UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS
DEPARTAMENTO DE FÍSICA E INFORMÁTICA

Paulino Ribeiro Villas Boas

Efeito da amostragem nas propriedades topológicas de redes complexas

SÃO CARLOS
2008

Livros Grátis

http://www.livrosgratis.com.br

Milhares de livros grátis para download.

Paulino Ribeiro Villas Boas

Efeito da amostragem nas propriedades topológicas de redes complexas

Tese apresentada ao Programa de Pós-Graduação em Física do Instituto de Física de São Carlos da Universidade de São Paulo, para obtenção do título de Doutor em Ciências.

Área de Concentração: Física Aplicada –
Opção Física Computacional
Orientador: *Prof. Dr. Luciano da Fontoura Costa*

SÃO CARLOS
2008

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO, PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Ficha catalográfica elaborada pelo Serviço de Biblioteca e Informação IFSC/USP

Villas Boas, Paulino R.
Efeito da amostragem nas propriedadesl topológicas de redes complexas./Paulino R. Villas Boas; orientador Luciano da Fontoura Costa.-- São Carlos, 2008.
114 p.

Tese (Doutorado em Ciências - Área de concentração: Física Aplicada) – Instituto de Física de São Carlos da Universidade de São Paulo.

1. Redes complexas 2. Grafos 3. Incompleteza.4. Amostragem. 5. Estrutura topológicas I. Título.

Resumo

VILLAS BOAS, P. R. **Efeito da amostragem nas propriedades topológicas de redes complexas**. 2008. 114pp. Tese (Doutorado) – Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2008.

Muitos sistemas complexos naturais ou construídos pelos seres humanos podem ser representados por redes complexas, uma teoria que une o estudo de grafos com a mecânica estatística. Esse tipo de representação, porém, pode ser comprometido pela maneira como os dados são obtidos. Em geral, os dados utilizados para representar tais sistemas nem sempre são precisos ou completos e correspondem a apenas amostras pequenas de redes maiores, como é o caso da teia mundial (WWW). Dessa forma, mesmo que as amostras sejam grandes, as suas propriedades são diretamente afetadas pela maneira como elas são obtidas e podem não corresponder com as de suas respectivas redes originais. Por exemplo, a amostragem mais utilizada para captura de roteadores da Internet, se empregada em redes aleatórias, tende a obter redes livres de escala como resultado. Em contrapartida, amostras de redes livres de escala não têm garantia de preservar essa estrutura. Por causa desses e outros problemas que possam ocorrer na amostragem das redes, é muito importante avaliar a variação das propriedades das redes a ruídos (para saber quais variam menos, sendo, portanto, mais adequadas para caracterizar redes com problemas de amostragem) e os efeitos da amostragem na caracterização, classificação e análise de redes complexas (pois redes amostradas podem não corresponder ao sistemas dos quais foram obtidas, tornando os resultados incorretos). Neste trabalho, foi investigada a influência de três tipos de perturbação (ruído): adição, remoção e troca aleatória de conexões nas propriedades de redes complexas, e as mais apropriadas para caracterizar redes amostradas foram identificadas. Além disso, foram definidas duas novas estruturas em redes complexas: árvores de borda e cadeias de vértices. A ocorrência dessas estruturas em redes mal amostradas tende a ser alta, indicando que existe uma relação com redes parcialmente amostradas. Para verificar tal hipótese, foi investigada a presença de cadeias de vértices em redes gradativamente amostradas por caminhadas aleatórias.

Palavras-chave: redes complexas, grafos, incompleteza, amostragem, estrutura topológica, medidas.

Abstract

VILLAS BOAS, P. R. **Sampling effect on the topological properties of complex networks**. 2008. 114pp. Thesis (PhD) – Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2008.

Several natural or human made complex systems can be represented by complex networks – a theory which integrates the study of graphs with statistical mechanics. This kind of representation, however, can be biased by the way in which the data is obtained. In general, the data used to represent such systems is not always accurate, as in the case of the World Wide Web (WWW). Therefore, even if the sampled networks are large, their properties are directly affected by the way in which they were obtained and may not correspond to those of their respective original networks. For instance, the most used sampling methodology for capturing routers of the Internet, if performed on random networks, tends to obtain scale-free networks as results. On the other hand, sampled scale-free networks are not guaranteed to have this property. Because of these and other problems which may occur during the network sampling, it is very important to evaluate the variation of the network properties with respect to noise (in order to know which of them have less variation, being therefore more suitable for the characterization of networks with sampling problems) and the effect of sampling in the characterization, classification, and analysis of complex networks. In this work, we investigated the effect of three types of perturbations (noise), namely, edge addition, removal, and rewiring on the respectively estimated complex network properties, and the most suitable properties to characterize sampled networks were identified. Furthermore, two novel structures in complex networks were defined, namely, border trees and chains of vertices, which are possibly related to sampling. The occurrence of these structures in poorly-sampled networks was found to be high, implying a relation with partially sampled networks. In order to investigate such a hypothesis, the presence of chains of vertices was investigated in networks which were gradually sampled by random walks.

Keywords: complex networks, graph, incompletness, sampling, topological structure, measurements.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

Quaisquer sistemas compostos por diversos elementos que interagem entre si podem ser modelados como *redes complexas* [6, 7, 8, 9, 10]. Tal termo se refere às redes (ou grafos como são chamados na literatura computacional [11]) que possuem características topológicas não triviais como acontece com redes simples. Estas últimas apresentam alta similaridade estrutural, não importando qual parte seja analisada, e são representadas tipicamente por redes aleatórias ou grades regulares. Entretanto a grande maioria das redes naturais ou criadas pelos seres humanos apresenta estruturas mais complexas que não podem ser representadas por meio de redes simples. Exemplos incluem redes biológicas, metabólicas, cadeia alimentar, Internet (Figura 1), teia mundial (WWW), redes de citação, redes sociais e muitas outras [9]. Essas redes são caracterizadas, entre outras, pelas propriedades: mundo pequeno [12] (a distância entre dois elementos da rede é pequena), distribuição de conexões livre de escala [13] (segue uma lei de potência) e estrutura modular [14] (partes da rede mais conectadas internamente do que externamente).

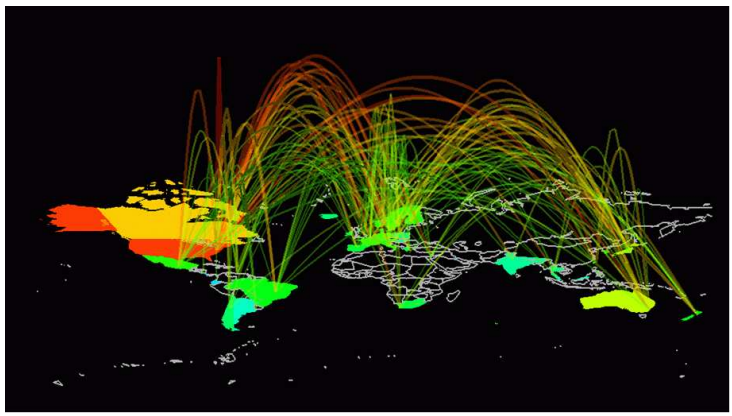


Figura 1: Um exemplo de rede complexa: a Internet. Fonte [1].

O grande sucesso na representação, caracterização e modelagem de sistemas reais por meios

de redes complexas foi obtido graças a dois fatores: aumento do poder computacional (velocidade de processamento e capacidade da memória) dos computadores e disponibilidade de várias bases de dados. Devido ao primeiro fator, redes complexas com milhares ou até mesmo milhões de vértices agora podem ser processadas, permitindo um estudo bastante amplo de vários tipos de rede. Além disso, inúmeros mapeamentos de redes naturais ou construídas pelo homem estão disponíveis desde a década de 90 do século passado, tornando possível a aplicação da teoria de redes complexas nos mais diversos ramos. Entretanto, em sua maioria esses mapeamentos ainda não estão completos, o que dificulta a análise das redes representadas e compromete a confiabilidade das propriedades e resultados obtidos [15]. Por causa disso, amostragem em redes complexas é um dos mais importantes problemas atuais, pois dados incompletos podem resultar em conclusões incorretas sobre os sistemas reais estudados.

No processo de amostragem, a estrutura das conexões das redes estudadas podem diferir substancialmente dos sistemas originais de onde elas foram obtidas. Esse efeito pode resultar em caracterização, classificação e modelagem incorretas dos sistemas reais estudados. Além disso, processos dinâmicos como, por exemplo, propagação de opiniões e doenças [9, 16], caminhadas aleatórias [17] e fluxo de informações [18], por serem altamente dependentes da estrutura da rede são profundamente afetados pelo processo de amostragem empregado.

O problema de amostragem em redes complexas não pode ser ignorado e recentemente tem sido levado em consideração nas diferentes abordagens de construção de redes de informação [19], de proteínas [20], tecnológicas [21] e sociais [22]. As interações de proteínas disponíveis atualmente, por exemplo, cobrem apenas uma pequena parte do mapa completo de interação de proteínas. Além disso, o processo de obtenção das interações de proteínas, como no caso da levedura, gera uma grande quantidade de falsos positivos, isto é, interações observadas no laboratório que nunca ocorrem nas células [23, 24]. Sprinzak e colaboradores [25] sugeriram que a confiabilidade do processo de obtenção das interações das proteínas da levedura é em torno de 50%. Assume-se também, em geral, que mapas incompletos podem ser extrapolados para o mapa completo de interações das proteínas e que suas propriedades topológicas são mantidas, apesar da amostragem pequena [26]. Essa suposição se baseia na propriedade livre de escala presente nas redes de proteínas. Entretanto, não é garantido que amostras de redes livres

de escala também sejam livres de escalas [15]. Além disso, pequenas amostragens podem gerar redes livres de escala independente da topologia da rede original [20, 27]. Para superar essas dificuldades, alguns esforços têm sido empregados para obter dados de interação de proteína mais confiáveis (veja Ref. [28]).

Não é diferente o que acontece com a teia mundial. Obter a rede completa de páginas da teia mundial é praticamente impossível por causa da velocidade com que ela muda (páginas são adicionadas e removidas a todo momento) e por causa do seu tamanho (da ordem de 20 bilhões de páginas). Desta forma, a única maneira de estudar esse tipo de rede é através de amostras de domínios específicos. Para obter essas amostras, são utilizados *web crawlers*, programas especialmente projetados para visitar páginas dentro de um certo domínio e capturar os *links* entre elas [29]. O *web crawler* utilizado e o domínio escolhido, portanto, influenciam diretamente as amostras obtidas [19, 30]. Além disso, alguns *web crawlers* tendem a sobreestimar o número médio de conexões em cada página. Uma maneira de minimizar esse problema é iniciar a procura partindo de várias páginas [30], mesmo assim não resolve completamente.

No caso da Internet, a amostragem é, em geral, realizada pelo comando *traceroute* (procura por rotas) que envia pacotes pela rede a fim de obter os endereços IP dos roteadores no caminho. No entanto, assume-se que esses pacotes seguem o menor caminho entre a fonte e o destino [31], fazendo com que uma grande quantidade de conexões seja perdida, devido à possível presença de *links* redundantes entre os roteadores. Adicionalmente, na estratégia empregada pelo *traceroute*, roteadores próximos à fonte tem maior visibilidade que os mais distantes [21], obtendo uma boa representação da vizinhança da fonte em prejuízo do resto da rede. Também foi observado que esse tipo de amostragem aplicado em redes aleatórias gera redes livres de escala [32].

Redes sociais também apresentam mapeamentos incompletos. Em geral, esse tipo de rede é restrita a apenas classes especiais de atividade humana (por exemplo: música, esportes, participação em filmes e colaboração em ciência) ou geradas a partir de relações humanas (por exemplo: amizade, relações sexuais e contatos profissionais). A maneira como esse tipo de rede é obtida geralmente pode gerar dados com os seguintes problemas: especificação de fronteira

(não inclusão de pessoas ou afiliações), inacessibilidade de sujeitos e imprecisão na aplicação de questionários [22]. Além disso, conforme o tipo de relação social considerada, torna-se bastante difícil definir as conexões entre os indivíduos. De certa forma, estimar os efeitos da falta de dados na redes sociais pode ser uma tarefa muito difícil.

1.1 Objetivos

Apesar da recente preocupação com os problemas de incompleteza e amostragem no mapeamento de redes naturais ou artificias, ainda existem questões as serem resolvidas, tais como determinar se a rede estudada apresenta problemas de amostragem, aplicar métodos para classificar redes amostradas e encontrar propriedades menos sensíveis a ruídos. Deste modo, o maior objetivo deste trabalho foi prover uma análise do efeito da amostragem na estrutura topológica de redes complexas, incluindo:

- encontrar estruturas topológicas relacionadas ao problema de amostragem nas redes complexas;
- analisar a ocorrência de tais estruturas em várias redes do mundo real;
- investigar a estrutura de redes amostradas por caminhada aleatória a partir de redes geradas por modelos teóricos;
- investigar o efeito da amostragem na classificação de redes complexas;
- analisar a sensibilidade e robustez das medidas topológicas de redes diante de ruídos (perturbações).

1.2 Contribuições

As principais contribuições deste trabalho são:

- **Caracterização de redes complexas.** Neste trabalho, foram realizados um levantamento da maioria das medidas de redes complexas e uma análise de como elas podem ser usadas

para classificação de redes reais em modelos teóricos tradicionais através de estatística multivariada e teoria de decisão Bayesiana [2] (ver capítulo 4). Além disso, foram introduzidos os conceitos de duas estruturas em redes complexas: cadeias de vértices [4] e árvores de borda [3].

- **Aplicações de redes complexas.** Embora não diretamente relacionado com este trabalho, também foi realizada, em colaboração com outros pesquisadores, uma revisão ampla do uso da teoria de redes complexas em diversas áreas [10]. Além de descrever as aplicações de cada área, as dificuldades e os principais resultados em redes complexas são apresentados.

- **Análise de redes amostradas.** Foi realizado uma análise da ocorrência de cadeias de vértices em redes amostradas que foram obtidas por caminhadas aleatórias em redes de modelos teóricos [4]. Os resultados indicaram que quanto menor o tamanho da amostra de rede (i.e. caminhada aleatória pequena), mais freqüente são as cadeias longas de vértices. Portanto, existe uma relação entre o tamanho da amostra com a ocorrência dessas estruturas.

- **Sensibilidade de medidas.** Também foi realizado neste trabalho um estudo sobre a sensibilidade e robustez de várias medidas de redes complexas diante de perturbações nas conexões dos vértices de redes introduzidos pela adição, remoção e troca aleatória de conexões [5].

1.3 Descrição dos capítulos

Esta tese inicia no Capítulo 2 com os conceitos básicos de redes complexas, uma breve história sobre o assunto e a descrição das bases de dados e dos modelos teóricos utilizados nas análises dos capítulos subseqüentes. Logo em seguida, nos Capítulos 3 e 4, são apresentadas, respectivamente, as principais medidas de redes complexas com ênfase àquelas que são utilizadas ao longo desta tese e a descrição dos métodos de estatística multivariada para classificar redes complexas. No final do Capítulo 4, são mostrados ainda os resultados de classificação de

duas redes reais e o efeito da amostragem na identificação de uma rede de *emails*. As estruturas: árvores de borda e cadeias de vértices são apresentadas em seguida, no Capítulo 5, assim como a análise realizada em várias redes do mundo real. Já o Capítulo 6 traz a análise da sensibilidade de diferentes medidas de redes complexas diante de três tipos de perturbações (ruídos): adição, remoção e troca aleatória das conexões em quatro modelos teóricos e cinco redes reais. Finalmente, o Capítulo 7 fecha esta tese com as principais conclusões deste trabalho e algumas perspectivas de trabalhos futuros.

2 Redes complexas: conceitos básicos, bases de dados e modelos

O que a teia mundial, interação de proteínas, ecossistemas, redes de amizade e propagação de doenças têm em comum? Até início de 1990, a resposta seria: muito pouco. Entretanto, apesar de origem distintas, sabe-se que estas redes e muitas outras possuem arquiteturas muito semelhantes devido às recentes descobertas na área de redes complexas [7, 9]. O estudo dessas arquiteturas assim como a sua evolução e os resultados de processos dinâmicos que ocorram nelas deu origem à teoria de redes complexas que tem sido largamente aplicada na caracterização de sistemas complexos naturais ou artificiais. Um sistema complexo é um sistema composto por uma quantidade grande de partes simples ou complexas interconectadas cujas interações entre si e com o meio externo levam a resultados complexos [33]. Um exemplo típico é a revoada de gansos, que apesar de serem simples como unidades, interagem entre si de forma complexa e também com o meio ambiente, pois eles se readaptam constantemente conforme mudanças no meio em que vivem (é por isso que eles migram). Além disso, o processo de migração também é complexo, pois eles se auto-reorganizam durante o vôo sem a necessidade de um líder.

Assim como as redes complexas, sistemas complexos apresentam propriedades particulares importantes, tais como: *emergência* – a complexidade do todo é maior do que a complexidade da soma das partes, *readaptação* – o sistema se auto-organiza conforme mudanças no meio e *universalidade* – sistemas pertencentes à mesma classe compartilham propriedades semelhantes [33]. Por serem formados por partes discretas que interagem entre si, sistemas complexos são modelados naturalmente em termos de redes complexas [7, 9, 33]. Por exemplo, a sociedade é um sistema complexo composto por pessoas ligadas por laços de amizade [34, 35]; a ciência, por cientistas conectados por colaboração em artigos científicos [36, 37]; a Internet, por

roteadores conectados por fibras óticas [38, 39]; a teia mundial, por páginas interligadas através de *hiperlinks* [29, 40, 41]; o cérebro, por neurônios conectados através de axônios e dendritos [12, 42, 43, 44]; as cadeias alimentares, por animais e plantas relacionados por relações de predação [45, 46]; o transporte aéreo, por aeroportos ligados por linhas aéreas [47, 48]; e a linguagem, por palavras ligadas por similaridade ou função [49, 50]. Além destas, existem inúmeras outras aplicações de redes complexas e uma revisão de muitas delas pode ser encontrada em [10]. Nas próximas seções, além de uma breve descrição da história de redes complexas, serão apresentados a definição formal de redes, maneiras de representá-las, algumas bases de dados de redes reais e alguns modelos teóricos que foram utilizados nas análises dos capítulos seguintes.

2.1 Introdução histórica

O estudo de redes complexas nasceu com a teoria dos grafos e os primeiros trabalhos dessa área datam de 1736, quando Leonhard Euler resolveu o problema das Sete Pontes de Königsberg, capital leste da Prússia (atual Kaliningrado, Rússia). A cidade foi construída em torno do Rio Pregel onde se junta com outro rio e, no meio, há uma ilha chamada Kneiphoff. Havia no total 7 pontes ligando as diferentes partes da cidade, conforme ilustra a Figura 2(a). Naquela época, os habitantes dessa cidade se perguntavam se era possível passar todas as pontes sem repetir uma sequer. Entretanto, Euler mostrou que tal feito não era possível. Para resolver tal problema ele representou a configuração das pontes como um grafo (Figura 2(b)), onde cada vértice correspondia a uma porção de terra e cada aresta, a uma ponte, criando provavelmente o primeiro grafo de que se tem notícia. Desde então, os avanços na teoria de grafos se restringiram a apenas grafos estáticos e com poucas aplicações práticas [51].

Se foi Euler o responsável pela criação da teoria dos grafos, Erdős foi quem, em grande parte, a "difundiu". Em 1959, Erdős e Rényi começaram a estudar grafos como objetos estocásticos ao invés de determinísticos [52]. O estudo dos grafos desta maneira conduziu Erdős e Rényi à introdução de grafos aleatórios (um conjunto de vértices conectados aleatoriamente) [52, 53, 54]. Além da definição desse tipo de rede, Erdős e Rényi analisaram diversas propriedades

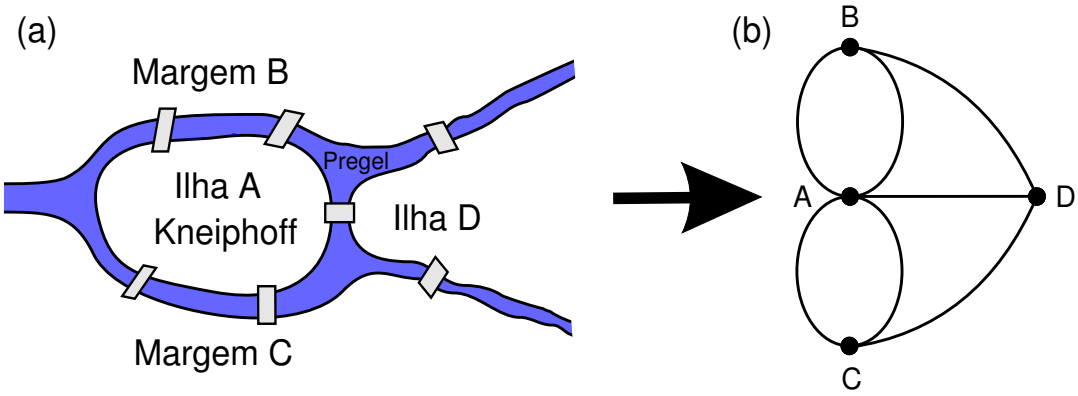


Figura 2: O problema das pontes de Königsberg. (a) Representação esquemática das pontes de Königsberg antes de 1875, com a ilha de Kneiphoff (A), a ilha (D) entre os dois braços do rio Pregel e as duas porções de terra (margens B e C) que circundam a ilha A. Euler representou essa configuração como um grafo (b) e provou que não há possibilidade de passar uma única vez por todas as pontes.

desses grafos e obtiveram importantes resultados, tais como a identificação do limite de percolação, isto é, o número de conexões necessárias para que o grafo correspondente esteja conectado (i.e. não existam partes isoladas da rede).

Ainda na década de 50, os sociólogos ao estudar redes de interações sociais [34] (por exemplo, amizade) utilizaram grafos para representá-las e criaram várias medidas que, mais tarde, se tornaram referências em redes complexas, como, por exemplo: grau de intermediação de um vértice e coeficiente de aglomeração (Seções 3.2 e 3.3, respectivamente). Também nessa mesma época, alguns pesquisadores começaram a usar os grafos como meio de simulação de processos dinâmicos, como a propagação de doenças e formação de opiniões [51].

Mais tarde, em 1967, Stanley Milgram, um pesquisador de sociologia em Harvad, Estados Unidos, investigou a hipótese do *mundo pequeno*, proposta pelo escritor húngaro Frigyes Karinthy [55]. Segundo tal escritor, entre duas pessoas quaisquer no mundo, há poucas pessoas os separando. Para testar tal hipótese, Milgram enviou centenas de cartas a várias pessoas residentes em Wichita, Kansas, e Omaha, Nebrasca, pedindo-lhes que as enviassem a conhecidos que pudessem submetê-las a um corretor de fundos públicos em Boston ou a mulher de um estudante de graduação em Sharon, Massachusetts. Milgran descobriu que as cartas que eventualmente chegaram ao seu destino passaram por seis pessoas em média, dando origem ao tão conhecido termo de “seis graus de separação” [55]. Hoje, acredita-se que esse número seja

menor [56]. Mais recentemente, em 1998, os pesquisadores Watts e Strogatz descobriram que o fenômeno de mundo pequeno também é observado em outras redes reais, como é o caso da rede de neurônio do *Caernohabditis elegans* e da rede de transmissão de energia elétrica dos Estados Unidos [12]. Além dessa propriedade, eles também constataram que essas redes possuem uma quantidade grande de ciclos (caminhos fechados) de ordem 3 em relação às redes aleatórias [12].

A descoberta de Watts e Strogatz serviu de motivação para outras investigações quanto à estruturas das redes complexas. Em 1999, Barabási e Albert [13] descobriram que a distribuição de grau dos vértices da teia mundial (grau de um vértice i, representado por k_i é seu o número de conexões) não é aleatória, mas segue uma lei de potência, ou seja, a distribuição de grau é livre de escala (*scale-free*) e é da forma $P(k) \sim k^{-\lambda}$. Além dessa rede, muitas outras seguem essa lei e, em geral, λ está entre 2 e 3 [7].

Em 2002, Girvan e Newman mostraram que a maioria das redes reais possuem estrutura modular, isto é, elas são compostas por grupos de vértices que possuem mais conexões entre si do que com o resto da rede [14]. Tais grupos ficaram conhecidos como *comunidades*. Bons exemplos de redes com essa estrutura são as redes sociais, uma vez que comunidades nessas redes representam grupos de pessoas que compartilham as mesmas opiniões e/ou interesses.

A partir da descoberta destas propriedades, a teoria de redes complexas se estabeleceu rapidamente. A rapidez com que ela se desenvolveu foi, em grande parte, devido ao advento da Internet com a disponibilização de diversas bases de dados de sistemas complexos naturais e artificiais e ao aumento do poder computacional dos computadores, que permitiu a análise de redes com mais de centenas de milhares de vértices (e.g. a teia mundial). Antes, o estudo de grafos, como acontecia com as redes sociais que possuíam apenas centenas de vértices (e.g. as redes obtidas por Milgram), era realizado através de inspeção visual dos vértices com as respectivas conexões. Entretanto, quando redes maiores puderam ser armazenadas e processadas, não foi possível obter padrões apenas através da visualização dos grafos correspondentes, sendo necessário a definição de medidas topológicas para melhor representar tais redes, como pode ser constatado no próximo capítulo.

2.2 Representação

Como já mencionado anteriormente, os sistemas complexos são naturalmente representados em termos de redes complexas. Uma rede pode ser representada por um conjunto de nós ou vértices – as pequenas partes que a compõem, e um conjunto de arcos ou arestas, que representam as interações (conexões) entre os vértices (e.g. [7]). Se as interações não tiverem direção são denominadas arestas, e a rede é *não-dirigida*, ou grafo como é conhecido na computação, mas se tiverem, como acontece com vôos de uma cidade para outra numa rede de aeroportos, são denominadas arcos e a rede é *dirigida*, também conhecida como dígrafo. Se a intensidade das interações for importante, por exemplo a quantidade de passageiros nos vôos, a rede é tida como rede *com peso*, caso contrário, *sem peso*. As redes ainda podem ter auto-arcos ou arestas (auto-conexões entre os vértices) e multi-arcos ou arestas (mais de um arco ou aresta entre os mesmos pares de vértices). Neste caso, a representação adequada é o multigrafo [11], mas geralmente não é utilizada para representar redes complexas. Se forem considerados apenas a direção e o peso, as redes podem ser divididas em quatro tipos: rede dirigida com peso, o caso mais geral; rede dirigida sem peso; rede não-dirigida com peso; e rede não-dirigida sem peso, o caso mais simples, conforme ilustrado na Figura 3. Dependendo do tipo de estudo realizado, redes dirigidas podem ser transformadas em não-dirigidas através da operação de *simetrização* e as redes com peso nas suas correspondentes versões sem peso através da operação de *limiarização* (e.g.[2]). Essas operações assim como a definição formal de redes é apresentada a seguir.

O caso mais geral, rede dirigida e com peso, é definida como $\mathcal{G} = \{V, E, \omega\}$, onde V é um conjunto de N vértices, E, um conjunto de M arcos e ω, um mapa de pesos tal que $\omega : E(\mathcal{G}) \mapsto \mathbb{R}$. Se existir uma conexão do vértice i para o vértice j, $\omega(i, j) \neq 0$ e o arco correspondente é representado pelo par (i, j). Neste trabalho, nem auto-conexões nem múltiplos arcos são levados em conta.

Numa rede sem peso, os arcos não possuem pesos associados e, portanto, o mapa ω não é necessário. Uma rede não-dirigida é aquela em que os arcos não possuem direção.

