma **probabilidade pequena** de coesão nas estruturas de comunidades.
  - **Clusters**:
    - O número de **clusters aglomerativos hierárquicos** foi **19**. O tamanho das comunidades variou entre o mínimo de **54** e o máximo de **567**.
  - **Assortatividade Nominal**:
    - A **Homofilia de Portabilidade** apresentou $\boldsymbol{r_a = 0.0121}$. Esta estatistica indicou que os clientes apresentaram correlação positiva fraca com a portabilidade de seus vizinhos.
  - **Assortatividade de Graus**:
    - A **Homofilia de Graus**, apresentou $\boldsymbol{r_a}$ = **-0.0074** . Esta estatística indicou que os clientes apresentaram correlação negativa fraca de graus com seus vizinhos.

Distribuição dos Graus

A distribuição dos graus representa a contagem do número de vértices em cada grau do grafo. Para ilustrar a forma da distribuição foi utilizado o gráfico do tipo histograma.

```
# Cálcula graus da rede
graus <- degree(rede)

# Desenha Histograma de Graus
ggplot(as_tibble(graus), aes(x = value))+
    geom_histogram(binwidth = 1,
                   fill = "blue",
                   col="grey",
                   position = 'dodge')+
    xlab("Graus")+
    ylab("Frequencia")+
    ggtitle("Histograma dos Graus")
```



Pode-se observar que a distribuição da estatística de graus dos vértices apresenta **amplitude** de mínimo **1** e máximo **14** graus, com **1º quartil = 3** e **3º quartil= 6**. A distribuição apresenta uma **cauda a direita ou positiva** indicando que existem poucas observações atípicas que apresentaram estatística de grau de incidência nos vértices maior que **10**.

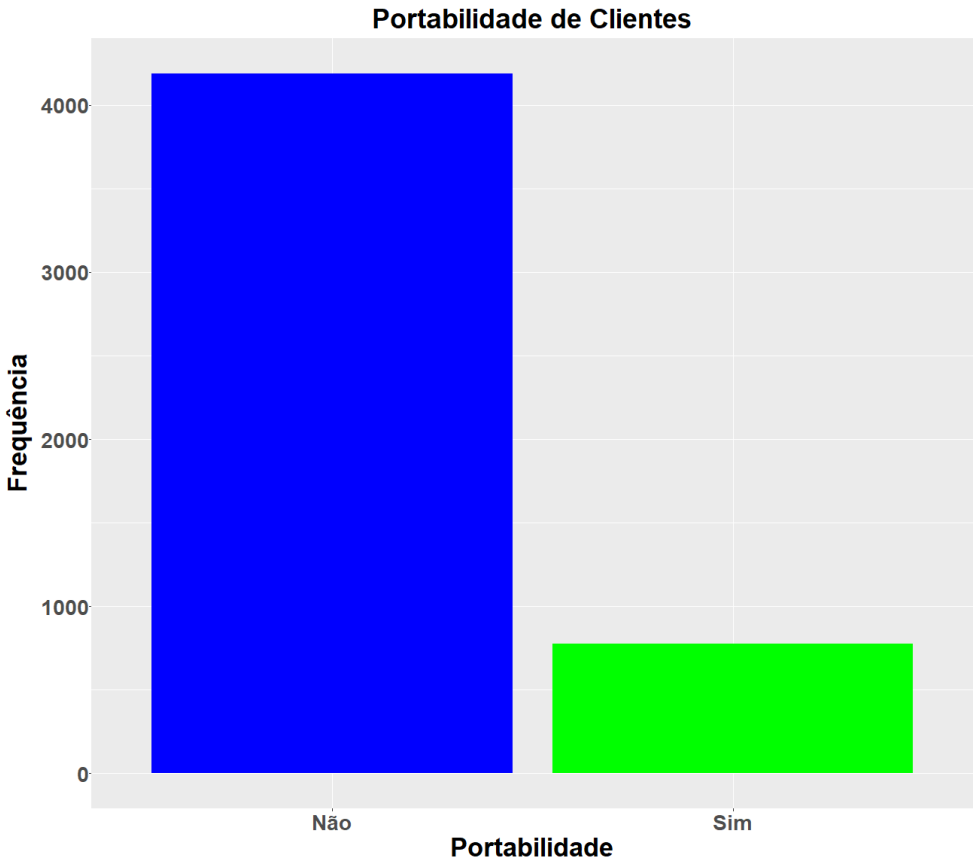
A aplicação do teste de Shapiro-Wilk indicou a rejeição da hipótese de normalidade da distribuição dos graus apresentando uma estatística **W = 0.96589** com P-Valor de **9.86e-33**.

