

Análise de Redes

RELATÓRIO DE PROJETO

RELATÓRIO TÉCNICO SOBRE O TRABALHO DE GRUPO DE ANÁLISE DE REDES

TRABALHO REALIZADO POR:

DANIEL FONSECA | 125158 | CD-PL

FRANCISCO GONÇALVES | 130649 | CD-PL

GUILHERME PIRES | 131658 | CD-PL

JOÃO FILIPE | 130665 | CD-PL

DOCENTES:

Maria João Frazão Lopes

Ana Catarina Nunes

Parte I

A rede apresentada é relativa a contactos entre pessoas numa festa, sendo que os **nodos representam pessoas** e as **ligações interações** entre as pessoas presentes na festa. Ao longo deste capítulo, abordaremos a rede como **Rede Festa**.

Análise da Rede da Festa

a) **Indique a dimensão e o número de ligações da rede. Determine a densidade e classifique a rede.**

A rede representa as interações entre **113 pessoas** - como demonstrado na Figura 1 - tendo sido registadas **2196 interações** no total. A densidade é 0,347, o que significa que **34,7%** das possíveis interações entre os convidados ocorreram. Apesar do número elevado de interações, a rede pode ser considerada relativamente esparsa, tendo em conta o contexto social em análise.

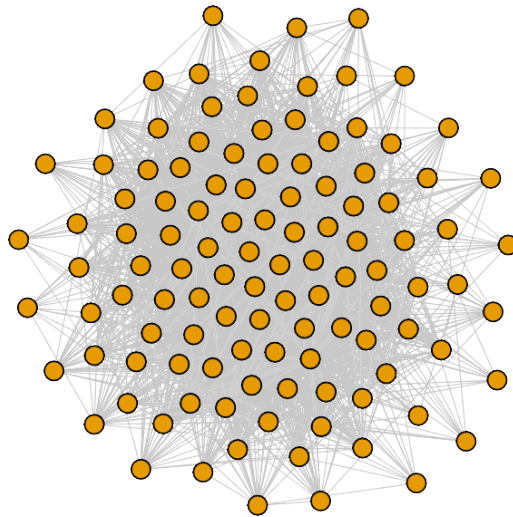


Figura 1: Rede da Festa

b) **Obtenha o grau médio e a distribuição de grau. Caracterize a distribuição de grau (mediana, quartis, etc.) e comente. Calcule o parâmetro de heterogeneidade. Indique o que pode concluir quanto à existência de hubs.**

- **Grau Médio:** 38.8672566
- **Mediana :** 37
- **Mínimo :** 1
- **Máximo :** 98
- **1º Quartil :** 27
- **3º Quartil :** 50
- **Heterogeneidade :** 1.222905

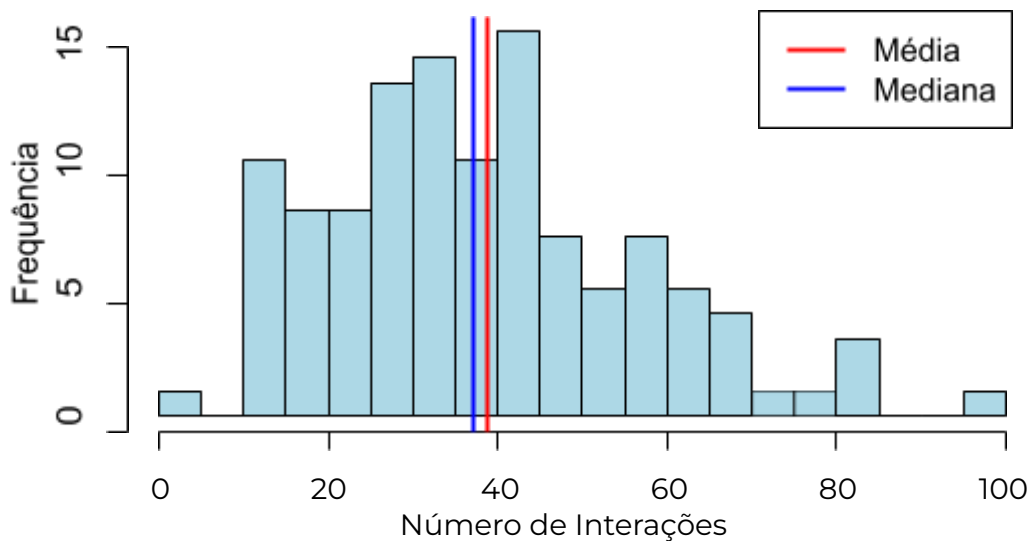


Figura 2: Distribuição dos graus

Através das estatísticas da distribuição de grau na **Rede Festa**, inferimos que a pessoa típica interagiu com **37** pessoas, o que representa uma festa bastante social.

Como a **mediana é menor que a média** - visível na Figura 2 - a distribuição é **assimétrica à direita**, o que indica que algumas pessoas interagiram com mais participantes do que o normal.

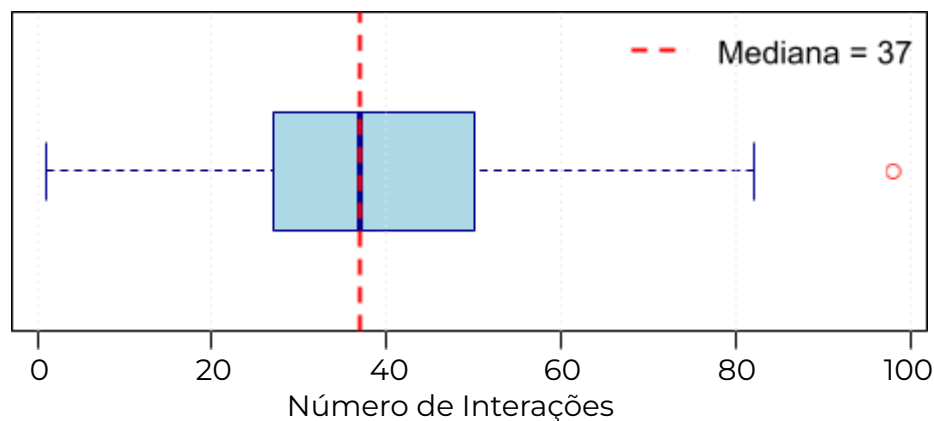


Figura 3: Boxplot da distribuição de graus na Festa

Também inferimos que a maioria das pessoas (intervalo entre Q1 e Q3 da Figura 3), teve entre **27** a **50** interações.

