Instituto de Ciências Matemáticas e de Computação

ISSN - 0103-2569

Redes Complexas: conceitos e aplicações

Jean Metz
Rodrigo Calvo
Eloize Rossi Marques Seno
Roseli A. F. Romero
Zhao Liang

Nº 290

RELATÓRIOS TÉCNICOS DO ICMC

São Carlos
janeiro/2007

Redes Complexas: conceitos e aplicações.

Jean Metz
Rodrigo Calvo
Eloize Rossi Marques Seno
Roseli A. F. Romero
Zhao Liang

Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Departamento de Ciências de Computação e Estatística
Laboratório de Inteligência Computacional
Caixa Postal 668, 13560-970 - São Carlos, SP, Brasil

Resumo: As redes complexas são um tipo de grafo que apresentam propriedades topográficas bastante particulares, não encontradas em grafos mais simples. Este relatório tem como objetivo apresentar aos leitores iniciantes da área alguns conceitos fundamentais para o entendimento dessas redes, bem como suas propriedades principais e alguns modelos mais comumente estudados. Além de conceitos introdutórios, apresentam-se também algumas aplicações reais envolvendo redes complexas.

Palavras-Chave: Teoria dos grafos, redes complexas, aplicações de redes complexas

---

janeiro/2007

Este documento foi preparado com o formatador de textos LaTeX. O sistema de citações de referências bibliográficas utiliza o padrão *Chicago* do sistema bibTeX.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

O estudo de redes complexas é um tema inter-disciplinar que abrange diversas áreas de conhecimento, tais como a ciência da computação, matemática, física, biologia e sociologia. O termo redes complexas refere-se a um grafo que apresenta uma estrutura topográfica não trivial, composto por um conjunto de vértices (nós) que são interligados por meio de arestas (Barabási, 2003). O estudo de redes na forma de grafos é um dos pilares da matemática discreta e teve início em 1735, quando Euler propôs uma solução para o problema das pontes de Königsberg, originando a teoria dos grafos.

Desse modo, diversos aspectos do mundo real podem ser representados por meio de redes complexas a partir de analogias para a resolução de problemas específicos. É possível, por exemplo, modelar toda a estrutura física de uma grande rede de computadores tal como a Internet. Nesse caso, os computadores conectados à Internet referem-se aos vértices da rede enquanto que os cabos e meios de tranmissão representam as arestas do grafo. Outras analogias pode ser também utilizadas, tais como o conteúdo de páginas *WEB — World Wide Web*, relações sociais entre grupos de pessoas, redes organizacionais ou de negócios entre companhias, redes neurais, redes metabólicas, cadeia alimentar, entre outras. Como ilustração da modelagem de redes complexas como grafos, considere as Figuras 1[1] e 2 mostram a estrutura da rede Internet e a estrutura da rede Web do Google, respectivamente.

Os estudos das redes complexas foram iniciados em meados de 1930, quando sociólogos utilizavam essas redes com a finalidade de estudar o comportamento da sociedade e a relação entre os indivíduos. Essas pesquisas eram baseadas em características muito peculiares das redes, como a centralidade (o vértice mais central) e a conectividade (vértices com maior número de conexões). As redes sociais eram constituídas por indivíduos, que representados por vértices, e pelas interações entre eles, as arestas. A centralidade e a conectividade eram usadas, por exemplo, para determinar os indivíduos que melhor se relacionavam com os demais ou para identificar os indivíduos mais influentes.

Com o avanço da tecnologia de informação e a disponibilidade de computadores e redes de comunicação que permitem a análise de dados em grandes quantidades, houve uma mudança significativa na área. As pesquisas, antes focadas nas pequenas redes e nas propriedades de vértices individuais ou arestas, passaram a considerar propriedades estatísticas em larga-escala. Atualmente, são comuns estudos com redes envolvendo milhões ou bilhões de vértices, as quais antes eram compostas por dezenas ou, em casos extremos,

[1] Disponível em: http://commons.wikimedia.org/wiki/Image:WorldWideWebAroundGoogle.png (último acesso em 27/06/06)

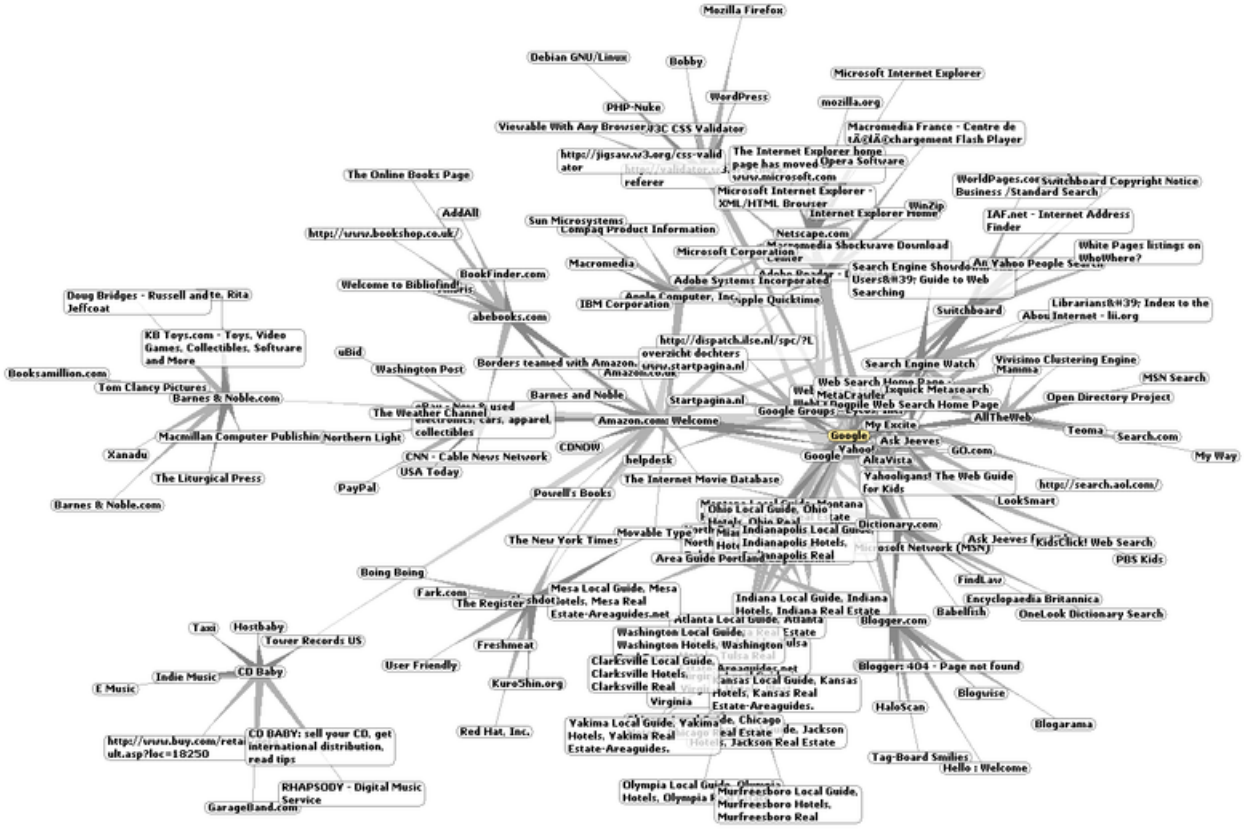


Figura 1: Representação da rede *Web* do Google.

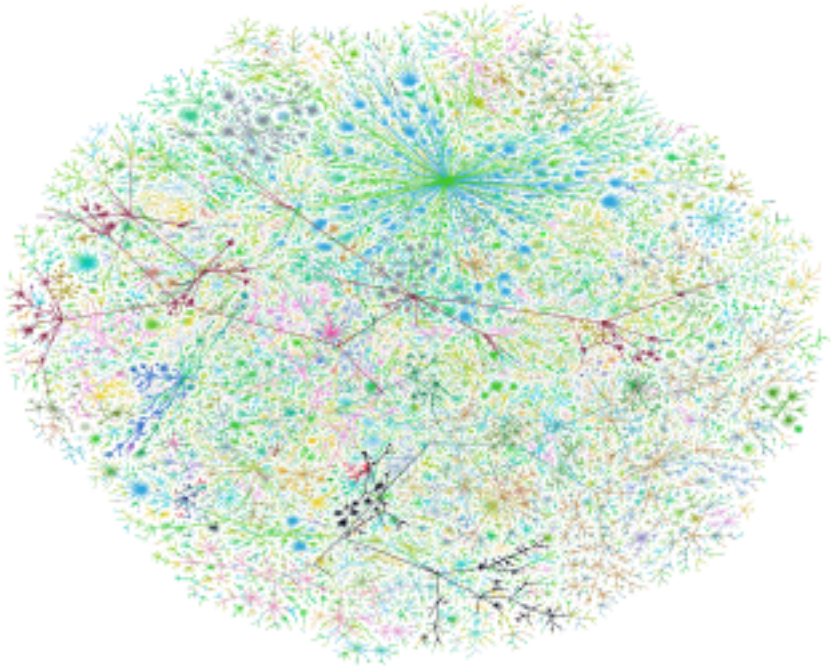


Figura 2: Representação da estrutura da Internet (Newman, 2003).

centenas de vértices. A mudança de paradigma revelou várias características que diferem substancialmente as redes do mundo real das redes aleatórias, tidas por muitos anos como o principal modelo de redes (Barabási, 2003; Newman, 2003). Descobriu-se que a topologia e a evolução das redes do mundo real apresentam propriedades organizacionais bastante robustas e distintas das redes aleatórias. Essa é a principal razão pela qual as redes passarem a ser chamadas de redes complexas.

De maneira simplificada, pode-se dizer que as redes complexas são estruturas que não seguem um padrão regular. No entanto, não há um consenso na literatura que identifique exatamente o que é um padrão regular. Nem tampouco, uma conceituação universalmente aceita sobre o que constituem essas redes. Embora não haja um concenso claro sobre a definição dessas redes, sabe-se que elas apresentam características prórias que não estão presentes em redes regulares. Essas características revelam como as redes são formadas e como suas estruturas podem ser explorada na análise de um determinado problema.

Neste trabalho o objetivo é fornecer um material introdutório sobre redes complexas, apresentando alguns conceitos fundamentais de maneira simples que possam situar pesquisadores iniciantes na área. Além desses conceitos básicos, são apresentados também exemplos de alguma aplicações envolvendo redes complexas. Vale ressaltar que grande parte das definições apresentadas neste relatório foram obtidas de duas fontes principais: (Newman, 2003) e (da F. Costa et al., 2005).