Outra maneira de representar redes é através de matrizes. Para o tipo mais geral de redes,

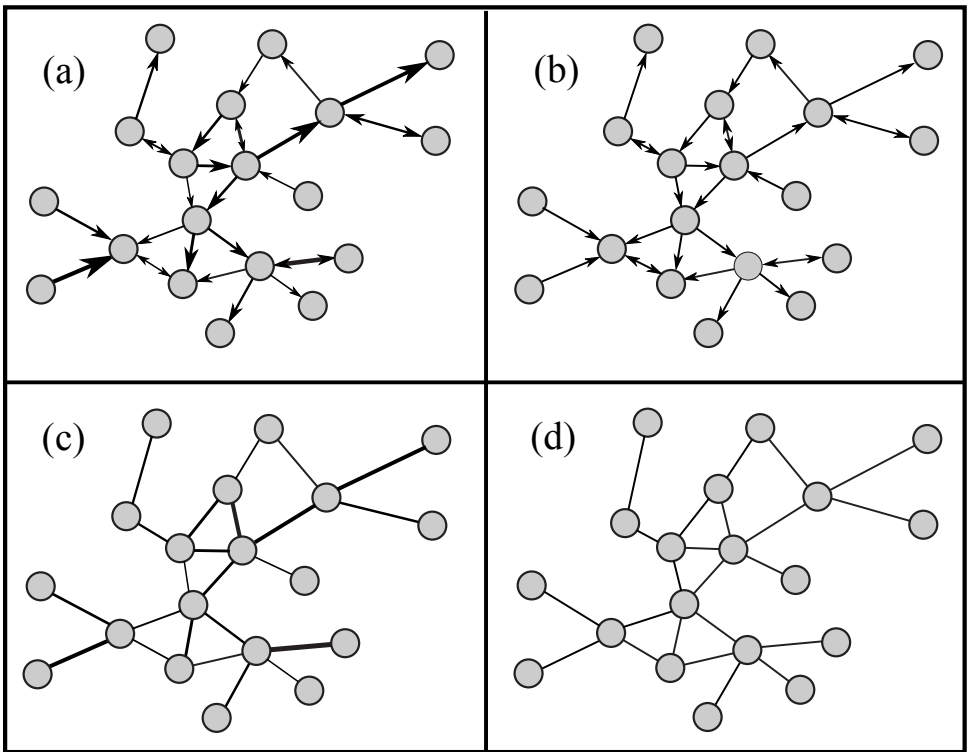


Figura 3: Os quatro tipos de rede complexa considerados: (a) dirigida e com peso, (b) dirigida e sem peso, (c) não-dirigida e com peso e (d) não-dirigida e sem peso. A rede (b) foi obtida através da operação de limiarização (ver texto) realizada na rede (a); a rede (c) é a versão simetrizada (ver texto) de (a); e a rede (d) é a versão simetrizada e sem peso de (a). O peso de cada arco é proporcional à sua espessura e o limiar utilizado na limiarização da rede (a) é o menor peso encontrado nessa rede.

uma matriz de peso W é utilizada, onde cada elemento $w_{ij} = \omega(i, j)$ expressa o peso da conexão do vértice i para o vértice j. Note que para redes sem direção $w_{ij} = w_{ji}$. A operação de *limiarização* pode ser aplicada nessas redes para obter a sua versão sem peso. Tal operação, representada por $\delta_T(W)$, é aplicada sobre todos os elementos w_{ij} resultando uma matriz binária A tal que $A = \delta_T(W)$. Os elementos a_{ij} da matriz A são tais que: $a_{ij} = 0$ se $w_{ij} < T$ ou $a_{ij} = 1$ caso contrário. A matriz resultante dessa operação é largamente conhecida na literatura como *matriz de adjacência* e é utilizada para representar redes dirigidas sem peso (e.g. [7]).

Para transformar redes dirigidas com ou sem peso nas correspondentes versões não-dirigidas, a operação de *simetrização* deve ser utilizada. Essa operação, representada por $\sigma(W)$, é dada por: $\sigma(W) = W + W^T$, onde W^T é a transposta da matriz W.

Numa rede não-dirigida, dois vértices são adjacentes ou vizinhos se existir um arco (i, j) tal que $\omega(i, j) \neq 0$. No caso de redes dirigidas, o conceito correspondente é o de sucessor e o

de predecessor. No arco (i, j) com $\omega(i, j) \neq 0$, o vértice i é predecessor de j e j é o sucessor de i. A partir do conceito de adjacência, define-se vizinhança $\mathcal{V}_i$ do vértice i como o conjunto de vértices adjacentes a i. Se a rede for dirigida, $\mathcal{V}_i = \mathcal{V}_i^{\text{in}} \cup \mathcal{V}_i^{\text{out}}$, onde $\mathcal{V}_i^{\text{in}}$ e $\mathcal{V}_i^{\text{out}}$ é o conjunto de predecessores e o de sucessores, respectivamente.

Apenas para uso na descrição das bases de dados e dos modelos teóricos deste capítulo, as seguintes propriedades básicas de redes complexas não-dirigidas e sem peso são definidas: grau médio, distribuição de grau $P(k)$, coeficiente de agrupamento médio e caminho mínimo médio. A primeira está relacionada com a densidade de conexões da rede (grau de um vértice corresponde ao seu número de vizinhos); a segunda, com probabilidade de encontrar um vértice com grau k; a terceira, com a coesão local entre os vizinhos de cada vértice da rede; e a quarta, com a distância média entre quaisquer vértices da rede. Maiores detalhes dessas medidas e propriedades podem ser encontrados Capítulo 3.

2.3 Bases de dados consideradas

O grande sucesso da teoria de redes complexas é devido à sua capacidade de representar vários sistemas discretos físicos ou abstratos do mundo real. De fato, há muitos exemplos na natureza, mas foi apenas recentemente que bases de dados dos mais variados tipos se tornaram disponíveis. Como existem inúmeras bases de dados de redes complexas atualmente, aqui serão cobertas apenas as utilizadas neste trabalho e serão subdividas em quatro classes: redes sociais, de informação, tecnológicas e biológicas, conforme divisão realizada em [7].

2.3.1 Redes sociais

Redes sociais são aquelas formadas por pessoas ou grupo de pessoas (e.g. empresas, times ou classes econômicas) conectadas por algum tipo de interação, como por exemplo: amizade, relação comercial entre companhias, colaborações científicas e participação em filmes [7]. Os tipos de redes sociais considerados nas análises dos capítulos seguintes são: redes de colaboração entre pesquisadores e redes de *email*.

As **redes de colaboração científica** são formadas por pesquisadores que estão conectados

se forem autores de um mesmo artigo. Nos capítulos subseqüentes foram consideradas apenas as redes de colaboração de astrofísica entre 1995 e 1999 [37], da matéria condensada entre 1995 a 2005 [37], física de alta energia entre 1995 a 1999 [57, 58] e de pesquisadores de redes complexas, denominada *netscience*. As três primeiras foram obtidas através da página `http://www.arxiv.org` e a última foi compilada de dois artigos de revisão sobre redes [7] e [9]. Tais redes estão disponíveis em [59].

As **redes de *email*** podem ser obtidas de duas formas: através de arquivos de *log* de servidores de *email*, ou por listas de *emails* de vários usuários de uma instituição específica [10]. No primeiro caso, os vértices são endereços de *email* e há um arco dirigido do vértice i para o vértice j se i enviou, pelo menos um *email* para j. Já, no segundo caso, os vértices também são endereços, mas há um arco de i para j se j está na lista de contatos de i. Alguns autores, porém, preferem analisar a versão não-dirigidas dessas redes, considerando apenas conexões entre endereços de emails que trocaram, pelo menos, um *email* entre si [60]. A rede de *email* considerada nos próximos capítulos foi obtida através do arquivo de *log* do servidor de *emails* da Universidade Rovira i Virgili (Tarragona) [61].

2.3.2 Redes de informação

Redes de informação, em alguns casos também conhecidas como "redes de conhecimento", são aquelas cujos vértices guardam algum tipo de informação [7]. Um exemplo típico deste tipo de rede são as de citações de artigos acadêmicos que refletem a estrutura de informação armazenada nos seus vértices [7]. As redes de informação analisadas nos capítulos subseqüentes são:

Rede do dicionário de Roget é construída associando cada vértice da rede a uma das 1022 categorias da edição de 1879 do dicionário de palavras e expressões inglesas de Peter Mark Roget, editado por John Lewis Roget [62]. Duas categorias i e j estão ligadas se Peter Roget fez uma referência para j entre todas as palavras e expressões de i, ou se uma das categorias está diretamente relacionada à outra no livro de Roget [62]. Tal rede está disponível nas bases de dados do Pajek [63].

Wordnet é uma rede semântica de palavras que geralmente é usada como uma forma de representação do conhecimento. Ela é um grafo direcionado cujos vértices representam conceitos e os arcos, as relações semânticas entre os conceitos. Esta rede também foi obtida através da base de dados do Pajek [63].

Teia mundial (WWW) é uma rede dirigida de páginas da teia mundial pertencentes ao domínio `nd.edu` que estão ligadas por meio de ponteiros (*hyperlinks*). Esta base de dados pode ser obtida através do *Center for Complex Network Research* do grupo de Barabási [64]

Rede de adjacência em livros. As adjacências de palavras em livros podem ser representadas como uma rede de palavras conectadas pela proximidade no texto [65]. Há um arco direcionado da palavra A para a palavra B se elas forem adjacentes e B for subsequente a A. O peso do arco correspondente é o número de vezes em que essas palavras aparecem no texto na mesma seqüência. Antes de construir uma rede desse tipo, o texto precisa ser pré-processado, isto é, todas as palavras freqüentes (e.g. artigos, preposições, conjunções, etc) são removidas e as restantes são colocadas na sua forma singular (isto é, palavras femininas são transformadas em masculinas e plural em singular) [65]. Nas análises seguintes, os livros utilizados foram: "*David Copperfield*" de Charles Dickens, "*Night and Day*" de Virginia Woolf, e "*On the Origin of Species*" de Charles Darwin. As representações de tais livros como redes complexas foram cedidas por Antiqueira [66].

2.3.3 Redes tecnológicas

Nesta classe de redes, estão as redes construídas pelos seres humanos, geralmente, destinadas ao transporte de produtos ou recursos, como por exemplo: eletricidade ou *bytes*. Existem muitos exemplos deste tipo de rede e a rede de transmissão de energia elétrica e a Internet são apenas alguns deles. As redes tecnológicas consideradas foram:

Internet no nível de sistemas autônomos. Um sistema autônomo é uma coleção de redes IP ou roteadores sob controle de uma entidade que apresenta uma política de roteamento comum para a Internet. Cada sistema autônomo representa um grande domínio de endereços IP que

geralmente pertencem a uma organização, como por exemplo, universidades, empresas ou provedores de Internet. Neste tipo de rede, cada vértice representa um sistema autônomo e dois vértices estão conectados se há, pelo menos, uma conexão física entre eles. Na análises dos próximos capítulos, Internet será sinônimo de sistemas autônomos, quando utilizada como base de dados. A rede considerada nos capítulos seguintes foi coletada por Newman em julho de 2006 [59].

Rede de transportes aéreos dos Estados Unidos é formada por aeroportos americanos de 1997 conectados por linhas aéreas. Tal rede pode ser encontrada na base de dados do Pajek [63].

Rede de transmissão de elétrica dos estados ocidentais dos Estados Unidos é composta por geradores, transformadores e subestações cujas conexões são as linhas de transmissão de alta-voltagem. Esta rede foi pioneiramente estudada por Watts e Strogatz em [12].

2.3.4 Redes biológicas

Existe um número muito grande de sistemas biológicos que podem ser modelados como redes complexas. Exemplos incluem: redes neurais, redes metabólicas, redes de interação de proteínas, cadeia alimentares, entre outros [9]. As redes biológicas utilizadas nos próximos capítulos são:

Rede neural de *Caenorhabditis elegans* é formada por neurônios de *Caenorhabditis elegans* conectados por meios de sinapses [67, 12].

Rede de transcrição genética de *Escherichia coli* é composta por *operons* (um *operon* é um grupo de genes contínuos que são transcritos em uma única molécula de mRNA). Os arcos são direcionados de um *operon*, aquele que codifica o fator de transcrição para outro o qual é regulado por esse fato de transcrição. Este tipo de rede é muito importante devido ao controle da expressão gênica [68].

Rede de interação proteína-proteína de *Saccharomyces cerevisiae* possui como vértices proteínas conectadas de acordo com interações físicas dirigidas [26]. Esta rede está disponível em [64]. Também foi utilizada outra base de dados para esta rede, que foi analisada em [25].

2.4 Modelos teóricos

As propriedades descritas na Seção 2.1 motivaram o desenvolvimento de modelos teóricos que fossem capazes de reproduzi-las. Alguns deles foram objetos de grande interesse, como grafos aleatórios, modelos de mundo pequeno, grafos aleatórios generalizados e modelos de redes livres de escala [7, 9]. Além destes, outros foram propostos com o intuito de estudar a topologia de redes com propriedades específicas, como por exemplo, as redes geográficas [9]. As próximas subseções apresentarão alguns modelos simples utilizados nos capítulos subseqüentes. No final desta seção, os resultados analíticos para algumas medidas de alguns modelos teóricos descritos aqui serão mostrados na Tabela 1.

2.4.1 Modelo aleatório de Erdős e Rényi

Este pode ser considerado o modelo de rede mais simples e foi desenvolvido por Rapoport [69, 70, 71] e independentemente por Erdős e Rényi [52, 53, 54]. Em 1959, Erdős e Rényi [52] introduziram o modelo aleatório consistindo de N vértices e M conexões. Começando com uma rede totalmente desconectada, tal modelo é construído realizando M conexões entre pares de vértices escolhidos aleatoriamente de forma a não gerar auto-conexões e múltiplas arestas. Outra maneira de construir tal modelo é também iniciar com N vértices totalmente desconectados e conectar cada par de vértices com probabilidade p. Quando N é grande e a conectividade média, dada como $\langle k\rangle = 2M/N$, é mantida constante, a distribuição de grau dos vértices tende a de Poisson, $P(k) = e^{-\langle k\rangle}\langle k\rangle^k/k!$. Além disso, o caminho mínimo médio desse modelo é pequeno, da ordem de $\ln N/\ln\langle k\rangle$.

A preocupação inicial de Erdős e Rényi era apenas em obter propriedades matemáticas dos grafos aleatórios e não em aplicar tal modelo para estudar redes do mundo real. Eles apenas mencionam no artigo [52] que a evolução de grafos poderia ser utilizada como modelos simples de certas redes, como a de estradas e a de ferrovias.

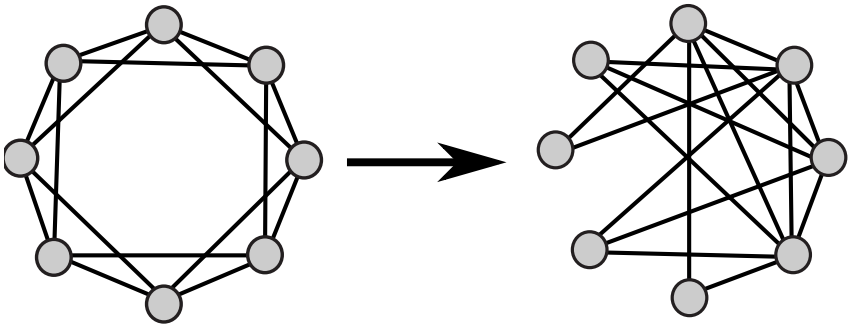


Figura 4: Método de construção da rede de pequeno-mundo de Watts e Strogatz: a partir de uma rede regular (a), cada conexão é reconectada aleatoriamente com uma certa probabilidade q, obtendo a rede (b). Se $q = 0$, a rede é totalmente regular, enquanto que se $q = 1$ a rede é aleatória. Watts e Strogatz mostraram que com valores intermediários de q redes são geradas com baixo caminho mínimo médio e alto coeficiente de aglomeração.

2.4.2 Redes de mundo pequeno de Watts e Strogatz

No final do século passado, os pesquisadores Watts e Strogatz ao estudar a rede de transmissão elétrica dos estados ocidentais dos Estados Unidos e a rede de neurônios do *Caenorhabditis elegans* [12], constataram que nessas redes o caminho mínimo médio é baixo e que há uma grande quantidade de ciclos de ordem 3 em relação às redes aleatórias. Para reproduzir essas descobertas, Watts e Strogatz propuseram um modelo que é um misto de redes regulares (por apresentarem uma grande quantidade de ciclos de ordem 3) com redes aleatórias (baixo caminho mínimo médio).

Segundo Watts e Strogatz [12], este tipo de rede é construído a partir de uma rede circular e regular de N vértices, cada um conectado com κ vizinhos imediatos de cada lado, totalizando 2κ conexões, como na Figura 4(a). Em seguida, cada conexão é refeita com probabilidade q, proporcionando caráter aleatório à rede, Figura 4(b). Se $q = 0$, a rede gerada é regular, ou seja, o coeficiente de aglomeração e o menor caminho médio são grandes. Em contrapartida, se $q = 1$, a rede é completamente aleatória com as duas medidas baixas. Em um estado intermediário, Watts e Strogatz [12] mostraram que o coeficiente de aglomeração médio é alto e o caminho mínimo médio é baixo.

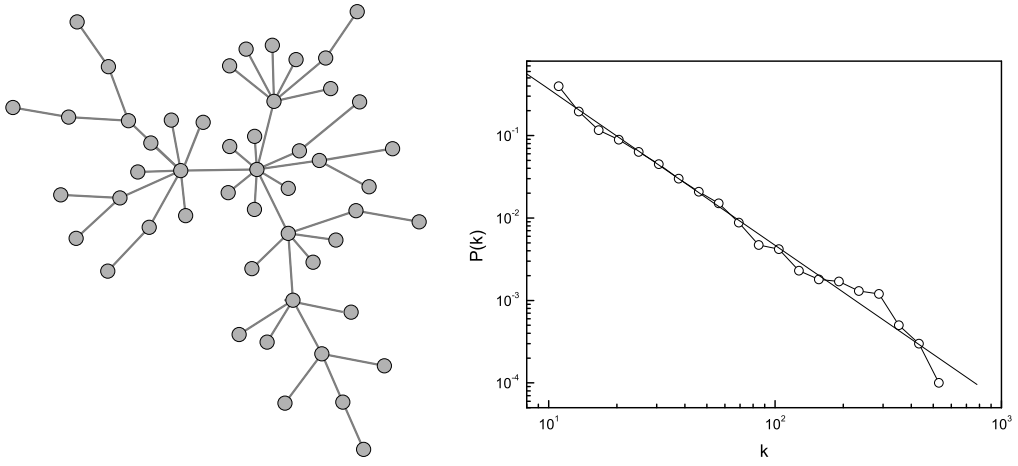
2.4.3 Redes livres de escala de Barabási e Albert

Apesar do modelo de mundo pequeno de Watts e Strogatz resolver o problema sobre formação de triângulos em redes, a sua distribuição de grau também segue a lei de Poisson, como acontece com as redes aleatórias. Entretanto, em 1999, Barabási e Albert [13] descobriram que as conexões da teia mundial não são homogêneas, mas seguem uma distribuição de graus na forma de lei de potência[1] ($P(k) \sim k^{-\lambda}$), isto é, poucos vértices altamente conectados e muitos com poucas conexões. No mesmo ano, os irmãos Faloutsos [38] também verificaram esse comportamento na Internet. Mais tarde, descobriu-se que leis de potência são comuns em muitas outras redes do mundo real, tais como: redes de colaboração científica, redes metabólicas e redes de interação de proteínas [56].

Para criar uma rede com tal propriedade, Barabási e Albert [13] propuseram um modelo com crescimento baseado em ligação preferencial. Neste modelo, a rede começa totalmente conectada com n_0 vértices e a cada passo um vértice i é adicionado com $m \leq n_0$ conexões. Para cada uma dessas conexões, a probabilidade de conexão do vértice i com qualquer vértice j da rede é dada por: $P_{i \to j} = k_j / \sum_u k_u$. Após n vértices adicionados, a rede terá $N = n_0 + n$ vértices e $M = mn + n_0(n_0 - 1)/2$ arestas, o que corresponde a um grau médio $\langle k \rangle \simeq 2m$, se $n \gg n_0$. Na Figura 5(a) é mostrado um exemplo de rede gerada a partir desse modelo e na Figura 5(b), a distribuição de graus para uma rede com 10000 vértices.

Nas redes geradas pelo modelo livre de escala de Barabási e Albert, os vértices mais conectados tendem a receber mais conexões (paradigma conhecido como "rico fica mais rico"). Em redes com este tipo de estrutura, há poucos vértices com muitas conexões, denominados *hubs*, e muitos com poucas conexões. Este comportamento define uma distribuição livre de escala.

[1] Leis de potência, também conhecidas como Lei de Zipf ou Distribuição de Pareto [72], são comuns na natureza e refletem invariância de escalas (pois não possuem nem média nem desvio característicos). Exemplos de tais distribuições incluem distribuição de riquezas, intensidade de terremotos, de tamanho de cidades, freqüência de palavras em livros e muitas outras [72].

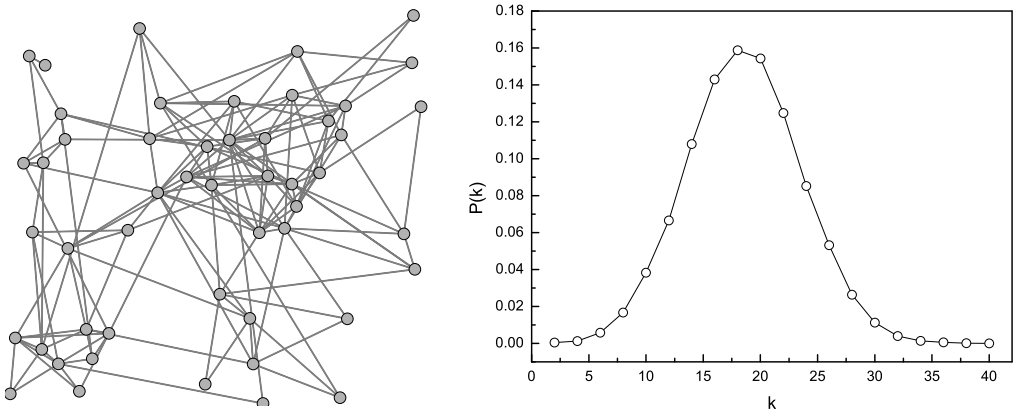


(a) **(b)**

Figura 5: O modelo livre de escala de Barabási e Albert: (a) exemplo com 50 vértices e $m = 1$ (veja que há poucos vértices com muitas conexões e muitos com poucas), e (b) a distribuição de graus média de 10 redes geradas a partir deste modelo com 10.000 vértices e $m = 10$ (grau médio 20). A distribuição de graus deste tipo de rede é da forma de $P(k) \sim k^{-\lambda}$.

2.4.4 Redes geográficas de Waxman

Em geral, imagina-se que as redes do mundo real estão situadas no espaço abstrato, em que a posição dos vértices não é importante. Entretanto, este não é o caso de alguns tipos de redes como as de rodovias e as de aeroportos, que são conhecidas como redes *espaciais* ou *geográficas* [73]. Em casos com este, a posição dos vértices pode ser fundamental para compreender o crescimento e a evolução da rede analisada e os processos dinâmicos que ocorrem nela. Redes geográficas podem ser construídas distribuindo aleatoriamente N vértices dentro de uma caixa quadrada de lado L e os conectando com probabilidade $p \sim e^{-\lambda D}$, onde D é a distância geográfica entre eles e λ é o parâmetro do modelo que permite o ajuste do grau médio desejado. O comprimento L da caixa é escolhido tal que a densidade de vértices é mantida constante, ou seja, $L = \sqrt{N}$. Este modelo foi proposto, em 1988, por Waxman [74] para explicar a topologia da Internet. Um exemplo de rede gerada por tal modelo é ilustrado na Figura 6(a) e a distribuição de graus para uma rede do mesmo tipo com 10000 vértices, na Figura 6(b).



(a) **(b)**

Figura 6: O modelo geográfico de Waxman: (a) exemplo com 50 vértices e $\lambda = 0,96$ (note que há mais conexões entre vértices mais próximos espacialmente do que entre vértices distantes), e (b) a distribuição de graus de uma rede gerada a partir deste modelo com 10000 vértices e $\lambda = 0,545$ (grau médio 20). A distribuição de graus deste tipo de rede também é da forma de Poisson.

2.4.5 Outros modelos

Há muitos outros modelos de redes complexas. Alguns são aprimoramentos dos modelos descritos nesta seção (e.g. modelo aleatório de configuração [75, 76], que gera modelos aleatórios a partir de uma certa distribuição de grau dos vértices, e modelo livre de escala de Barabási e Albert com ligação preferencial caracterizada por três fatores: idade, capacidade e semelhança dos vértices [77]) e outros foram desenvolvidos para reproduzir redes específicas (e.g. GdTang [78] e Inet [79, 80] para reproduzir a Internet no nível de sistemas autônomos). No entanto, a intenção neste capítulo não foi prover uma revisão detalhada dos modelos de redes complexas, mas sim apresentar aqueles que foram utilizados nos próximos capítulos. Para mais informações sobre modelos teóricos de redes complexas é sugerida a leitura de [9].

Tabela 1: Resultados analíticos para o grau médio $\langle k \rangle$, distribuição de grau $P(k)$, coeficiente de aglomeração médio C e caminho mínimo médio ℓ de três modelos teóricos: aleatório de Erdős e Rényi, de mundo pequeno de Watts e Strogatz e livre de escala de Barabási e Albert. Tabela extraída de [2].

Erdős e Rényi	Watts e Strogatz	Barabási e Albert
$\langle k \rangle = p(N-1)$	$\langle k \rangle = 2\kappa^{\star}$	$\langle k \rangle = 2m$
$P(k) = \frac{e^{-\langle k \rangle}\langle k \rangle^k}{k!}$	$P(k) = \sum_{i=1}^{min(k-\kappa,\kappa)} \begin{pmatrix} \kappa \\ i \end{pmatrix} (1-p)^i p^{\kappa-i} \frac{(p\kappa)^{k-\kappa-i}}{(k-\kappa-i)!} e^{-p\kappa}$	$P(k) \sim k^{-3}$
$C = p$	$C(p) \sim \frac{3(\kappa-1)}{2(2\kappa-1)}(1-p)^3$	$C \sim N^{-0.75}$
$\ell \sim \frac{\ln N}{\ln\langle k \rangle}$	$\ell(N,p) \sim p^{\tau} f(Np^{\tau})^{*}$	$\ell \sim \frac{\log N}{\log(\log N)}$

$\star$ κ é o número de vizinhos de cada vértice na configuração inicial do modelo de pequeno mundo de Watts e Strogatz (como na Figura 4, onde $\kappa = 4$).
$*$ A função $f(u)$ é uma constante se $u \ll 1$ ou $f(u) = \ln(u)/u$ se $u \gg 1$.

3 Medidas de redes complexas

As primeiras redes complexas a serem estudadas eram redes sociais com apenas centenas de vértices e podiam ser analisadas com simples inspeção visual dos grafos correspondentes. No final do século passado, porém, bases de dados de outras redes reais (algumas com centenas de milhares de vértices, como a teia mundial) se tornaram disponíveis, e o estudo de tais redes exigiu outras maneiras de caracterizá-las. A solução encontrada, neste caso, foi mapear as redes em termos de medidas da sua topologia, como ilustra a Figura 7. Se o mapeamento for inversível, ou seja, dado um conjunto de medidas a rede original puder ser obtida, o mapeamento é uma representação [81]. Um exemplo de representação de redes complexas sem multi-arcos é a matriz de peso W.

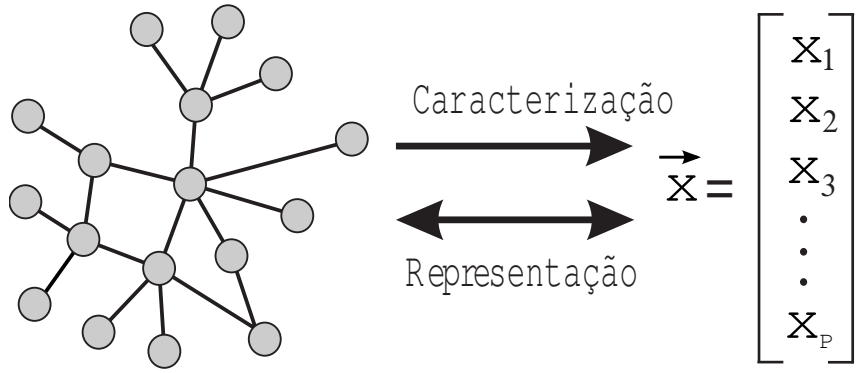


Figura 7: Caracterização de uma rede complexa em termos do correspondente vetor de atributos, $\vec{x} = \{x_1, x_2, \ldots, x_P\}$, onde x_i, $i = 1, 2, \ldots, P$ representam medidas da rede. Se o mapeamento for inversível, ou seja, a partir do vetor de atributos, a rede original puder ser obtida, o mapeamento é uma representação. Um exemplo de mapeamento inversível para uma redes sem peso e sem multi-arcos é a matriz de adjacência A. Figura adaptada de [2].