Há valores elevados, sendo o **máximo 98 interações**, ou seja, alguém interagiu com quase todos os participantes, logo é um candidato para *hub*. Em contraste, alguém esteve isolado na festa, com apenas 1 interação.

A rede é considerada heterogênea devido ao seu valor de heterogeneidade - **1.223**. Inferimos que há alguma variância nas interações registadas, e, devido à distribuição assimétrica à direita, concluímos que **existem hubs**, por exemplo, a pessoa com 98 interações.

c) **Estude a associação de grau e indique o que poderá concluir-se.**

A associação de grau é uma medida que permite avaliar se existe uma correlação entre os graus dos nodos - participantes da festa. É um coeficiente que varia de -1 a 1.

- **Valor positivo** indica que nodos com grau alto tendem a ligar-se a outros nodos de grau alto.
- **Valor negativo** indica que nodos com grau alto tendem a ligar-se a nodos de grau baixo.
- E um **valor igual a 0** indica que não há nenhuma correlação.

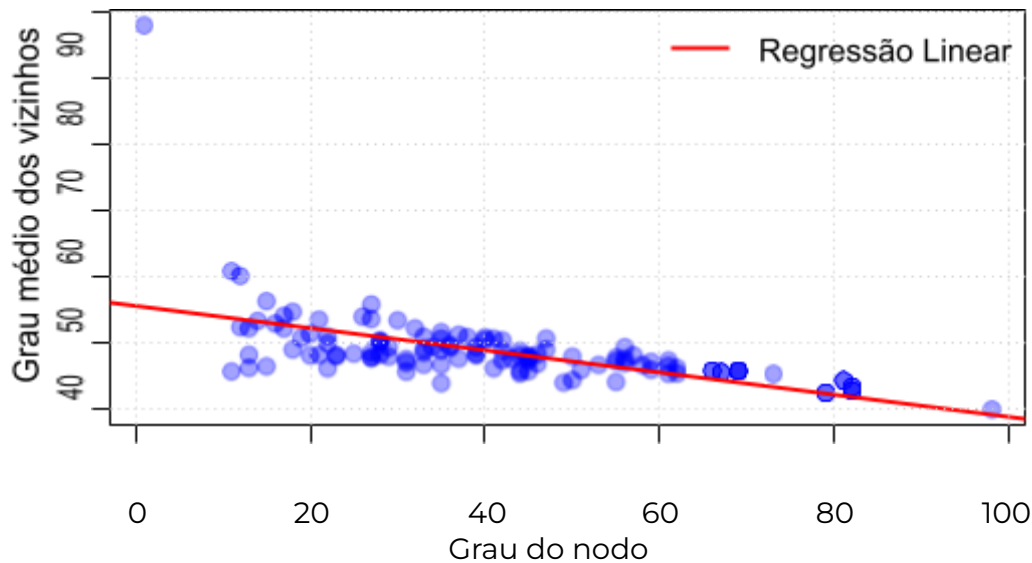


Figura 4: Grau do nó vs Grau médio dos vizinhos

A Figura 4 permite perceber se as pessoas com **mais interações** tendem a interagir com outras igualmente populares ou com pessoas menos conectadas.

A assortatividade ligeiramente negativa (**-0.12**) indica que houve uma tendência: pessoas muito sociáveis tendem a interagir com pessoas menos sociáveis - Tabela 1.

Tabela 1: Nodos com grau 1 e respetivo kNN

nodo	grau	knn
74	1	98
110	1	98

d) **Determine a média dos comprimentos dos caminhos mais curtos. Indique o que pode concluir-se quanto à distância média.**

A média dos comprimentos dos caminhos mais curtos na rede foi: **1.64**.

Este valor indica o **número médio de passos necessários para que uma pessoa possa chegar a outra através de uma cadeia de interações**.

No contexto da festa, isto significa que, em média, qualquer pessoa está apenas a cerca de **1,64 interações de distância de qualquer outra**. Este valor sugere um comportamento típico de redes

small-world: mesmo com mais de uma centena de participantes, a distância média entre dois indivíduos permanece bastante reduzida. Para referência, a teoria dos “seis graus de separação” afirma que qualquer pessoa está, no máximo, a seis ligações de qualquer outra no mundo. Assim, na festa, teoricamente, para chegar ao indivíduo X seriam necessários apenas cerca de 1,642 “passos”!

e) Determine os coeficientes de clustering dos nodos e da rede. Diga o que pode concluir-se quanto à existência de triângulos.

O coeficiente de **clustering local** mede quanto os amigos de uma pessoa também são amigos entre si, ou no contexto da rede, se eu interagi com os indivíduos X e Y, o clustering local verifica se X e Y também interagiram um com o outro. Se sim, forma-se um triângulo entre “eu, X e Y”. O **clustering global**, da rede, avalia se a festa teve muitos grupos de 3 pessoas completamente conectados.

Desta forma, a Rede da Festa, teve um clustering local de **0.535**, o que é considerado bastante elevado para uma rede social, que revela como os participantes agiram na festa: em grupos coesos, com muitas interações dentro de pequenos círculos sociais, por exemplo amigos que vieram juntos, colegas de trabalho ou conhecidos antes da festa.