Este relatório está organizado da seguinte maneira: na Seção 2 são apresentados os fundamentos teóricos refente às redes complexas, como suas propriedades e principais tipos de rede. Na Seção 3 são apresentadas algumas aplicações que se baseiam em redes complexas para a resolução de problemas específicos. Por fim, na Seção 4 são apresentadas as considerações finais.

2 Fundamentos teóricos

Uma rede é um grafo no qual há um conjunto de vértices (ou nós) e um conjunto de arestas (ou arcos) que conectam esses vértices. As arestas estabelecem algum tipo de relação entre dois vértices de acordo com o problema modelado. Além disso, o grafo pode ser direcionado ou não. Em um grafo direcionado (dígrafo), cada aresta tem um sentido (direção) que conecta um vértice origem à um vértice destino. Exemplos de dígrafos são aqueles usados para representar chamadas telefônicas e mensagens de *e-mails*, nos quais as mensagens são direcionadas de uma pessoa para outra. Os dígrafos pode ser cíclicos, quando há um caminho de um vértice para ele mesmo, ou acíclicos quando não existe esse caminho.

É importante lembrar que nem todo grafo pode ser considerado uma rede complexa, pois essa classificação só é possível se o grafo apresentar algumas propriedades topográficas que não estão presentes em grafos simples. Algumas dessas propriedades são descritas brevemente a seguir.

2.1 Propriedades das Redes

As redes complexas apresentam algumas propriedades que podem ser úteis nas análises dos mais diversos aspectos das redes e com os mais variados propósitos. Nesta seção, são apresentadas algumas propriedades principais que têm recebido muita atenção na literatura.

Coeficiente de aglomeração: os agrupamentos intrínsecos às redes são quatificados por meio do coeficiente de aglomeração, também conhecido como fenômeno de transitividade. Esse fenômeno ocorre quando um vértice A está conectado a um vértice B, e o vértice B está conectado a um vértice C, aumentando as chances do vértice A também estar conectado ao vértice C. Em outras palavras, a transitividade indica a presença de um número elevado de triângulos na rede, *i.e.*, conjuntos de três vértices conectados uns aos outros. Para enteder melhor, considere a analogia com uma rede social. Nesse caso, pode-se dizer que se A é amigo de B e B é amigo de C, existem grandes chances de A e C também serem amigos.

O coeficiente de aglomeração CA de uma rede é obtido a partir da Equação 1, onde $\#\triangle$ refere-se ao número de triângulos na rede e, $\#v$ representa o número de "vértices triplamente conectados", *i.e.*, vértices com arestas não direcionadas para o outro par de nós. O fator 3 no numerador refere-se ao fato de que cada triângulo apresenta três triplas e também para garantir que o coeficiente de aglomeração seja um valor entre 0 (zero) e 1 (um).

$$CA = \frac{3 \times \#\triangle}{\#v} \tag{1}$$

Distribuição de Graus: o grau de um vértice qualquer em uma rede define o número de arestas que incidem (conectam) aquele vértice. Desse modo, a distribuição de graus é uma função de distribuição probabilística que indica a probabilidade de um determinado vértice ter grau fixo. Uma maneira de quantificar essa distribuição é por meio de uma função de distribuição cumulativa (Equação 2), onde $p_{k'}$ é a fração de nós da rede com grau k e P_k é a função cumulativa de distribuição de probabilidades.

$$P_k = \sum_{k'=k}^{\infty} p_{k'} \tag{2}$$

Em um dígrafo, por outro lado, cada vértice tem um grau de entrada e de saída, acarretando em uma equação diferente para o cálculo da distribuição de graus. Essa nova equação é escritas em função de p_{jk} com duas variáveis, representando a fração de vértices que têm, simultaneamente, um grau de entrada j e um grau de saída k.

A distribuição de graus nas redes aleatórias segue a distribuição de Poisson. No entanto, em muitas redes reais a distribuição de graus segue a Lei de Potência, em que $p_{k'} \sim k^{-\alpha}$ para uma constante α qualquer.

Resistência: indica a capacidae de resistênca da rede quanto às remoções de alguns vértices, sem que haja perda de sua funcionalidade. Essa propriedade está diretamente relacionada com a distribuição de graus dos vértices, pois a remoção de vértices pode resultar na perda de conexão entre pares de vértices ou, ainda, aumentar significativamente o caminho de um vértice a outro.

Misturas de Padrões: alguns tipos de redes apresentam uma mistura de padrões diferentes onde os vértices pode representar diferentes tipos de objetos. Nas redes de cadeias alimentares, por exemplo, existem vértices que representam plantas, animais herbívoros e animais carnívoros. Em geral, a probabilidade de conexão entre esses vértices é dependente do seu tipo. Nesse caso específico, existem arestas conectando os herbívoros às plantas e os herbívoros aos carnívoros. Por outro lado, existem poucas conexões entre herbívoros e herbívoros ou entre animais carnívoros e plantas.

As redes de relações sociais também apresentam essa propriedade, pois são constituídas por vértices de representam pessoas de diferentes etnias. Nesse tipo de rede, há uma tendência de existirem mais conexões

entre vértices do mesmo tipo, uma vez que as pessoas estão mais propensas a se relacionarem com outras pessoas da mesma etnia (Newman, 2003). Uma curiosidade também observada por Newman (2003) é que, essencialmente, todas as redes sociais apresentam essas variações de padrões, enquanto outros tipos de redes não.

Correlação de Graus: indica se as arestas em uma rede associam vértices com graus parecidos. Essa correlação é usada, principalmente, em redes com variações de padrões, para investigar a probabilidade de conexão dos vértices de diferentes tipos.

2.2 Tipos de Redes

Nesta seção são brevemente descritos os três principais modelos de redes complexas: redes aleatórias, redes pequeno-mundo e redes livres de escala.

Redes Aleatórias: porposto por Erdös e Rény, esse é o modelo mais simples que uma rede complexa pode assumir. Nesse modelo, arestas não direcionadas são adicionadas aleatóriamente entre um número fixo de N vértices. Cada aresta é independentemente representada com base em alguma probabilidade p. O número de arestas que conectam cada vértice na rede, denomidado grau do vértice, segue a distribuição de Poisson com um limite máximo N. O grau esperado de um vértice qualquer é definido pela Equação 3, onde p é a probabilidade de um vértice se concetar a um outro vértice qualquer, N representa o número de vértices da rede e k é o total de arestas que incidem em um determinado vértice.

$$\langle k \rangle = p(N-1) \tag{3}$$

Esse modelo gera grafos aleatórios com N vértices e k arestas, denominados grafo aleatório ER, definido como $G^{ER}_{N,k}$. Inicialmente com N vértices desconectados, o modelo ER é obtido conectando-se os vértices selecionados aleatóriamente até o número de arestas do grafo ser igual a k.

Acredita-se que o processo de construção da rede seja aleatório no sentido de que vértices se agregam aleatoriamente. Com base nessa premissa, Erdös e Rény concluíram que todos os vértices de uma determinada rede têm aproximadamente a mesma quantidade de conexões e as mesmas chances de receberem novas ligações (Barabasi and Albert, 1999a). Segundo os autores, quanto mais complexa for a rede, maiores serão as chances dela ser aleatória.

Uma alternativa para o modelo ER de grafos aleatórios é concetar cada par de vértices comprobabilidade $0 < p < 1$. Esse procedimento define um conjunto representado como $G_{N_p}^{ER}$ e formado por grados com diferentes número de arestas. Grafos com k arestas aparecem no conjunto com uma probabilidade $p^k(1-p)^{N(N-1)/2-k}$. Nota-se que o limite $N \to \infty$ é fixado em $\langle k \rangle$, que corresponde a $2k/N$, no primeiro modelo e $p(N-1)$, no segundo modelo.

Redes Pequeno-mundo: Segundo Watts and Strogatz (1998), muitas redes apresentam padrões altamente concetados, tendendo a formar pequenas quantidades de conexões em cada vértice. Assim, eles propuseram um modelo semelhante ao de Erdös e Rény, no qual grande parte das conexões são estabelecidas entre vérices mais próximos, apresentando-se como um mundo pequeno. Nesse modelo, a distância média entre quaisquer dois vértices de uma rede muito grande não ultrapassa um número pequeno de vértices. Para isso, basta que algumas conexões aleatórias entre grupos sejam estabelecidas (Buchanan, 2002). Na Figura 3 é apresentado um exemplo de rede pequeno mundo.

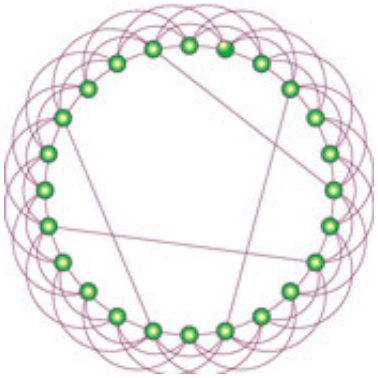


Figura 3: Rede complexa pequeno-mundo (Strogatz, 2001).

O efeito pequeno-mundo é observado nas redes em que a maioria dos vértices se conecta a outros através de um caminho mínimo. O caminho mínimo, também chamado de caminho geodésico ou distância geodésica, é aquele formado pelo menor número de arestas que conectam um vértice origem e um vértice destino. Para melhor ilustrar esse efeito, considere os indivíduos de uma sociedade qualquer. De acordo com o experimento conduzido por Stanley Milgram em 1960, se uma carta fosse entregue a um indivíduo, que não fosse o destinatário, e ele a repassasse a um outro e, assim, por diante, em aproximadamente seis passagens ela chegaria ao destinatário. Esse resultado é uma demonstração direta do efeito pequeno-mundo, em que o caminho percorrido pela carta, partindo de um indivíduo qualquer até o destinatário, é mínimo. O comprimento

do caminho mínimo médio CM entre pares de vértices em um grafo não direcionado é dada pela Equação 4,onde d_{ij} é a distância geodésica do vértice i até o vértice j.

$$l = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij} \tag{4}$$

Essa definição apresenta problemas nas redes com mais de um componente. Um componente é representado por um único vértice ou por um conjunto de vértices e de arestas que conectam os pares de vértices. Nas redes com mais de um componente não há um caminho conectando um vértice qualquer de um componente com um outro vértice qualquer de outro componente. Em outras palavras, há um subconjunto de vértices interconectados entre si, mas sem qualquer conexão com um outro subconjunto da rede. Para evitar problemas no cálculo da distância média geodésica, são considerados apenas os pares de nós em que há um caminho entre eles.