Gráfico de Barras da Portabilidade

Para inspecionar a quantidade de clientes que optaram pela portabilidade foi produzido um gráfico de barras com duas barras. A primeira representa a frequência dos clientes que permaneceram na plataforma e a segunda representa os clientes que optaram pela portabilidade.

Os clientes que não optaram pela portabilidade receberam a cor azul e os clientes que optaram pela portabilidade receberam a cor verde compatíveis com o grafo da rede.

```
# Desenha gráfico de barras da portabilidade de clientes
ggplot(clientes, aes(x = as.factor(portabilidade),
                     fill = as.factor(portabilidade) )) +
    geom_bar() +
    xlab("Portabilidade") +
    ylab("Frequencia") +
    scale_fill_manual("Portabilidade",
                      values = c("blue", "green"))
```



- O total de clientes que optaram por permanecer na plataforma foi de **4190**.
- O total de clientes que optaram pela portabilidade foi de **774**.
- A taxa de conversão foi $\approx$ **16 %**.

Modelos de Grafos Aleatórios Exponênciais

Para verificar se existe cooperação ou ligação entre os clientes em função da portabilidade e outras estruturas de subgrafos foi ajustado o modelo de Grafos Aleatórios Exponenciais.

Especificações do Modelo

- Número de arestas:
  - $S_1 = \sum_{i,j} y_{i,j}$
- Solução de Snijders de restrição paramétrica para alternância da estatística k-estrelas:
  - $AKS_\lambda(y) = \sum_{k=2}^{N_v - 1} (-1)^k \frac{S_k(y)}{\lambda^{k-2}}$
- Generalização de estruturas triádicas baseadas na soma alternada de triângulos:
  - $AKT_\lambda(y) = 3T_1 + \sum_{k=2}^{N_v - 2} (-1)^{k+1} \frac{T_k(y)}{\lambda^{k-1}}$
- Estatística do atributo portabilidade:
  - $g(y, x) = \sum_{1<i<j<N_v} y_{i,j} h(x_i, x_j)$
- Modelo Completo:
  - $P_{\theta,\beta}(Y = y|X = x) = (\frac{1}{k(\theta,\beta)}) exp(\theta_1 S_1(y) + \theta_2 AKS_\lambda(y) + \theta_3 AKT_\lambda(y) + \beta g(x, y))$

```
# Define rede do tipo network a partir de matriz de adjacências
rede.s <- network::as.network( as.matrix( get.adjacency(rede) ), directed = F)

# Define atributo de portabilidade
network::set.vertex.attribute(rede.s, "Portabilidade", clientes$churn)

# Especifica modelo
minha.ergm <- formula(rede.s ~ edges + kstar(2) + kstar(3) + triangle +
                      nodemain("Portabilidade") + match("Portabilidade"))

# Recupera resumo da especificação
especs <- t(t(summary(minha.ergm)))

# Declara data frame de resultados
especs <- tibble(Estatisticas = row.names(especs),
                 Contagem = especs[,1] %>%  as.vector())

# Imprime tabela
kable(especs)
```

Estatisticas	Contagem
edges	12491
kstar2	62622
kstar3	104539
triangle	22
nodecov.Portabilidade	3972
nodematch.Portabilidade	9191

Ajuste do Modelo

O modelo foi ajustado usando métodos numéricos para estimar o log da máxima verossimilhança dos coeficientes $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$ e $\hat{\beta}$ utilizando **Monte Carlo** com **Cadeias de Markov** .

```
# Indica Semente
set.seed(10)

# Ajusta modelo
modelo <- ergm(minha.ergm, set.control.ergm = control.ergm(MCMC.burnin = 1e5))

# Imprime tabela de Resumo do modelo
kable(summary(modelo)$coefficients)
```

Atributos	Estimativa	Erro Padrão	MCMC %	Valor z	Pr(>|z|)
Arestas	-6.8273396	0.0974004	0	-70.0956087	0.0000000
Estrela-2	-0.0199553	0.0165505	0	-1.2057199	0.2279255
Estrela-3	0.0030715	0.0031541	0	0.9738131	0.3301493
Triângulo	0.0295085	0.2138504	0	0.1379868	0.8902508
nodcov.Portabilidade	0.0607622	0.0294942	1	2.0601411	0.0393851
nodmatch.Portabilidade	0.0517534	0.0328415	1	1.5758541	0.1150594

O coeficiente **edges** que representa a probabilidade homogênea de ligação (**ties**) entre vértices apresentou **P-valor** significativo de **1e-10**. O log da chance de ocorrer um vínculo (**tie**) é **-6.8273** $\times$ **1**. Considerando todos os vínculos $\frac{exp(-6.8273)}{1+exp(-6.8273)} \approx$ **0.0011**. Esse valor corresponde a densidade **0.001** observada na etapa das estatísticas descritivas da rede.