Observou-se alguns nodos com coeficientes de clustering elevados (**0.7 e 0.89**). Estes encontram-se inseridos em grupos altamente coesos, nos quais as pessoas interagem intensamente entre si. Estes nodos podem ser interpretados como pertencendo a ‘grupos fechados’ ou círculos sociais muito compactos dentro da festa.

f) Efetue a decomposição de core da rede. Indique o número de conchas existentes e a dimensão de cada uma. Indique o que poderá concluir-se.

O **core da rede** ajuda a identificar núcleos densamente conectados dentro de uma rede. O **k-core** de uma rede é um subgrafo que cada nodo tem pelo menos **k ligações** dentro desse subgrafo. No contexto da rede, um k=3 (3-core), só inclui as pessoas que têm **pelo menos 3** interações dentro desse subgrupo.

O core da rede corresponde ao k-core de maior ordem, ou seja, à concha com o maior valor de k: **28-core** (visível na Figura 5). Este núcleo contém **60 nodos**, todos com pelo menos **28 ligações** internas no subgrafo correspondente, indicando que a rede tem um núcleo muito denso e conectado.

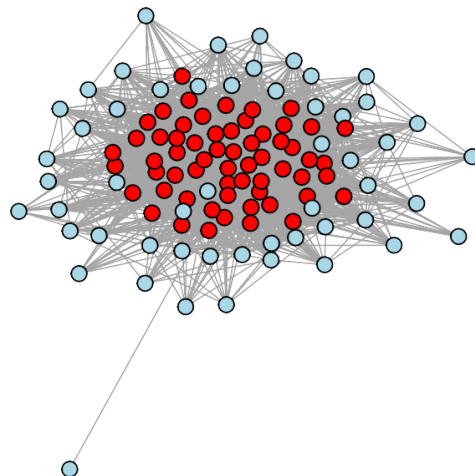


Figura 5: 28-core da Rede da Festa

Relativamente ao **número de conchas**, existem **17**:

Tabela 2: Distribuição de *k*-cores na rede

k-core	1	11	12	13	14	15	17	18	19	20	21	22	23	25	26	27	28
Nº de nodos	1	2	2	3	1	3	3	2	2	4	2	4	1	14	5	4	60

Interessante que, se “salta” do *k*-core 1 diretamente o *k*-core 11. Na **Rede da Festa** não existe nenhum conjunto de nodos que possam formar uma concha onde existam nodos com grau 2. Por isso, não existe 2-core, 3-core, etc.

Concluimos que a rede é bastante **densa** no seu centro, dado pertencerem **60 nodos ao 28-core**. Existe uma estrutura *core periphery*, com um **núcleo denso** e uma periferia que se liga ao núcleo mas não é tão interligada internamente. Existe uma variação devido às 17 conchas, o que remete para a heterogeneidade da rede.

Os 60 nodos do 28-core são os mais influentes em termos de interações. Assim, remover ou isolar o núcleo pode **fragmentar significativamente a rede**, enquanto remover nodos da periferia teria menos impacto.

g) Elabore um pequeno texto com comentários adicionais que considere pertinentes.

A rede analisada apresenta uma estrutura heterogénea, com uma densidade relativamente elevada (~35%) e uma **média** de grau de aproximadamente **39 ligações** por nó. A distribuição de grau revela que a maioria dos indivíduos está moderadamente conectada, mas existem nodos com grau muito elevado, sugerindo a presença de alguns *hubs*.

A heterogeneidade ligeira sugere que, embora existam *hubs*, a maioria dos indivíduos participa de forma equilibrada nas interações, o que indica uma certa igualdade na participação social.

O coeficiente de *clustering* global (~0,50) e os valores de *clustering* local demonstram a presença de numerosos triângulos, refletindo que amigos de um mesmo indivíduo tendem a conhecer-se entre si.

A estrutura da rede é propícia a formar “pequenos mundos”: nodos periféricos conseguem aceder rapidamente ao *core* através de poucos passos, favorecendo dinâmicas sociais.

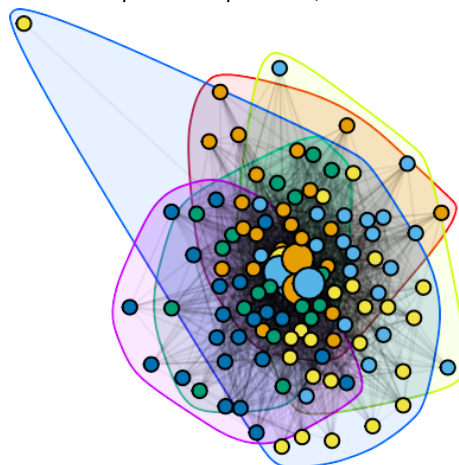


Figura 6: Comunidades da Rede da Festa

Na Figura 6 - Gráfico de Louvain - é possível observar as comunidades.

Cada cor representa uma comunidade detectada pelo algoritmo Louvain, sendo que os **nodos de cada comunidade estão mais conectados entre si**.

Por sua vez, os **nodos mais próximos uns dos outros tendem a ter mais ligações entre si**.

Os **nodos mais próximos das fronteiras** de cada comunidade são considerados 'nodos-pontes' dado facilitarem as interações entre diversas comunidades. Desta forma, é possível observar, de forma visual, o que fora abordado até agora, os *hubs*, os nodos centrais e os grupos mais coesos.

Em suma, trata-se de uma rede social relativamente densa e estruturada, com núcleos centrais influentes e uma periferia mais dispersa, permitindo identificar os indivíduos mais centrais e os padrões de ligação predominantes.

h) Tendo em conta o contexto, identifique algumas características/o papel de alguns dos presentes na festa. Justifique.