As medidas de redes complexas também são importantes na discriminação e modelagem de diferentes tipos de redes [2]. No primeiro caso, como cada rede possui características topológi-

cas específicas de conectividade, as medidas podem ser usadas para determinar as classes a que elas pertencem. No segundo caso, para validar os modelos desenvolvidos para determinados tipos de rede, é importante que suas medidas sejam comparadas com as das redes correspondentes. Além disso, as medidas são importantes no estudo das relações entre estrutura e função das redes complexas e na análise da influência da topologia nos processos dinâmicos, tais como por exemplo: fluxo de informações [18].

Nas próximas seções, as medidas de redes complexas serão apresentadas de acordo com o seu tipo: relacionadas ao grau dos vértices, à distância geodésica, a ciclos (coeficiente de aglomeração), às estruturas de comunidade e à hierarquia de vizinhança. Não é intenção neste capítulo cobrir todas as medidas de redes complexas, mas apresentar aquelas que serão utilizadas nos capítulos seguintes. Para uma revisão de medidas de redes complexas, recomenda-se a leitura de [2].

3.1 Grau e medidas correlacionadas

Grau, também conhecido na física como conectividade [8], corresponde ao número de conexões de um dado vértice, isto é, a cardinalidade[1] de sua vizinhança. Para redes não-dirigidas, o grau do vértice i, k_i, é dado pela cardinalidade de $\mathcal{V}_i$ ou, em termos da matriz de adjacência,

$$k_i = \sum_j a_{ij}, \tag{3.1}$$

Se a rede for dirigida, o grau do vértice i é composto de duas partes: grau de entrada e de saída. O primeiro, representado por k_i^{in}, corresponde à quantidade de arcos que chega ao vértice i, isto é, $k_i^{\text{in}} = |\mathcal{V}_i^{\text{in}}|$; e o segundo, por sua vez, é representado por k_i^{out} e indica o número de arcos que sai do vértice i, isto é, $k_i^{\text{out}} = |\mathcal{V}_i^{\text{out}}|$. Em termos da matriz de adjacência,

$$k_i^{\text{out}} = \sum_j a_{ij}, \tag{3.2}$$

$$k_i^{\text{in}} = \sum_j a_{ji}. \tag{3.3}$$

[1] Número de elementos de um conjunto. Note que $|\mathcal{X}|$ representa a cardinalidade do conjunto $\mathcal{X}$.

O grau total do vértice i é então definido como $k_i = k_i^{\text{in}} + k_i^{\text{out}}$. Como as redes são tidas como isoladas (nenhuma conexão com vértices externos), as médias $\langle k^{\text{in}} \rangle$ e $\langle k^{\text{out}} \rangle$ são iguais,

$$\langle k^{\text{out}} \rangle = \langle k^{\text{in}} \rangle = \frac{1}{N} \sum_{ij} a_{ij}. \tag{3.4}$$

A medida global associada à conectividade dos vértices é o *grau médio*, $\langle k \rangle$, e é calculado a partir da média de todos vértices da rede,

$$\langle k \rangle = \frac{1}{N} \sum_{i} k_i. \tag{3.5}$$

Nas redes com peso, não apenas a quantidade de conexões de um vértice é importante mas também a intensidade delas. Daí surge o conceito de *força* de um dado vértice que é definida como a soma dos pesos de seus arcos. Se a rede não for dirigida, a força s_i do vértice i é dado por [82]:

$$s_i = \sum_{j} w_{ij}, \tag{3.6}$$

No entanto, caso a rede com peso seja dirigida são definidas as quantidades [82]:

$$s_i^{\text{in}} = \sum_{j} w_{ji}, \tag{3.7}$$

$$s_i^{\text{out}} = \sum_{j} w_{ij}, \tag{3.8}$$

onde s_i^{in} e s_i^{out} é a força de entrada e a de saída do vértice i, respectivamente. Note que $s_i = s_i^{\text{in}} + s_i^{\text{out}}$ e que:

$$\langle s^{\text{out}} \rangle = \langle s^{\text{in}} \rangle = \frac{1}{N} \sum_{ij} w_{ij}. \tag{3.9}$$

A força média $\langle s \rangle$ de uma rede com peso é dada por:

$$\langle s \rangle = \frac{1}{N} \sum_{i} s_i. \tag{3.10}$$

Informações adicionais podem ser obtidas através da distribuição de graus, $P(k)$, que representa a probabilidade de um vértice escolhido aleatoriamente ter grau k. Definições similares podem ser obtidas para a distribuição de grau de entrada, $P(k^{\text{in}})$, e a de grau de saída, $P(k^{\text{out}})$,

para redes dirigidas; e a distribuição de força, $P(s)$, a de força de entrada, $P(s^{\text{in}})$, e a de força de saída, $P(s^{\text{out}})$, para redes com peso. Uma propriedade importante de muitas redes reais é a distribuição de graus seguir uma lei de potência [13], e uma forma eficiente de quantificá-la é através da medida de retidão, st – coeficiente de Pearson dos pontos da distribuição de graus na escala logarítmica. Para as coordenadas x e y de um conjunto com n pontos, tal coeficiente pode ser calculado pela expressão [83, 81]:

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\sqrt{\sum_{i=1}^{n} (x_i - \langle x \rangle)^2} \sqrt{\sum_{i=1}^{n} (y_i - \langle y \rangle)^2}} \tag{3.11}$$

onde x_i e y_i são as coordenadas do ponto i e $\langle x \rangle$ e $\langle y \rangle$ são as médias de x e de y, respectivamente. Tal coeficiente mede a correlação entre as variáveis x e y e se for igual a 1, as variáveis são correlacionadas; se for -1, são anti-correlacionadas; e se for 0, não há correlação entre as variáveis. Substituindo x por $\log(k)$ e y por $\log[P(k)]$ na expressão acima, o resultado é a medida de retidão. Quanto mais próxima estiver de -1, mais próxima de seguir uma lei de potência com coeficiente negativo de inclinação está a distribuição de graus da rede correspondente.

Uma propriedade importante de redes complexas é a *assortatividade*, que determina a correlação de grau dos vértices. Uma rede é *assortativa* se *hubs* (os vértices mais conectados de uma rede) se conectarem mais entre si do que com vértices de grau menor, enquanto os vértices com poucas conexões forem mais conectados entre si do que com os *hubs*. Se for o inverso disso, *hubs* tenderem a se conectar com vértices com poucas conexões, a rede é *disassortativa*. Caso não haja nenhuma relação entre os graus dos vértices, a rede é *não-assortativa*. Uma maneira de determinar essa propriedade em uma rede sem peso e sem direção é através do *coeficiente de assortatividade*, que é o coeficiente de correlação de Pearson dos graus dos vértices das extremidades dos arcos, dada por [84]:

$$r = \frac{\frac{1}{M}\sum_{j>i} k_i k_j a_{ij} - \left[\frac{1}{M}\sum_{j>i} \frac{1}{2}(k_i + k_j) a_{ij}\right]^2}{\frac{1}{M}\sum_{j>i} \frac{1}{2}(k_i^2 + k_j^2) a_{ij} - \left[\frac{1}{M}\sum_{j>i} \frac{1}{2}(k_i + k_j) a_{ij}\right]^2}, \tag{3.12}$$

onde M é o número de arcos da rede. Se $r > 0$, a rede é assortativa; se $r < 0$, a rede é disassortativa; e se $r = 0$ a rede é não-assortativa.

3.2 Medidas relacionadas à distância

Nem sempre dois vértices de uma rede estão conectados diretamente. Na maioria das vezes só é possível sair de um e chegar em outro através de um *caminho* com mais de um vértice entre eles, exceto quando a rede apresentar partes desconectadas ou vértices isolados. Em uma rede sem peso, o número de arcos existentes no caminho que conecta dois vértices representa seu *comprimento*. O *menor caminho* (também conhecido como *caminho geodésico*) entre os vértices i e j é aquele cujo comprimento é o menor de todos os caminhos possíveis entre esses vértices. A distância d_{ij} entre os vértices i e j é então definida como o comprimento do menor caminho entre eles.

Se a rede tiver um peso associado às conexões, há duas possibilidades para o cálculo do comprimento dos caminhos. A primeira delas parte do pressuposto de que os pesos dos arcos representam algum tipo de distância física como, por exemplo, a distância entre cidades numa rede de rodovias. Neste caso, o comprimento de um caminho é a soma dos pesos dos arcos contidos nesse caminho. Na outra possibilidade, os pesos dos arcos podem refletir a capacidade de conexão entre dois vértices. Por exemplo, numa rede de computadores o peso dos arcos corresponde à largura de banda das conexões entre os computadores. Portanto, neste caso, o comprimento do caminho entre dois vértices é a soma do inverso dos pesos dos arcos ao longo desse caminho.

No caso de não existir caminho entre os vértices i e j, a distância entre eles é definida como $d_{ij} = \infty$. Para redes dirigidas em geral $d_{ij} \neq d_{ji}$, pois o menor caminho indo do vértice i para o vértice j pode não ser igual ao menor caminho indo do vértice j para o vértice i.

A maior distância entre qualquer par de vértices de uma rede é chamada de *diâmetro* da rede e a distância média, de *caminho mínimo médio*, que é representado por l. Nos próximos capítulos, o caminho mínimo médio utilizado não é calculado simplesmente como a média dos menores caminhos (pois, se houver algum $d_{ij} = \infty$ para quaisquer vértices i e j, a média diverge) mas sim como:

$$l = \frac{1}{E}, \tag{3.13}$$

onde E é a *eficiência global* da rede [85] e é definida pela expressão:

$$E = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{ij}}, \tag{3.14}$$

Em redes que envolvam algum tipo de transporte, vértices que possuam muitos caminhos passando por eles tendem a receber a maior parte do tráfego. Uma maneira de quantificar a importância de um vértice segundo esse princípio é através de seu *grau de intermediação* (*betweenness*), definido como [86]:

$$B_u = \sum_{ij} \frac{\sigma(i,u,j)}{\sigma(i,j)}, \tag{3.15}$$

onde $\sigma(i,u,j)$ é o número de menores caminhos entre os vértices i e j que passam através do *vértice* u, $\sigma(i,j)$ é o número total de menores caminhos entre i e j, e a soma é tomada sobre todos os pares distintos i,j da rede. O grau de intermediação médio de uma rede é dado por $\langle B \rangle = \sum_i B_i/N$. O mesmo raciocínio pode ser usado para definir *grau de intermediação de arco*. Há também outra medida global relacionada ao grau de intermediação de vértice e é chamada de *dominância de ponto central*, definida como:

$$CPD = \frac{1}{N-1} \sum_i (B_{\max} - B_i), \tag{3.16}$$

onde $B_{\max}$ é o maior valor do grau de intermediação na rede. A dominância de ponto central é 0 para redes totalmente conectadas e 1 para redes na forma de estrela, em que existe um vértice central por onde todos os caminhos passam.

3.3 Medidas relacionadas a ciclos

A maioria das redes do mundo real apresentam uma grande quantidade de ciclos de ordem três (vértices conectados na forma de um triângulo) [12]. Essa propriedade, conhecida como aglomeração em redes sociais, indica a freqüência com que dois amigos quaisquer tenham um terceiro amigo em comum. Em redes não-dirigidas e sem peso, o coeficiente de aglomeração do vértice i é dado por [12]:

$$c_i = \frac{2l_i}{k_i(k_i-1)}, \tag{3.17}$$

onde l_i é o número de conexões entre os vizinhos de i e $k_i(k_i-1)/2$, a quantidade máxima de possíveis conexões entre os vizinhos de i. O coeficiente de aglomeração médio de uma rede não-dirigida e sem peso é a média:

$$C = \frac{1}{N}\sum_i c_i. \tag{3.18}$$

Para redes com peso, Barthélemy e colaboradores [82] introduziram o conceito de *coeficiente de aglomeração com peso* de um vértice como:

$$c_i^w = \frac{1}{s_i(k_i-1)}\sum_{j,k}\frac{w_{ij}+w_{ik}}{2}a_{ij}a_{ik}a_{jk}, \tag{3.19}$$

onde o fator $s_i(k_i-1)$ normaliza c_i^w tal que $0 \leq c_i^w \leq 1$. O coeficiente de aglomeração de uma rede com peso é dado por:

$$C^w = \frac{1}{N}\sum_i c_i^w. \tag{3.20}$$

Além dos valores médios, é possível também obter a distribuição do coeficiente de aglomeração $C(k)$ em função do grau k. Para algumas redes reais, foi observado que $C(k) \sim k^{-\alpha}$, onde α é conhecido como *expoente hierárquico*, pois a forma de $C(k)$ está associada com a estrutura hierárquica da rede [87].

Para redes sem peso, Bianconi e colaboradores [88] mostraram que o número de ciclos de ordem 3, 4 e 5 pode ser obtido, em termos da matriz de adjacência, pelas respectivas expressões:

$$N_3 = \frac{1}{6}\sum_i (A^3)_{ii}, \tag{3.21}$$

$$N_4 = \frac{1}{8}\left[\sum_i (A^4)_{ii} - 2\sum_i (A^2)_{ii}(A^2)_{ii} + \sum_i (A^2)_{ii}\right] \tag{3.22}$$

e

$$N_5 = \frac{1}{10}\left[\sum_i (A^5)_{ii} - 5\sum_i (A^2)_{ii}(A^3)_{ii} + 5\sum_i (A^3)_{ii}\right]. \tag{3.23}$$

Outras medidas relacionadas a ciclos incluem o *coeficiente cíclico* [89] que determina o quanto a rede é cíclica (valor 0 indica que a rede é da forma de uma árvore) e o coeficiente *rich-club* [90] que mede a tendência dos *hubs* se conectarem, formando comunidades.

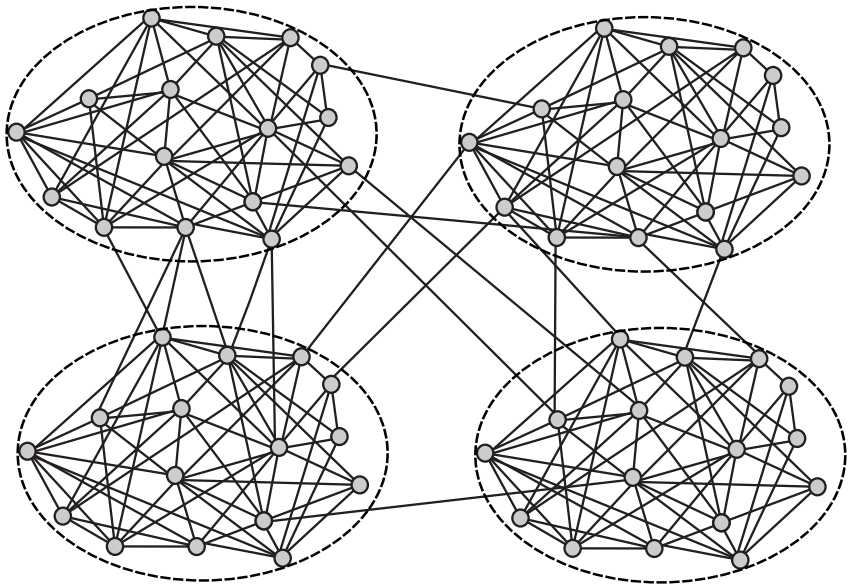


Figura 8: Exemplo de rede não-dirigida e sem peso com 4 comunidades. Há mais conexões entre vértices internos de cada comunidade do que com os das outras.

3.4 Detecção de comunidades e medidas

A maioria das redes reais possui estrutura modular, isto é, as redes são compostas por grupos cujos vértices são muito mais conectados entre si do que com o resto da rede (veja exemplo de uma rede com estrutura modular na Figura 8). Tais grupos recebem a denominação de comunidades e são resultados da interação local dos vértices. Exemplos de redes que possuem tal estrutura modular incluem: redes sociais [91, 92], redes metabólicas [93], redes de transporte aéreo [48], entre outras.

A identificação de comunidades é, portanto, uma tarefa importante, pois ajuda a encontrar vértices que compartilham propriedades semelhantes. A teia mundial é um bom exemplo de rede com esse tipo de estrutura em que páginas pertencentes a um determinado assunto geralmente formam um grupo. Neste caso, a identificação dessas comunidades pode ajudar na busca por informação na teia mundial.

Apesar de sua importância, ainda não existe um método ótimo para identificação de comunidades, já que não é possível saber de antemão quantas comunidades uma determinada rede possui nem mesmo o tamanho delas. Além disso, pode haver subdivisão entre as comunidades definindo hierarquias entre elas. E, ainda, não existe nem mesmo um consenso a respeito da

definição do que é comunidade. Uma proposta inituitiva foi dada por Radichi e colaboradores [94] e se baseia na densidade de arcos entre vértices. Segundo Radichi e colaboradores, há duas maneiras de definir comunidades: no *sentido forte* ou *fraco*. No primeiro, um subgrafo é uma comunidade se todos os seus vértices possuírem mais conexões entre eles do que com o resto da rede. Já no segundo, um subgrafo é uma comunidade se a soma dos graus de seus vértices considerando apenas conexões internas for maior do que fora do subgrafo. Um problema com tais definições é que a união de comunidades também é uma comunidade. Para superar essa limitação, Reichardt e Bornholdt [95] propuseram uma definição em que hierarquia entre as comunidades pode ser assumida *a priori*.

Além do problema da definição, não menos importante é saber como melhor dividir uma rede em comunidades. Como, a princípio, nenhuma informação está disponível sobre a quantidade de comunidades, Newman e Girvan [96] propuseram uma medida conhecida como *modularidade* para quantificar o quão eficiente é uma divisão específica. Seja c o número de comunidades obtidas numa determinada divisão e a matrix simétrica E de dimensão $c \times c$ cujos elementos e_{ij} representam a fração de conexões que existem entre os vértices da comunidade i e os da comunidade j. Os elementos e_{ii} representam o número de conexões internas à comunidade i. A modularidade, representada por Q, pode então ser calculada por:

$$Q = \sum_i \left[e_{ii} - \left(\sum_j e_{ij} \right)^2 \right] = \mathrm{Tr}E - ||E^2||, \tag{3.24}$$

Se $Q = 1$, a rede é formada apenas por partes totalmente desconexas. Embora simples, esta medida tem sido amplamente utilizada nos algoritmos de detecção de comunidades . Para maiores detalhes sobre algoritmos de detecção de comunidades veja [2, 97, 98].

3.5 Medidas hierárquicas

Medidas hierárquicas são uma extensão das medidas tradicionais, como grau e coeficiente de aglomeração, que não mais são calculadas apenas no nível seguinte imediato, ou seja, usando os vizinhos de um vértice de referência, mas sim no níveis subseqüentes de vizinhança [99, 100, 101, 102, 103]. Assim, usando conceitos de morfologia matemática em grafos [104, 105, 101],

anel hierárquico de nível r de um vértice i é definido como sendo o subgrafo[2] composto pelos vértices a uma distância r de i. Portanto, os vértices à distância igual a 1 de um vértice i são os vizinhos de i, os de distância 2 são os vizinhos dos vizinhos de i desconsiderando o próprio i, e assim por diante.

Para redes não-dirigidas, o grau hierárquico de nível d do vértice i, representado por $hk_d(i)$, é igual ao número de arestas que os vértices do anel d fazem com o resto da rede desconsiderando as arestas que ligam os vértices do mesmo nível e os dos níveis inferiores [101].

Também para redes não-dirigidas, o coeficiente de aglomeração hierárquico entres os níveis r e s do vértice i, representado por $hc_{rs}(i)$, é dado pela razão do número de arestas dentro do anel rs pelo número máximo de arestas possíveis dentro desse anel; o anel rs é o subgrafo definido pelos vértices localizados entre as distâncias $r-1$ e s.

Além dessas medidas, outras foram criadas como razão de convergência e razão de divergência [101]. A primeira, calculada a um distância d do vértice i e representada por $cr_d(i)$, corresponde ao quociente entre $hk_d(i)$ e o número de vértices do anel d. A segunda, também calculada a uma distância d, é representada por $dr_d(i)$ e é igual ao recíproco de $cr_d(i)$.

3.6 Outras medidas e estruturas

Há muitas outras medidas de redes complexas, tais como: grau médio dos vizinhos, vulnerabilidade, motivos, reciprocidade dos arcos de uma rede dirigida, medidas relacionadas à entropia e à energia, outras medidas de centralidade (além de grau de intermediação), medidas espectrais e medidas de complexidade. Estas medidas e maiores detalhes sobre as descritas neste capítulo podem ser encontradas em [2].

3.7 Correlação entre medidas

Mesmo com uma quantidade muito grande de medidas para caracterizar a estrutura de conectividade de redes complexas, muitas não podem ser utilizadas em conjunto, pois apresentam

[2]Um grafo g é um *subgrafo* ou *subrede* da rede $\mathcal{G}$ se $V(g) \subseteq V$, $E(g) \subseteq E$ e $\omega(g) \subseteq \omega$.

redundância entre si [2], isto é, são correlacionadas. Para determinar se duas medidas x e y são correlacionadas, a expressão (3.11) é empregada e x_i e y_i representam, respectivamente, a medida x e y da i-ésima rede do conjunto considerado, e $\langle x \rangle$ e $\langle y \rangle$ são as respectivas médias.

Para quantificar o grau de correlação entre pares de algumas medidas, o coeficiente acima foi calculado para redes geradas com 1000 vértices e grau médio 4 a partir dos modelos: aleatório de Erdős e Rényi (ER), livre de escala de Barabási e Albert (BA) e geográfico de Waxman (GW). Os resultados estão na Tabela 2, que também mostra o caso em que todos os modelos foram considerados juntos.

Os resultados apresentados na Tabela 2 indicam que os valores mais altos de correlação entre medidas foram observados para o modelo livre de escala de Barabási e Albert. Além disso, as correlações obtidas para cada modelo em separado são diferentes das obtidas para os modelos analisados em conjunto. Devido a esse fato, a análise das correlações entre medidas não é determinada de maneira simples, mas envolve os tipos de modelos considerados.

O tipo de aplicação desejada é que determina a escolha das medidas a serem utilizadas. Se o intuito for apenas caracterização de uma rede, deve-se escolher medidas que quantifiquem as propriedades esperadas. Por exemplo, para determinar se uma dada rede possui a propriedade de mundo pequeno, só é necessário calcular o caminho mínimo médio da rede em questão. No caso de classificação de redes (próximo capítulo), o maior número de medidas não correlacionadas (ver Tabela 2) deve ser considerado, pois, do contrário, a análise pode se tornar incompleta, proporcionando resultados equivocados. Até mesmo no caso de modelagem de redes do mundo real, a análise de medidas não deve ser ignorada, pois, se poucas medidas forem consideradas, as demais podem não estar sendo reproduzidas pelos modelos desenvolvidos.

Tabela 2: Correlações entre medidas obtidas para os modelos: aleatório de Erdős e Rényi (ER), livre de escala de Barabási e Albert (BA) e geográfico de Waxman (GW) e todos juntos. Os resultados foram estimados de 1000 realizações para cada modelo com 1000 vértices e grau médio 4 cada. As medidas consideradas foram: coeficiente de correlação de Pearson da distribuição de pontos na escala *log-log*, st; coeficiente de assortatividade, r; coeficiente de aglomeração médio, C; caminho mínimo médio, ℓ; dominância de ponto central, CPD; grau hierárquico médio de nível 2, $\langle hk_2(i)\rangle$; razão de convergência média de nível 2, $\langle cr_2(i)\rangle$; e razão de divergência média de nível 3, $\langle dr_3(i)\rangle$. Tabela extraída de [2].

		st	r	C	ℓ	CPD	$\langle hk_2(i)\rangle$	$\langle cr_2(i)\rangle$	$\langle dr_3(i)\rangle$
st	BA	1.00							
	ER	1.00							
	GW	1.00							
	Todos	1.00							
r	BA	–0.22	1.00						
	ER	–0.01	1.00						
	GW	–0.13	1.00						
	Todos	0.71	1.00						
C	BA	0.06	–0.29	1.00					
	ER	–0.01	0.07	1.00					
	GW	0.04	–0.00	1.00					
	Todos	0.31	0.82	1.00					
ℓ	BA	–0.01	0.38	–0.63	1.00				
	ER	–0.06	0.04	–0.08	1.00				
	GW	–0.10	0.02	0.03	1.00				
	Todos	0.69	0.96	0.88	1.00				
CPD	BA	–0.09	0.23	0.39	–0.58	1.00			
	ER	–0.61	0.10	0.03	0.07	1.00			
	GW	–0.05	–0.02	0.03	0.23	1.00			
	Todos	–0.87	–0.44	0.02	–0.41	1.00			
$\langle hk_2(i)\rangle$	BA	0.01	–0.30	0.63	–0.99	0.60	1.00		
	ER	0.04	0.03	0.08	–0.90	–0.06	1.00		
	GW	0.08	0.28	–0.02	–0.65	–0.13	1.00		
	Todos	–0.96	–0.80	–0.43	–0.79	0.85	1.00		
$\langle cr_2(i)\rangle$	BA	0.02	0.02	0.58	–0.74	0.59	0.76	1.00	
	ER	–0.03	0.04	0.45	–0.16	0.02	0.19	1.00	
	GW	–0.00	0.09	0.59	0.18	0.07	–0.11	1.00	
	Todos	0.37	0.86	0.99	0.91	–0.05	–0.49	1.00	
$\langle dr_3(i)\rangle$	BA	0.01	0.26	–0.57	0.91	–0.52	–0.94	–0.69	1.00
	ER	0.03	–0.10	–0.01	–0.25	–0.01	–0.16	–0.04	1.00
	GW	–0.02	–0.28	–0.09	–0.03	–0.00	–0.50	–0.21	1.00
	Todos	–0.14	–0.74	–0.97	–0.79	–0.18	0.27	–0.96	1.00

4 Classificação de redes complexas

Além da caracterização, as medidas podem ser utilizadas para classificar redes complexas. Classificar significa atribuir grupos, classes ou categorias a elementos que compartilham propriedades e/ou características comuns [81]. A classificação pode ainda apresentar hierarquias definidas a partir de especialização de classes mais gerais. Um exemplo típico é a taxonomia dos seres vivos que possui ao todo sete níveis de classificação: reino, filo, classe, família, gênero e espécie. Devido ao seu caráter geral, qualquer conjunto de elementos, desde objetos inanimados e lugares até conceitos e relações, pode ser classificado por esquemas semelhantes ao da taxonomia dos seres vivos. No caso das redes complexas não é diferente. Há diversos sistemas na natureza e muitos outros construídos pelos seres humanos cujas representações por rede apresentam propriedades estruturais e dinâmicas comuns. Devido a esse fato, nada mais natural que categorizar tais sistemas de acordo com essas propriedades. Para alcançar esse propósito, são necessários dois passos: extração de características das redes complexas e uso de um classificador. O primeiro passo consiste em obter o maior número de medidas, como as descritas no Capítulo 3, para melhor caracterizar as redes analisadas. Para cada rede, é então obtido um vetor de atributos $\vec{x}$, onde cada elemento x_i corresponde a uma medida distinta. O segundo passo é, a partir desses vetores, classificar as respectivas redes. Esse processo pode ser realizado de duas formas: *supervisionada* em que as classes são conhecidas *a priori* ou *não-supervisionada*, quando não há conhecimento algum sobre a divisão das redes [106]. A classificação supervisionada envolve dois estágios. O primeiro é o treinamento do modelo de classificação em que os métodos utilizados são treinados com redes cujas classes são conhecidas. Já o segundo estágio corresponde a atribuir classes às redes desconhecidas. No caso da classificação não-supervisionada, uma maneira de realizá-la é através do método conhecido

como *agrupamento* (*clustering*) [106, 81], que maximiza tanto a similaridade entre as medidas de redes que pertençam à mesma classe quanto a diferença de medidas de redes de classes distintas. Utilizando este método, classes de redes são obtidas sem nenhum conhecimento prévio, podendo ser empregado para obter a taxonomia das redes complexas, mas não foi realizado neste trabalho.