O coeficiente de efeito positivo **nodcov.Portabilidade** apresentou **P-valor** = **0.03** significativo que pode ser interpretado como **log da razão da chance** de cooperação entre clientes condicionado na portabilidade. O coeficiente exponenciado **exp(0.061)** $\approx$ **1.0626462** indicou que os clientes que optaram pela portabilidade apresentaram chance de cooperação de aproximadamente **6 %** mais do que os clientes que não optaram pela portabilidade.

Análise de Variância

A análise de variância foi utilizada para comparação com o modelo sem nenhuma variável e verificar o quanto da variação de conectividade da rede foi explicada pelas variáveis.

```
# Executa Anova e recupera resultados da comparação ente modelos
kable(tidy(anova(modelo)) %>% rename("Pvalor" = "Pr...Chisq..") %>%
        replace_na(list(df = "-", Deviance = "-", Pvalor = "-")))
```

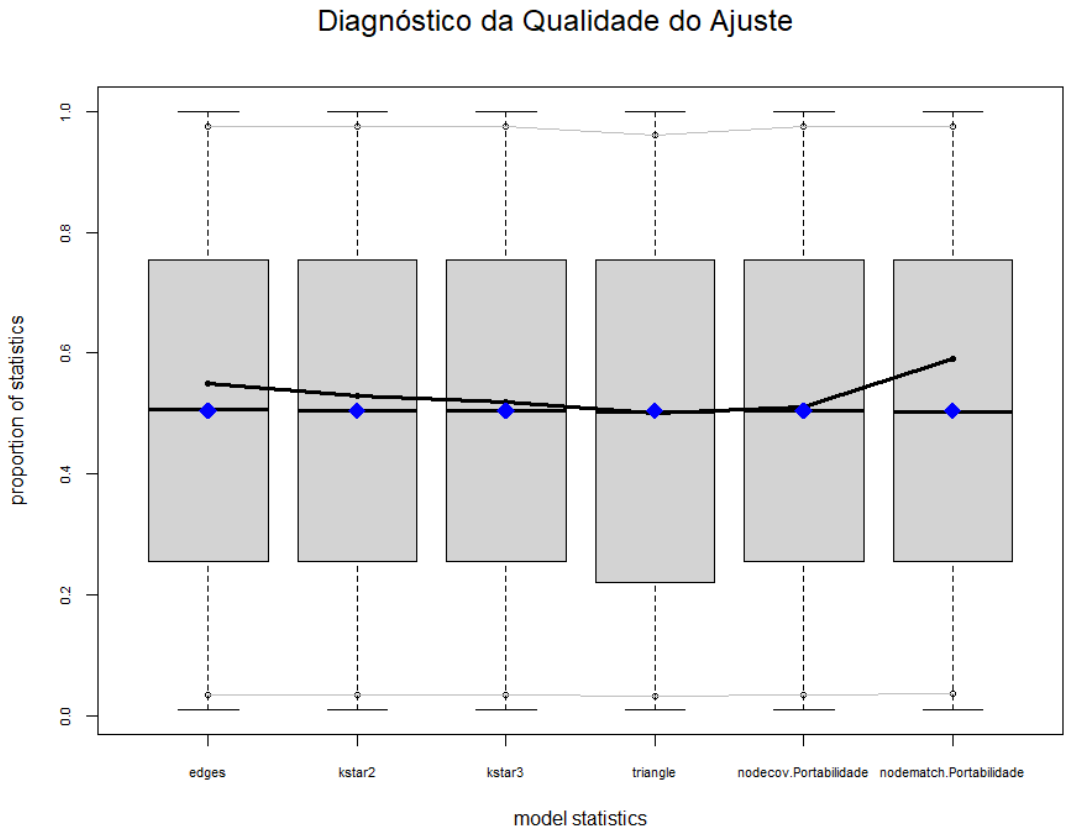
Modelo	Gl	Desvio	Resíduo Gl	Desvio Resíduos	Pvalor
Nulo	-	-	12318166	17076604.1	-
Especificado	6	16879417.9939368	12318160	197186.1	0

O resultado da análise de variância indicou uma forte evidência de que o modelo especificado explica uma parte da variabilidade das conexões da rede se comparado ao modelo nulo com **P-valor** de **2.22e-16**.