Pessoas com muitas interações (grau elevado) são os mais sociáveis e/ou populares na festa, servindo como *hubs* de interação, conectando várias pessoas que, de outra forma, poderiam não interagir. Quem pertence ao 28-core - núcleo mais interno - desempenha um papel central na coesão do grupo, podendo ser considerados '**a alma da festa**'.

A **betweenness** mede quantas vezes um nó está no caminho mais curto entre outros dois nodos. Pessoas com **betweenness alta** são consideradas **pontes**, - intermediários - que ligam grupos diferentes, e com **betweenness baixa**, estão integradas num grupo, mas **não** ligam grupos entre si.

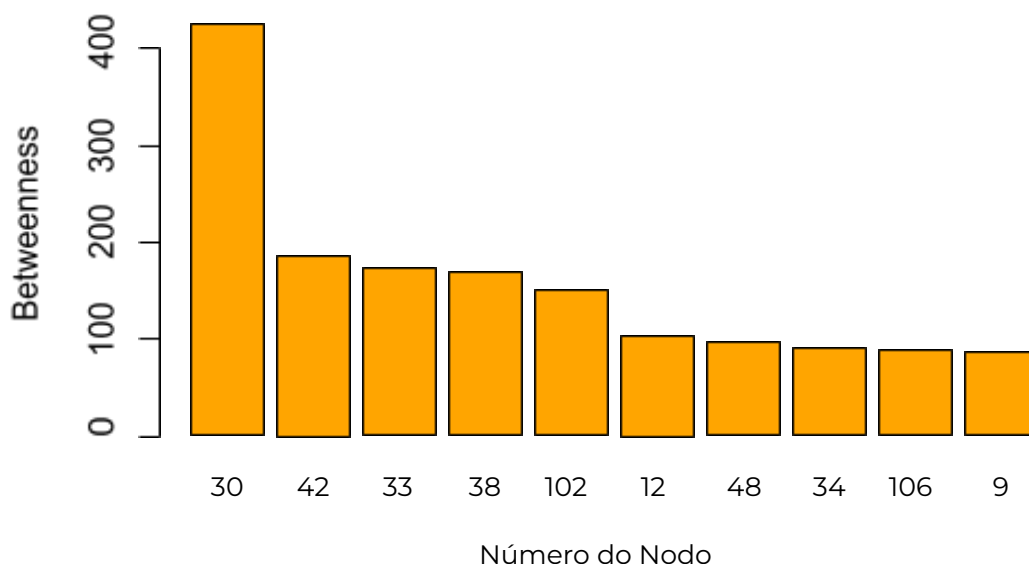


Figura 7: Os 10 nodos com mais intermediação

Assim, de acordo com as Figuras 7 e 8, o Nodo 30 apresenta a maior *betweenness* na rede, indicando que atuou como um intermediário crucial entre diferentes grupos na festa, facilitando a comunicação e as interações entre participantes que, de outra forma, estariam menos conectados.

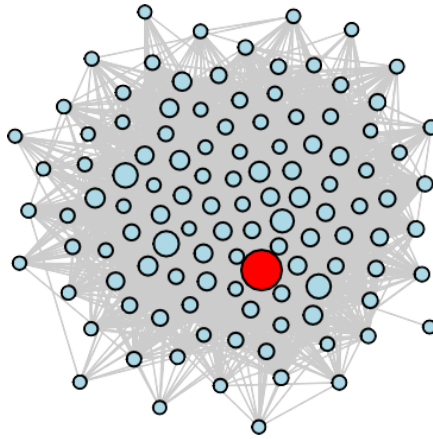


Figure 8: O nodo mais central da Rede da Festa

Participantes periféricos (de grau baixo e conchas pequenas) têm poucas ligações e podem representar pessoas menos integradas e/ou mais introvertidas, interagindo apenas com pequenos subgrupos.

Já os **núcleos intermediários**, podem ser grupos de amigos próximos ou *cliques*, que interagem entre si e, em alguns casos, ligam-se ao núcleo central.

Nodos com **clustering local** elevado - indivíduos cujos vizinhos estão muito conectados entre si - podem desempenhar o papel de **'intermediadores'**, mantendo pequenas "bolhas" de interação dentro da festa vivas.

O **nodo** com o maior grau é considerado o mais social devido ao elevado número de interações, podendo ter o papel de organizador da festa ou - se tal for o caso - o **aniversariante**.

Parte II - Q1

Foi pedida a criação de uma rede aleatória através do pacote *Igraph*, usando o modelo *Preferential Attachment* (doravante **PA**) na qual devem existir **500 nodos** e cerca de **1000 ligações**. Este modelo gera uma componente gigante que integra a maioria dos nodos o que reflete a estrutura de muitas redes reais, como **redes sociais**.

Para tal, gerou-se a rede através da função `rede_q1 <- sample_pa(n = 500, m = 2, directed = FALSE)`.

- O parâmetro $n = 500$ para os 500 nodos ;
- No modelo PA, cada novo nodos n liga-se a m nodos já existentes. Como o número total de arestas aproximado é: $arestas = n * m$, então $m = 2$;
- `directed = FALSE` dado que a rede é não direcionada.

a) **Determine a densidade e classifique a rede**

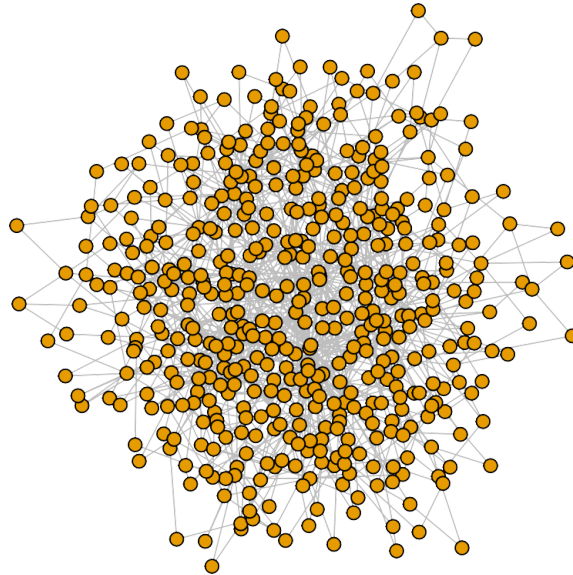


Figura 9: Rede Aleatória Q1, gerada pelo Modelo Preferential Attachment

A densidade da rede gerada pelo modelo PA criada é de **0.008**, isto é, dentro desta rede apenas existem **0.8%** das ligações possíveis.