O efeito pequeno-mundo tem implicações óbvias na dinâmica de processos em redes. Por exemplo, um boato pode se espalhar muito mais rápido se, ao invés de mil passos, levarem apenas seis para chegar de um indivíduo qualquer a outro.

Redes Livres de Escala: Barabasi and Albert (1999a) demonstraram que algumas redes apresentam uma ordem na dinâmica de estruturação, com características bem específicas. Uma das principais características, denominada conexão preferencial, é a tendencia de uma novo vértice se concetar a um vértice da rede que tem um grau elevado de conexões. Essa característica implica em redes com poucos vértices altamente conectados, denominados *hubs*, e muito vértices com poucas conexões. As redes com essas características são denominadas livres de escala devido à representação matemática da rede. Ela segue uma função $f(x)$ que permanece inalterada com um fator multiplicativo sob um re-escalonamento da variável independente x. Em outras palavras, isso significa que as redes livres de escalas são aquelas em que a distribuição de graus segue a Lei de Potência, desde que exista uma solução somente para $f(ax) = bf(x)$. Conforme apresentado em (Newman, 2003), essas redes têm sido observadas em vários sistemas, por exemplo, na internet, na *Web*, em redes de metabolismos e em redes de citações de artigos científicos. Na Figura 4 é apresentado um exemplo de rede livre de escala.

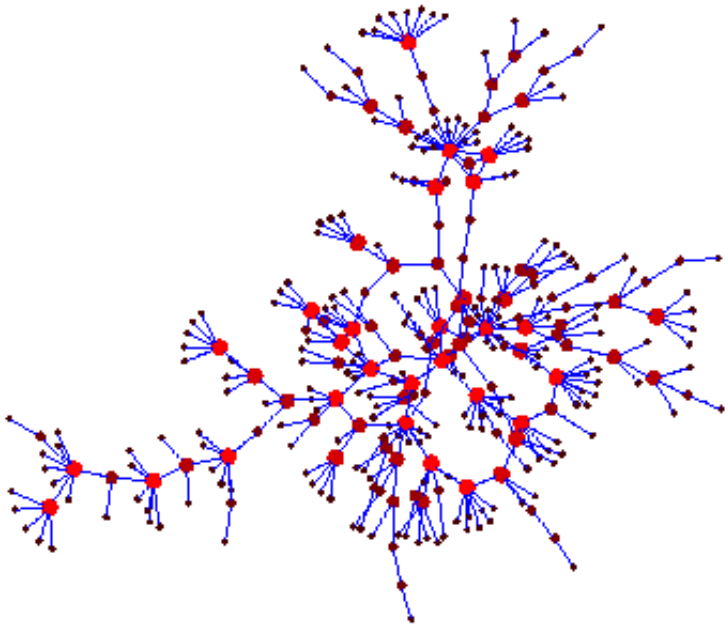


Figura 4: Rede complexa livre de escala (Strogatz, 2001).

3 Aplicações de Redes Complexas

As redes complexas têm sido aplicadas nas mais diversas áreas, para a resolução dos mais variados tipos de problemas. Por exemplo, na avaliação da qualidade de textos (Antiqueira et al., 2005b,a), na avaliação de sistemas de sumarização automática (Pardo et al., 2006b,a), na construção de sistemas de sumarização (Antiqueira, 2006), citar aplicações em outras áreas. As subseções a seguir apresentam brevemente algumas aplicações.

3.1 Avaliação da Qualidade de Textos

Antiqueira et al. (2005b) modelaram textos como redes complexas e usaram essa modelagem para avaliar sua qualidade. Em seu modelo, um texto é representado por uma rede complexa, na qual cada palavra é um vértice e cada aresta representa uma relação de adjacência entre dois vértices, ou seja, para cada par de palavras consecutivas, existe uma aresta direcionada correspondente na rede. Cada aresta contém um peso que indica o número de vezes que as respectivas associações de palavras ocorrem no texto. O objetivo dessa representação é codificar as relações entre os conceitos de um texto. Para isso, antes de serem representados como redes complexas, os textos foram pré-processados em duas etapas iniciais: a) remoção de palavras pouco significativas(*stopwords*) como, preposições e conjunções; e b) lematização das palavras restantes, para o agrupamento de conceitos que tinham a mesma forma canônica, mas apresentavam flexões diferentes (por exemplo, "fizeram"e "fazem"que correspondem ao lema "fazer").

Para melhor ilustrar a representação de um texto como rede complexa, na Figura 5 é apresentada a rede referente ao poema de Carlos Drummond de Andrade, representado na Figura 6.

No meio do caminho tinha uma pedra
tinha uma pedra no meio do caminho
tinha uma pedra.

Nunca me esquecerei desse acontecimento
na vida de minhas retinas tão fatigadas.
Nunca me esquecerei que no meio do caminho
tinha uma pedra
tinha uma pedra no meio do caminho
no meio do caminho tinha uma pedra.

Figura 5: Poema "No meio do caminho" (Antiqueira et al., 2005b).

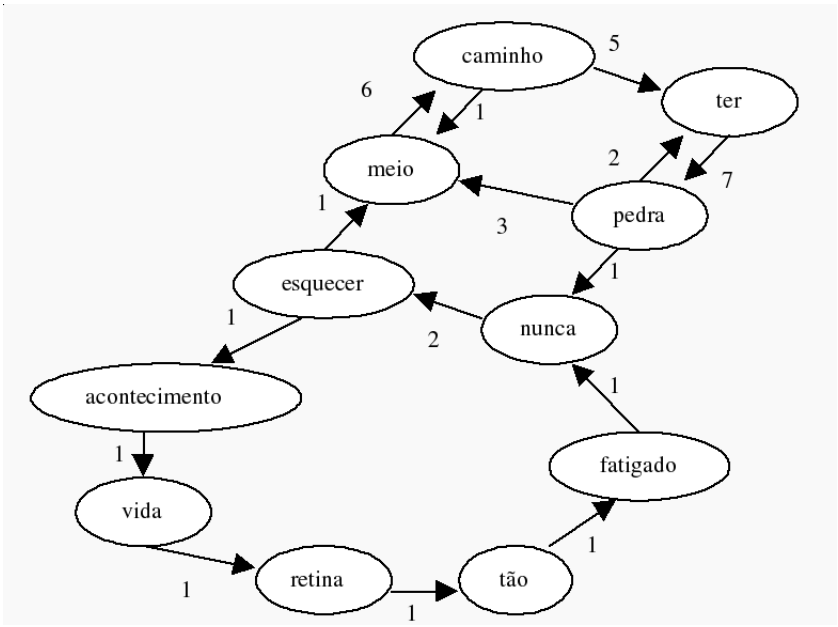


Figura 6: Rede complexa subjacente ao poema da "No meio do caminho" (Antiqueira et al., 2005b).

Após o pré-processamento, a rede derivada de um texto é representada por uma matriz de adjacência W de dimensão $N \times N$, onde N corresponde ao número de palavras distintas após o pré-processamento. Inicialmente, todos os elementos da matriz são iguais. À medida em que cada par de palavras (i, j) era lido do texto, incrementava-se o peso da aresta $i \rightarrow j$ com $W(i, j) = W(i, j) + 1$.

Com o propósito de avaliar a potencialidade das redes na avaliação da qualidade de textos, várias medidas estatísticas foram computadas. São elas: a média dos graus de saída dos vértices[2], média do coeficiente de aglomeração

[2]Dado que em um dígrafo a média dos graus de entrada e saída são iguais, somente a segunda foi calculada.

de cada nó e o caminho mínimo médio entre todos os pares de nós da rede (exceto das auto-conexões).

As medidas foram calculadas com base em dois conjuntos de textos diferentes. O primeiro deles composto por 10 textos do gênero informativo, os quais foram produzidos por estudantes do curso de Letras. O segundo, com 10 redações produzidas por vestibulandos da Fuvest. Os 20 textos foram avaliados por outros 6 alunos do curso de Letras, que atribuíram notas de 0 a 10 para cada um deles. Os textos do primeiro conjunto obtiveram notas acima da média e foram classificados como bons, enquanto que os do segundo conjunto obtiveram as piores notas, sendo julgados como ruins.

Após, foram comparadas as notas atribuídas pelos humanos com as medidas extraídas da rede. Essa comparação revelou fenômenos bastante interessantes: quando considerados os textos dos dois conjuntos, a qualidade tende a diminuir na mesma proporção em que aumentam os graus de saída dos vértices. Entretanto, observa-se que, ao considerar apenas os textos bons, a qualidade praticamente independe dos graus de saída. Por outro lado, considerar somente os textos ruins, nota-se uma melhora da qualidade na medida em que aumentavam os graus de saída. Percebeu-se ainda, que a média dos graus de saída dos textos bons foi menor do que a encontrada nos textos ruins. Além do mais, os textos ruins apresentaram um maior número de arestas, sendo que, dentro dessa classe, aqueles que obtiveram melhores notas têm um grau de saída maior. Em relação ao coeficiente de aglomeração, observou-se que a qualidade dos textos diminui à medida que o coeficiente reduz. Por fim, quando compara as notas com as medidas baseadas no caminho mínimo, concluiu-se que a qualidade é prejudicada quando o caminho mínimo médio é maior. Os autores acreditam que isso possa estar relacionado ao fato de que escritores inexperientes têm maior dificuldade em estabelecer conexões entre conceitos mais distantes no texto.

Um experimento adicional foi realizado em (Antiqueira et al., 2005a), com o propósito de verificar o comportamento da rede em relação ao tempo, ou seja, o crescimento dinâmico da rede. A dinâmica de crescimento foi calculada com base no número de componentes conexos na rede em um dado instante de tempo em que uma nova associação de palavras era encontrada no texto. Inicialmente, em um instante t_0, a rede era composta por N componentes, representados pelas N diferentes palavras do texto. No instante de tempo subseqüente, t_1, quando uma associação era encontrada entre duas palavras subjacentes w_1 e w_2, havia $N-1$ componentes, isto é, o componente formado por w_1 e w_2 e os $N-2$ componentes que restaram sem qualquer ligação entre eles. Esse procedimento foi repetido para cada nova associação de palavra encontrada até obter um único componente representando o texto todo. Ao

projetar o número de componentes da rede versus o tempo, durante a inserção de uma nova associação, observou-se que os textos de boa qualidade representavam uma reta, enquanto que o desvio aumentava na medida em que a qualidade deteriorava. O desvio foi calculado com base na Equação 5:

$$desvio = \sum_{m=1}^{A} \frac{|f(M) - g(M)|/N}{A} \tag{5}$$

onde $f(M)$ é uma função que determina o número de componentes para as M associações de palavras, $g(M)$ é uma função que determina a variação linear dos componentes para as M associações, N é o número de palavras diferentes no texto e A é o total de associações encontradas.