Uma aplicação prática da metodologia de classificação em redes complexas é classificar as redes do mundo real nos modelos teóricos existentes. Esse processo é muito importante pois permite validar e aperfeiçoar os modelos teóricos de forma a representar melhor a realidade dos sistemas analisados. Neste caso, é possível aplicar a classificação supervisionada pois as classes (isto é, os modelos teóricos) são conhecidos. Entretanto, há algumas dificuldades a serem superadas. Uma delas consiste em saber quais modelos são mais adequados para representar a rede real em questão. Outra dificuldade encontrada é determinar quais medidas são mais apropriadas. Finalmente, saber qual o melhor método de classificação a ser utilizado.

Existem vários modelos teóricos para redes complexas (alguns deles estão descritos na Seção 2.4) e cada um reproduz uma ou mais propriedades encontradas nas redes naturais ou artificiais estudadas atualmente. Além dessa gama enorme de opções, há também modelos, que dependendo dos parâmetros utilizados, podem gerar redes com propriedades semelhantes às geradas por outros modelos (e.g. o modelo de mundo pequeno de Watts e Strogatz pode ser regular se a probabilidade de reconexão for nula, ou aleatória se igual a 1). Tanto a quantidade de modelos existentes quanto os parâmetros que cada um possui impõem dificuldades no processo de classificação da redes reais, pois pode haver sobreposição entre eles no espaço de medidas e separá-los pode não ser possível. Uma maneira de evitar ou minimizar tal problema é ajustar os parâmetros dos modelos de acordo com a rede estudada.

No caso das medidas, a escolha das mais adequadas a um determinado tipo de rede pode também ser uma tarefa difícil. Muitas são específicas a determinados tipos de estruturas e, se utilizadas sozinhas, não oferecem uma representação completa da rede estudada. Outras são correlacionadas (como pode ser comprovado na Tabela 2), prejudicando a classificação. As melhores soluções para este tipo de problema incluem: escolha de medidas conforme os

resultados desejados e/ou uso de métodos de estatística multivariada para reduzir a redundância entre as medidas (assunto da próxima seção).

Escolher o melhor método de classificação de redes naturais ou artificiais pode também ser difícil, mas, neste caso específico em que redes reais são classificadas em modelos teóricos conhecidos, não é, pois se trata de classificação supervisionada e o melhor metodologia conhecida é o critério de decisão Bayesiana [106]. Maiores detalhes desta metodologia estão na Seção 4.2.

4.1 Estatística multivariada

Tanto o processo de amostragens de redes reais quanto as simulações de modelos teóricos geram padrões de conectividade com certa variabilidade estatística. Por causa disso, devem ser levados em conta não apenas os valores médios das medidas, mas também as suas respectivas variabilidades, tais como a variância. Por exemplo, várias realizações do modelo livre de escala de Barabási e Albert com parâmetros fixos possuem propriedades estruturais equivalentes, mas não idênticas, como ilustra a Figura 9, onde são mostradas 1000 redes com $N = 1000$ e $m = 3$ com relação a três medidas: coeficiente de assortatividade, coeficiente de aglomeração e caminho mínimo médios. Embora o conjunto de pontos obtidos formem uma nuvem em torno das médias de cada medida, há uma certa correlação entre o coeficiente de assortatividade e o caminho mínimo médio, implicando em redundância dessas medidas. Uma maneira de eliminar tal redundância é através da análise dos componentes principais (PCA), descrita abaixo.

4.1.1 Análise dos componentes principais

Cada elemento de um conjunto de R redes complexas pode ser descrito em termos de P medidas escalares x_i, $i = 1, 2, \ldots, P$, organizadas na forma de um vetor de atributos $\vec{x} = (x_1, x_2, \ldots, x_P)^T$. A matriz de covariância desse conjunto pode ser obtida através da seguinte expressão:

$$K = \frac{(\vec{x} - \langle\vec{x}\rangle)(\vec{x} - \langle\vec{x}\rangle)^T}{R}, \tag{4.1}$$

onde $\langle\vec{x}\rangle$ é ao vetor de atributos médios, cujos elementos correspondem às respectivas médias de cada medida. A expressão acima resulta em uma matriz real e simétrica com dimensões $P \times P$.

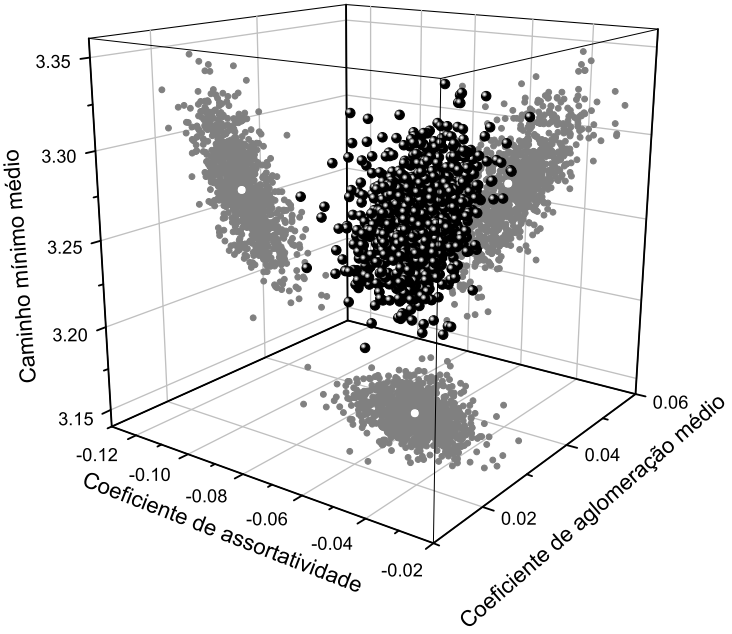


Figura 9: Distribuição espacial de 1000 realizações do modelo livre de escala de Barabási e Albert para $N = 1000$ e $m = 3$. Para melhor visualização, a nuvem central de pontos foi projetada nos três planos principais (regiões em cinza). As médias de cada medida (pontos brancos) estão projetadas nesses planos. Fonte [2].

Devido a esse fato, os P autovalores λ_i dessa matriz são reais. Se, além disso, os autovalores λ_i forem distintos, os correspondentes autovetores $\vec{v_i}$ são ortogonais, caso contrário, ainda é possível obter autovetores ortogonais a partir dos autovalores repetidos [81]. Estes autovetores podem, então, ser ordenados para obter a matriz de transformação:

$$T = \begin{bmatrix} \leftarrow \vec{v_1} \rightarrow \\ \leftarrow \vec{v_2} \rightarrow \\ \cdots \\ \leftarrow \vec{v_P} \rightarrow \end{bmatrix}, \tag{4.2}$$

onde $\vec{v_1}$ corresponde ao maior autovalor, $\vec{v_i}$, ao segundo, e assim por diante (os autovalores estão ordenados na seguinte forma: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_P$). A partir dessa matriz de transformação, o vetor original de atributos $\vec{x}$ de cada rede pode ser transformado em um novo sistema de coordenadas através da seguinte transformação linear:

$$\vec{X} = T\vec{x}, \tag{4.3}$$

que corresponde a uma rotação de eixos e define a *projeção dos componentes principais*.

A distribuição de pontos obtidas para o novo espaço definido pela transformação acima é de tal forma que as maiores variações são observadas ao longo dos primeiros eixos, seguido pelos eixos subseqüentes em ordem decrescente, sendo que os primeiros eixos são chamados de *principais* [107]. Devido à tal propriedade, ao considerar apenas os eixos principais, é possível projetar a distribuição inicial de pontos em um número menor de dimensões p – metodologia esta conhecida como *redução de dimensionalidade*. Para obter tal resultado, somente os primeiros p autovetores são utilizados e a matriz de transformação T_p resultante é:

$$T_p = \begin{bmatrix} \leftarrow \vec{v_1} \rightarrow \\ \leftarrow \vec{v_2} \rightarrow \\ \cdots \\ \leftarrow \vec{v_p} \rightarrow \end{bmatrix}. \tag{4.4}$$

Um exemplo de como tal metodologia pode se empregada num caso real é mostrado na Figura 10, em que a nuvem de pontos da Figura 9 é projetada nos dois primeiros componentes principais, eliminando a redundância existente entre as medidas. Apesar de ser útil na redução de dimensionalidade quando várias medidas são utilizadas, tal metodologia apresenta uma limitação por não considerar categorias de grupos diferentes de rede. Para superar esta limitação, a análise de variáveis canônicas, descrita abaixo, pode ser empregada.

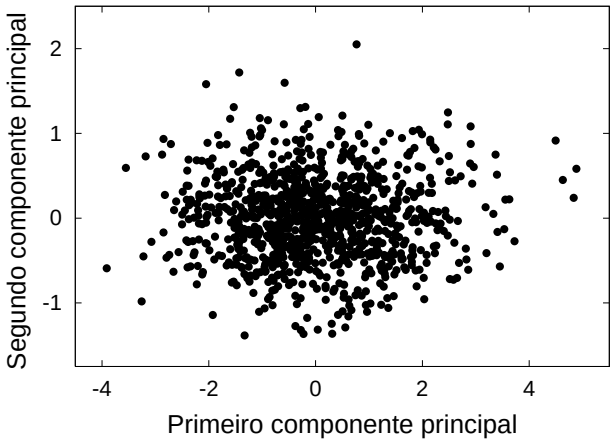


Figura 10: Projeções da nuvem de pontos da Figura 9 obtida através da análise de componentes principais. As medidas foram normalizadas antes de aplicar tal metodologia devido às diferenças dos valores absolutos entre elas. A normalização de uma medida corresponde a subtrair do valor de cada elemento da nuvem a média e dividir pelo desvio padrão. Fonte [2].

4.1.2 Análise de variáveis canônicas

A análise de variáveis canônicas é uma extensão da análise dos componentes principais através da qual as projeções são obtidas de forma a otimizar a separação entre objetos de classes distintas. No caso da análise de variáveis canônicas, são necessárias duas matrizes, uma para quantificar a variação dentro dos grupos, denominada S_{intra} e outra para quantificar a variação entre grupos, denominada S_{inter}. A formalização desse método é realizada a seguir.

Considere um conjunto de R redes, separadas em N_C classes, cada uma com N_i elementos, identificada por C_i, $i = 1, 2, \ldots, N_C$ e que cada elemento ξ de R é representado por seu respectivo vetor de medidas $\vec{x_\xi} = (x_1, x_2, \ldots, x_P)^T$. Define-se a matriz: S_i – a *matriz de dispersão* para cada classe C_i, como [106]:

$$S_i = \sum_{\xi \in C_i} \left(\vec{x_\xi} - \langle \vec{x} \rangle_i\right) \left(\vec{x_\xi} - \langle \vec{x} \rangle_i\right)^T, \tag{4.5}$$

onde $\langle \vec{x} \rangle_i$ representa o vetor da média das medidas dos elementos da classe C_i.

As matrizes S_{intra} e S_{inter} são, então, definidas como, respectivamente:

$$S_{\text{intra}} = \sum_{i=1}^{N_c} S_i. \tag{4.6}$$

e

$$S_{\text{inter}} = \sum_{i=1}^{N_c} N_i \left(\langle \vec{x} \rangle_i - \langle \vec{x} \rangle\right) \left(\langle \vec{x} \rangle_i - \langle \vec{x} \rangle\right)^T. \tag{4.7}$$

Para obter a máxima dispersão entre classes e a mínima dispersão dentro de cada classe, a seguinte transformação linear deve ser empregada [108]:

$$\vec{X_\xi} = \Gamma \vec{x_\xi}, \tag{4.8}$$

onde $\Gamma = [\vec{\gamma_1}, \vec{\gamma_2}, \ldots, \vec{\gamma_P}]^T$ é escolhido de forma que o vetor $\vec{\gamma_1}$ maximiza a relação

$$\frac{\vec{\gamma_1}^T S_{\text{inter}} \vec{\gamma_1}}{\vec{\gamma_1}^T S_{\text{intra}} \vec{\gamma_1}}, \tag{4.9}$$

enquanto que os demais vetores $\vec{\gamma_j}$, $j = 2, 3, \ldots, p$, maximizem relações semelhantes e

$$\vec{\gamma_j}^T S_{\text{intra}} \vec{\gamma_j} = 0. \tag{4.10}$$

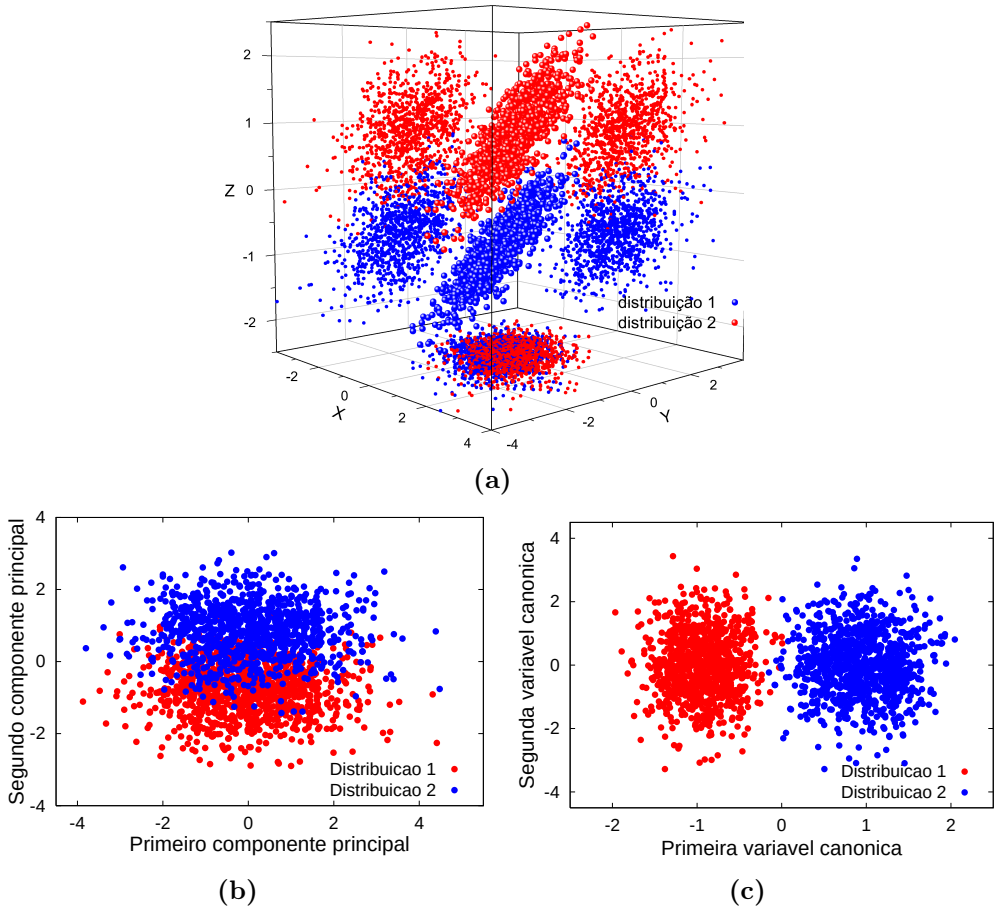


Figura 11: Duas nuvens distintas e arbitrárias de pontos em três dimensões (a) não puderam ser separadas usando análise de componentes principais (b). Isso só foi obtido pela análise de variáveis canônicas (c) que maximiza a separação entre grupos de pontos distintos. Figura adaptada de [2].

Os vetores $\vec{\gamma_1}, \vec{\gamma_2} \ldots, \vec{\gamma_P}$ são os autovetores da matriz $S_{\text{intra}}^{-1} S_{\text{inter}}$ [108].

Para mostrar o potencial da análise de variáveis canônicas, duas nuvens arbitrárias foram geradas em três dimensões de modo que suas projeções nos respectivos planos principais não pudessem ser distinguidas, Figura 11(a). Aplicando a análise dos componentes principais nessas nuvens não foi possível separá-las, Figura 11(b). Isso só foi possível através da análise por variáveis canônicas, Figura 11(c).

4.2 Teoria de decisão Bayesiana para classificação de redes

A maior dificuldade encontrada na classificação consiste em minimizar o erro de atribuir classes aos objetos. Uma das melhores soluções, neste caso, é utilizar a *teoria de decisão Baye-*

siana [106]. Definindo P_i como a probabilidade de um objeto pertencer à classe C_i, assume-se que P_i e a densidade de probabilidade condicional $p(\vec{x_\xi}|C_i)$ são conhecidas ou podem ser devidamente estimadas através de métodos paramétricos ou não-paramétricos [106, 109, 81]). A probabilidade P_i pode ser obtida através da freqüência relativa. Se a função densidade de probabilidade condicional for conhecida e apenas seus parâmetros tiverem que ser determinados, tal estimação é denominada *paramétrica*. Em contrapartida, caso nem mesmo essa função seja conhecida, torna-se necessário determiná-la. Neste caso, a estimação da função de densidade de probabilidade é chamada de *não-paramétrica* e ela pode ser determinada através do método de Parzen, em que cada ponto é representado como uma função, denominada *núcleo de Parzen*, no espaço definido pelas medidas [106]. O critério de decisão Bayesiana pode ser expresso da seguinte forma: se $p(\vec{x_\xi}|i)P(i) = \max_{b=1,C}\{p(\vec{x_\xi}|b)P(b)\}$ então o elemento ξ pertence à classe i [2].

A Figura 12 mostra como os conceitos discutidos acima podem ser empregados para encontrar a região de separação entre três categorias de redes complexas: modelo aleatório de Erdős e Rényi, modelo de mundo pequeno de Watts e Strogatz e modelo geográfico de Waxman, definidos no espaço das medidas coeficiente de assortatividade e o caminho mínimo médio.

4.3 Classificação de redes do mundo real

De um modo geral, o processo de classificação envolve dois estágios: *treinamento* e *identificação* [81]. O primeiro estágio corresponde a utilizar dados de objetos conhecidos para treinar o método de classificação a obter as classes correspondentes. No caso de classificação por teoria de decisão Bayesiana, este estágio equivale a obter as regiões de separação, ou seja, determinar, a partir das distribuições de objetos conhecidos, os parâmetros e até mesmo a própria função de densidade de probabilidade. O segundo estágio, por sua vez, é a classificação em si, em que objetos desconhecidos são identificados conforme a região a que eles pertençam no espaço de medidas.

No caso da classificação de redes do mundo real em modelos, antes do treinamento é necessário escolher o conjunto de medidas e os tipos de modelos teóricos a serem empregados. Se

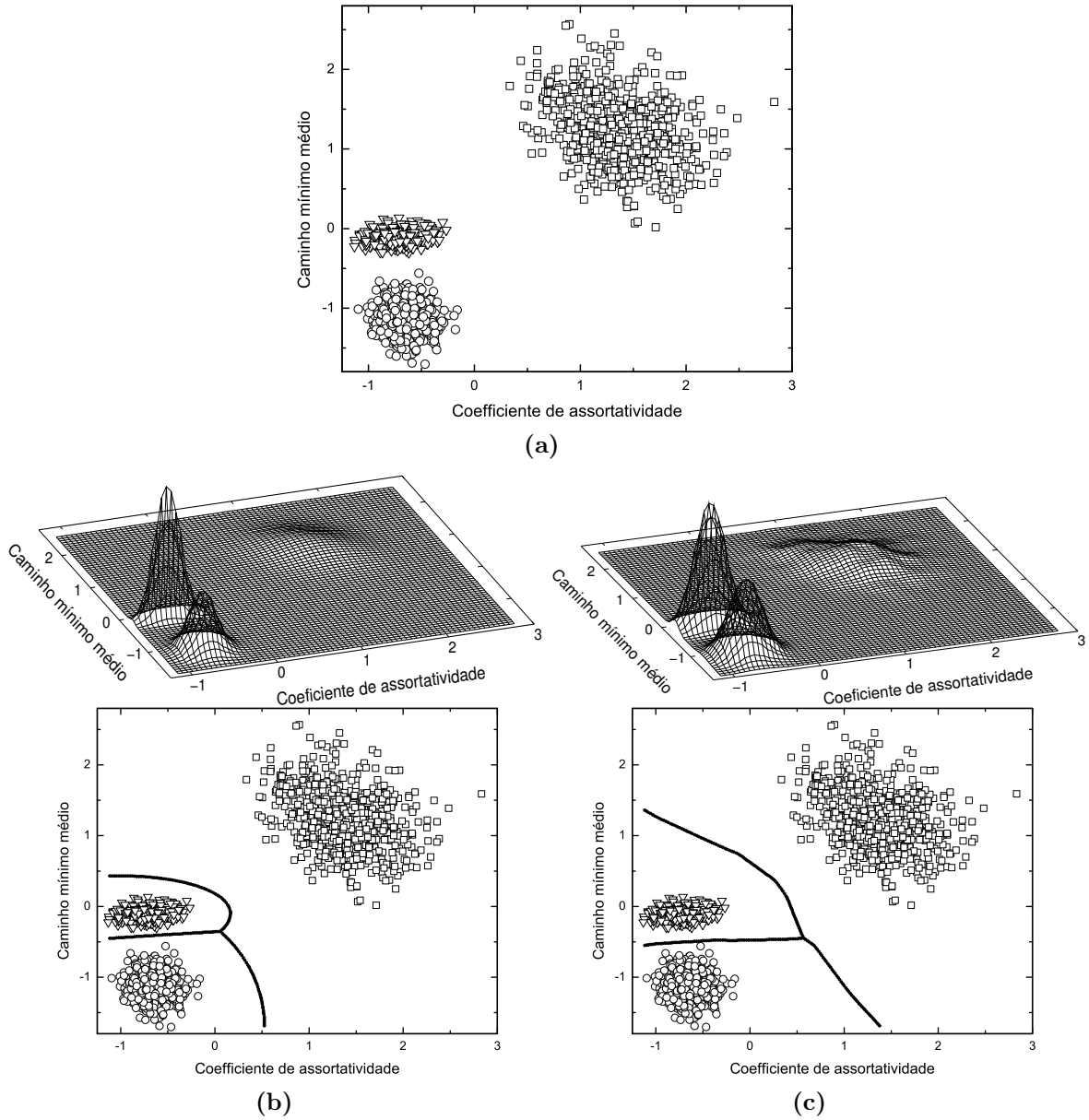


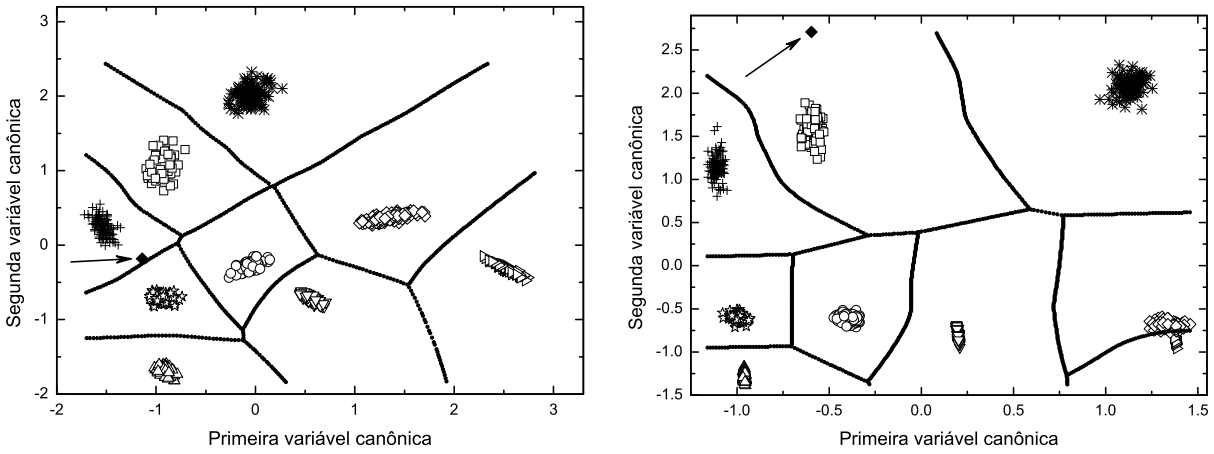
Figura 12: Espaço definido pelas medidas normalizadas: coeficiente de assortatividade e o caminho mínimo médio para redes obtidas a partir de três modelos: aleatório de Erdős e Rényi (◯), de mundo pequeno de Watts e Strogatz (▽) e geográfico de Waxman (□). Logo abaixo, são mostradas as correspondentes funções gaussianas e as regiões de decisão considerando estimação paramétrica (b) e não-paramétrica (c). Ao todo são 1000 redes para cada modelo com 250 vértices e grau médio 20 cada. A probabilidade de reconexão no modelo de mundo pequeno foi de 0.4. Figura extraída de [2].

muitas medidas forem utilizadas, aquelas que apresentarem redundância entre si podem prejudicar a classificação [81], além de não permitir a visualização dos dados (quando forem utilizadas 4 medidas ou mais). Uma maneira eficiente de resolver tal problema é utilizar a análise de variáveis canônicas e reduzir o conjunto de medidas em três ou menos projeções, permitindo a visualização dos dados. No caso dos modelos, é importante que eles tenham correspondência

com a rede real a ser classificada, como mesmo número de vértices e mesmo grau médio. Após essas considerações iniciais, o treinamento é realizado considerando um número grande de redes obtidas de modelos teóricos, três ou menos projeções resultantes da redução por análise de variáveis canônicas e teoria de decisão Bayesiana por estimação não paramétrica da função densidade de probabilidade, já que sua forma não é conhecida *a priori*. Obtidas as regiões de separação, a rede real é então projetada no espaço definido pelas variáveis canônicas e classificada segundo a região a que pertencer. Dois exemplos de aplicação de tal metodologia podem ser visualizados na Figura 13, onde a rede de transcrição genética do *Escherichia coli* e a rede de interação de proteínas do *Saccharomyces cerevisiae* (ambas descritas na Seção 2.3) foram classificadas considerando três modelos: aleatório de Erdős e Rényi, livre de escala de Barabási e Albert e o geográfico de Waxman e duas projeções obtidas pela análise de variáveis canônicas de nove medidas normalizadas: grau médio, coeficiente de assortatividade, coeficiente de Pearson da distribuição de graus na escala log-log, coeficiente de aglomeração médio, caminho mínimo médio, dominância de ponto central, grau hierárquico médio de nível 2, coeficiente de aglomeração hierárquico médio de nível 2 e razão de divergência de nível 3.

A rede de transcrição genética do *Escherichia coli* e a rede de interações de proteínas do *Saccharomyces cerevisiae*, que apresentam a lei de potência na distribuição da conectividade, foram classificadas como redes geográficas. Tal resultado é surpreendente, já que o modelo que gera redes geográficas produz redes com distribuição de Poisson. Entretanto, se fossem consideradas apenas as medidas que geralmente são utilizadas para a caracterização de redes, como a distribuição das conexões e o caminho mínimo médio, as redes de transcrição genética e de interação de proteínas seriam associadas ao modelo de Barabási e Albert, o que é um resultado incompleto, pois tal modelo não reproduz grande parte das propriedades estruturais destas redes. Um dos fatores que influenciaram na classificação apresentada na Figura 13 é que, apesar do modelo de Barabási e Albert gerar a lei de potência na distribuição das conexões, e por isso ser largamente utilizado como modelo de redes reais, ele não gera redes com alta ocorrência de ciclos de ordem 3, que é uma característica presente nas maioria das redes reais e nas redes geradas pelo modelo geográfico considerado.