Como a densidade da rede é muito próxima de 0, a rede é classificada como **rede esparsa**.

Esta baixa densidade é típica de redes geradas por modelos de crescimento preferencial, dado que **poucos nodos acumulam muitas ligações** enquanto a maioria permanece pouco conectada, resultando numa estrutura global muito esparsa.

b) **Indique se a rede é conexa;**

Uma rede é conexa se, a partir de qualquer nodo na rede, seja possível alcançar qualquer outro nodo da rede através das suas ligações. Este facto verifica-se para a rede criada, através do comando `is_connected(rede_q1)`, o que retorna `TRUE`, logo a rede é conexa.

c) **Obtenha o grau médio e a distribuição de grau. Caracterize a distribuição de grau (mediana, quartis, etc.) e comente;**

O grau médio da rede é de **3.99**, ou seja, em média cada nodo possui cerca de **4** ligações.

A distribuição de grau mostra valores entre **2** e **47**, com quartis **Q1=2**, **Q2(Mediana)=3** e **Q3=4**, o que indica que a maioria dos nodos possui entre 2 e 4 ligações, permitindo observar a dispersão dos graus, evidenciando a heterogeneidade típica do modelo Preferential Attachment.

As estatísticas descritivas mostram que a **mediana do grau é menor que a média**, refletindo que a maioria dos nodos possui poucas ligações, enquanto alguns nodos (*hubs*) apresentam um número elevado de ligações.

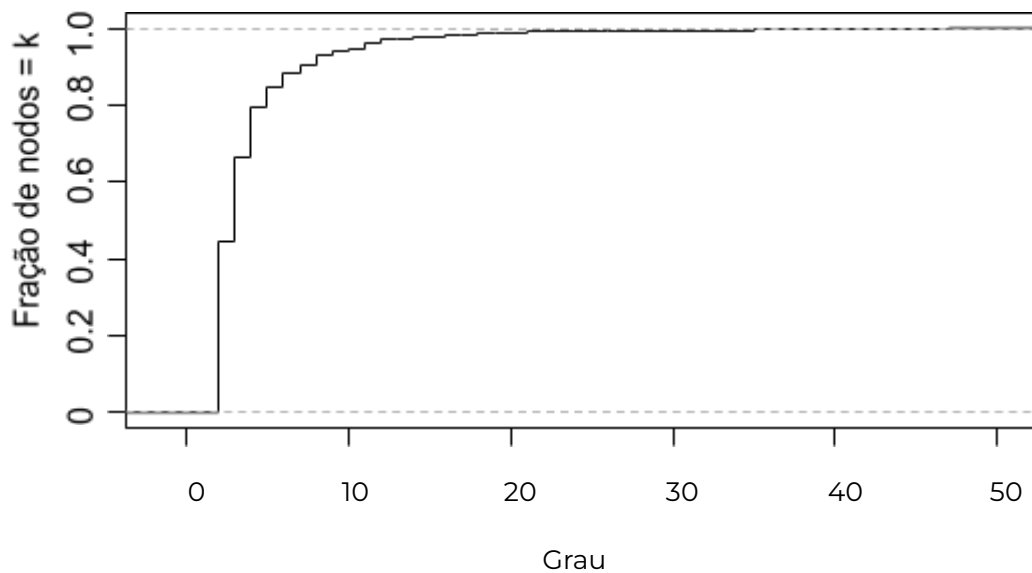


Figura 10: Distribuição Acumulada de Grau

A Figura 10 confirma que a rede apresenta muitos nodos com baixo grau e poucos nodos altamente conectados.

Esta distribuição de grau é característica do modelo **PA** e é importante para compreender a presença de hubs e as potenciais influências e propagação de informação dentro da rede.

d) Calcule o parâmetro de heterogeneidade. Indique o que pode concluir quanto à existência de hubs;

O parâmetro de heterogeneidade, calculado através da divisão do grau quadrado médio pelo grau médio ao quadrado, mede o quão desigual é a distribuição de graus da rede. Para a rede gerada, o valor obtido foi de **2.129**.

Como este valor é superior a 1, a rede é **heterogénea**, ou seja, existe uma grande variação no número de ligações entre os nodos. Sendo que o número é relativamente elevado, podemos ainda concluir que há a existência de hubs.

Assim, a maioria dos nodos apresenta poucas ligações, enquanto um pequeno número de nodos concentra um número muito elevado de ligações, característica do modelo de *preferential attachment*, onde nodos mais antigos tendem a acumular mais ligações ao longo do tempo.

Na figura 11, podemos observar os nodos com número de ligações acima do 3o Quartil, que consideramos os **hubs** da rede.