Qualidade do Ajuste

Para verificar a qualidade do ajuste foram implementadas simulações Monte Carlo de grafos aleatórios semelhantes para comparar características como distribuição dos graus, comprimento das geodésicas e número de vizinhos compartilhados por um par de vértices.

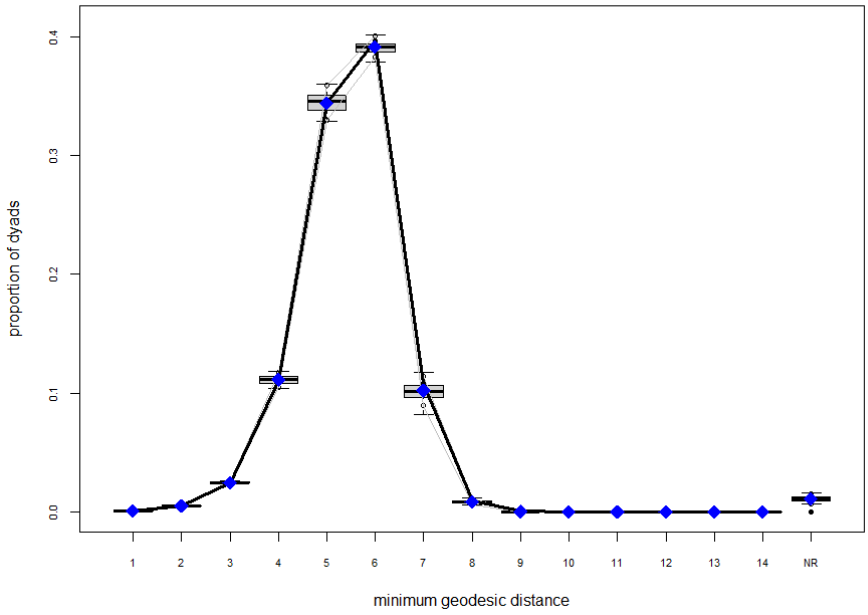
```
# Simula grafos aleatórios para comparar ajuste
bondade_ajuste <- gof(modelo)

# Desenha gráficos
plot(bondade_ajuste, main = "Diagnóstico da Qualidade do Ajuste")
```



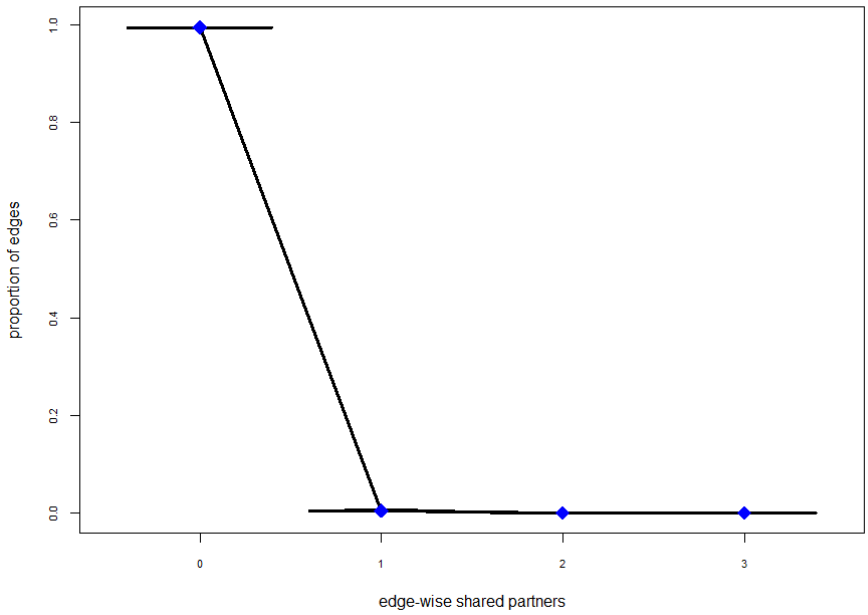
A proporção das estatísticas de contagem do número de **arestas**, e **nodematch.Portabilidade** do modelo ajustado foi superior a **mediana** das simulações de **grafos aleatórios MCMC**. Porém estas estatísticas estão localizadas dentro dos intervalos entre o **1º** e **3º** quartis indicando bom ajuste do modelo referente a estas estatísticas.

A proporção das estatística de contagem de ***kstar2*** e **kstar3** do modelo ajustado apresentaram estimativas muito próximas da **mediana** das simulações de **grafos aleatórios MCMC**. Tal fato indicou ótimo ajuste do modelo em relação a estas estatísticas.