Uma análise mais completa de classificação de redes reais nesses três modelos teóricos,



(a) **(b)**

Figura 13: Classificação de redes reais em modelos teóricos utilizando redução de dimensionalidade via análise das variáveis canônicas para um conjunto de nove medidas normalizadas e teoria de decisão Bayesiana. As redes reais consideradas (indicadas por setas e representadas por ⧫) foram: (a) a rede de transcrição genética do *Escherichia coli* com grau médio 2.45 e (b) a rede de interações de proteínas do *Saccharomyces cerevisiae* com grau médio 3.03. Tanto em (a) quanto em (b) os modelos utilizados foram: aleatório de Erdős e Rényi com grau médio 2 (☆), 2.45 para (a) e 3 para (b) (○) e 6 (◊), livre de escala de Barabási e Albert com grau médio 2 (△), 4 (▽) e 6 (▷) e o geográfico de Waxman com grau médio 2 (+), 2.45 para (a) e 3 para (b) (□) e 6 (∗). Embora essas redes sejam geralmente associadas aos modelos de Barabási e Albert, se apenas a distribuição de graus e o caminho mínimo médio são considerados, com 9 medidas elas são classificadas como geográficas. Figura extraída de [2].

utilizando combinações diferentes dessas nove medidas, pode ser encontrada em [2]. Os resultados apresentados em [2] evidenciam que para uma compreensão mais precisa da estrutura das redes complexas é necessário utilizar um conjunto maior de medidas e de modelos, pois, do contrário, os resultados da classificação de redes podem ser incorretos.

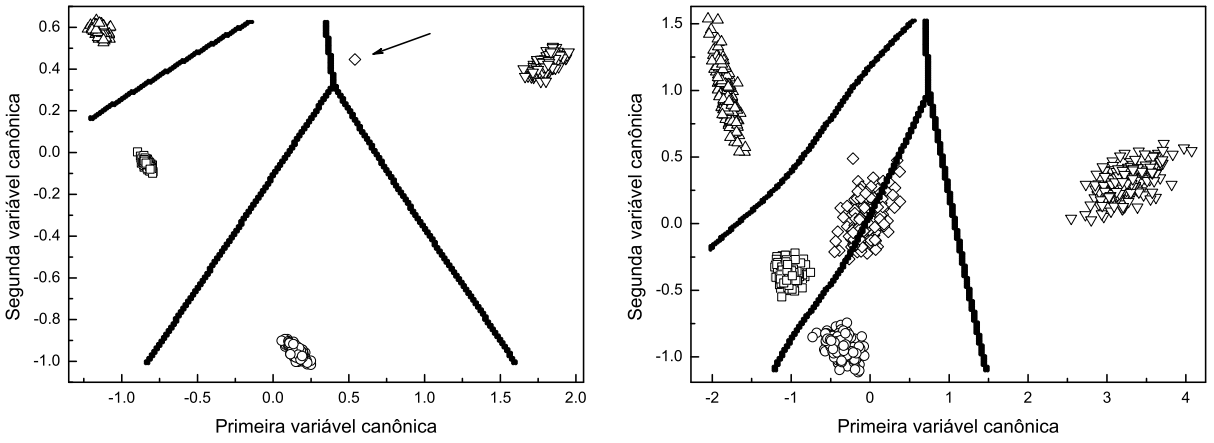
Utilizando a mesma metodologia de classificação apresentada neste capítulo, em [110] é apresentada a classificação da Internet, definida no nível de sistemas autônomos, considerando um conjunto mais amplo de medidas diferentes (num total de 21) e vários modelos (7 ao todo), dos quais alguns foram especialmente desenvolvidos para representar esse tipo de rede, como por exemplo: o modelo geográfico de Waxman (Seção 2.4), GdTang – modelo de rede dirigida da topologia da Internet baseado em crescimento preferencial e geográfico [78] e Inet – gerador de topologia da Internet [79, 80]. Os resultados de [110] indicaram que nenhum desses modelos foi capaz de reproduzir com precisão todas as medidas da Internet. Desta forma, métodos de classificação, como o apresentado neste capítulo, devem ser considerados na formulação de

novos modelos.

4.4 Efeito da amostragem na classificação

A fim de verificar o efeito da amostragem na classificação de redes, a metodologia descrita nas seções anteriores foi empregada para classificar a versão completa e amostras da rede de *emails* da Universidade Rovira i Virgili (Tarragona) [61] (Seção 2.3.1). As amostras foram obtidas através de caminhadas aleatórias [111, 17] na rede original. Uma caminhada aleatória é aquela em que, assumindo que no passo t ela está no vértice i, o próximo passo é pular para um dos vizinhos de i com probabilidade $1/k_i$. O vértice inicial da caminhada é também escolhido aleatoriamente, e a caminhada termina quando a quantidade desejada de vértices diferentes for alcançada. Note que este tipo de caminhada permite passar por vértices e arestas que já foram visitados. A rede amostrada por esse método é então obtida pelos vértices e arestas diferentes encontradas durante a caminhada.

Esse tipo de amostragem foi aplicado na rede de *emails* e 100 amostras foram obtidas. A rede original e as amostras têm, respectivamente: 1133 vértices e 626 (13); grau médio 9.62 e 4.00 (0.08); coeficiente de aglomeração médio 0.22 e 0.076 (0.009); e caminho mínimo médio 3.33 e 4.43 (0.08). Os valores apresentados para as amostras são os valores médios e entre parêntesis estão os respectivos desvios padrões. Note que, embora o número de vértice não tenha diminuído muito, essas medidas mudaram bastante. A Figura 14 mostra o resultado da classificação da rede original e as versões amostradas considerando quatro modelos: aleatório de Erdős e Rényi, de mundo pequeno de Watts e Strogatz, livre de escala de Barabási e Albert e geográfico de Waxman e 9 medidas: coeficiente de assortatividade, coeficiente de Pearson da distribuição de graus na escala log-log, coeficiente de aglomeração médio, caminho mínimo médio, grau de intermediação médio, ponto de dominância central, grau hierárquico médio de nível 2, coeficiente de aglomeração hierárquico médio de nível 2 e razão de divergência de nível 2. Os parâmetros dos modelos foram escolhidos de tal forma a serem compatíveis com os dois casos analisados. No caso do modelo de mundo pequeno de Watts e Strogatz, a probabilidade de reconexão foi escolhida de forma que o coeficiente de aglomeração e o caminho mínimo médios



(a) **(b)**

Figura 14: Efeito da amostragem na classificação da rede de *email* ($\Diamond$) utilizando o mesmo método aplicado na Figura 13. Os modelos utilizados foram: aleatório de Erdős e Rényi ($\Box$), de mundo pequeno de Watts Strogatz ($\bigcirc$), livre de escala de Barabási e Albert ($\triangle$), e geográfico de Waxman ($\bigtriangledown$). A probabilidade de reconexão no modelo de mundo pequeno de Watts Strogatz foi escolhido de tal forma a ter coeficiente de aglomeração e caminho mínimo médios similares aos encontrados na rede original e nas versões amostradas. A rede original (indicada por seta) foi identificada como pertencente ao modelo geográfico em (a), porém, quando foi parcialmente amostrada, a classificação não se manteve e as respectivas amostras foram indentificadas como pertencentes a outros dois modelos: o aleatório e o de pequeno mundo.

fossem os mais próximos possíveis tanto para a rede original quanto para as redes amostradas.

Os resultados mostrados na Figura 14 indicam que a amostragem modificou completamente a classificação da rede de *emails*. No primeiro caso, a rede original foi classificada como rede geográfica de Waxman. Entretanto, quando foi amostrada, as amostras resultantes foram classificadas como redes aleatórias de Erdős e Rényi e redes de mundo pequeno de Watts e Strogatz. A partir desses resultados, fica evidente que considerar amostras menores de redes complexas pode levar a conclusões erradas.

O Capítulo 5 apresentará estruturas que têm grande incidência em redes que tiveram problemas de amostragem, e o capítulo 6 mostrará o efeito de perturbações na conectividade das redes e quais propriedades são mais adequadas para caracterizar redes com problemas de amostragem.

5 Estruturas comuns em redes amostradas

Apesar de existirem inúmeras medidas de redes complexas disponíveis na literatura, seu uso sempre estará relacionado com aplicações específicas, pois não existe um conjunto completo de medidas para caracterização de redes complexas (Capítulo 3). Por exemplo, se o interesse for somente em saber se determinada rede possui a propriedade de mundo pequeno, é necessário apenas realizar o cálculo de distância entre os vértices da rede em questão. Neste sentido, o estudo de problemas específicos continuará motivando a descoberta de novas estruturas e maneiras de quantificá-las. Este é o caso das estruturas (motivos) que serão apresentadas neste capítulo: árvores de borda e cadeias de vértices, pois são frutos do estudo do problema de amostragem em redes complexas.

5.1 Árvores de borda

Árvores de borda [3] podem ser compreendidas como um tipo especial de motivos – subgrafos cuja probabilidade de ocorrerem em redes reais é maior do que nas redes aleatórias correspondentes (e.g. [68]). No entanto, diferentemente dos motivos tradicionais que podem ocorrer em qualquer parte da rede, as árvores de borda, como o próprio nome revela, estão situadas apenas na borda da rede, região esta que é definida como os vértices de grau unitário (que, por sua vez, são considerados as partes mais sensíveis da rede). As árvores de borda (assim como outros motivos periféricos) podem estar relacionados com o crescimento de redes, como as geográficas, ou seja, as redes podem evoluir como uma árvore, onde cada ramo de vértices parte do maior componente conectado para fora da rede. Além disso, as árvores de borda

estão relacionadas com a decomposição de núcleo-k (e.g. [112]) das camadas mais externas. Este tipo de decomposição é uma metodologia utilizada para descrever a topologia das redes complexas em termos de subgrafos. O núcleo-k de uma rede é obtido através da remoção de todos os vértices com grau menor que k. Depois de tal remoção, os vértices restantes que apresentarem grau menor que k são também removidos. O processo continua até tal remoção não ser mais possível, resultando no núcleo-k da rede original. A este processo é dado o nome de *decomposição de núcleo-k*. Define-se ainda camada-k de uma rede como o conjunto de vértices removidos no processo de decomposição de núcleo-k. Muitas aplicações de tal metodologia foram realizadas em redes reais, como por exemplo: na Internet [113] e na redes de interação de proteínas [114, 115]. Essa metodologia pode ser usada inclusive para encontrar os vértices que pertencem às árvores de borda, pois a camada-1 corresponde aos vértices que não participam de ciclos. Entretanto, esta não foi a abordagem utilizada na análise de árvores de borda em redes complexas, pois mais informações sobre essas estruturas foram necessárias. Um algoritmo para encontrar tais estruturas será apresentado abaixo, depois da definição formal de árvores de borda.

Na teoria de grafos, uma *árvore* é definida como um grafo conectado no qual existe somente um caminho entre dois vértices (e.g. [11, 116]). Em outras palavras, qualquer grafo conectado que não possua ciclo é uma árvore. Com base nessas definições, o núcleo de uma rede é definido como sendo o maior componente conectado cujos vértices participem de pelo menos um ciclo. A *árvore de borda* é, portanto, uma árvore que possui um vértice em comum com o núcleo, que é chamado de *raiz* da árvore de borda e está situado na periferia do núcleo (alguns exemplos de árvores de borda podem ser visualizados na Figura 15). Definem-se ainda: as *folhas* de uma árvore de borda como os vértices que estão nas extremidades da árvore, ou seja, os vértices de grau um; a *profundidade* como sendo a maior distância entre a raiz e as folhas; e o *número de ramos* como o número de caminhos partindo da raiz até as folhas (i.e. o número de ramos é igual ao número de folhas). Os vértices que estão o mais distante possível da raiz pertencem ao nível 0 da árvore de borda. No primeiro nível estão os vértices cuja distância à raiz é a profundidade menos um, no segundo nível estão os vértices cuja distância à raiz é a profundidade menos 2, e assim por diante. No nível mais alto está a raiz.

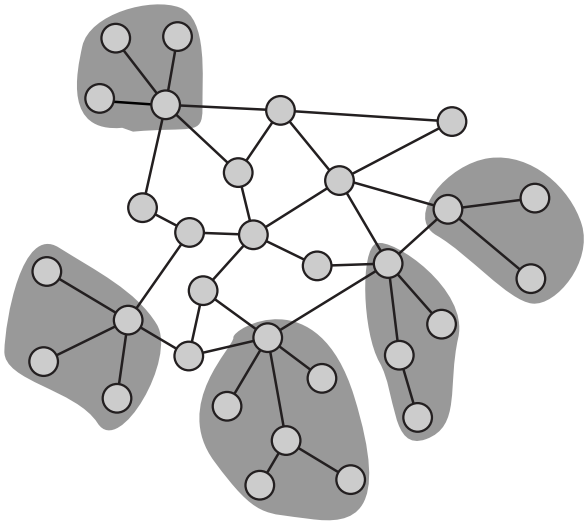


Figura 15: Árvores de borda (regiões em cinza escuro) em uma rede pequena. Figura extraída de [3].

5.1.1 Algoritmo

O algoritmo utilizado para encontrar as árvores de borda em uma dada rede é baseado na topologia de fora para dentro, ou seja, começa pelos vértices da borda, vai para seus vizinhos de níveis mais elevados até encontrar a raiz (vértice que pertence ao núcleo). Inicialmente é necessário encontrar todos os vértices de grau 1. Para cada um desses vértices, uma árvore (na forma de estrutura de dados) é criada. Desta forma, cada árvore tem apenas um vértice. O próximo passo corresponde a verificar recursivamente se o vértice no topo de cada árvore tem mais de um vizinho, ignorando aqueles que estão nos níveis inferiores. Se possuir mais de um vizinho, a árvore correspondente é mantida em uma lista de espera. Mas se possuir apenas um vizinho, ele é adicionado à árvore juntamente com qualquer outra árvore que esteja na lista de espera que possua esse vértice como raiz (note que a estrutura da rede é mantida). O algoritmo termina quando todas as árvores estão na lista de espera e não podem mais ser juntadas. As árvores isoladas (que não têm um vértice no núcleo da rede) são ignoradas.

5.1.2 Análise em redes teóricas e redes reais

Na análise das árvores de borda foram utilizados os modelos (Seção 2.4): aleatório de Erdős e Rényi (ER), de mundo pequeno de Watts e Strogatz (WS), livre de escala de Barabási e Albert

(BA), e geográfico de Waxman (GW) e 16 redes reais[1], das quais 4 são redes de colaboração: astrofísica, *netscience* (pesquisadores na área de redes complexas), matéria condensada, física de altas energias; 3 são de informação: Roget, Wordnet e WWW; 3 são de adjacência em livros: "David Copperfield", "Night and day" e "On the origin of species"; 3 são tecnológicas: Internet (no nível de sistemas autônomos), rede de aeroportos dos Estados Unidos e rede de transmissão elétrica dos estados ocidentais dos Estados Unidos; e 3 são biológicas: rede neural do *C. elegans*, rede de transcrição genética do *E. coli* e a rede de interação de proteínas do *S. cerevisiae*. No caso dos modelos, foram consideradas 100 realizações e todos tiveram $N = 1000$ vértices e graus médios 2, 4 e 6 e a probabilidade reconexão foi 0.2 para o modelo WS. No caso das redes reais dirigidas, elas foram transformadas nas versões não-dirigidas utilizando a operação de simetrização (conforme descrita na Seção 2.2), pois a direção dos arcos não foi levada em conta nas análises seguintes.

Para verificar a significância das árvores de borda nas redes consideradas, para cada uma foram geradas 1000 redes aleatórias através do método de reconexão de arestas, conhecido como *rewiring*[2]. Para os modelos teóricos, porém, foram geradas 1000 redes aleatórias por esse método para cada uma das 100 realizações de cada modelo. A importância das árvores de borda, em termos estatísticos, são mostradas na Tabela 3. Alguns modelos não foram incluídos nessa tabelas porque não apresentavam árvores de borda. Nas redes de colaboração, as árvores de borda possuem de 11.2 a 25.2% do número total de vértices e, em geral, não são profundas nem possuem muitas folhas. No caso das redes de informação, em particular a Wordnet e WWW, mais da metade dos vértices pertencem às arvores de borda – uma quantidade expressiva e reflete a importância de tais motivos na estrutura dessas redes. Situações similares são observadas para as redes biológicas tais como transcrição genética e de interação de proteínas e para as redes tecnológicas, tais como a Internet e a rede de transmissão elétrica. No caso das redes de adjacência em livros, uma situação totalmente diferente é observada: os vértices das árvores de

[1]Note que não foram utilizados as mesmas categorias de redes reais da Seção 2.3. Isso foi feito para facilitar as análises seguintes

[2]*Rewiring* é um processo de construção de redes que, a partir de uma determinada rede, gera a sua versão aleatória sem alterar a distribuição de graus (e.g. [117]). Tal processo começa com uma rede com uma dada distribuição de graus, escolhe iterativamente duas arestas e troca os correspondentes vértices de suas extremidades. Para garantir que a rede gerada seja completamente aleatória (sem correlação de graus), o número de passos deve ser 100 vezes o número total de arestas da rede original [117].

Tabela 3: Estatística das árvores de borda na redes consideradas, onde: N é o tamanho da rede; P_{v}, a porcentagem de vértices nas árvores de borda; N_{a}, número de árvores de borda; N_{v}, número médio de vértices nas árvores de borda; H, a profundidade média das árvores de borda; N_{f}, a média de folhas das árvores de borda; BA, modelo livre de escala de Barabási e Albert; ER, modelo aleatório de Erdős e Rényi; WS, modelo de mundo pequeno de Watts e Strogatz; e GW, modelo geográfico de Waxman. Os números entre parêntesis representam os desvios padrões obtidos para as versões aleatórias de cada rede analisada. Tabela extraída de [3].

Rede	N	P_{v}	N_{a}	N_{v}	H	N_{f}
BA $\langle k\rangle = 2$	1000	100.0%	1	1000 (0.0)	9.06 (0.01)	663.9 (0.1)
ER $\langle k\rangle = 2$	1000	51.0%	166	3.1 (0.0)	1.62 (0.01)	1.4 (0.0)
$\langle k\rangle = 4$	1000	14.4%	68	2.1 (0.0)	1.08 (0.00)	1.0 (0.0)
$\langle k\rangle = 6$	1000	2.9%	14	2 (0.0)	1.01 (0.00)	1.0 (0.0)
WS $\langle k\rangle = 2$	1000	95.2%	21	44.6 (1.1)	13.89 (0.20)	7.6 (0.2)
GW $\langle k\rangle = 2$	1000	47.0%	144	3.3 (0.0)	1.69 (0.01)	1.5 (0.0)
$\langle k\rangle = 4$	1000	16.0%	71	2.3 (0.0)	1.18 (0.00)	1.1 (0.0)
$\langle k\rangle = 6$	1000	3.6%	17	2.1 (0.0)	1.07 (0.00)	1.0 (0.0)
Astrofísica	16706	11.2%	797	2.3 (0.8)	1.06 (0.26)	1.3 (0.7)
Netscience	1461	12.7%	81	2.3 (0.6)	1.01 (0.11)	1.3 (0.5)
Matéria condensada	40421	12.6%	2095	2.4 (0.8)	1.06 (0.25)	1.4 (0.8)
Física de alta energia	8361	25.2%	828	2.5 (0.9)	1.15 (0.43)	1.4 (0.7)
Roget	1022	9.9%	42	2.4 (0.7)	1.29 (0.51)	1.1 (0.3)
Wordnet	82670	60.6%	9248	5.4 (7.5)	1.25 (0.55)	4.1 (6.8)
WWW	325729	56.2%	17070	10.7 (61)	1.13 (0.47)	9.5 (60.0)
David Copperfield	11378	0.6%	30	2.2 (0.5)	1.00 (0.00)	1.2 (0.5)
Night and day	7959	0.4%	14	2.4 (0.7)	1.00 (0.00)	1.4 (0.7)
On the origin of species	6973	0.4%	12	2.1 (0.3)	1.00 (0.00)	1.1 (0.3)
Internet	22963	42.3%	1712	5.7 (16)	1.07 (0.27)	4.6 (15.8)
Aeroportos dos EUA	332	24.4%	26	3.1 (2.3)	1.00 (0.00)	2.1 (2.3)
Transmissão elétrica	4941	48.4%	805	3 (2.0)	1.39 (0.80)	1.5 (1.2)
Rede neural	297	6.1%	3	6 (4.6)	1.00 (0.00)	5.0 (4.6)
Transcrição genética	423	63.8%	44	6.1 (5.4)	1.30 (0.55)	4.6 (5.0)
Interação de proteínas	4135	49.5%	570	3.6 (5.3)	1.25 (0.52)	2.3 (4.6)

borda representam apenas uma pequena fração do total. Este efeito é conseqüência do processo seqüencial de que tais redes são geradas.

As Figuras 16 e 17 apresentam as distribuições da profundidade e do número de folhas encontradas para as redes consideradas e as respectivas versões aleatórias (pontos em cinza). Um resultado interessante dessas figuras é que a maioria das distribuições seguem leis de potência com um truncamento exponencial, $P(x) \approx (x+x_0)^{\gamma} e^{-(x+x_0)/x_c}$ [118]. As redes: WWW, Wordnet, interação de proteínas e Internet apresentaram distribuições de leis de potência da forma $P(x) \approx (x + x_0)^{\gamma}$ para a profundidade e para o número de folhas. Este resultado indica que árvores de borda profundas e com muitas folhas não são muito comuns.

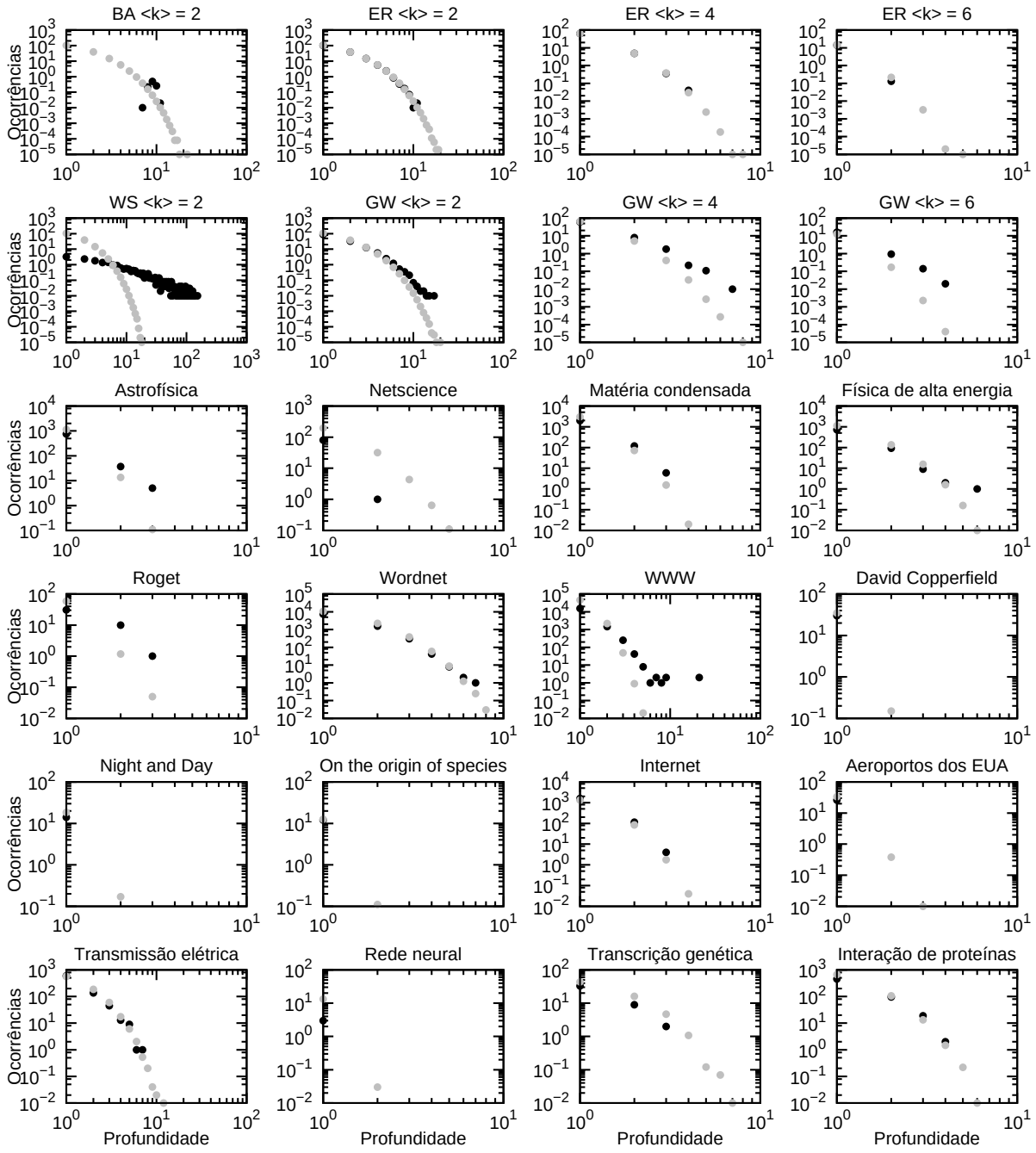


Figura 16: Distribuição da profundidade das árvores de borda obtidas para cada rede considerada (pontos pretos) e da média das respectivas versões aleatórias (pontos em cinza). Figura extraída de [3].

Através das Figuras 16 e 17, é possível ainda verificar em quais intervalos de profundidade ou do número de folhas as redes consideradas possuem quantidade significativa de árvores de borda em relação às suas versões aleatórias. Para obter tal resultado, a variação de profundidade (o mesmo aplica para o número de folhas) foi dividida em intervalos nos quais a distribuição da rede considerada está acima ou abaixo da distribuição da média das respectivas versões

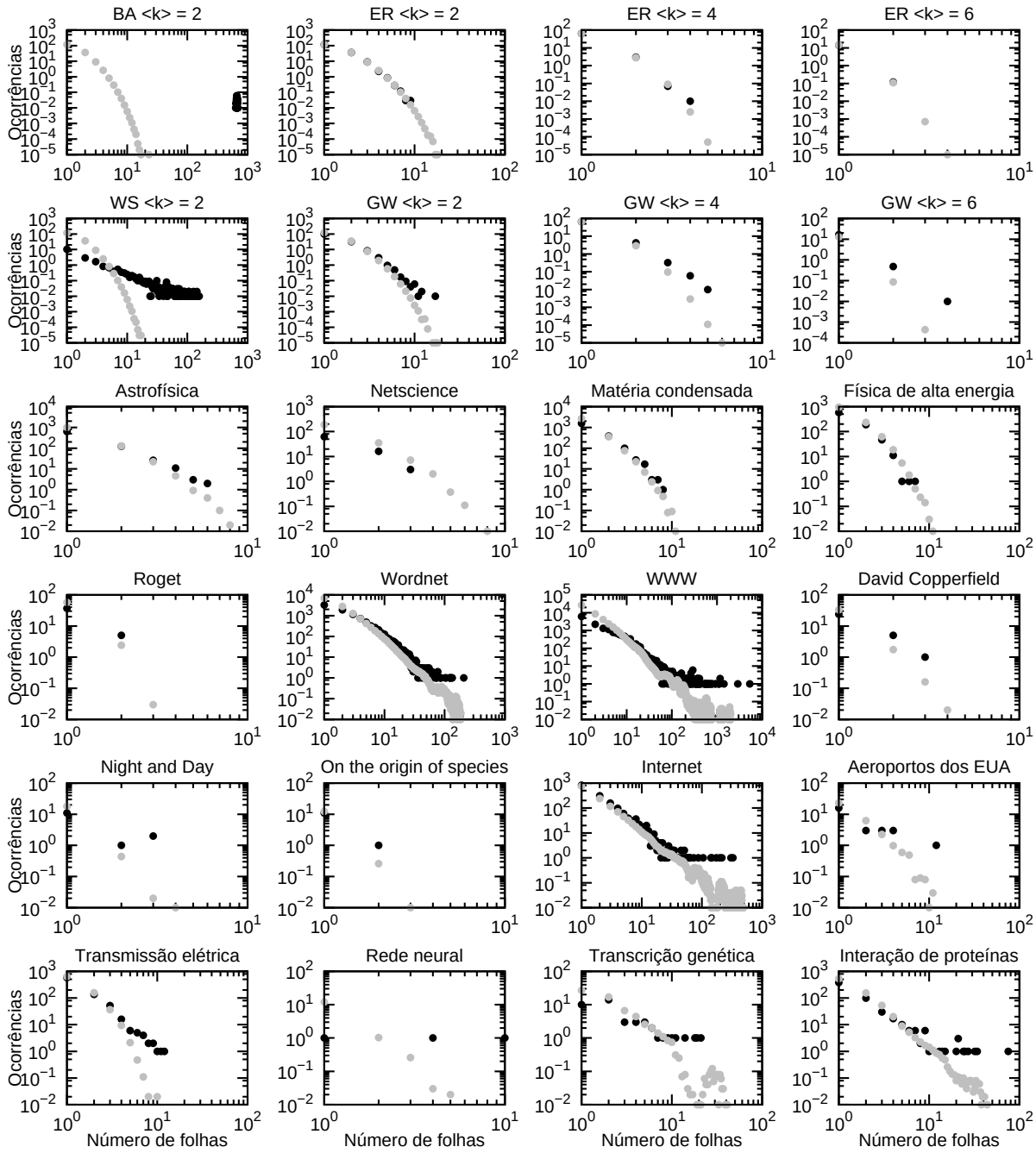


Figura 17: Distribuição do número de folhas das árvores de borda obtidas para cada rede considerada (pontos pretos) e da média das respectivas versões aleatórias (pontos em cinza). Figura extraída de [3].

aleatórias. No caso do modelo WS com grau médio 2 da Figura 16, dois intervalos foram considerados: um cuja profundidade vai de 1 até 5 e outro cuja profundidade vai de 6 até 152. No primeiro intervalo considerado a distribuição do modelo está abaixo da distribuição da média das respectivas versões aleatórias. O efeito inverso ocorre no segundo intervalo. Para determinar o grau de significância da presença das árvores de borda em termos do intervalo de

profundidade e do intervalo do número de folhas, a seguinte medida estatística foi utilizada:

$$Z = \frac{X_{\text{real}} - \langle X \rangle}{\sigma}, \tag{5.1}$$

onde Z é o *Z-score*, X_{real} é o número de árvores de borda existentes dentro de um certo intervalo de profundidade (o mesmo para o número de folhas), e $\langle X \rangle$ e σ são, respectivamente, a média e o desvio padrão das versões aleatórias para o mesmo intervalo de profundidade. Valores positivos do *Z-score* indicam que a rede de interesse possui mais árvores de borda do que suas versões aleatórias no intervalo considerado; o caso oposto é observado para valores negativos. Os resultados do *Z-score* da profundidade e do número de folhas das redes consideradas estão na Tabela 4.