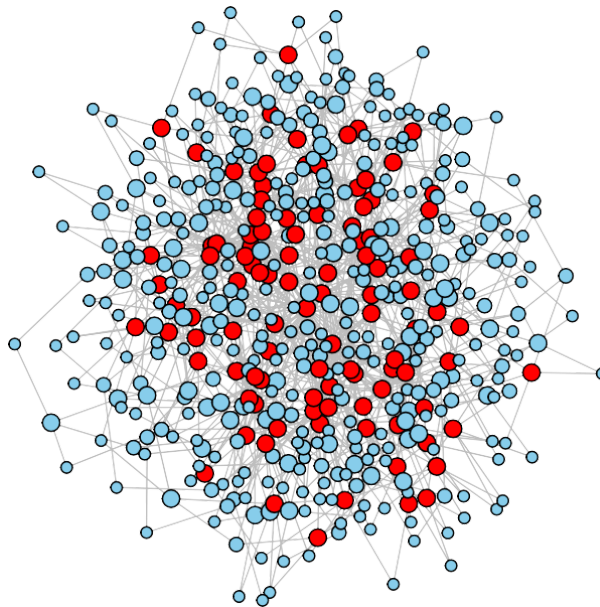


Figura 11: Nodos acima de 3o Quartil

e) Estude a associação de grau e indique o que poderá concluir-se;

A associação de grau avalia se nodos com muitas ligações tendem a ligar-se a outros nodos também com várias ligações - **rede assortativa** - ou se, pelo contrário, se ligam a nodos com poucas ligações - **rede desassortativa**.

A associação de grau foi calculada através da função `assortativity_degree(rede_q1, directed=FALSE)` tendo sido obtido o valor de **-0.064**.

Como o valor é **negativo**, conclui-se que a rede é **desassortativa**, significando que os nodos com grau mais elevado tendem a ligar-se a nodos com grau mais baixo, enquanto a maioria destes últimos não se conecta preferencialmente entre si, como se pode comprovar pela tendência negativa da Figura 12.

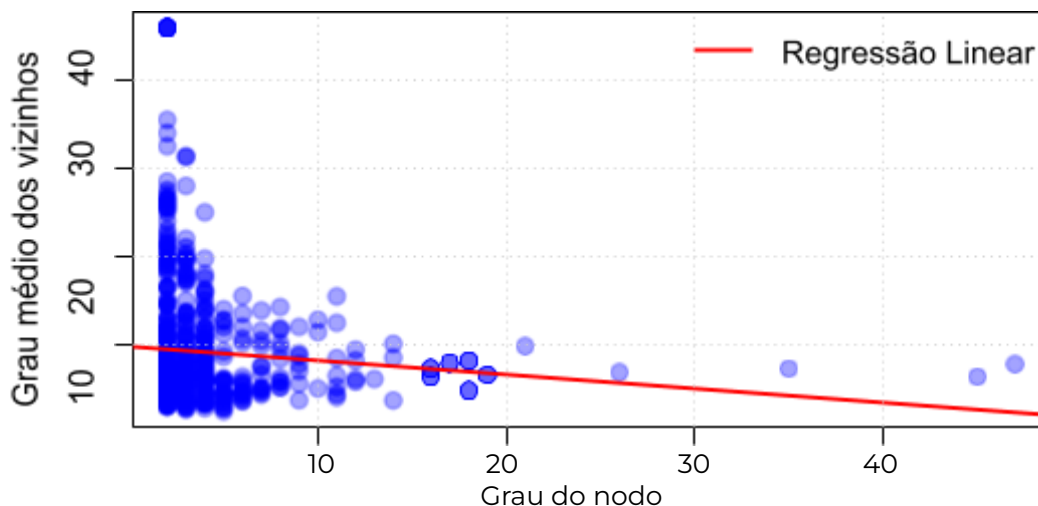


Figura 12: Grau do nodo vs Grau médio dos vizinhos rede Q1

Este comportamento é típico de redes geradas pelo modelo de *preferential attachment*, nas quais existem alguns nodos altamente conectados (*hubs*) que funcionam como **pontos centrais de ligação** para muitos nodos periféricos com poucas ligações. A disassortatividade sugere que a rede é eficiente na conectividade geral e espalha de informação, mas também que os *hubs* são pontos **críticos** cuja falha pode fragmentar a rede.

f) Determine a média dos comprimentos dos caminhos mais curtos e o diâmetro da rede. Indique o que pode concluir-se quanto à distância média;

Calculou-se o comprimento médio dos caminhos mais curtos usando a função *mean_distance()* obtendo-se um valor de **3.94**, enquanto o diâmetro da rede foi de **8**. Como visto anteriormente, a rede é conexa, logo o diâmetro é calculado sobre a rede inteira. Caso não fosse, era preciso calcular a componente gigante e consequentemente, o seu diâmetro.

Em média, são necessários cerca de **3.94** passos para ligar dois nodos da rede, indicando um comportamento *small-world*. Apesar do grande número de nodos, qualquer par de nodos está separado por um pequeno número de passos, devido à presença de *hubs* - característica do modelo PA. O diâmetro indica o maior caminho dos caminhos mais curtos, e confirma que mesmo os nodos mais distantes estão separados por um número reduzido de passos, ilustrado na Figura 13.

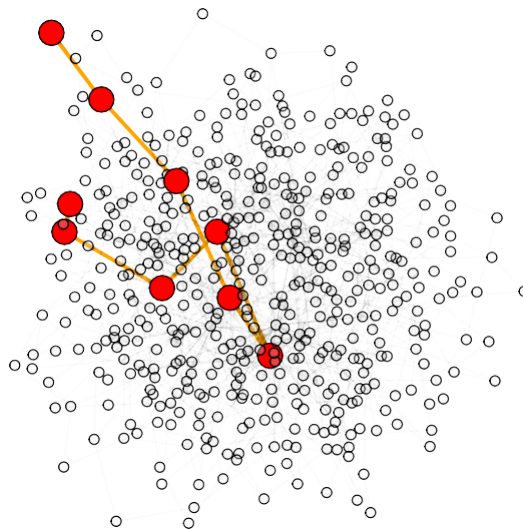


Figura 13: Diâmetro da Rede Q1

Assim, conclui-se que a rede apresenta uma elevada eficiência em termos de conectividade, o que por si só mostra características típicas de redes do tipo **small-world**.

g) Determine os coeficientes de clustering da rede. Diga o que pode concluir-se quanto à existência de triângulos.