O experimento revelou que a variação do número de componentes da rede também pode ser usada para distingüir textos de boa e má qualidade.

Em resumo, os resultados obtidos em todos os experimentos mostraram que os parâmetros das redes complexas apresentam forte correlação com a qualidade dos textos e, portanto, são potencialmente úteis para a análise de textos.

3.2 Avaliação de Sumários

Com base na modelagem proposta em (Antiqueira et al., 2005a,b), Pardo et al. (2006a,b) propõem um modelo para a avaliação de sumários produzidos automaticamente baseado em cinco diferentes representações de redes complexas. Em todas elas, os textos foram previamente processados em duas etapas: (i) eliminação de *stopwords* e (ii) lematização de palavras.

A primeira representação é semelhante à proposta por Antiqueira et al. (2005b), em que cada vértice corresponde a uma palavra e as arestas direcionadas estabelecem as associações entre elas. Cada associação é determinada por uma simples relação de adjacência, ou seja, para cada par de palavras adjacentes no sumário, há uma aresta na rede apontando da primeira para a segunda palavra. As arestas contêm pesos que representam o número de vezes que as palavras adjacentes correspondentes são encontradas no sumário. Essa representação é denominada Markov-1, pois representa o modelo de Markov de um estado, no qual cada palavra está relacionada apenas a palavra imediata anterior no texto.

Esse modelo especifica como a determinação de um estado depende da observação de estados anteriores. Nesse caso, cada estado é representado por uma palavra no sumário. As quatro representações restantes, denominadas Markov-2, Markov-3, Markov-4 e Markov-5, são simplesmente variações da Markov-1. Elas diferem apenas no número de palavras anteriores que se relacionam com cada palavra do sumário. Por exemplo, em Markov-2, para

	Sumários Manuais	GEI	GistSumm	SuPor
Markov-1	0.03045	0.03538	0.03673	0.04373
Markov-2	0.03045	0.03538	0.03673	0.04374
Markov-3	0.03174	0.03657	0.03833	0.04489
Markov-4	0.03350	0.03807	0.04046	0.04643
Markov-5	0.03537	0.03977	0.04262	0.04808

Tabela 1: Resultados do experimento 1: Desvio de crescimento dinâmico das redes.

uma seqüência de palavras w_1, w_2, w_3 no sumário, há uma aresta de w_1 para w_3 e outra de w_2 para w_3, indicando que w_3 está relacionada com as duas palavras anteriores.

As mesmas medidas utilizadas por Antiqueira et al. (2005b) foram calculadas a partir das redes: a) médias dos graus de saída, b) coeficiente de aglomeração e c) dinâmica de crescimento linear (ou desvio). Essas medidas foram obtidas para dois conjuntos de sumários diferentes, produzidos a partir de um conjunto de 100 textos jornalísticos[3]: (i) conjunto de sumários manuais escritos por um profissional humano e (ii) conjunto de sumários automáticos gerados por três sumarizadores do português, denominados GistSumm (Pardo et al., 2003), SuPor (Módolo, 2003) e GEI (Pardo and Rino, 2004).

De acordo com Pardo et al. (2006b,a), os sumários manuais são reconhecidamente melhores que os sumários automáticos. Entre os automáticos, aqueles produzidos pelo sistema GEI foram considerados melhores que os produzidos pelos sistemas GistSumm e SuPor, pois foram construídos com base em sumários manuais. Comparando os dois últimos sistemas, em um experimento anterior realizado com o mesmo conjunto de textos, o SuPor apresentou desempenho melhor do que o GistSumm.

A fim de verificar se as medidas extraídas das redes apresentavam alguma correlação com esse *ranking* de sumários, foram realizados dois experimentos. No primeiro, cada sumário foi representado com os cinco tipos de redes e, para cada uma delas, calcularam o desvio de crescimento dinâmico, à medida em que uma nova associação de palavras era incluída na rede. No segundo experimento foram utilizadas somente as redes baseadas nos modelos de Markov-1 e Markov-2 e calculadas as medidas de grau de saída e coeficiente de aglomeração. Na Tabela 1 são apresentados os resultados obtidos com os sumários manuais e com cada sistema, para o primeiro experimento.

Por meio da Tabela 1, observa-se que os sumários manuais obtiveram os menores desvios em todos os tipos de redes. Vale lembrar que, segundo Antiqueira et al. (2005b), o desvio diminui na medida em que aumenta a qualidade

[3]Esses textos compõem o Corpus TeMário disponível em: http://www.linguateca.pt/Repositorio/TeMario (último acesso em 21/06/06).

	Markov-1		Markov-2	
	Grau de Saída	Coef. Aglom.	Grau de Saída	Coef. Aglom.
Sumários Manuais	1.23065	0.00267	2.44927	0.44933
GEI	1.28568	0.00395	2.56037	0.44594
GistSumm	1.27730	0.00447	2.54034	0.44846
SuPor	1.35283	0.00522	2.69500	0.44299

Tabela 2: Resultados do experimento 2: Grau de saída dos vértices e coeficiente de aglomeração.

dos textos, sendo que o crescimento dinâmico dos melhores textos é uma reta. Em relação aos três sistemas, os menores desvios foram obtidos pelo GEI. Esses resultados são correlatos à hipótese dos autores de que os sumários manuais são melhores que os automáticos e que, entre esses, os do GEI são os melhores. Por outro lado, o SuPor teve um desempenho pior do que o GistSumm, ao contrário do que se esperava. Os autores especulam que isso possa ser conseqüência de uma influência positiva da rede no modo como o GistSumm constrói os sumários.

O primeiro experimento também mostrou que não há diferença entre os resultados obtidos pelas redes Markov-1 e Markov-2, enquanto que, para as outras representações, o desvio aumenta consistentemente, embora a tendência se mantenha. Por essa razão, somente as duas primeiras representações foram consideradas no segundo experimento. Os resultados obtidos são mostrados na Tabela 2 a partir da qual observou-se que, novamente, os sumários manuais são melhores do que os sumários automáticos, uma vez que obtiveram os menores graus de saída e coeficientes de aglomeração.

Nota-se, também, que o coeficiente de aglomeração das redes de Markov-2 praticamente não variou com os diferentes conjuntos de sumários. Pardo et al. (2006b) concluem que as medidas extraídas das redes complexas apresentam correlação com a qualidade de textos sugerida em (Antiqueira et al., 2005b) e, portanto, podem ser usadas na avaliação da qualidade de sumários.

3.3 Deteção de Comunidades

O processo de Mineração de Dados é comumente utilizado para descobrir conhecimento sobre determinado domínio de aplicação. Para isso, podem ser utilizadas diversas tecnologias como as ferramentas de aprendizado de máquina (AM), uma sub-área da Inteligência Artificial (IA), cujo o objetivo é a construção de sistemas capazes de adquirir conhecimento útil de maneira automática ou semi-automática (Monard and Baranauskas, 2003). O aprendizado de máquina pode ser dividido em supervisionado, não-supervisionado e semi-supervisionado. Essa classificação depende da disponibilidade e características dos dados utilizados na execução dos algoritmos de AM.

O *clustering* é uma das técnicas freqüentemente utilizadas para análise e exploração de dados não-supervisionados. Essa técnica tem sido aplicada no contexto de redes complexas, em diferentes temas, tais como análise do comportamento social, análise da estrutura física da Internet, de páginas Web, problemas de epidemiologia e outros relacionados à bioinformática. Dentro da nomenclatura utilizada pelos pesquisadores de redes complexas, o *clustering* é usualmente denominado detecção de comunidades. É importante observar que o *clustering* em AM não é exatamente o mesmo que detecção de comunidades aplicada sobre redes complexas e, portanto, não devem ser confundidos apesar de apresentarem diversas caracterísitas em comum.

A semelhança entre essas duas técnicas possibilita que algoritmos implementados para uma possa ser facilmente adaptado para outra e vice-versa. Por exemplo, um conjunto de dados em alta dimensão pode ser representado por meio de uma rede complexa, adicionando arestas entre os vértices similares, para aplicação de um algoritmo de detecção de comunidade. Entretando, essa adaptação, em geral, apresenta resultados piores que os algoritmos já existentes para a resolução de problemas específicos.

O estudo de detecção de comunidades está altamente correlacionado com os conceitos da teoria dos grafos e com a abordagem de *clustering* hierárquico. Essa correlação vem da utilização de métodos de particionamento da rede em sub-grafos que representam individualmente cada comunidade presente na rede. O particionamento em si não é suficiente para análise e entendimento da estrutura de dados mapeados na rede, pois não se conhece *a priori* "se" e "como" a rede separa os vértices em comunidades, nem tampouco o número e o tamanho das possíveis comunidades. Por outro lado, o *clustering* hierárquico é utilizado para de descobrir divisões naturais na rede. Essa técnica é normalmente baseada em métricas de similaridade ou força das conexões entre vértices. Os aloritmos de *clustering* hierárquico são classificados em duas abordagens (Jain and Dubes, 1988): aglomerativa e divisiva. Na primeira, cada exemplo (vértice da rede) é considerado um cluster unitário. Em seguida, arestas são iterativamente adicionadas ao grafo, para a união dos sub-grafos até que todos os vértices pertençam a apenas um grafo (cluster). A abordagem divisiva faz o oposto, ela inicia com apenas um grafo contendo todos os vértices e procede dividindo esse grafo em sub-grafos cada vez menores, até que cada vértice seja um grafo isolado ou até que se alcance algum critério de parada, freqüentemente o número de sub-grafos desejados (Murtagh, 1983).

Em alguns casos, os algoritmos de *clustering* são capazases de encontrar as divisões naturais da rede, pois a métrica utilizada pelo algoritmo corresponde à métrica interna da rede. Em outros casos, o algoritmo pode não ser capaz de identificar essa estrutura, pois a rede não tem uma descrição métrica

natural. Nesses casos, outras medidas podem ser utilizadas na identificação dos clusters, tais como coeficientes de correlação, comprimento de caminhos entre vértices, fluxo maximo (Ahuja et al., 1993) e operações sobre matrizes (Newman and Girvan, 2004).