Como pode ser visto na Tabela 4, as redes aleatórias obtidas possuem muitas árvores de borda baixas e com poucas ramificações, mas poucas altas e com muitas ramificações. Isso ocorre porque o processo de reconexão tende a eliminar as grandes estruturas da rede, como as árvores de borda grandes, dando origem à árvores de borda menores.

Dentre os modelos analisados, apenas BA e WS com grau 2 e todas as redes geradas para o modelo GW apresentaram um número significativo de árvores de borda. Interessante notar que o modelo BA com grau 2 possui muitas folhas enquanto que o modelo WS possui as árvores de borda mais profundas. Estes resultados já eram esperados pois o modelo BA com grau 2 é uma árvore (como ilustra a Figura 5(a)) e o modelo WS com grau 2 é um círculo que foi quebrado em partes menores devido ao seu processo de reconexão[3], dando origem a muitas árvores de borda longas com apenas um ramo (tais árvores são também *caudas* de vértices, conforme será mostrado na Seção 5.2). Já no caso do modelo GW, tanto a profundidade quanto o número de ramos tende a cair à medida que o grau médio dos vértices aumenta. A presença de árvores de borda grande nesse modelo é fruto da sua evolução que leva em conta a distribuição espacial dos vértices. Neste caso, vértices que estão próximos à borda da caixa tendem a estabelecer poucas conexões. As árvores de borda aparecem nesse modelo como uma conseqüência desse efeito e quanto menor o grau médio, maior a chance de aparecer árvores de borda grandes e com muitas folhas. À medida que o grau médio aumenta, também aumenta o coeficiente de aglomeração

[3]Note que este processo de reconexão é do próprio modelo e não o que torna as redes aleatórias.

Tabela 4: Caracterização das árvores de borda das redes consideradas em termos do *Z-score* (Z) calculado para intervalos específicos de profundidade e do número de folhas. O número de árvores encontradas em cada rede considerada no intervalo analisado é indicado por N_{real}; $N_{\text{aleatoria}}$ é o correspondente para a média das árvores de borda nas versões aleatórias de cada rede (o número entre parêntesis é o desvio padrão). Os intervalos foram obtidos de forma a garantir que a distribuição de profundidade ou do número de folhas está acima ou abaixo da distribuição da média das respectivas versões aleatórias. Tabela extraída de [3].

	Profundidade				**Número de folhas**			
Redes	**Intervalo**	N_{real}	$N_{\text{aleatoria}}$	Z	**Intervalo**	N_{real}	$N_{\text{aleatoria}}$	Z
BA $\langle k\rangle = 2$	1–7	0	166.0 (0.8)	-201.1	1–634	0	166.3 (0.8)	-203.1
	8–11	1	0.2 (0.1)	14.6	635–684	1	0.0 (0.0)	ND
WS $\langle k\rangle = 2$	1–5	10	164.7 (0.8)	-186.3	1–5	16	165.8 (0.8)	-184.6
	6–152	11	1.6 (0.1)	77.9	6–157	5	0.5 (0.1)	65.0
GW $\langle k\rangle = 2$	1–3	134	159.1 (0.9)	-28.3	1–2	131	156.4 (0.9)	-27.5
	4–20	11	7.8 (0.3)	11.0	3–18	13	10.5 (0.3)	8.7
GW $\langle k\rangle = 4$	1–1	61	63.8 (0.7)	-4.6	1–1	66	66.3 (0.7)	-0.4
	2–8	10	5.5 (0.2)	20.6				
GW $\langle k\rangle = 6$	1–4	17	12.9 (0.3)	11.8	1–4	17	12.9 (0.3)	11.8
Astrofísica	1	755	1142 (13)	-30.8	1–2	755	1128 (16)	-23.8
	2–3	42	13.7 (3.4)	8.3	3–6	42	28.2 (4.8)	2.9
Netscience	1–5	81	233.1 (7.4)	-20.5	1–8	81	233.1 (7.4)	-20.5
Cond-mat	1	1971	3062 (23)	-47.6	1	1561	2666 (31)	-36.2
	2–3	124	73.2 (8.6)	5.9	2–8	534	468.5 (13.2)	5.0
Física de alta	1–5	827	1277 (16)	-29.1	1–6	827	1276 (16)	-28.8
energia	6–6	1	0.0 (0.1)	9.9	7–7	1	0.5 (0.7)	0.7
Roget	1	31	58.7 (2.1)	-13.5	1	37	57.5 (2.9)	-7.1
	2–3	11	1.2 (1.1)	9.0	2	5	2.4 (1.3)	2.0
Wordnet	1–5	9245	13304 (51)	-79.2	1–3	6323	11041 (62)	-76.2
	6–7	3	1.5 (1.3)	1.2	4–208	2925	2264 (23)	28.7
WWW	1–2	16762	48108 (108)	-290.9	1–8	13402	45449 (115)	-278
	3–21	308	49.6 (6.2)	41.9	9–5324	3668	2709 (24)	40.3
David Copperfield	1–2	30	34.9 (1.5)	-3.3	1	24	33.0 (2.8)	-3.2
					2–3	6	1.9 (1.3)	3.0
Night and Day	1–2	14	18.5 (0.7)	-6.3	1	11	18.0 (1.3)	-5.2
					2–3	3	0.5 (0.6)	4.0
On the origin of	1–2	12	12.7 (0.6)	-1.3	1	11	12.4 (1.1)	-1.3
species					2	1	0.3 (0.5)	1.5
Internet	1–3	1712	1427 (22)	12.9	1–214	1710	1423 (22)	13.0
					215–492	2	4.2 (0.8)	-2.9
Aeroportos dos	1–3	26	34.0 (2.8)	-2.8	1–2	19	29.4 (3.7)	-2.8
EUA					3–12	7	4.6 (1.3)	1.8
Transmissão	1–4	794	836 (14)	-3.1	1–2	715	797 (16)	-5.3
elétrica	5–5	9	6.1 (2.5)	1.2	3–12	90	48.5 (5.6)	7.4
Rede neural	1–2	3	13.3 (1.4)	-7.5	1–3	1	13.2 (1.5)	-8.4
					4–10	2	0.1 (0.2)	8.9
Transcrição	1–7	44	65.3 (4.3)	-5.0	1–8	37	62.0 (4.6)	-5.5
genética					9–21	7	2.4 (1.2)	3.9
					22–41	0	1.0 (0.2)	-5.7
Interação de	1–2	549	771 (15)	-14.7	1–19	560	784 (14)	-15.7
proteínas	3–4	21	14.7 (3.4)	1.9	20–75	10	1.3 (0.5)	16.1

que, por sua vez, faz diminuir a chance de aparecer árvores de borda profundas e com muitas ramificações.

A maioria das redes reais exibe mais árvores de borda profundas e muito ramificadas do que suas versões aleatórias. Essa situação é observada para as redes de colaboração de astrofísica, de matéria condensada e física de alta energia. Entretanto as árvores de borda de tais redes não

são nem muito profundas nem muito ramificadas. Uma possível explicação para tal descoberta é que a maioria dos pesquisadores das áreas consideradas nessas redes colaboram com alguns pesquisadores de outras áreas. Como os pesquisadores de outras áreas provavelmente publicaram poucos artigos nessas redes, eles possuem poucas conexões com outros pesquisadores, podendo inclusive ser folhas das árvores de borda. Em outras palavras, a falta de inclusão de dados pode ter gerado essas estruturas de árvores nas bordas dessas redes, o que leva a acreditar que tais redes são incompletas.

No caso das redes de informação há um número considerável de árvores de borda profundas e com muitas ramificações, embora numa escala menor para a rede de Roget. A rede WWW chama atenção devido à quantidade de árvores de borda com profundidades elevadas e número expressivo de ramos. Tal estrutura de conectividade é provável ser uma conseqüência da amostragem pequena da teia mundial (como já identificado em [4]), em que as folhas podem ter conexões com outras páginas fora da rede analisada.

As redes de livros não apresentam quantidades significativas de árvores de borda por causa da maneira como foram construídas. Essas redes são construídas de maneira seqüencial o que favorece a geração de ciclos e dificulta o aparecimento de árvores de borda.

Já as redes tecnológicas são bem distintas das demais, pois são as únicas que possuem quantidades significativas de árvores de borda pequenas e com muitas ramificações. No caso da Internet esse fenômeno pode ser explicado pela estrutura de *água-viva* desta rede, que possui muitos vértices conectados ao núcleo central [38]. No caso da rede de transmissão de energia elétrica, há muitas árvores de borda pequenas, mas com muitas ramificações. Esse efeito deve estar relacionado com o fato de que novos vértices necessários para cobrir uma determinada região tendem a se conectar ao vértice de distribuição de energia elétrica mais próximo. Situação semelhante é observada para a rede de aeroportos em que pequenos aeroportos se conectam na sua maioria com aeroportos maiores.

No caso das redes biológicas, elas apresentam uma quantidade relativamente grande de folhas. A explicação para tal efeito na rede neural considerada é similar àquela da modelo teórico de rede geográfica. Na rede de transcrição genética, a grande quantidade de pequenas árvo-

res de borda é uma conseqüência do fato de que alguns genes participam da regulação de uma grande quantidade de outros. Uma situação semelhante é observada para rede de interação de proteínas, em que proteínas mais específicas mantém conexão com aquela da qual foi derivada, dando origem a uma grande quantidade de árvores de borda com muitas ramificações.

Observe ainda que muitas redes, tais como WWW, Wordnet e modelos BA e WS com grau médio 2, apresentaram valores do *Z-score* negativos de módulo elevado para árvores de borda pouco profundas e com poucas folhas. Este efeito pode ser explicado pelo fato de que, nas redes originais havia muitas árvores de borda grandes (profundas e com muitas folhas) que foram divididas em árvores menores pelo processo de reconexão.

5.1.3 Análise local

Em algumas das redes consideradas, os vértices são identificados por *rótulos* de modo que torna possível realizar uma análise funcional das árvores de borda. Tais redes são: Roget, Wordnet, a rede de aeroportos dos EUA, e a rede de interação de proteínas.

No caso da rede Roget, duas palavras i e j estão conectadas se elas estiverem diretamente relacionadas. Desta maneira, as palavras das posições mais inferiores das árvores de borda tendem a ser mais específicas. Além disso, palavras que pertencerem ao mesmo ramo estão associadas. Por exemplo, há uma árvore de borda cuja raiz é a palavra "*demon*". Os filhos dessa palavra são "*Jupiter*" e "*Satan*" que não possuem conexão entre si. "*Satan* " ainda possui um filho que é "*Angel*" que também não possui conexão com "*Jupiter*". A maioria das árvores de borda são árvores sem ramificações (apenas um ramo). Tais árvores também são conhecidas como caudas de vértices (Seção 5.2).

A Wordnet é outro tipo de rede de informação, cujas árvores de borda apresentam estruturas similares às da rede Roget. Um bom exemplo nesta rede é a árvore de borda com a palavra "*sport* " como raiz. No nível abaixo da raiz estão as palavras "*archery*", "*team sport*", "*cycling*", "*nonresident*", "*sledding*", "*skating*" e "*racing*". Estas palavras não possuem relação entre si, mas estão associadas à palavra "*sport*". Entre as ramificações, a palavra "*cycling*" possui ligação com "*bicycling*", "*motorcycling*" e "*dune cycling*"; a palavra "*skating*" está conectada a "*roller*

skating", "*skateboarding*" e "*ice skating*". Como pode ser notado, estas últimas palavras, que pertencem a ramos diferentes, não possuem relações semânticas entre si e nem com as outras de ramos diferentes. Este exemplo reforça a observação de que níveis mais baixos da árvore de borda correspondem a palavras mais específicas.

A rede de transporte aéreo dos EUA é composta basicamente por três tipos de aeroportos: internacionais, regionais e pequenos. Como observado nas análises anteriores (Tabela 4), esta rede possui árvores de borda com muitas folhas e não muito profundas (toda árvore de borda possui profundidade unitária). A análise dos *rótulos* dos aeroportos que compõem tais árvores revelou que a maioria delas possui como raiz os aeroportos internacionais (ao todo são 20 de 26 árvores de borda) e como folhas, aeroportos pequenos e regionais (de 55 folhas, apenas 5 são internacionais). Entre as exceções de aeroportos pequenos ou regionais que são raízes de árvores de borda está o aeroporto Bethel, localizado no Alaska, mas que possui propriedades de aeroporto internacional, pois possui conexão com outros estados dos EUA. Desta análise local, conclui-se que há uma alta relação entre a importância do aeroporto com a sua posição na árvore de borda.

No caso da rede de interação de proteínas do *S. cerevisiae*, as árvores de borda são compostas por proteínas que possuem funções similares acordo com a sua posição na árvore de borda, onde proteínas com funções parecidas tendem a estar conectadas, como sugerido pela regra da maioria [119]. Por outro lado, proteínas que pertencem a ramos diferentes tendem a ter funções diferentes. Além disso, a proteína na raiz da árvore de borda tende a desempenhar funções mais gerais que as outras [3]. As folhas são, portanto, as mais específicas. Dois exemplos de tais estruturas podem ser visualizados na Figura 18.

Os resultados apresentados nesta seção indicam que as folhas apresentam funções mais específicas do que as raízes das árvores de borda. Portanto, enquanto *hubs* desempenham funções mais gerais na rede, as folhas das árvores de borda correspondem a uma situação oposta.

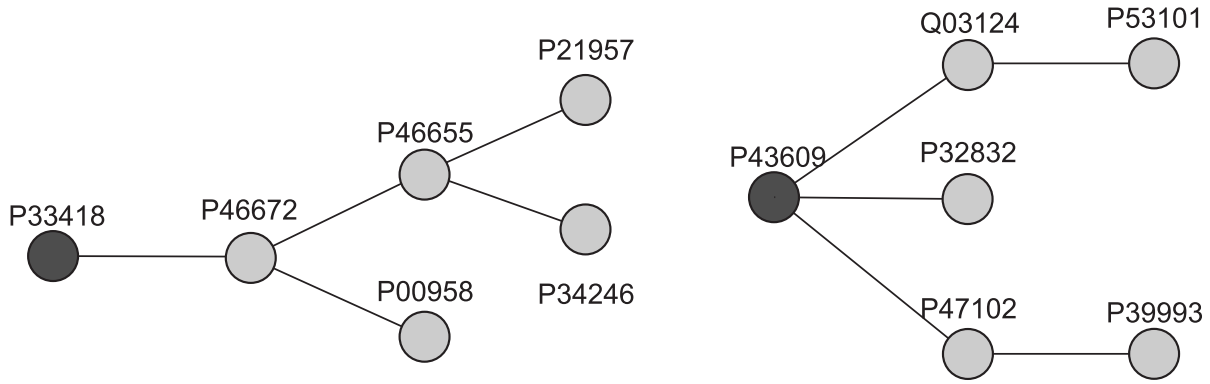


Figura 18: Exemplos de árvores de borda na rede de interação de proteínas do *S. cerevisiae*. As raízes (indicadas por cinza escuro) são proteínas com funções mais gerais. Figura extraída de [3].

5.2 Cadeias de vértices

Cadeias de vértices são motivos compostos por vértices conectados de forma seqüencial, em que os vértices internos possuem apenas dois vizinhos. Em outras palavras, a seqüência de vértices $(n_1, n_2, \ldots, n_{m+1})$ é considerada uma cadeia de tamanho m se ela tiver as seguintes propriedades:

1. Há uma aresta entre os vértices n_i e n_{i+1}, $1 \leq i \leq m$;
2. Os vértices n_1 e n_{m+1} possuem grau diferente de 2; e
3. Vértices intermediários n_i, $2 \leq i \leq m$, se existirem, possuem grau 2.

Os vértices n_1 e n_{m+1} são chamados de *extremidades* da cadeia.

Quanto às suas extremidades, as cadeias de vértices podem se classificadas em quatro categorias (k_{n_i} é o grau do vértice n_i):

Cordões são cadeias com $k_{n_1} = 1$ e $k_{n_{m+1}} = 1$.

Caudas são cadeias com $k_{n_1} = 1$ e $k_{n_{m+1}} > 2$ (ou equivalentemente $k_{n_1} > 2$ e $k_{n_{m+1}} = 1$).

Alças são cadeias com $k_{n_1} > 2$ e $k_{n_{m+1}} > 2$.

Anéis (de comprimento m) são seqüências $(n_1, n_2, \ldots, n_m)$ de m vértices onde o grau de

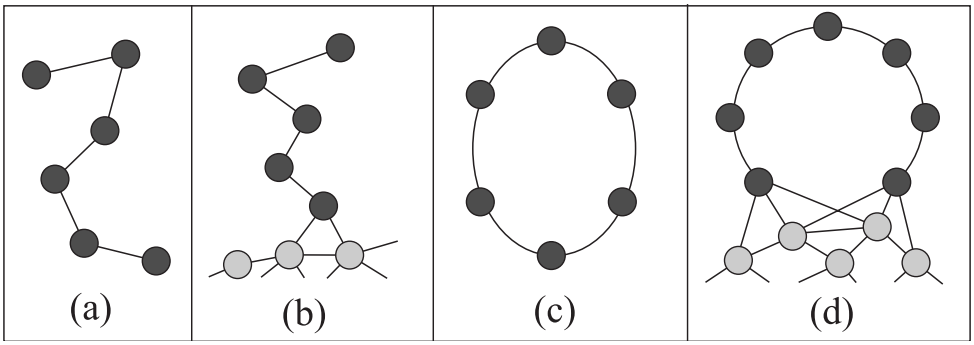


Figura 19: As cadeias de vértices (vértices mais escuros) podem ser classificadas em diferentes classes conforme o tipo de extremidade. Aqui é mostrado 4 tipos de cadeias de vértices: (a) um cordão, (b) uma cauda, (c) um anel, e (d) uma alça. Figura adaptada de [4].

cada vértice é $k_{n_i} = 2,\ 1 \leq n \leq m$, n_i é adjacente a n_{i+1} (para $1 \leq i \leq m-1$), e n_m é adjacente a n_1.

Anéis são casos especiais e raros de cadeias nos quais não há extremidade e foram incluídos na classificação de cadeias apenas por completeza, mas não serão considerados nas próximas análises, pois não ocorrem com freqüência em redes reais. Outros motivos baseados em cadeias de vértices podem ser definidos, como por exemplo n-alças, $n \geq 2$ alças conectadas ao restante da rede pelas mesmas extremidades. O mesmo vale para as n-caudas. Entretanto as n-alças e as n-caudas não serão consideradas neste trabalho. A Figura 19 ilustra os tipos básicos de cadeias de vértices.

Incluindo o caso trivial de $m = 1$, é fácil notar que cada vértice de grau 1 é uma extremidade de um cordão ou de uma cauda e que cada vértice de grau maior do que 2 está na extremidade de uma cauda ou de uma alça. Note que esta definição inclui o caso degenerado em que as extremidades são as mesmas: $n_1 = n_{m+1}$.

5.2.1 Algoritmo

O algoritmo para encontrar cadeias de vértices em redes complexas envolve duas etapas: uma para encontrar cadeias de vértices maiores que 1 e outra para encontrar as cadeias de tamanho unitário. A primeira etapa está ilustrada na Figura 20 e é esquematizada no algoritmo 1.

A lista L obtida no algoritmo 1 contém todas as cadeias de tamanho maior ou igual a 2. De

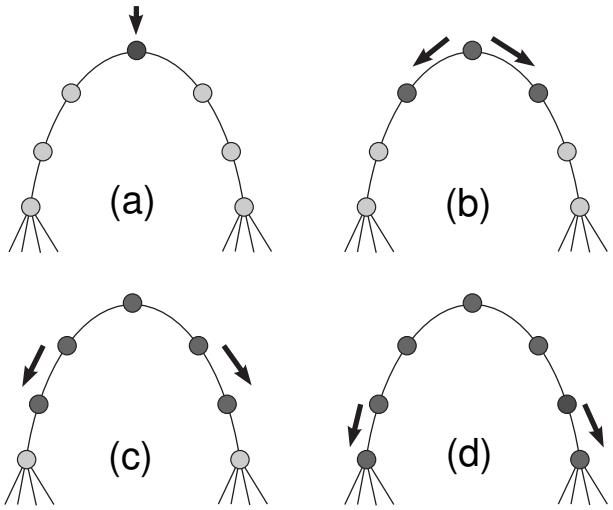


Figura 20: A identificação de alças de tamanho maior que 2 envolve os seguintes passos: (a) escolher arbitrariamente um vértice de grau 2 (vértice em cinza escuro) e adicioná-lo a uma lista; (b) ir até os seus vizinhos e também adicioná-los a essa lista se eles tiverem grau 2; (c) continuar com os próximos vizinhos, excluindo os que já foram visitados, e adicioná-los se eles também tiverem grau 2; e (d) parar de adicionar vértices quando encontrar dois vértices com grau maior que 2. No caso ilustrado, o tamanho da alça é 6. O mesmo procedimento pode ser utilizado para encontrar caudas e cordões, mas, pelo menos, uma das extremidades precisa ter grau igual a 1. Figura adaptada de [4].

acordo com o grau dos vértices nas extremidades, as cadeias de vértices da lista L podem ser classificadas em cordões (as duas extremidades têm grau 1), caudas (uma das extremidades tem grau 1 e a outra grau maior que 2), alças (as duas extremidades têm grau maior que 2) e anéis (se as extremidades forem o mesmo vértice).

A segunda etapa, necessária para encontrar as cadeias de tamanho unitário, está esquematizada no algoritmo 2. Os resultados desse algoritmo: C_1, T_1 e H_1 contém todas os cordões, caudas e alças de tamanho unitário da rede analisada.

5.2.2 Previsão teórica

Seja N_C, N_T, N_H, e N_R o número total de cordões, de caudas, de alças e de anéis, respectivamente, e $N(k)$ o número de vértices com grau k, as seguintes expressões são obtidas:

$$N(1) = 2N_C + N_T, \tag{5.2}$$

$$\sum_{k>2} kN(k) = 2N_H + N_T. \tag{5.3}$$

Entrada: Rede complexa R
Saída: Lista L contendo todas as cadeias de vértices maiores ou iguais a 2
Calcule o grau de todos os vértices de R e os guarde em uma lista K
Encontre os vértices i tais que $k_i = 2$, $k_i \in K$, e os guarde em uma pilha Q_2
para todo *elemento de* Q_2 **faça**
  Remova o primeiro elemento de Q_2 e insira seu primeiro vizinho, em seguida ele mesmo, e depois o segundo vizinho em uma fila P, mantendo essa ordem
  enquanto *primeiro e último elementos de* P *não tiverem grau igual a* 2 *ou não forem o mesmo vértice* **faça**
    se *primeiro elemento de* P *tiver grau 2* **então**
      Seja A e B os vizinhos do primeiro elemento de P
      se $A \notin P$ **então**
        Inclua A na primeira posição de P
        se $A \in Q_2$ **então**
          Remova A de Q_2
        fim
      fim
      senão
        Inclua B na primeira posição de P
        se $B \in Q_2$ **então**
          Remova B de Q_2
        fim
      fim
    fim
    se *último elemento de* P *tiver grau 2* **então**
      Seja C e D os vizinhos do último elemento de P
      se $C \notin P$ **então**
        Inclua C na última posição de P
        se $C \in Q_2$ **então**
          Remova C de Q_2
        fim
      fim
      senão
        Inclua D na última posição de P
        se $D \in Q_2$ **então**
          Remova D de Q_2
        fim
      fim
    fim
  fim
  Insira P em uma lista L e apague P
fim

Algoritmo 1: Algoritmo para encontrar cadeias de vértices maiores ou iguais a 2.

Para calcular o número de vértices com grau 2, são introduzidas as seguintes notações: $N_C(m)$ para o número de cordões de tamanho m e similarmente $N_T(m)$ para caudas, $N_H(m)$ para alças e $N_R(m)$ para anéis. Cada cadeia de vértices de tamanho m tem $m-1$ vértices de

Entrada: Rede complexa G, lista K do grau dos vértices e lista L com as cadeias de vértices de tamanho maior ou igual a 2
Saída: Lista de cordões C_1, caudas T_1 e alças H_1 de tamanho 1
Encontre todos os vértices de grau igual a 1 e guarde em uma pilha Q_1 somente aqueles que não estão em L
para todo *elemento de* Q_1 **faça**
  Remova o primeiro elemento A de Q_1 e o insira em uma fila P_1
  se *o vizinho de* A *também tiver grau* 1 **então**
    Remova-o de Q_1, insira-o em P_1 e insira P_1 em uma lista C_1
  fim
  senão
    Insira-o em P_1 e insira P_1 na lista T_1
  fim
  Apague P_1
fim
Inclua todos os pares de vértices conectados que não estão em L, C_1 ou T_1 na lista H_1

Algoritmo 2: Algoritmo para encontrar alças, caudas e cordões de tamanho unitário.

grau 2 enquanto que cada anel tem m vértices de grau 2. Portanto,

$$N(2) = \sum_{m=1}^{\infty} \left[mN_R(m) + (m-1)\left(N_C(m) + N_H(m) + N_T(m)\right)\right] \tag{5.4}$$

Sabendo as correlações de grau de uma rede, $P(k,k')$, as distribuições de grau $P(k)$ e a condicional $P(k'|k)$ (i.e. a probabilidade de que um vizinho de um vértice com grau k tenha grau k') podem ser calculadas como:

$$P(k) = \frac{\sum_{k'} P(k,k')/k}{\sum_{k',k''} P(k',k'')/k'}, \tag{5.5}$$

$$P(k'|k) = \frac{\langle k \rangle P(k,k')}{kP(k)}, \tag{5.6}$$

onde $\langle k \rangle = \sum_k kP(k)$ é o grau médio da rede.

5.2.2.1 Anéis

Um anel de tamanho m é formado por apenas vértices com grau 2. A probabilidade de transição entre os seus vértices, exceto entre o último e o primeiro que fecham o anel, é igual

a $P(2|2)$. Para fechar o anel, o último vértice tem que se conectar com um vértice de grau 2, cuja probabilidade é a mesma que nos outros casos, e este vértice tem que ser exatamente o primeiro. A probabilidade disso acontecer é $1/NP(2)$. Se todos os vértices forem considerados como início, cada anel será contado m vezes, e, portanto:

$$N_R(m) = \frac{1}{m} P(2|2)^m. \tag{5.7}$$

Tal expressão é válida apenas quando m é pequeno e N é grande de maneira que os vértices incluídos nos anéis não alterem significativamente as probabilidades condicionais. Esta aproximação será utilizadas nas análises seguintes. Repare que tal expressão resulta em valores muitos pequenos, que, na maioria dos casos, podem ser ignorados.