O coeficiente de *clustering* indica a tendência dos nodos da rede formarem triângulos, ou seja, situações em que dois nodos ligados a um terceiro também estão ligados entre si.

Os coeficientes - local e global - da rede foram calculados através da função *transitivity()*, tendo sido obtido o valor de clustering médio de **0.038**.

Também se realizou a contagem direta de triângulos, em percentagem, o que retornou o valor

de **13.8%**.

Com base nos valores baixos de transitividade obtidos, inferimos que os triângulos são pouco frequentes na rede. Os vizinhos de um nodo têm **baixa tendência** a estar ligados entre si. Este padrão é coerente com o modelo **PA**, que tende a gerar redes com *hubs*, reduzindo a formação de triângulos.

Assim, a rede Q1 favorece caminhos curtos através de *hubs*, no entanto, impede a formação de comunidades interligadas.

Em termos práticos, influenciar-se-ia processos de redundância de informação como, por exemplo, **informação que poderá propagar-se rapidamente via hubs mas sem redundância local** garantida por triângulos, ou seja, como os vizinhos de um *hub* não estão muito ligados entre si, se uma aresta falha, a informação perde um possível caminho alternativo e a transmissão fica mais dependente dos *hubs*. Logo, concluímos que a informação chega rapidamente a muitas partes da rede, devido aos hubs, mas não existem pequenos ciclos (triângulos) que reforcem a conectividade local.

Parte II - Q2

a) Indique o número de ligações removidas;

Alterámos a rede obtida em Q1, atribuindo uma probabilidade de **0,1** (`prob_rem <- 0.1`) a cada uma das ligações desta ser removida. As ligações a serem removidas foram selecionadas através da criação de uma distribuição uniforme - 0.9 probabilidade de permanecerem na rede e 0.1 probabilidade de serem removidas - usando a função `runif(ecount(rede_q1)) < prob_rem`.

Verificou-se uma redução de **108 ligações** na nova rede.

b) Determine a densidade e classifique a rede;

A densidade da nova rede criada é de **0.007**, ou seja, dentro da nova rede apenas existem **0.713%** das ligações possíveis.

A nova rede é também **muito esparsa**, sendo ainda mais esparsa do que a original porque o valor da densidade aproxima-se ainda mais de 0.

c) Indique se a rede é conexa;

Devido à remoção dos nodos - e consequentemente a perda de **108** ligações - a nova rede deixou de ser conexa, pois pelo menos um nodo perdeu todas as ligações, ficando isolado da rede. A função `is_connected(rede_q2)` comprova este facto, devolvendo **FALSE**.

d) **Obtenha o grau médio e a distribuição de grau. Caracterize a distribuição de grau (mediana, quartis, etc.) e comente;**

O grau médio da rede é de **3.56**, ou seja, em média cada nodo possui cerca de **4** ligações.

- **Grau Médio:** 3.56
- **Mediana :** 2
- **Mínimo :** 0
- **Máximo :** 41
- **1º Quartil :** 2
- **3º Quartil :** 4

Como é possível observar na figura 14 a distribuição é bastante assimétrica à direita, o que indica que a grande maioria dos nodos tem poucas ligações.