Freqüentemente são desenvolvidos estudos para avaliação de interações sociais entre indivíduos pela comunidade científica (Zachary, 1997). Essas interações são representadas por meio de redes complexas para a detecção de comunidades que auxiliem os pesquisadores na interpretação do comportamento dos indivíduos. Essas comunidades são obtidas com o agrupamento de vértices que contêm alta densidade de arestas entre eles e baixa densidade de arestas que interligam grupos distintos. Esse comportamento pode ser verificado quando ocorre a divisão de pessoas em grupos de interesse, ocupação ou faixa etária, por exemplo (Newman, 2003; Hopcroft et al., 2003). A sub-divisão das áreas de conhecimento e suas sub-áreas é outro exemplo claro de existência de comunidades. Nesse caso, pode-se analisar aspectos como a cooperação entre pesquisados de uma determinada área na elaboração de trabalhos, a inter-disciplinaridade das linhas de pesquisa e também as citações entre tabalhos de diferentes autores.

Outras aplicações da detecção de comunidades em redes complexas podem ser citadas. Por exemplo, o estudo da estrutura de redes *Webs* desenvolvido por Virtanen (2003). Nesse trabalho, o autores realizam o *clustering* sobre um grafo que representa um sub-conjunto da *Web*, restrito às páginas chilenas. A identificação de clusters nesse sub-conjunto pode auxiliar nas estratégias de indexação das páginas e, também, na extração de conhecimentos semânticos a respeito da estrutura da rede.

Em (Newman and Girvan, 2004), foi proposto um algoritmo para identificação e avaliação de comunidades em redes complexas. As melhorias propostas pelo autor para esse algoritmo foram posteriormente implementadas em (Newman, 2004). Outros trabalhos foram desenvolvidos utilizando as idéias apresentadas nesse algoritmo e aplicando o conceito de *betweenness* para o cálculo do caminho mínimo entre vértices do grafo (Girvan and Newman, 2001; Holme et al., 2003). O *betweenness* é uma medida utilizada para identificar arestas que conectam comunidades, apresentando valores altos para essas arestas e penalizando as arestas que conectam vértices de um mesmo sub-grafo. Para entender a idéia dessa medida, considere duas comunidades que são ligadas por um conjunto pequeno de arestas. Neste caso, todos os caminhos da rede com origem em um vértice de uma comunidade e destino em um vértice da outra comunidade devem passar por alguma dessas arestas que conectam as duas comunidades. Com isso, pode-se mensurar a importância de cada aresta par a junção dessas comunidades com base no número de caminhos

que utilizam cada uma das arestas. A medida de *betweenness* é, portanto, baseada no caminho mínimo.

O algoritmo de detcção de comunidades proposto por Newman and Girvan (2004) segue a abordagem de *clustering* divisiva. Entretanto, ele utiliza uma estratégia diferente das adotadas por algoritmos propostos anteriormente, pois ao invéz de procurar por pares de vértices com menor similaridade e remover a aresta que os une, esse algoritmo remove as arestas responsáveis pela conexão entre sub-grafos. Essas arestas não são necessariamente fracas no contexto de similaridade entre vértices, mas são arestas que determinam o aparecimento das comunidades quando removidas da rede, pois contêm o maior valor de *betweenness*. Além disso, a cada remoção de uma aresta, o algoritmo atualiza o *betweenness* das arestas que permanecem no grafo. No caso dos outros algoritmos, o valor de *betweenness* é calculado apenas uma vez e, a partir desses valores, as arestas são removidas da rede em ordem decrescente de *betweenness* para a construção do dendograma. Porém, uma vez que uma aresta é removida da rede, eses valores não refletem mais o seu estado atual, o que pode apresentar resultados indesejados, resultando em uma estrutura que não pertence à rede. Devido a esse fato, o *betweenness* é reclaculado a cada iteração do algoritmo. Os passos realizados por esse algoritmo são:

1. Cálculo do *betweenness* para todas as arestas da rede;
2. Busca e remoção da aresta que maximiza o *betweenness*;
3. Recálculo do *betweenness* para as arestas restantes;
4. Retorno ao passo 2.

A complexidade do cálculo do *betweenness* depende da quantidade de arestas e vértices da rede. Em um grafo com M arestas e N vértices, o cálculo do caminho mínimo entre um par específico de vértices pode ser feito utilizando o procedimento de busca em largura com tempo de execução na ordem $O(M)$. Assim, como existem $O(N^2)$ pares de vértices, a complexidade total para o cálculo do *betweenness* está na ordem de $O(MN^2)$. Entretanto, Newman (2001) propôs um algoritmo que executa esse cálculo de maneira mais eficiente e consome tempo em ordem linear ($O(MN)$).

O algoritmo proposto por Newman and Girvan (2004) apresentou bons resultados quando aplicado sobre redes aleatórias e reais com estruturas conhecidas. No entanto, em situações reais, dificilmente a estrutura da rede é conhecida *a priori*, o que acarreta a necessidade de algum método para validar a estrutura recuperada pelo algoritmo. Isto ocorre porque todo algoritmo aglomerativo ou divisivo sempre produz uma divisão da rede em sub-grafos

(comunidades), mesmo em redes completamente aleatórias que não possuam comunidades significativas. Devido a isso, os autores do algoritmo criaram uma medida capaz de mensurar a qualidade da divisão feita na rede. Essa medida foi denominada modularidade (Newman, 2006).

Como ilustração do funcionamento dessa medida, considere a divisão de uma rede em k partições. Considere também, uma matriz simétria E de ordem k, cujos elementos e_{ij} são a fração de arestas na rede que conectam vértices presentes na comunidade i aos vértices presentes na comunidade j. O traço dessa matriz $Tr\ E = \sum_i e_{ii}$ representa a fração de arestas que conectam vértices dentro da mesma comunidade. Claramente, uma boa divisão entre comunidades deve apresentar valores altos para o traço da matriz. Porém, somente esse valor não é um bom indicador de qualidade da divisão da rede, pois se todos os vértices estiverem presentes em uma única comunidade o valor do traço será máximo, *i.e.*, $Tr\ E\ = 1$.

Para resolver essa limitação, é utilizado o somatório dos elementos das linhas (ou colunas) da matriz, o qual representa a fração de arestas que conectam dois vértices presentes em uma comunidade. Assim, $a_i = \sum_j e_{ij}$ indica a fração para a comunidade i. Com isso, a modularidade pode ser definida por meio da Equação 6:

$$Q = \sum_i (e_{ij} - a_i^2) = Tr\ E - \|E^2\| \tag{6}$$

onde, $\|E\|$ é a soma dos elementos da matriz E.

Valores próximos a 1 para Q indicam a forte presença de estrutura na rede. Na prática, os valores para a modularidade variam entre 0.3 e $.7$. Em alguns casos podem chegar mais próximo de 1, mas são raros os casos em que isso acontece. Para identificar a melhor divisão da rede, Q é usualmente calculado para cada conjunto de sub-grafos. A partir do dendograma do algoritmo hierárquico divisivo, calcula-se o valor de Q para cada nível de agrupamento, seguindo a abordagem *top-down*. O máximo valor alcançado indica a melhor divisão encontrada pelo algoritmo para a rede analisada.

Esse algoritmo foi utilizado para avaliar a estrutura de diversas redes, entre elas uma rede artificial, criada especificamente para a avaliação do algoritmo de detecção de comunidades, em um experimento controlado, e outras redes amplamente estudas pela comunidade científica. Das aplicações apresentada no artigo original de Newman and Girvan (2004), duas são descritas brevemente neste trabalho: rede artificial e análise de uma rede social.

3.3.1 Rede artificial

Para executar um experimento controlado, foi criada uma rede artificial com estrutura de comunidade conhecida. O objetivo do experimento foi avaliar se o algoritmo proposto é capaz de identificar essa estrutura. Essa rede é composta por 128 vértices divididos em 4 comunidades de tamanho uniforme, 32 vértices em cada comunidade. Foram considerados valores de probabilidades para a coneção entre vértices: p_{in} para arestas que conectam vértices de uma mesma comunidade e p_{out} para arestas que conectam vértices de diferentes comunidades. Cada vértice da rede possui grau igual a 16. O dendograma[4] obtido para essa rede é apresentado na Figura 7 na qual é apresentado, também, o gráfico do valor de modularidade construiído em função do corte no dendograma. Como pode ser observado no gráfico, há um pico bem definido, que indica a divisão do dendograma nas 4 comunidades conhecidas. O valor da modularidade obtido para essa divisão do dendograma está dentro da faixa de valores típicos qye variam entre 0.3 e 0.7.

[4]Para melhor visualização, é apresentado o dendograma de apenas 64 vértices da rede

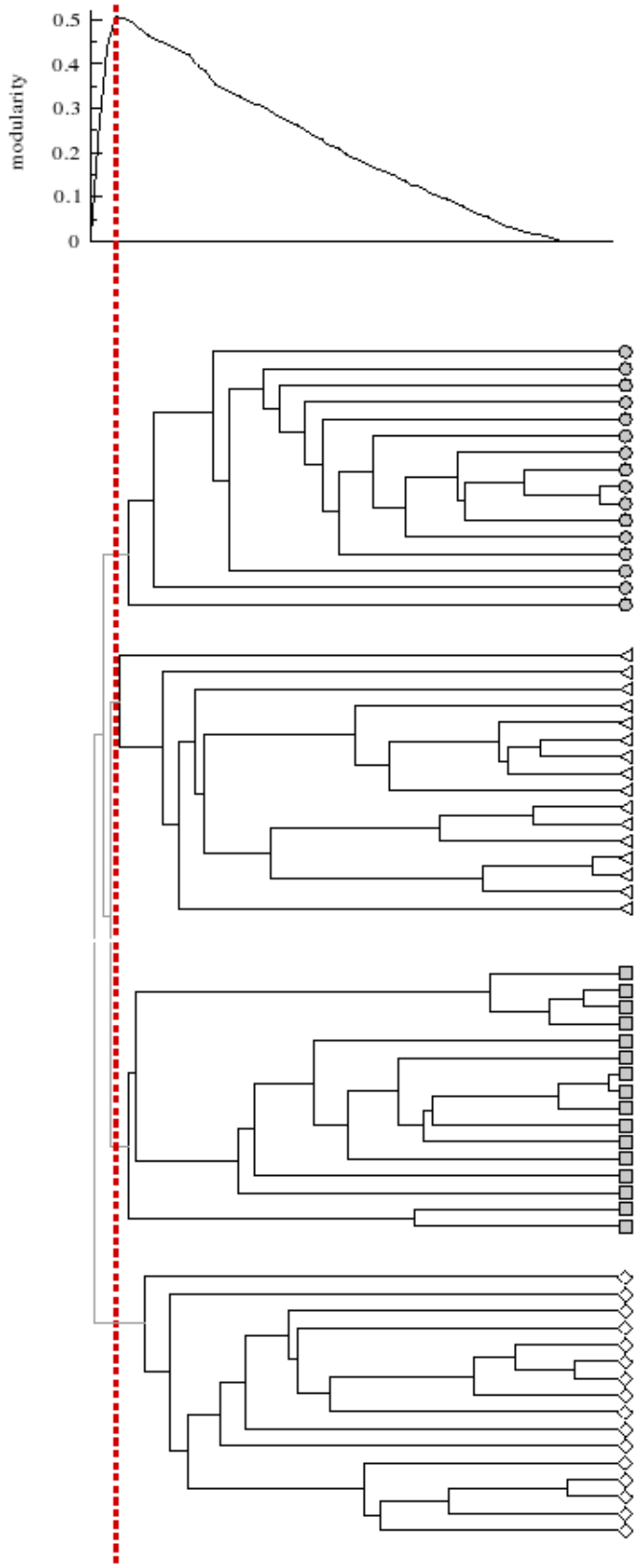


Figura 7: Dendograma e modularidade da rede artificial (Newman and Girvan, 2004).