5.2.2.2 Cordões

Cordões são cadeias de vértices que começam e terminam com vértices de grau unitário, podendo ou não ter vértices no meio. O número de cordões de tamanho m é dado pela expressão:

$$N_C(m) = \begin{cases} \frac{1}{2} NP(1)P(1|1) & \text{if } m = 1, \\ \frac{1}{2} NP(1)P(2|1)P(2|2)^{m-2}P(1|2) & \text{if } m > 1. \end{cases} \tag{5.8}$$

5.2.2.3 Caudas

A cauda é como um cordão, mas o vértice de uma de suas extremidades tem grau maior que 2. Similarmente ao cálculo dos cordões, o número de caudas de tamanho m é dado por:

$$N_T(m) = \begin{cases} NP(1)P(>2|1) & \text{if } m = 1, \\ NP(1)P(2|1)P(2|2)^{m-2}P(>2|2) & \text{if } m > 1, \end{cases} \tag{5.9}$$

onde $P(>2|k) = \sum_{k'>2} P(k'|k)$.

5.2.2.4 Alças

As alças são cadeias de vértices que começam e terminam em vértices com grau maior que 2, podendo ou não ter vértices no meio. Começando em um dos $NP(k)$ vértices de grau $k > 2$, há k possibilidades de seguir uma cadeia, cada uma sendo uma seqüência de vértices

de grau 2, até chegar em um vértice de grau $k' > 2$. Alças de tamanho unitário são dadas por $NkP(k)P(> 2|k)$ e as alças de tamanho $m > 1$, por $NkP(k)P(2|k)P(2|2)^{m-2}P(> 2|2)$. Somando para todos os valores de $k > 2$, usando $\sum_k kP(k)P(k'|k) = k'P(k')$, que pode ser deduzida das expressões (5.5) e (5.6), e considerando que cada alça é contada duas vezes quando todos os vértices com grau maior do que 2 são levados em conta, o número de alças de tamanho m é dado por:

$$N_H(m) = \begin{cases} \frac{1}{2}N\left\{\langle k\rangle - P(1)[2 - P(1|1) - P(2|1)] - P(2)[4 - P(1|2) - P(2|2)]\right\} & \text{if } m = 1, \\ \frac{1}{2}N[2P(2) - P(1)P(2|1) - 2P(2)P(2|2)]P(2|2)^{m-2}P(> 2|2) & \text{if } m > 1. \end{cases} \tag{5.10}$$

Juntando as Equações (5.8), (5.9), e (5.10), a seguinte expressão é obtida:

$$\sum_{m=1}^{\infty} \left[(m-1)\left(N_C(m) + N_H(m) + N_T(m)\right)\right] = N(2).$$

Note que o número de anéis não aparece nesta expressão, em contraste com a Equação (5.4), pois foi ignorado o fato de que a presença de anéis na rede diminui o número de cadeias no total. Entretanto, esse efeito pode ser ignorado em redes grandes, validando a aproximação. Note também que o tamanho das cadeias são proporcionais a $P(2|2)^m$. Por causa disso, cadeias grandes de vértices devem ser exponencialmente raras, a menos que sejam favorecidas de alguma forma, como no processo de evolução da rede.

5.2.3 Análise teórica para redes não correlacionadas

No caso de redes sem correlação de grau, a distribuição $P(k, k')$ pode ser expressa como:

$$P(k, k') = \frac{kP(k)k'P(k')}{\langle k\rangle^2}, \tag{5.11}$$

enquanto que a probabilidade condicional simplificada a:

$$P(k'|k) = \frac{k'P(k')}{\langle k\rangle}. \tag{5.12}$$

Com a última expressão, obtêm-se para redes sem correlação de grau, as relações:

$$N_R(m) \;=\; \frac{1}{m}\left[\frac{2P(2)}{\langle k\rangle}\right]^m \tag{5.13}$$

$$N_C(m) = \frac{2^{m-2}NP(1)^2P(2)^{m-1}}{\langle k\rangle^m} \quad (5.14)$$

$$N_T(m) = NP(1)\left[\frac{2P(2)}{\langle k\rangle}\right]^{m-1}\alpha \quad (5.15)$$

$$N_H(m) = \frac{N\langle k\rangle}{2}\left[\frac{2P(2)}{\langle k\rangle}\right]^{m-1}\alpha^2. \quad (5.16)$$

onde $\alpha = \left[1 - \frac{P(1)}{\langle k\rangle} - \frac{2P(2)}{\langle k\rangle}\right]$.

5.2.3.1 Redes aleatórias de Erdős e Rényi

Este tipo de rede não possui correlação de grau e tem distribuição de graus dada por:

$$P(k) = \frac{e^{-\langle k\rangle}\langle k\rangle^k}{k!}. \quad (5.17)$$

Utilizando esta expressão nas Equações (5.13), (5.14), (5.15) e (5.16), obtêm-se, respectivamente, para o número de anéis, de cordões, de caudas e de alças:

$$N_R(m) = \frac{\langle k\rangle^m e^{-m\langle k\rangle}}{m} \quad (5.18)$$

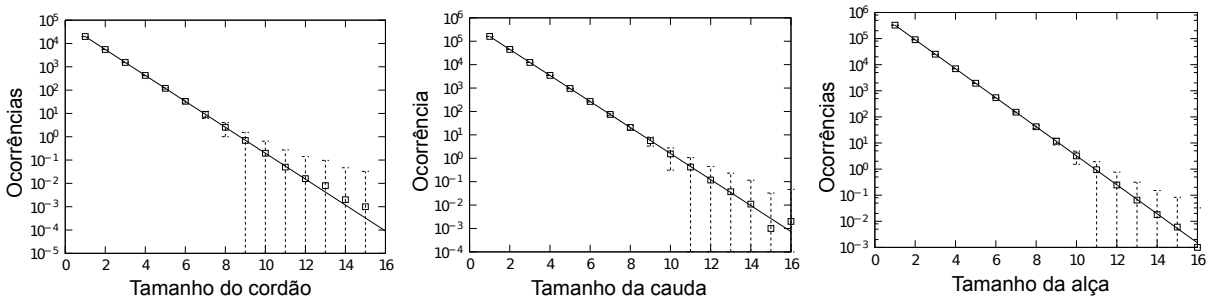
$$N_C(m) = \frac{N}{2}\langle k\rangle^m e^{-(m+1)\langle k\rangle} \quad (5.19)$$

$$N_T(m) = N\langle k\rangle^m e^{-(m+1)\langle k\rangle}\varepsilon \quad (5.20)$$

$$N_H(m) = \frac{N}{2}\langle k\rangle^m e^{-(m+1)\langle k\rangle}\varepsilon^2 \quad (5.21)$$

onde $\varepsilon = \left(e^{\langle k\rangle} - \langle k\rangle - 1\right)$. A comparação dos resultados analíticos com os simulados para redes aleatórias de Erdős e Rényi com $N = 10^6$ vértices e $\langle k\rangle = 1.95$ pode ser visualizada na Figura 21. Foram geradas 1000 redes desse modelo para calcular as médias e os desvios padrões. A Figura 21 não mostra nenhum resultado para os anéis, porque eles não foram encontrados.

Os resultados apresentados nesta seção serviram apenas para validar a teoria para este modelo teórico. Nas próximas seções, a teoria será investigada em redes do mundo real.



(a) **(b)** **(c)**

Figura 21: Número de cordões (a), caudas (b) e alças (c) de diferentes tamanhos em redes geradas pelo modelo aleatório de Erdős e Rényi. Os pontos representam médias dos valores medidos para 1000 realizações e as barras de erros são os correspondentes desvios padrões. Os valores analíticos são as retas. Note que o aumento abrupto da largura das barras de erro é uma conseqüência da escala logarítmica..

5.2.4 Análise de redes reais

As redes reais foram analisadas de duas formas: uma através da comparação dos valores obtidos para essas redes com modelos nulos (gerados a partir do processo de reconexão, como descrito na Seção 5.1.2) e outra através da comparação com a teoria proposta na Seção 5.2.2.

No primeiro caso, foi utilizado o mesmo procedimento para análise das árvores de borda nas redes consideradas. Para cada rede real, foi gerado um conjunto de 1000 redes aleatórias pelo processo de reconexão e a quantidade de cordões, caudas e alças obtidas para a rede considerada foi comparada com a de suas redes aleatórias através do "*Z-score*", Equação (5.1). Neste caso, X_{Real} é o número de cordões, caudas ou alças da rede original analisada com um tamanho específico e $\langle X \rangle$ e σ são, respectivamente, a média e o desvio padrão dos valores correspondentes das suas versões aleatórias. Um valor nulo do "*Z-score*" indica que não há diferença estatística entre o número de cordões, caudas ou alças da rede considerada e de suas versões randomizadas. Os valores de "*Z-score*" podem ser visualizados na Figura 22. Valor de "*Z-score*" acima ou abaixo de 0 indica a rede real possui mais cadeias de vértices ou menos que suas versões aleatórias. Os casos em que os valores de "*Z-score*" não são definidos ($\sigma = 0$) são desconsiderados.

A maioria dos resultados apresentados na Figura 22 pode ser explicada pelo fato de que o processo de reconexão tende a uniformizar a distribuição dos tamanhos de cordões, caudas e

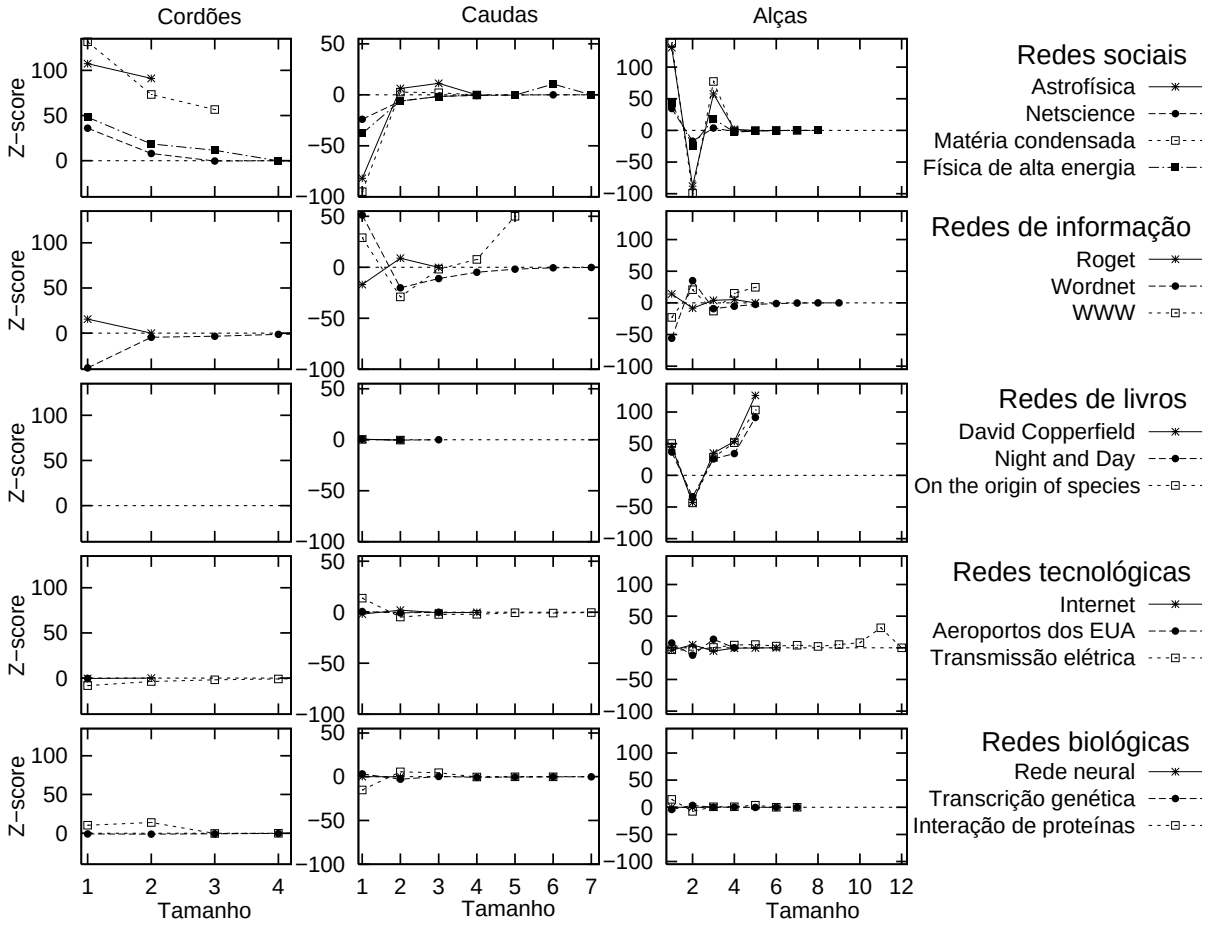


Figura 22: Valores de "*Z-score*" do número de cordões, de caudas e de alças para cada comprimento. O número de redes aleatórias geradas foi de 1000 para todas as redes consideradas, exceto para a rede WWW que foi 100 (devido ao tamanho desta rede). Figura extraída de [4].

alças. Desta forma, o excesso de uma dessas estruturas nas redes reais tenderá a diminuir nas suas versões aleatórias. Por exemplo, se uma rede possui várias alças grandes, as suas versões aleatórias apresentarão poucas dessas estruturas, mas várias pequenas. As próximas discussões não levarão em conta a forma das distribuições de cadeias de vértices, mas sim apenas alguns resultados mais relevantes.

No caso das redes de colaboração, existe uma quantidade grande de cordões. Estes resultados são conseqüência do fato de haver pesquisadores que publicaram artigos com apenas um, dois ou três outros autores. Isto corresponde ao caso em que estes pesquisadores são de outras áreas não incluídas na rede analisada e, portanto, não têm conexões com o restante da rede. Se outras áreas de pesquisa fossem adicionadas à rede analisada, provavelmente o resultado seria diferente. Isto leva a concluir que a presença de cordões nessas redes é resultado da falta de

dados, ou seja, incompleteza da rede analisada.

As redes de informação não apresentam um padrão bem definido como ocorre com as redes de colaboração. A rede de dicionário Roget é diferente das demais, porém seus resultados não são muito expressivos para serem discutidos. Importante notar que nas redes Wordnet e WWW, há um quantidade grande de caudas de tamanho unitário. No caso da Wordnet, que é uma rede de palavras, isso acontece por causa das palavras específicas que só possuem ligação com palavras mais gerais, e estas possuem conexões entre si. Já no caso da rede WWW, essa estrutura é uma conseqüência de existir páginas mais específicas dentro da teia mundial. Além dessas caudas pequenas, a rede WWW possui caudas e alças longas. Esse resultado pode estar relacionado com a maneira como tal rede foi obtida, que foi através de um *web crawler* [29], um programa para captura de páginas e dos *links* entre elas. À medida que as páginas são visitadas por por este programa, o caminho que ele percorre pode originar cadeias de vértices. Se o programa não for executado durante um período suficientemente grande, longas cadeias de vértices podem aparecer. Portanto, a presença de caudas e alças longas nessas redes pode ter sido causada pela maneira como a rede WWW foi obtida, logo, ela pode estar incompleta (ver discussão de incompleteza na Seção 5.2.5). Além disso, pelo mesmo motivo, a rede WWW não possui componentes isolados, logo ela não possui cordões.

As redes de adjacência de palavras em livros apresentou um padrão de cadeias de vértices bem definido: nenhum cordão de qualquer tamanho, a mesma quantidade de caudas de tamanho 1, 2 e 3 que as respectivas versões aleatórias e muita alças de tamanho 1, 3, 4 e 5. Como essas redes possuem muitas alças longas, o processo de reconexão quebra essas estruturas em partes menores dando origem às alças de tamanho 2. Por causa do aumento de alças de tamanho 2 nas redes aleatórias, muitos vértices que participavam de alças de tamanho unitário são utilizadas na alças de tamanho 2. Isto explica porque as redes reais de livros possuem mais alças de tamanho unitário que suas respectivas versões aleatórias. Já as alças longas são uma conseqüência direta do método seqüencial de captura das palavras e de suas conexões.

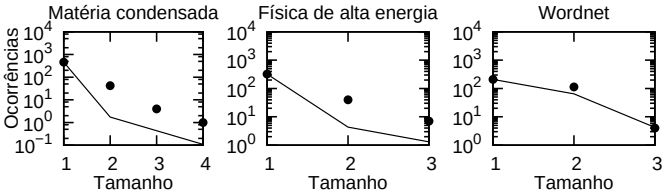
Os únicos resultados interessantes com relação às redes tecnológicas são os obtidos para a rede de transmissão de energia elétrica. Esta rede apresenta uma quantidade relativamente

grande de caudas de tamanho 1 e de alças de tamanho 11. Enquanto o primeiro resultado parece estar relacionado ao fato de que novos vértices necessários para cobrir uma nova região tendem a se conectar com o mais próximo vértice, o segundo pode ter sido originado devido às restrições geográficas, (e.g. os vértices da rede de transmissão podem ter sido colocados em pontos estratégicos de forma a contornar uma montanha ou um lago).

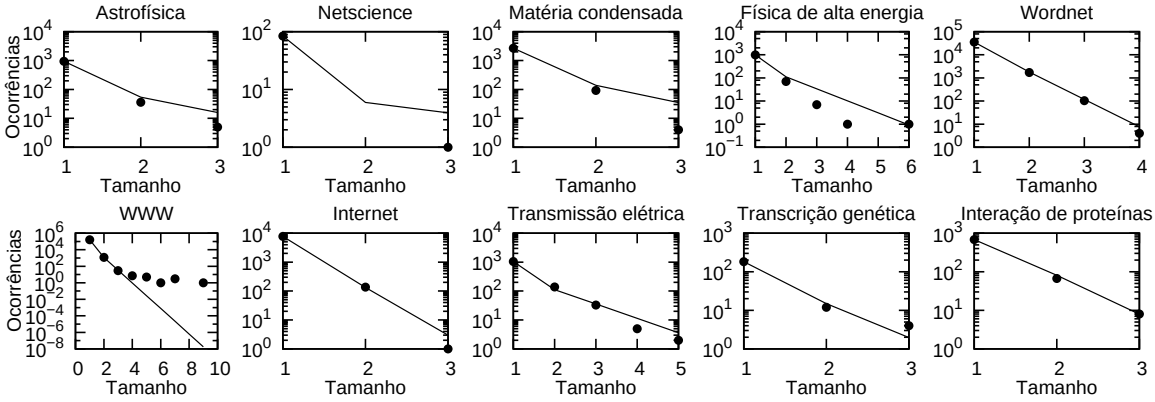
Os resultados para as redes biológicas não foram muito expressivos, porém a rede de interação de proteína de *S. cerevisiae* merece atenção, pois possui uma quantidade relativamente grande de cordões de tamanhos 1 e 2. A presença de cordões pequenos nesta rede é uma conseqüência de cadeias isoladas de proteínas, que só possuem interações com um número pequeno de outras proteínas. A falta de conexões entre as proteínas presentes nos cordões e as outras pode ter sido originada por não considerar outras interações de proteínas [20]. Portanto, esta rede também pode estar incompleta.

A segunda forma de análise das cadeias de vértices foi realizada com a comparação da previsão teórica apresentada na Seção 5.2.2 com os valores obtidos para as redes consideradas. Para cada rede analisada, foi obtida a respectiva correlação de graus e o número esperado de cordões, caudas e alças foi calculado usando, respectivamente, as Equações (5.8), (5.9) e (5.10). O número de anéis foi desconsiderado por causa da baixa probabilidade de ocorrerem em redes reais. Os resultados dessa comparação são mostrados na Figura 23. Os casos não mostrados são aqueles que tiveram menos de dois pontos para serem mostrados. Devido à baixa probabilidade de encontrar cordões nas redes analisada, apenas três tiveram dados para serem mostrados na Figura 23(a): as redes de colaboração da matéria condensada e da física de alta energia e a Wordnet. As previsões teóricas para o número de cordões não corresponderam aos valores observados para essas redes, exceto para a Wordnet. Uma situação oposta foi observada para o número de caudas e alças, indicados, respectivamente, pelas Figuras 23 (b) e (c). No entanto, há mais caudas e alças grandes nas redes analisadas do que a teoria prevê, exceto nas redes de colaboração de astrofísica, matéria condensada e física de alta energia.

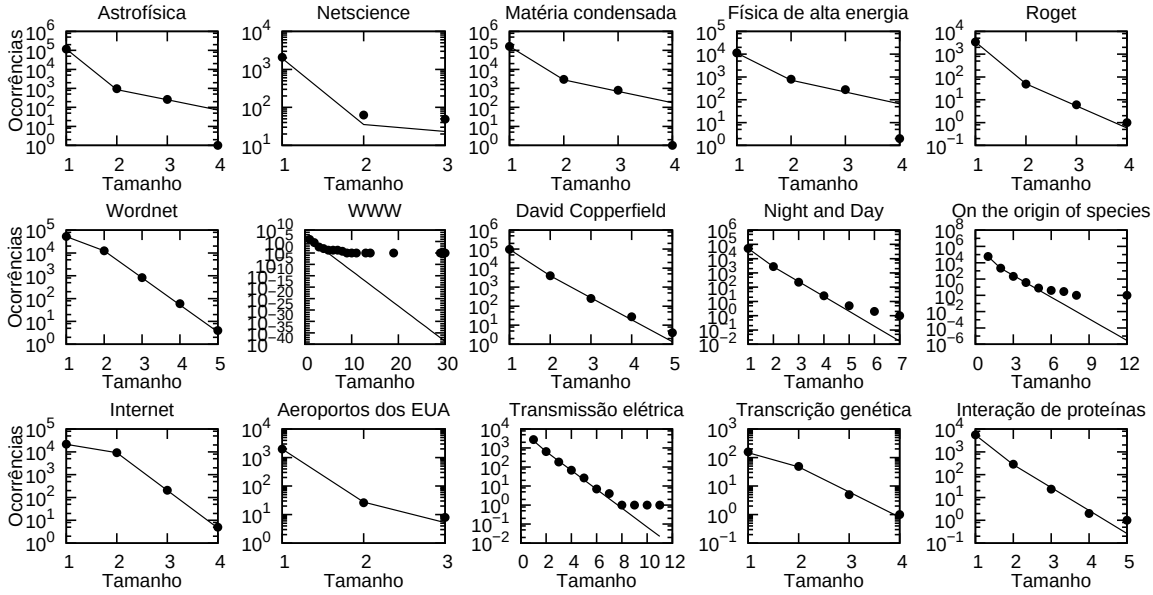
Apesar da diferença entre o número de cordões, caudas e alças pequenas observadas para as redes reais e os valores obtidos para as respectivas versões aleatórias (veja Figura 22) ser



(a) Número de cordões.



(b) Número de caudas.



(c) Número de alças.

Figura 23: Número de cordões (a), de caudas (b) e alças (c) para as redes reais analisadas. Pontos correspondem aos valores das redes consideradas e as linhas contínuas, às predições teóricas. Figura extraída de [4].

muito grande em alguns casos, a previsão teórica foi relativamente precisa para muitos casos, exceto para rede de colaboração de astrofísica (alças), rede de colaboração de física de alta energia (cordões, caudas e alças), WWW (caudas e alças), o livro "On the oririn of species" (alças) e rede de transmissão elétrica (alças) (veja Figura 23). Isso mostra que nessas exceções, os resultados obtidos na Figura 22 não são mera conseqüência da correlação de grau, mas sim

houve algum favorecimento no processo de construção das redes correspondentes. Os casos em que o número de cadeias grandes de vértices (cordões, caudas ou alças) é maior que o valor previsto pela teoria, também é maior do que o valor encontrado para as versões aleatórias. Exemplos incluem: alças e caudas para WWW e para a rede de transmissão de energia elétrica.

5.2.5 Amostragem por caminhada aleatória

A fim de verificar se redes incompletas apresentam caudas e alças grandes, foram obtidas amostras de rede de dois modelos teóricos através de caminhadas aleatórias [111, 17], e as distribuições dessas estruturas nessas redes foram analisadas. Os modelos escolhidos foram: o aleatório de Erdős e Rényi (ER) e o de livre de escala de Barabási e Albert (BA) com 10^5 vértices e grau médio 6. A escolha por caminhada aleatória ocorreu por causa de sua simplicidade e por não requerer qualquer restrição. Além disso, imagina-se que as amostragens de redes reais são realizadas por processos semelhantes às buscas por rotas entre dois vértices como aconteceu com a Internet através do comando `traceroute` [20]. A rede resultante da amostragem é dada pela seqüência de vértices e arestas capturadas no caminho da caminhada aleatória. Os resultados desta análise podem ser visualizados na Figura 24. Cada ponto da grade é uma média dos valores de 1000 realizações para cada comprimento da caminhada aleatória.

Os resultados são muito semelhantes para os dois modelos, exceto quando a caminhada é muito grande (em torno de 10^7) e o número de caudas das redes amostradas do modelo BA tende a 0. Isso ocorre porque a rede original de BA com grau médio 6 não possui caudas de qualquer tamanho e alças maiores que 2. O mesmo efeito não é observado para as redes amostradas do modelo ER pois a rede original apresenta uma quantidade significativa de caudas de tamanhos 1 e 2 (137 caudas de tamanho 1 e 2 de tamanho 2).

Pode ser concluído ainda através da Figura 24 que existem muitas caudas e alças nas redes amostradas quando o comprimento das caminhadas é pequeno, sendo que uma boa parte delas são grandes, especialmente as alças. À medida que é aumentado o comprimento das caminhadas a quantidade destas estruturas diminui, devido ao aumento da probabilidade de quebrar caudas e alças grandes em partes menores. Quanto maior ainda for o tamanho dessas caminhadas (da

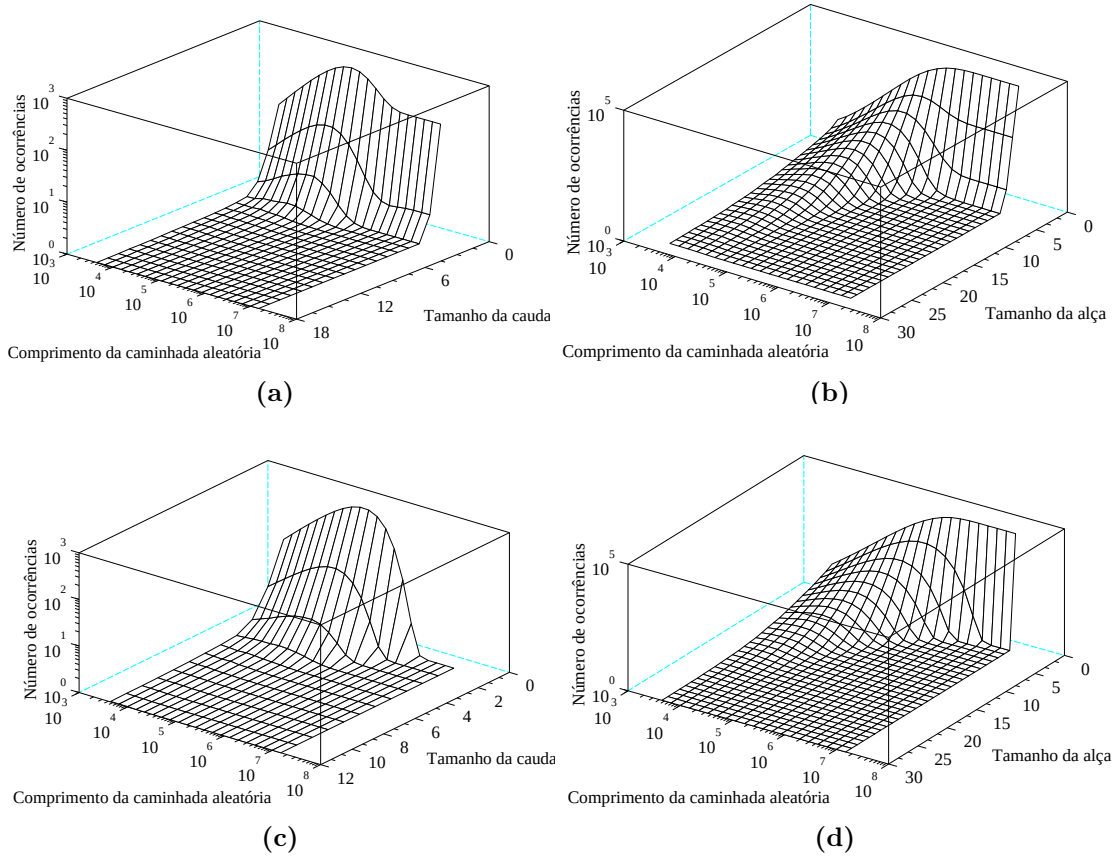


Figura 24: Número de caudas e de alças de diferentes tamanhos no modelo de Erdős e Rényi (a) e (b), respectivamente. Número de caudas e de alças no modelo de livre de escala de Barabási e Albert (c) e (d), respectivamente. Cada ponto da malha do gráfico corresponde ao valor médio de 1 000 realizações para cada comprimento da caminhada aleatória. Figura extraída de [4].

ordem de 10^6 a 10^7), a chance de encontrar essas estruturas tende a 0, exceto quando estiver presente na rede original. A partir de um certo valor para o tamanho da caminhada aleatória (nestes casos foi da ordem de 10^7), a rede original é obtida e suas propriedades recuperadas.