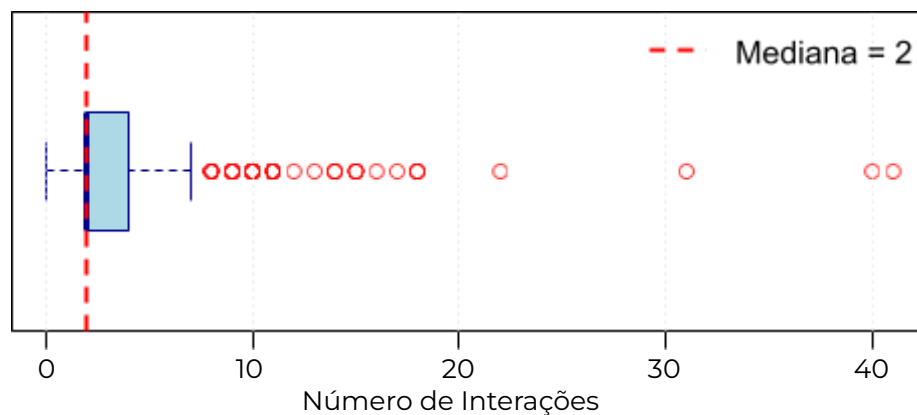


Figura 14: Boxplot da distribuição de graus na Nova Rede

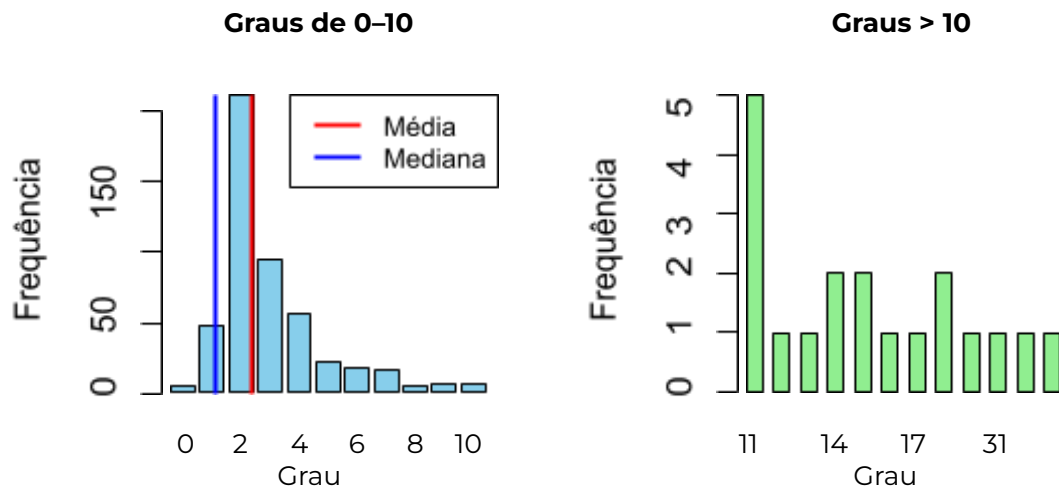


Figura 15: Distribuição dos graus

Apesar de haver nodos com graus bastante elevado, o valor da média é mais aproximado da mediana dado haver uma frequência mais elevada de nodos com graus perto do valor da mediana.

- e) **Calcule o parâmetro de heterogeneidade. Indique o que pode concluir quanto à existência de hubs;**

O parâmetro de heterogeneidade calculado foi de **2.15**.

Como este valor continua superior a 1, então a nova rede mantém-se heterogênea, ou seja, continua a existir uma grande variação no número de ligações entre os nodos. Conclui-se ainda que existem, muito provavelmente, os mesmos hubs.

- f) **Estude a associação de grau e indique o que poderá concluir-se;**

A associação de grau foi novamente calculada, através da função *assortativity_degree(rede_q2, directed=FALSE)* tendo sido obtido o valor de **-0.0571539**. O valor continua **negativo**, logo a rede continua **desassortativa**. Conclui-se portanto que nodos “diferentes” se ligam entre si, isto é, nodos com grau elevado tendem a ligar-se a nodos com grau menor.

- g) **Determine a média dos comprimentos dos caminhos mais curtos e o diâmetro da rede. Indique o que pode concluir-se quanto à distância média;**

Calculou-se o comprimento médio dos caminhos mais curtos usando a função *mean_distance(rede_q2, directed = FALSE)* obtendo-se um valor de **4.237**, enquanto o diâmetro da rede foi de **9**.

A nova rede - como referido anteriormente - **não é conexa**. Desta forma, o seu diâmetro é equivalente ao **diâmetro da componente gigante**, cálculo este que é efetuado automaticamente pelo pacote *igraph*.

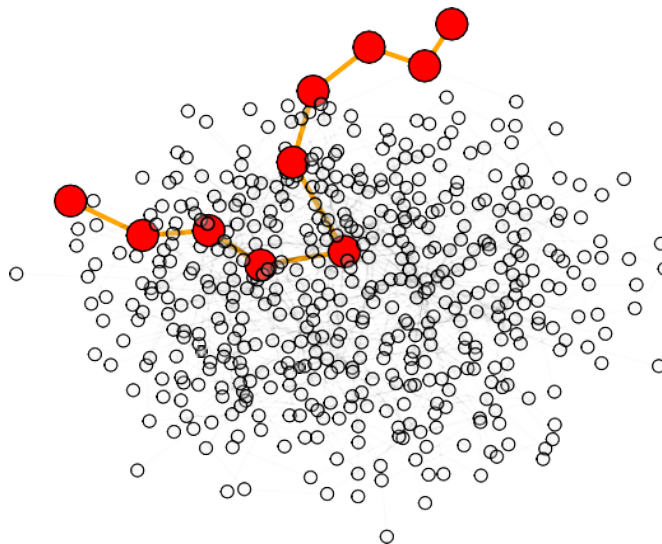


Figura 16: Diâmetro da Rede Q1

Na figura 16, é possível observar o maior caminho dos caminhos mais curtos (diâmetro).

h) Determine o coeficiente de clustering da rede. Diga o que pode concluir-se quanto à existência de triângulos;

O coeficientes global da nova rede foram calculados através da função *transitivity(rede_q2, type = "global")*, tendo sido obtido o valor de clustering médio de **0.0213**.

A nova rede cria percursos curtos graças aos hubs, mas dificulta a formação de comunidades ligadas entre si. A informação propaga-se rapidamente através dos hubs, porém sem redundância local: como os vizinhos dos hubs quase não se ligam entre si, a falha de uma aresta pode quebrar caminhos alternativos. Assim, a rede é rápida a difundir informação, mas carece de pequenos ciclos que reforcem a conectividade local.

i) Compare os resultados obtidos antes e após a remoção de ligações.

	Densidade	Conexa	Grau Médio	Min	Max
rede Q1	0.008	Sim	3.99	2	47
rede Q2	0.007	Não	3.56	0	41

	Mediana	Heterogeneidade	Associação de grau	Diâmetro	Coeficiente de clustering
rede Q1	3	2.13	-0.064	8	0.038
rede Q2	2	2.15	-0.057	9	0.021

A comparação entre as redes Q1 e Q2 evidencia o impacto da remoção de ligações na estrutura e nas propriedades topológicas do grafo. Apesar de ambas apresentarem densidades muito próximas (0,008 em Q1 e 0,007 em Q2), a pequena redução observada já é suficiente para alterar características estruturais relevantes.

Um dos efeitos mais imediatos é a perda de conectividade: enquanto a rede Q1 é conexa, a rede Q2 deixa de o ser, indicando que a eliminação de arestas fragmentou a rede em componentes menores.

No conjunto, as alterações observadas entre Q1 e Q2 são consistentes com a remoção de arestas: a rede torna-se menos conectada, mais dispersa, com menor clustering e com caminhos mais longos entre os nodos.

Conclui-se, portanto, que embora a estrutura geral se mantenha semelhante, estas mudanças revelam uma fragilização da coesão e da eficiência estrutural da rede após a redução das ligações.