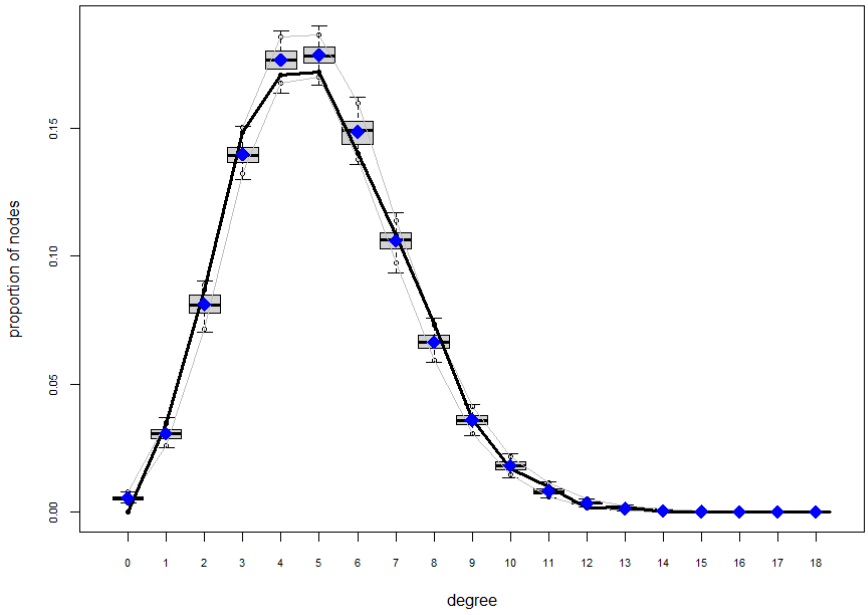
A proporção das estatística de contagem de **triângulos** e **nodecov.Portabilidade** do modelo apresentaram estimativas praticamente **iguais a mediana** das simulações de **grafos aleatórios MCMC**. Tal fato indicou excelente ajuste do modelo para estas estatísticas.



A qualidade do ajuste da proporção de **dyads**, ou seja, estrutura de subgrafo da ordem **2** onde existem **2** vértices, independente de existir conexão entre eles, em relação a cada **distância geodésica mínima**, apresentou valores muito próximos da **mediana** da distribuição de **dyads** da simulação de **grafos aleatórios MCMC** para todas as distâncias. Tal fato indicou um bom ajuste deste tipo de estrutura.



A proporção de **arestas** para o número de **vizinhos compartilhados por um par de vértices** parece apresentar valores semelhantes à simulação **grafos aleatórios MCMC**. Porém as distribuições da estatística desta proporção apresentaram variância muito pequena não sendo possível verificar os quartis.



A proporção de **vértices** em cada **grau** só apresentou compatibilidade com as simulações para graus **inferiores** a **4** e **superiores** a **5**. Tal fato indicou que o modelo ajustou uma proporção inferior de vértice com graus **4** e **5** nos **dados da rede de portabilidade** se comparado as **simulações aleatórias MCMC**. Nestes graus a proporção de vértices não pertence ao intervalo entre o **1º** e o **3º** quartil das simulações.

Conclusões

O modelo apresentou **ajuste significativo** para explicar a **conectividade da rede de clientes**. A qualidade do ajuste indicou que o modelo é **similar** aos grafos simulados aleatoriamente com a mesma estrutura de coeficientes. Somente na estatística **proporção de vértices por grau**, o modelo parece **distoar** das simulações na região central da distribuição.

O coeficiente **nodecov.Portabilidade** indicou que os grupos que optaram pela **portabilidade cooperavam** mais na rede. Tal evidência poderia sugerir que clientes **mais conectados** tenderiam a **conhecer novas plataformas** e a trocar de plataforma com mais frequência. Outra razão seria supor que clientes mais engajados tenderiam a **completar seus estudos** e migrar para plataformas diferentes.

A compreensão da variabilidade na conexão entre os indivíduos poderia ser melhor examinada com a **mineração de mais dados** sobre o comportamento dos clientes. No que diz respeito a implementações futuras, **modelos de classificação binária** poderiam ser considerados para tentar **inferir a portabilidade de clientes** considerando como atributos as estatísticas das estruturas desta rede.

Referências

- Kolaczyk, E. D. (2010). Statistical analysis of network data: Methods and models. New York: Springer.
- statnet
- Predictive Analytics Using Networked Data in R