3.3.2 Rede social sobre dados reais

Esse experimento foi realizado sobre uma rede amplamente estudada pela comunidade, cujo objetivo é a análise de interações sociais. Durante dois anos na década de 1970, Wayne Zachary observou interações entre membros de uma academia de karatê de uma universidade Norte Americana (Zachary, 1997). A partir dessa observação, ele construiu uma rede que representa as interações sociais entre membros dessa academia, dentro e fora dela. Zachary observou que existiam dois grupos distintos de pessoas, um que seguia o professor principal da academia e outro que seguia o administrador.

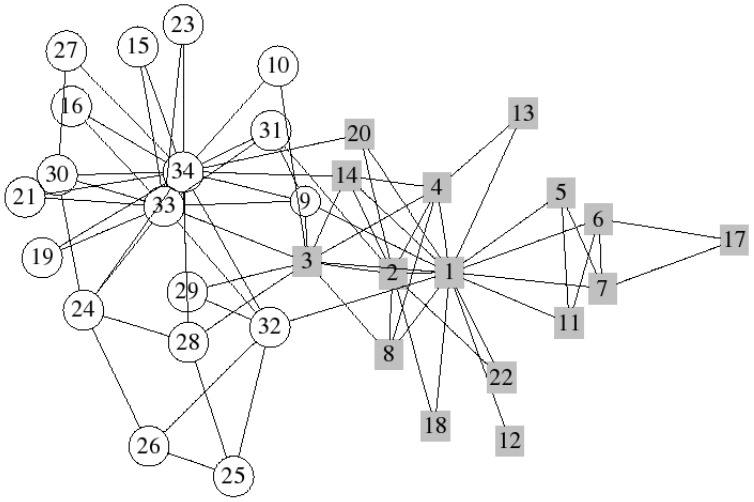


Figura 8: Rede de interação social dos membros da academia de karatê (Newman and Girvan, 2004).

Na Figura 8 é apresentada a estrutura da rede identificada por Zachary. Utilizando essa rede como entrada do algoritmo para detecção de comunidades, foi construído o dendograma apresentado na Figura 9. A partir desse dendograma, observa-se que o valor de modularidade obtido a partir da divisão da rede em dois sub-grafos é relativamente alto, indicando a existência de uma divisão natural desses vértices na rede. Além disso, a divisão nesse ponto é quase perfeita em relação a divisão verdadeira, pois apenas um vértice está presente no sub-grafo, o qual não pertence na divisão feita por Zachary.

Na Figura 10 é apresentado o resultado obtido sobre essa mesma rede, sem considerar a atualizaçàо do valor de *betweenness* das arestas restantes a cada iteração do algoritmo.

3.4 Congestionamento em redes

Em trabalho recente, Liang et al. (2005) abordaram o congestionamento em redes de comunicação através de uma modelagem matemática baseada na teoria de Redes Complexas. O principal objetivo do estudo foi analisar como

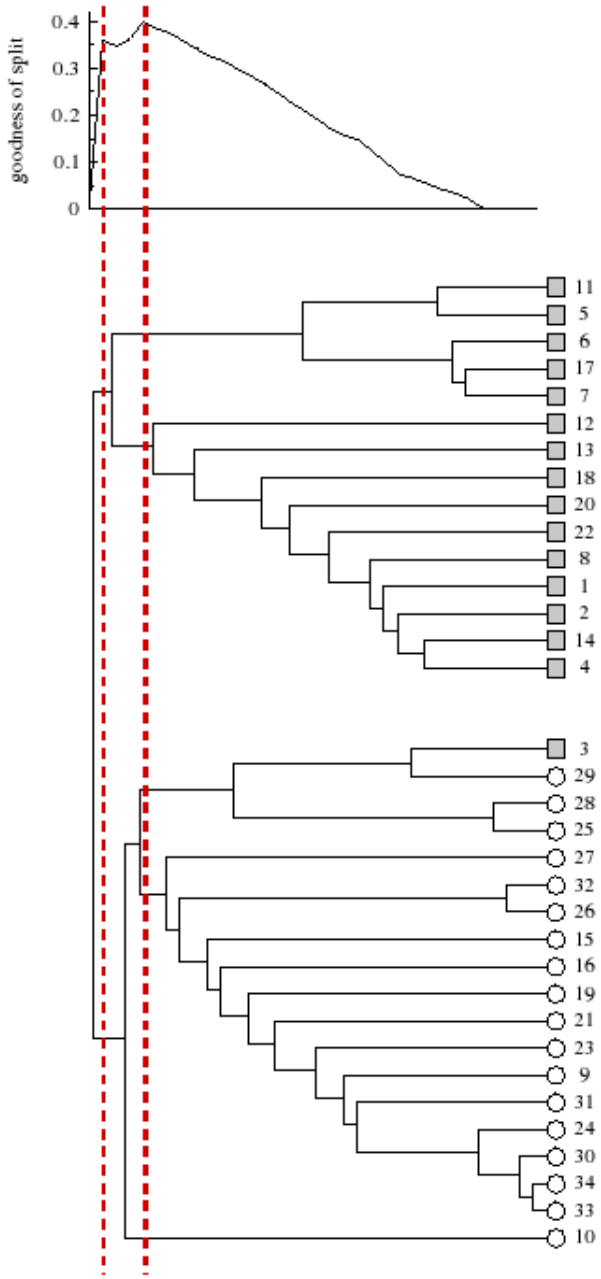


Figura 9: Dendograma e modularidade da rede social (Newman and Girvan, 2004).

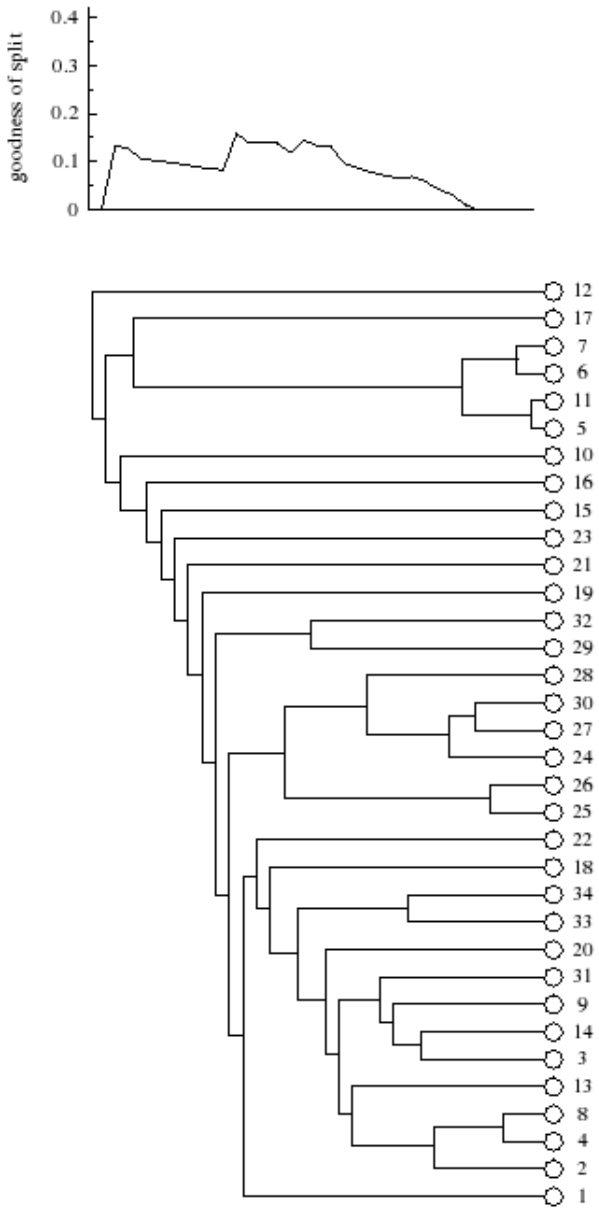


Figura 10: Dendograma e modularidade da rede social sem atualização de *betweenness* (Newman and Girvan, 2004).

a topologia de rede influencia no tráfego de pacotese, a partir dessa análise, explorar maneiras de minimizar seus efeitos.

Nas redes de tráfego de pacotes, há dois elementos computacionais: os *hosts* e os rotadores. Os primeiros criam pacotes com endereço de destinatário e recebem pacotes vindos de outros *hosts*. Os segundos são responsáveis por encontrar o melhor caminho entre os *hosts* remetente e destinatário e encaminhar os pacotes por este caminho ao longo do tempo. O melhor caminho é definido pelo menor número de *hosts* visitados para um pacote sair de sua origem e atingir seu destino.

Dois modelos foram criados para estudar o fenômeno e se baseiam na forma como os dados são transmitidos na Internet. Alguns estudos sugerem que a rede Internet possui características de redes Livres de Escala e Pequeno-Mundo (Barabasi and Albert, 1999b), (Albert and Barabási, 2002). Os modelos desenvolvidos trabalho, utilizam dois parâmetros: uma taxa de criação de pacotes λ e um parâmetro β , que controla a cpacidade de processamento de pacotes de cada vértice. No primeiro modelo, a capacidade de entrega de pacotes de um vértice é proporcional ao seu grau. No segundo modelo, é proporcional à quantidade de caminhos mínimos que passam pelo vértice (*betweenness*). A quantidade de de pacotes congestionados na rede, λ_c, é medida pelo número de pacotes criados na rede em um dado instante de tempo em que ocorre uma transição de estado do fluxo de tráfego livre para congestionado. Se $\lambda < \lambda_c$, então a razão de pacotes criados e entregues é equilibrada e a rede está em estado de fluxo livre. Se $\lambda > \lambda_c$, então há um desequilíbrio entre pacotes criados e entregues, resultando em congestionamento. Como o λ_c depende da topologia da rede, os autores estudaram sua relação com redes complexas do tipo Árvores de Cayley e algumas redes, tais como as Regulares, Livres de Escala e Aleatórias.

Os modelos de tráfego de pacotes são descritos pelos passos a seguir:

1. a cada iteração, o *host* i gera um pacote com probabilidade λ e entrega C_i pacotes na direção do caminho ótimo definido pelo roteador. Para o modelo 1, $C - I = (1 + int[\beta k_i])$, onde $0 < \beta < 1$ é um parâmetro de controle e k_i é o grau do vértice i. Para o modelo 2, $C_i = 1 + int[\beta B_i/N]$, onde B_i é o *beteweenness*. Uma vez que o pacote alcança seu destino, ele é retirado do tráfego.

2. Depois de ser criado, o pacote é posto no final fila do *host* que o criou ou é entregue se a fila de pacotes estiver vazia. Cada uma das ações de criação e entrega de um pacote ocorre em uma iteração. Portanto, um pacote não pode ser criado e entregue na mesma iteração. A entrega de um pacote ocorre no mínimo uma iteração após a sua criação. Uma vez criado o pacote, escolhe-se aleatoriamente um vértice destino para ser

enregue. O roteador encontra o menor caminho entre o *host* gerador e o vértice destino. Assim, o pacote gerado é transmitido para a rede ao longo do seu caminho durante os passos seguintes. Se existir mais de um caminho mínimo entre o *host* gerador e o vértice destino, escolhe-se o caminho cujo o próximo vértice, a partir do *host* gerador, possui o menor tamanho para a fila de pacotes.

3. A cada iteração, os C_i primeiros pacotes da fila de pacotes do vértice i são transmitidos para a rede em direção ao destino de cada um dos pacotes e, então, são postos no final da fila de pacotes do vértice escolhido (vizinho do vértice i). Caso o vértice i possua mais de C_i pacotes na fila, então os pacotes que permanecem na fila são reposicionados C_i posições. Dessa maneira, o tempo de entrega de um pacote não é contabilizado somente pela distância (número de iterações) entre o *host* gerador e o vértice destino, mas também pelo número de pacotes existentes ao longo de seu caminho, ou mais especificamente, pelo tamanho da fila de pacotes dos vértices intermediários que o pacote do vértice i visitou.

Sendo que N seja o número de vértices da rede, o número total de pacotes criados em uma iteração é λN e o número total de pacotes entregues a cada iteração é aproximadamente $\sum_{i=1}^{N} C_i$, se todo vértice tem uma quantidade suficiente de pacotes, que é maior que o número total de pacotes criados na rede, portanto $\lambda < 1$. Em se tratando de redes complexas, torna-se provável que os pacotes, antes de chegarem em seus destinos, sejam transmitidos para os vértices com altos valores de *betweenness*, pois apresentam os caminhos mínimos entre qualquer par de vértices. Este fato resulta em um possível congestionamento de pacotes.

3.5 Controle do Congestionamento de Pacotes

Considerando o problema de congestionamento de pacotes citado na seção anterior, é proposta uma abordagem para evitar o congestionamento. Para modelar uma rede de computadores é utilizada a represntação de uma rede complexa, em que os vértices representam os computadores (*hosts* e roteadores) e as arestas representam os *links* entre os computadores. A modelagem com rede complexa foi adotada por suportar grande quantidade de vértices como é o caso desse problema.

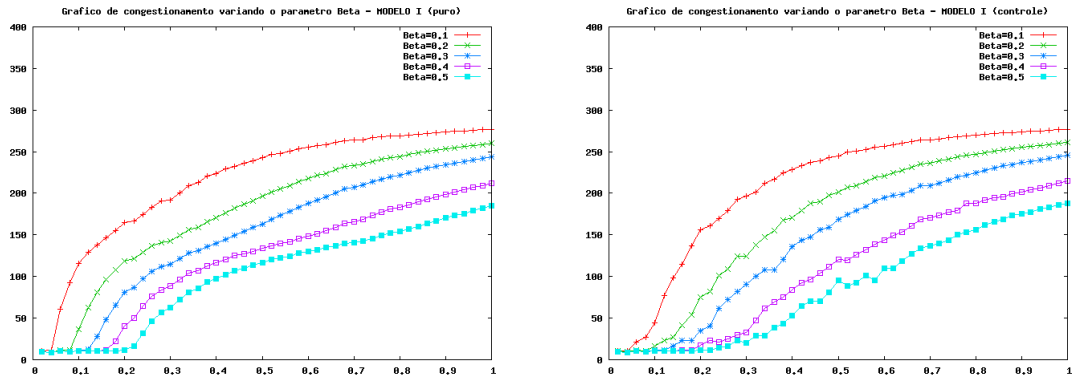
A rede complexa é do tipo aleatória com grau médio 4. Por ser aleatória, os vértices da rede podem ser conectados com quaisquer outros com probabilidade ρ. No entanto, algumas restrições foram impostas. A primeira restrição é a de que o grau médio da rede deve ser 4, ou seja, o número médio de conexões de cada vértices deve ser 4. A segunda refere-se ao fato de impedir a criação

de componentes isolados. Após a criação da rede, todos os vértices devem ser alcançados partindo de qualquer outro vértice da rede. Assim, é possível que um pacote gerado em qualquer vértice possa ser entregue ao seu destino.

Cada nó, representando um computador da rede, é responsável pela criação dos pacotes e pela hospedagem dos mesmos quando são provenientes de outros vértices. Para isso, todo vértice da rede rede possui uma fila de pacotes a ser entregues a seus destinos. À medida em que os pacotes chegam em um vértice, eles são armazenados nessa fila. A quantidade de pacotes entregues por um vértice em uma iteração é definida pelo valor de sua capaciade de entrega (ou processamento) de pacotes. A cada iteração, os vértices são capazes de gerarem um pacote com probabilidade λ. No instante da criação de um pacote é atrbuído a ele um vértice destino aleatoriamente e, então, o pacote é colocado no final da fila do vértice gerador para que, na próxima iteração, seja entregue ou deslocado na fila de pacotes. Ao chegar ao seu destino, o pacote é removido da rede. Além da fila de pacotes, cada vértice possui uma tabela de roteameto (criada através do algoritmo de Dijkstra) que contém o próximo vértice do caminho mínimo entre aquele vértice e o destino, para o qual o pacote será enviado. Em casos em que há mais de um caminho entre dois vértices, a tabela de roteamento é capaz de armazenar os próximos vértices de caminhos mínimos encontrados. Nesse caso, a escolha do vértice dependerá do tamanho da fila de pacotes. O vértice que apresentar o menor tamanho para a fila de pacotes é escolhido para receber o pacote.

Inicialmente, o algoritmo de Dijkstra é aplicado para determinar o caminho mínimo entre todos os pares de vértices da rede. Para isso, foram definidos custos unitários para todas as arestas da rede. Ao longo das iterações, pacotes são gerados e passam a circular na rede, sendo removidos quando chegam ao destino. Dessa maneira, é possível que os pacotes congestionem o tráfego na rede, sobrecarregando, principalmente, os vértices com maior valor para o *betweenness*, pois esses apresentam os menores caminhos, e portanto, são mais solicitados para intermediar a entrega de um pacote ao seu destino.

Com o objetivo de amenizar o congestionamento na rede, foi definido um intervalo de iterações para que o algoritmo de Dijkstra fosse aplicado a todos os vértices da rede novamente. Exceto na primeira aplicação do algoritmos, antes mesmo da criação dos pacotes e o envio dos mesmos, o algoritmo para encontrar o caminho mínimo entre dois vértices considera o tamanho da fila e a capacidade de entrega de pacotes dos vértices da rede. Quanto maior o tamanho da fila de pacotes, maior é o custo para que um pacote chegue ao seu destino. Por outro lado, quanto menor a capacidade de entrega de pacotes de um vértice, menor é o custo para que um pacote chegue ao destino. A alta capacidade de entrega de um vértice induz a um tamanho reduzido da fila



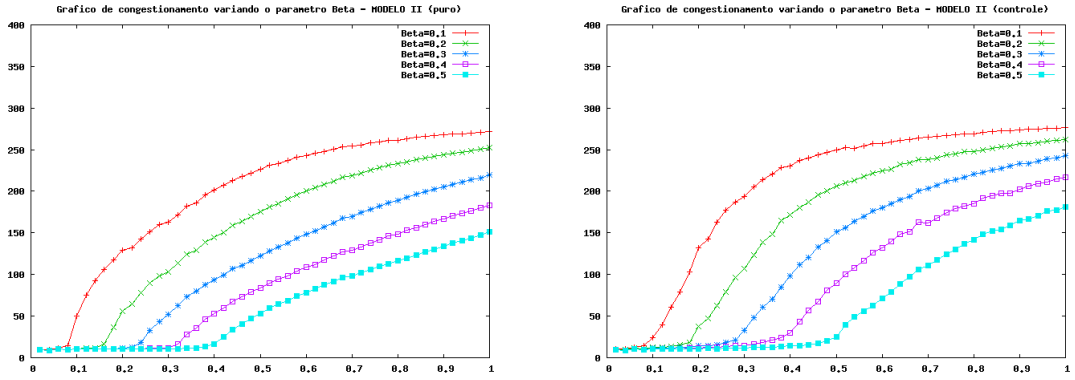
(a) (b)

Figura 11: Gráfico do congestionamento de pacotes com variação de β para o modelo 1.

de pacotes evitando que um pacote fique armazenado por muitas iterações. Dessa forma, as rotas entre dois vértices da rede podem ser alteradas devido aos fatores que influenciam no custo das arestas da rede. Dessa maneira, o tempo de entrega de um nó é influenciado não somente pela quantidade de vértices presentes no caminho até alcançar seu destino,mas também pelo tamanho da fila de pacotes e da capacidade de entrega dos vértices da rede.

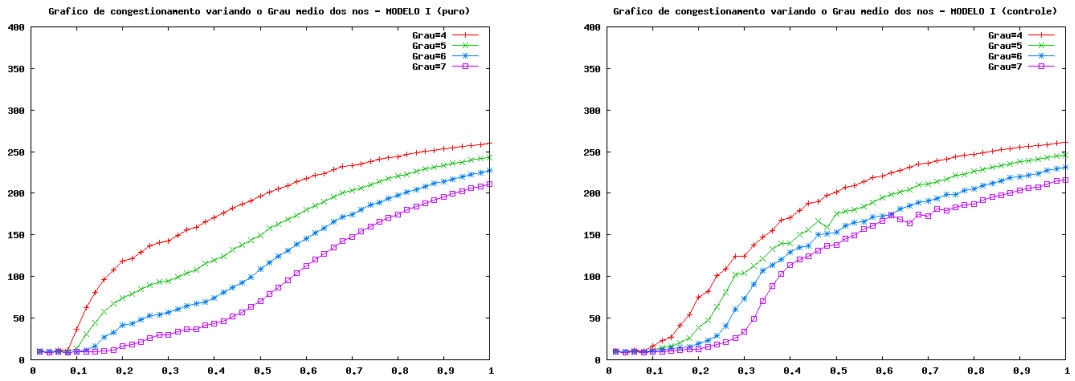
Um experimento foi realizado usando uma rede composta por 300 vértices com grau médio 4. O algoritmo de Dijkstra foi aplicado a cada 20 iterações para amenizar um possível congestionamento, e assim, criar rotas alternativas para a entrega de pacotes. Durante o experimento, cinco simulações foram realizadas, analisando-se a quantidade média de pacotes que circulavam na rede, denotada por η. Além disso, foi feita uma análise comparativa do tráfego de pacotes sem controle de congestionamento e com a aplicação do controle de congestionamento. Os resultados obtidos são sintetizados nas Figuras 11, 12, 13 e 14. Tais figuras mostram a relação entre a probabilidade de geração de pacotes (λ) e a quantidade de vértices existentes na rede (η).

As Figuras 11(a) e 11(b) apresentam os resultados da simulação variando-se os valores de $\beta = \{0.1, 0.2, 0.3, 0.4, 0.5\}$ para o modelo 1. Nota-se nas figuras que, praticamente, não houve controle de congestionamento. Na Figura 11(a), o controle é realizado pela aplicação esporádica do algoritmo de Dijkstra considerando o tamanho das filas de pacotes e de cada vértice da rede, conforme explicado anteriormente. Percebe-se que, em geral, o controle de congestionamento foi bem sucedido. As curvas dos η variando-se os valores de λ sofreram um pequeno deslocamento para a direita, indicando que o congestionamento ocorreu de forma tardia. Isso quer dizer que o número de pacotes aumentou. Estas considerações também podem ser aplicadas às Figuras 12(a) e 12(b), que ilustram os resultados dos experimentos baseados no modelo2.



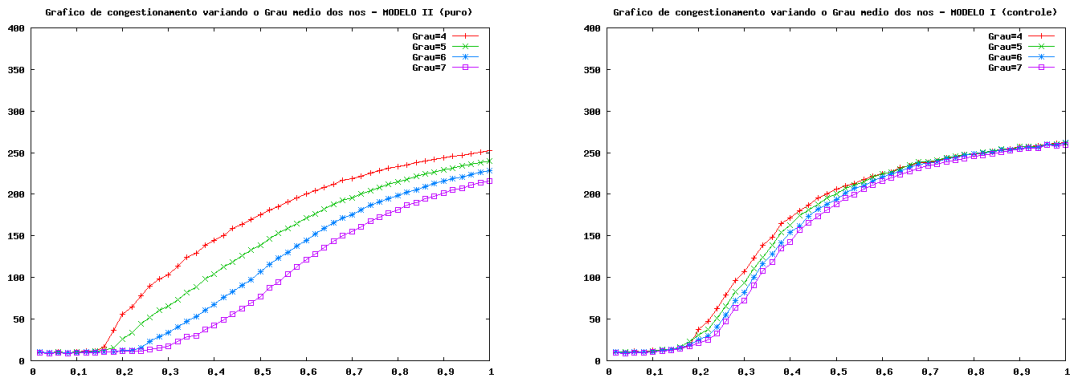
(a) (b)

Figura 12: Gráfico do congestionamento de pacotes com variação de β para o modelo 2.



(a) (b)

Figura 13: Gráfico do congestionamento de pacotes com variação do grau para o modelo 1.



(a) (b)

Figura 14: Gráfico do congestionamento de pacotes com variação do grau para o modelo 2.

As Figuras 13 e 14 mostram os resultados variando-se o grau médio de cada nó da rede. No caso do modelo 1 (Figuras 13(a) e 13(b)), o controle de congestionamento retardou o crescimento de η. Além disso, as curvas de crescimento ficaram mais suaves. As Figuras 14(a) e 14(b) mostram que o controle de congestionamento apresentou desempenho inconclusivo no modelo 2. Embora o congestionamento tenha ocorrido precocemente, as curvas curvas de crescimento de η foram suavizadas, sinalizado um aumento de congestionamento tardio.

4 Considerações Finais

Este relatório apresentou conceitos fundamentais sobre as redes complexas, assim como algumas de suas princioais propriedades e alguns dos modelos mais comumente utilizados. Aplicações reais envolvendo esses conceitos também foram descritas, como análise da qualidade de textos e sumários automáticos, detecção de comunidades em redes sociais e controle de tráfego de pacotes em redes de comunicação.

Referências Bibliográficas

Ahuja, R. K., Magnanti, T. L., and Orlin, J. B. (1993). *Network Flows: Theory, algorithms and Applications*. Prentice Hall. 16

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74(1):47–97. 24

Antiqueira, L. (2006). Desenvolvimento de técnicas baseadas em redes complexas para sumarização extrativa de textos (monografia de qualificação). Master's thesis, USP, ICMC. 9

Antiqueira, L., Nunes, M. G. V., Jr., O. N. O., and Costa, L. F. (2005a). Complex networks in the assessment of quality text. In *Physics*, 0504033. Physics. 9, 11, 12

Antiqueira, L., Nunes, M. G. V., Jr., O. N. O., and Costa, L. F. (2005b). Modelando textos como redes complexas. In *XXV Congresso da Sociedade Brasileira de Computação (III Workshop em Tecnologia da Informação e da Linguagem Humana - TIL 2005)*, São Leopoldo - RS, Brasil. In Anais do III Workshop em Tecnologia da Informação e da Linguagem Humana - TIL. 9, 10, 12, 13, 14

Barabasi, A. L. and Albert, R. (1999a). Emergence of scaling in random networks. *Science*, pages 286–509. 6, 8

Barabasi, A.-L. and Albert, R. (1999b). Emergence of scaling in random networks. *Science*, 286:509. 24

Barabási, A. L. (2003). *Linked: How everything is connected to everything else and what it means for business, science and everyday life*. Plume. 1, 3

Buchanan, M. (2002). *Nexus - small world and the groundbreaking science of network*. W. W. Norton Company. 7

da F. Costa, L., Rodrigues, F. A., Travieso, G., and Boas, P. R. V. (2005). Characterization of complex networks: A survey of measurements. 3

Girvan, M. and Newman, M. E. J. (2001). Community structure in social and biological networks. In *Proceedings of National Academy of Sciences*, number 99, pages 8271–8276, USA. 16

Holme, P., Huss, M., and Jeong, H. (2003). Subnetwork hierarchies of biochemical pathways. *Bioinformatics*, 19:532. 16

Hopcroft, J., Khan, O., Kulis, B., and Selman, B. (2003). Natural communities in large linked networks. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 541–546, New York, NY, USA. ACM Press. 16

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data*. 15

Liang, Z., Lai, Y.-C., Park, K., and Ye, N. (2005). Onset of traffic congestion in complex networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 71(2):026125. 21

Monard, M. C. and Baranauskas, J. A. (2003). *Conceitos sobre aprendizado de máquina*, volume 1 of *1*, chapter 4, pages 89–114. Barueri, SP, Brasil, 1 edition. 14

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26(40):354–359. 15

Módolo, M. (2003). Supor: Um ambiente para exploração de métodos extrativos para a sumarização automática de textos em português. Master's thesis, UFSCAR, Departamento de Computação. 13

Newman, M. (2003). The structure and function of complex networks. volume 45, pages 167–256. SIAM Review. 2, 3, 6, 8, 16

Newman, M. E. J. (2001). Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64(1):016132. 17

Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E*, 69:066133. 16

Newman, M. E. J. (2006). Modularity and community structure in networks. *PROC.NATL.ACAD.SCI.USA*, 103. 18

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 69(2):026113. iii, 16, 17, 18, 20, 21, 22, 23

Pardo, T. A. . S., Antiqueira, L., Nunes, M. G. V., Jr., O. N. O., and Costa, L. F. (2006a). Modeling and evaluation summaries using complex networks. In *7th Workshop on Computational Processing of Written and Spoken Portuguese - Propor*, Itatiaia - RJ, Brasil. 9, 12, 13

Pardo, T. A. . S., Antiqueira, L., Nunes, M. G. V., Jr., O. N. O., and Costa, L. F. (2006b). Using complex networks for language processing: The case of summary evaluation. In *4th International Conference on Communications, Circuits and Systems — ICCCAS*, Guilin, China. 9, 12, 13, 14

Pardo, T. A. . S. and Rino, L. H. M. (2004). Descrição do gei - gerador de extratos ideais para o português do brasil. Technical Report NILC-TR-04-07, Série de Relatórios do NILC. 13

Pardo, T. A. S., Rino, L. H. M., and Nunes, M. G. V. (2003). Gistsumm: A summarization tool based on a new extractive method. In *6th Workshop on Computational Processing of the Portuguese Language Written and Spoken - Propor*, pages 210–218. In Proceedings of the 6th Workshop on Computational Processing of the Portuguese Language Written and Spoken - Propor. 13

Strogatz, S. H. (2001). Exploring complex networks. *Nature*, 410:268–276. http://dx.doi.org/10.1038/35065725. 7, 9

Virtanen, S. E. (2003). Clustering the chilean web. In *Proceedings of the First Latin American Web Congress*, pages 229–231. IEEE Computer Society. 16

Watts, D. J. and Strogatz, S. H. (1998). Colletive dynamics of small-world networks. *Nature*, (393):440–442. 7

Zachary, W. W. (1997). An information flow model for conflict and fission in small groups. *Anthropological Research*, pages 452–473. 16, 21