Através dessa análise, conclui-se que redes amostradas tendem a apresentar esse tipo de estruturas, e que, se for por caminhada aleatória, quanto menor for o tamanho da caminhada, maior é a chance de ser encontradas caudas e alças. Desta forma, redes amostradas por métodos como esse apresentam caudas e alças, porém redes com esse tipo de estrutura não necessariamente são incompletas, pois esta estrutura pode estar relacionada com a sua evolução particular.

6 Sensibilidade de medidas

Nos Capítulos 3 e 4, foi mostrado como redes complexas podem ser caracterizadas e classificadas em termos de medidas. Se forem utilizadas várias medidas pouco correlacionadas entre si, mais precisa é a representação de uma dada rede complexa e, portanto, melhor é a sua caracterização. Situação semelhante ocorre com a classificação de redes, já que mais atributos pouco correlacionados contribuem para uma melhor separação dos modelos teóricos, permitindo uma identificação mais precisa da rede analisada. Entretanto, quando não há certeza sobre os dados utilizados para construir as redes do mundo real (seja por falta de dados ou por amostragens incompletas), tanto a caracterização quando a classificação ficam comprometidas. Desta forma, é muito importante saber quais medidas são mais apropriadas a tratar redes com esse tipo de problema[1] de forma a minimizar o erro de caracterização e classificação de redes reais. Neste capítulo, será apresentada uma metodologia para testar a sensibilidade de algumas medidas de redes complexas frente a perturbações nas arestas, tais como: adição, remoção e reconexão aleatória de arestas [4]. Além disso, as medidas serão classificadas conforme o grau de robustez em relação a esse tipo de perturbação e o quanto elas são eficientes para discriminar as redes analisadas.

6.1 Tipos de perturbação

Basicamente, há duas formas de perturbar redes: a primeira corresponde a remover ou adicionar vértices e a segunda, a remover, adicionar ou reconectar arestas. No primeiro caso, porém, é difícil desenvolver um método para adição de vértices, pois não se sabe, *a priori*, o

[1]Note que a maioria das redes do mundo real utilizadas atualmente não estão completas, conforme discutido no Capítulo 1.

mecanismo de formação das redes. Desta forma, somente a perturbação de arestas será coberta aqui. Os três tipos básicos de perturbação de arestas são:

- **Remoção de arestas:** arestas são selecionadas aleatoriamente e removidas da rede;
- **Adição de arestas:** dois vértices são escolhidos aleatoriamente, e, se não estiverem conectados, uma aresta é estabelecida entre eles;
- **Reconexão de arestas:** duas arestas são escolhidas aleatoriamente e, se a troca de extremidades não originar auto e multi-arestas, têm suas extremidades trocadas. Este tipo de perturbação corresponde ao método de reconexão (conhecido como *rewiring*) em que redes aleatórias são obtidas enquanto a distribuição de graus da rede original é mantida (e.g. [117]).

Nas análises seguintes, também foi considerado uma combinação aleatória de tais tipos de perturbações. Todas as perturbações consideradas variaram de 0 a 10% do número total de arestas da rede analisada. No caso da reconexão de arestas, como cada troca de arestas corresponde à perturbação de duas arestas por vez, o número total de passos necessários para completar 10% do número total de arestas da rede foi metade em relação aos outros dois tipos de perturbação.

6.2 Análise em modelos teóricos e redes reais

As perturbações descritas acima foram realizadas tanto em redes de modelos teóricos como em redes do mundo real. Nesta análise, foram considerados quatro modelos: aleatório de Erdős e Rényi, de mundo pequeno de Watts e Strogatz, livre de escala de Barabási e Albert e geográfico de Waxman (maiores detalhes destes modelos podem ser encontrados na Seção 2.4); e cinco redes reais: rede de *emails* da Universidade de Rovira i Virgili (Tarragona), rede de aeroportos dos Estados Unidos, rede de transmissão elétrica dos estados ocidentais dos Estados Unidos, rede neural do *Caernohabditis elegans* e a rede de interação de proteínas do *Saccharomyces cerevisiae* (a descrição destas redes está na Seção 2.3). No caso dos modelos teóricos, os parâmetros utilizados foram: $N = 2000$ vértices, grau médio 6 e probabilidade de reconexão das arestas 0.3 para o modelo de mundo pequeno de Watts e Strogatz. Maiores detalhes

Tabela 5: Propriedades topológicas das redes consideradas antes das perturbações, onde N é o número de vértices; $\langle k \rangle$, o grau médio; $\langle hk_2 \rangle$, grau hierárquico médio de nível 2; r, coeficiente de assortatividade; C, coeficiente de aglomeração médio; $\langle hc_2 \rangle$, coeficiente de aglomeração hierárquico médio de nível 2; $\langle dr_2 \rangle$, razão de divergência média de nível 2; ℓ, caminho mínimo médio; $\langle B \rangle$, grau de intermediação médio; CPD, ponto de dominância central; ER, modelo aleatório de Erdős e Rényi; WS, modelo de mundo pequeno de Watts e Strogatz com probabilidade de reconexão 0.3; BA, modelo livre de escala de Barabási e Albert; e GW, modelo geográfico de Waxman. Tabela extraída de [5].

Redes consideradas	N	$\langle k \rangle$	$\langle hk_2 \rangle$	r	C	$\langle hc_2 \rangle$	dr_2	ℓ	B	CPD
ER	2000	6.0	208	-0.02	0.003	0.003	0.99	4.2	$1.7\ 10^{-3}$	0.008
WS	2000	6.0	97	-0.04	0.213	0.037	0.88	4.9	$2.1\ 10^{-3}$	0.007
BA	2000	6.0	831	-0.05	0.017	0.012	0.98	3.5	$1.4\ 10^{-3}$	0.145
GW	2000	6.0	91	0.20	0.151	0.079	0.81	7.8	$4.3\ 10^{-3}$	0.080
Email	1133	9.6	934	0.08	0.220	0.078	0.80	3.3	$2.3\ 10^{-3}$	0.037
Aeroportos	332	12.8	449	-0.21	0.625	0.176	0.61	2.5	$5.2\ 10^{-3}$	0.202
Transmissão elétrica	4941	2.7	13	0.00	0.080	0.062	0.97	15.9	$3.6\ 10^{-3}$	0.285
Rede neural	297	14.5	618	-0.16	0.292	0.072	0.58	2.2	$4.9\ 10^{-3}$	0.299
Interação de proteínas	4134	4.2	498	-0.13	0.082	0.033	0.93	4.8	$0.8\ 10^{-3}$	0.173

sobre as propriedades das redes analisadas estão na Tabela 5. A análise da perturbação em tais redes não levou em conta nem o peso nem a direção das arestas, logo as redes reais dirigidas foram transformadas nas suas respectivas versões não-dirigidas pela operação de simetrização, descrita na Seção 2.2 (esta transformação equivale a ignorar a direção dos arcos) e as redes reais com peso, nas correspondentes versões sem peso, pela operação de limiarização, também descrita na Seção 2.2, cujo limiar utilizado foi o menor peso encontrado nos arcos da referida rede (esta operação corresponde a ignorar o peso dos arcos). Note que os modelos teóricos não precisam de tais operações, pois já são não-dirigidos e sem peso.

As perturbações foram realizadas de 0.2% até 10% do número total de arestas de cada rede em passos de 0.2%. No total, um conjunto de 50 redes perturbadas para cada um desses passos foi obtida assim como as suas respectivas medidas. As trajetórias definidas pela evolução das perturbações podem ser visualizadas no topo de cada sub-figura (que corresponde a uma medida) das Figuras 25 e 26. Para cada medida, um conjunto de nove gráficos foi obtido, sendo que a primeira linha corresponde aos valores absolutos (média das 50 redes geradas em cada passo de perturbação); a segunda linha é a correspondente variação da primeira linha em relação ao valor original (sem perturbação); e a terceira linha corresponde às projeções dos componentes principais obtidas por PCA (como essas projeções foram obtidas será descrito abaixo). As

Tabela 6: Média e desvio padrão (número em parênteses) da porcentagem de variação de cada medida para os modelos teóricos e para 10% de arestas perturbadas. Os símbolos e os acrônimos são os mesmos da Tabela 5. Tabela extraída de [5].

Medida	Adição	Reconexão	Remoção	Todas
$\langle hk_2 \rangle$	41.8 (24.6)	13.4 (17.2)	23.3 (3.3)	11.5 (12.6)
r	36.2 (33.5)	10.2 (5.6)	11.6 (8.6)	14.9 (16.8)
C	13.4 (3.6)	14.2 (13.4)	10.1 (1.9)	10.8 (10.8)
$\langle hc_2 \rangle$	19.6 (14.4)	18.4 (17.2)	6.4 (2.4)	14.6 (13.6)
$\langle dr_2 \rangle$	1.8 (2.0)	2.7 (3.3)	1.4 (1.6)	2.2 (2.6)
ℓ	13.3 (16.7)	9.5 (16.9)	5.4 (0.6)	9.1 (15.9)
$\langle B \rangle$	17.7 (23.1)	13.4 (23.8)	6.3 (1.2)	12.9 (22.8)
CPD	26.0 (39.0)	23.4 (39.8)	14.3 (8.5)	22.4 (38.4)

colunas para cada medida correspondem aos tipos de perturbações (adição, remoção, reconexão ou a combinação aleatória das três).

As projeções dos componentes principais via PCA (Seção 4.1.1) foram obtidas da seguinte forma: cada passo de perturbação corresponde a uma variável do vetor de atributos e todas as redes perturbadas como pertencentes ao conjunto de amostras. No total foram geradas 200 redes de modelos teóricos (50 de cada modelo considerado) e 250 das redes reais (50 de cada rede real). Como as perturbações variaram de 0.2% até 10% em passos de 0.2%, ao todo foram obtidas 50 variáveis. Os gráficos de PCA (parte inferior de cada sub-figura das Figuras 25 e 26) indicam, portanto, as projeções que correspondem aos eixos de maiores variações (dois primeiros autovalores). A metodologia de redução de dimensionalidade via PCA foi usada porque ela permite observar os efeitos das perturbações após eliminar as correlações entre as variáveis. Através das projeções obtidas via PCA, é possível identificar as medidas que discriminam melhor as categorias de redes analisadas.

As variações das medidas apresentadas nas Figuras 25 e 26 estão resumidas nas Tabelas 6 e 7, que mostram as médias e os desvios padrões da porcentagem de variação de todas as medidas no final de cada perturbação (que corresponde a 10% do total do número de arestas de cada rede) para todos os modelos teóricos (Tabela 6) e para todas redes reais (Tabela 7).

A análise dos resultados apresentados nas Figuras 25 e 26 e nas Tabelas 6 e 7 levam as seguintes conclusões:

Tabela 7: Média e desvio padrão (número em parênteses) da porcentagem de variação de cada medida para as redes reais e para 10% de arestas perturbadas. Os símbolos são os mesmos da Tabela 5. Tabela extraída de [5].

Medida	Adição	Reconexão	Remoção	Todas
$\langle hk_2 \rangle$	23.3 (12.8)	12.4 (5.0)	15.2 (9.7)	8.6 (3.2)
r	23.8 (13.3)	21.7 (24.7)	8.0 (16.7)	4.4 (2.8)
C	16.9 (5.5)	21.5 (3.7)	13.7 (2.9)	19.5 (2.8)
$\langle hc_2 \rangle$	16.1 (9.8)	15.1 (10.9)	8.5 (5.6)	14.3 (9.0)
$\langle dr_2 \rangle$	2.1 (1.3)	2.0 (1.5)	3.4 (2.1)	1.3 (1.3)
ℓ	11.9 (16.8)	8.7 (14.5)	11.4 (10.2)	7.3 (13.7)
$\langle B \rangle$	15.1 (19.9)	12.0 (17.9)	2.3 (2.3)	11.3 (18.0)
CPD	27.5 (34.4)	23.0 (33.3)	6.2 (2.5)	22.6 (33.1)

- A variação relativa das medidas é, em geral, muito maior que a porcentagem da variação de arestas para todas redes e em qualquer tipo de perturbação realizada. Desta forma, conclui-se que a maioria das medidas analisadas são sensíveis a pequenas perturbações na conexão das redes.

- Dentre os três tipos de perturbação (incluindo também a combinação aleatória das três), a remoção de arestas foi a que apresentou as menores variações nas medidas tanto para os modelos teóricos quanto para as redes reais (ver Tabelas 6 e 7). Uma conclusão imediata desse resultado é que arestas incertas não devem ser incluídas na rede, pois adicionar arestas inexistentes implica em maiores variações nas medidas do que a remoção de uma existente.

- Redes geográficas, como por exemplo: o modelo de Waxman e a rede de transmissão elétrica, apresentaram grandes variações nas medidas relacionadas à distância quando a perturbação empregada foi a adição ou reconexão de arestas (ver Tabelas 6 e 7). Esse resultado é uma conseqüência de que a adição ou remoção aleatória de arestas tende a conectar vértices cujas distâncias geodésicas sejam grandes[2], reduzindo o caminho mínimo entre eles e afetando as outras medidas relacionadas à distância.

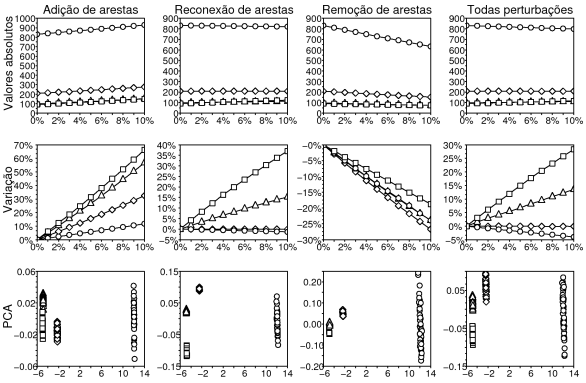
Além dessas conclusões, também é possível inferir a partir dos resultados apresentados nas Figuras 25 e 26 e nas Tabelas 6 e 7 quais medidas são mais estáveis às perturbações (i.e. pouca

[2]Neste caso, as distâncias espaciais também devem ser grandes, mas não são consideradas aqui pois toda análise de perturbação não leva em conta a distância espacial entre os vértices.

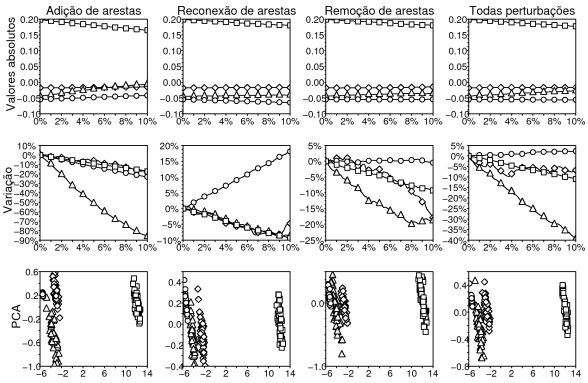
variação) enquanto capazes de discriminar eficientemente modelos teóricos e redes reais (i.e. não sobreposição dos modelos teóricos ou das redes reais no espaço definido pelas projeções dos componentes principais obtidos no PCA). Segundo estes critérios, as melhores medidas para análise de redes com problemas de amostragem são, em ordem de melhor para pior: razão de divergência de nível 2, caminho mínimo médio, ponto de intermediação médio e coeficiente de aglomeração hierárquico médio de nível 2. Por haver sobreposição entre modelos ou entre redes reais, as demais medidas não foram consideradas nesta classificação. Interessante notar que as melhores medidas são aquelas que levam em conta porções maiores da rede, como a razão de divergência e o caminho mínimo médio de forma a não ser muito afetadas pelas alterações locais da conectividade dos vértices. Convém mencionar ainda que a razão de divergência média de nível 2, além de permitir boa discriminação entre os modelos e entre as redes reais, apresentou variação menor do que a de arestas nas redes analisadas.

6.3 Principais conclusões

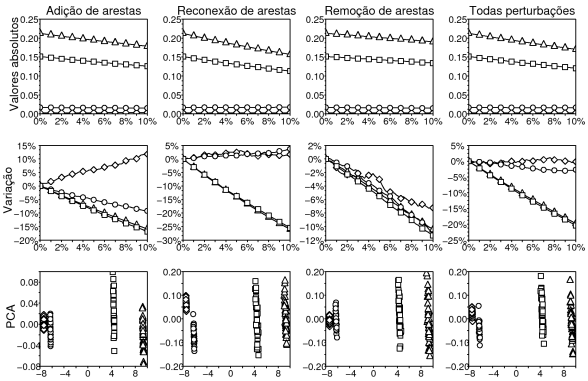
A análise da sensibilidade das medidas topológicas de redes complexas realizada neste capítulo indicou que pequenas variações na conectividade dos vértices podem causar grandes variações nas medidas. Desta forma, ao analisar redes com problemas de amostragem, é necessário tomar cuidado com as medidas a serem utilizadas, pois podem não condizer com os sistemas caracterizados por elas. A melhor solução para este problema seria obter amostragens mais precisas, mas se não for possível, o melhor é escolher medidas que ofereçam maior robustez a perturbações, de preferência aquelas que utilizem parcelas maiores da vizinhança de cada vértice, como por exemplo: razão de divergência média de nível 2. Uma maneira simples, porém, de reduzir o erro na construção das redes é evitar que arestas incertas sejam adicionadas, pois não afetam tanto as medidas, conforme indicam os resultados da seção anterior.



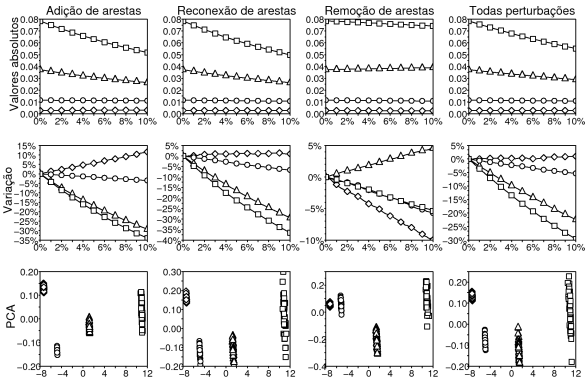
(a) Grau hierárquico médio de nível 2



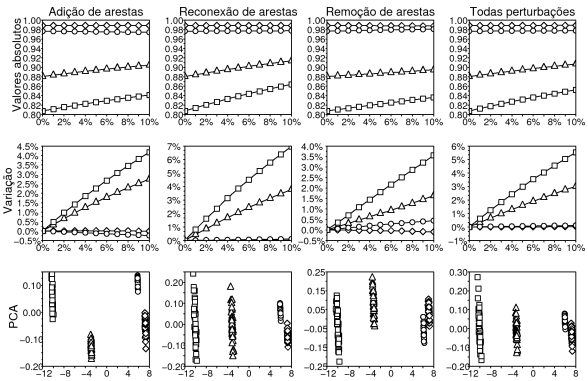
(b) Coeficiente de assortatividade



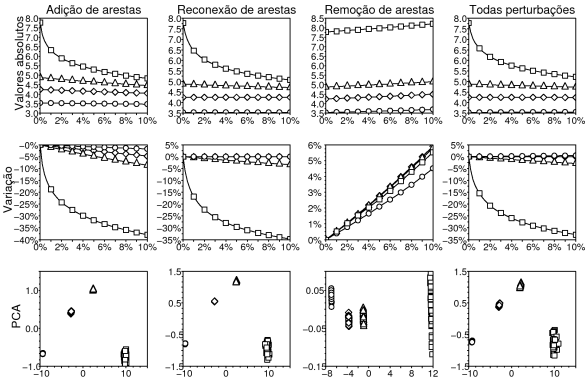
(c) Coeficiente de aglomeração médio



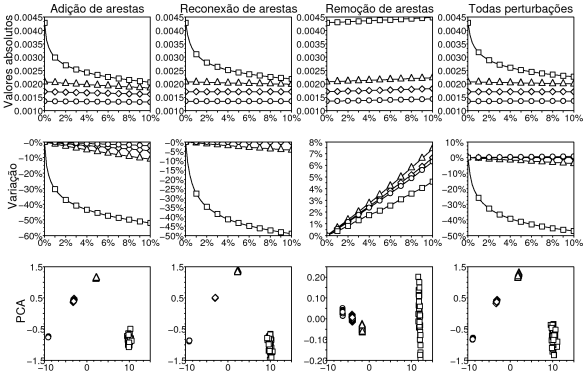
(d) Coeficiente de aglomeração hierárquico médio de nível 2



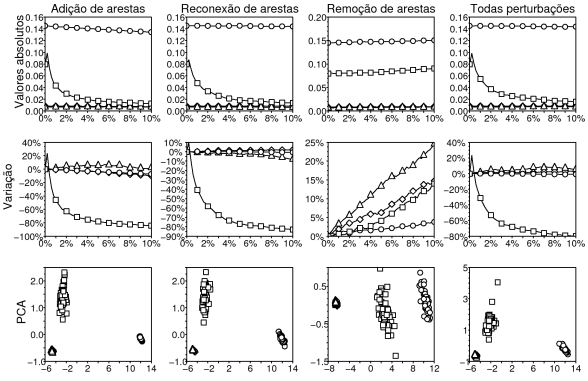
(e) Razão de divergência média de nível 2



(f) Caminho mínimo médio

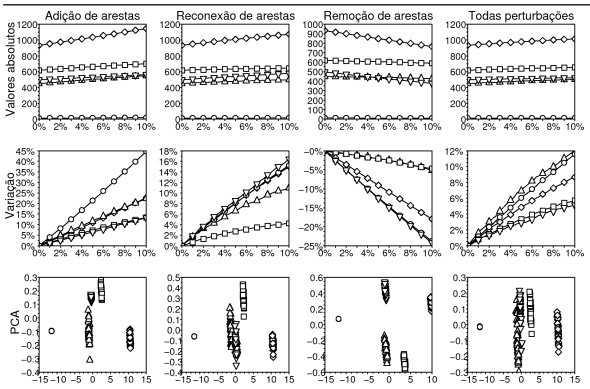


(g) Grau de intermediação médio

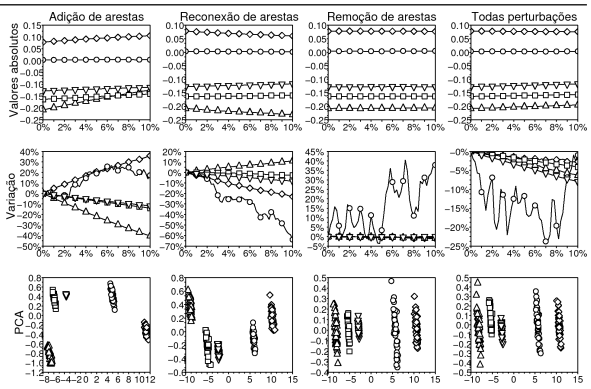


(h) Ponto de dominância central

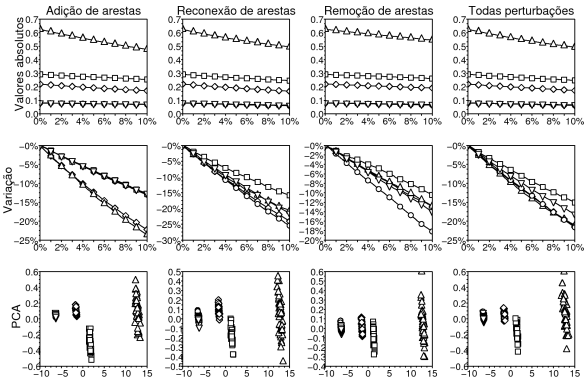
Figura 25: Perturbações das medidas acima para os modelos: aleatório de Erdős e Rényi (◊), de mundo pequeno de Watts e Strogatz (△); livre de escala de Barabási e Albert (○) e geográfico de Waxman (□). Para as duas primeiras linhas, o eixo x é a porcentagem de arestas adicionadas, removidas ou reconectadas.



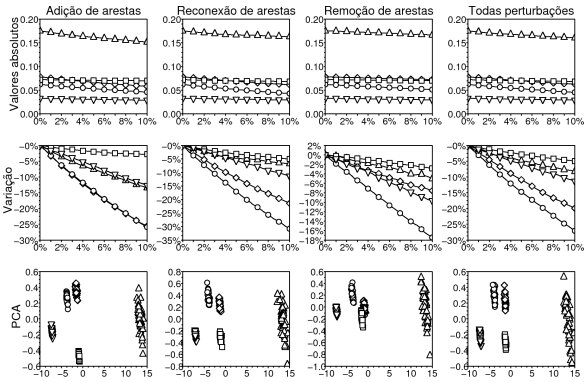
(a) Grau hierárquico médio de nível 2



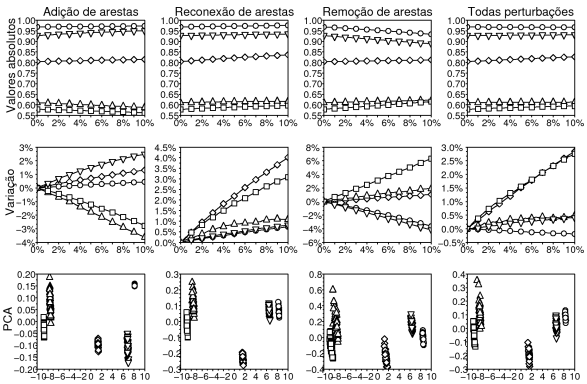
(b) Coeficiente de assortatividade



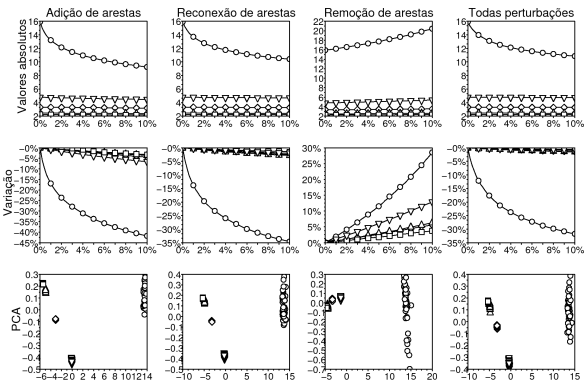
(c) Coeficiente de aglomeração médio



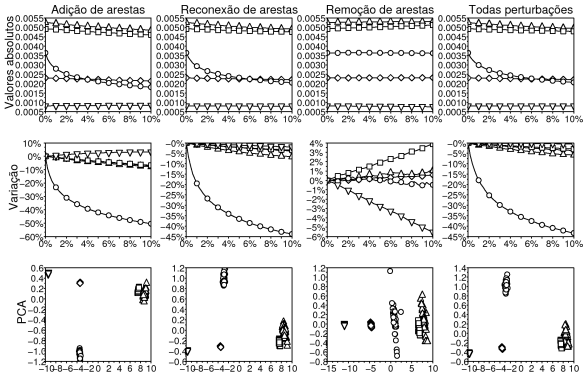
(d) Coeficiente de aglomeração hierárquico médio de nível 2



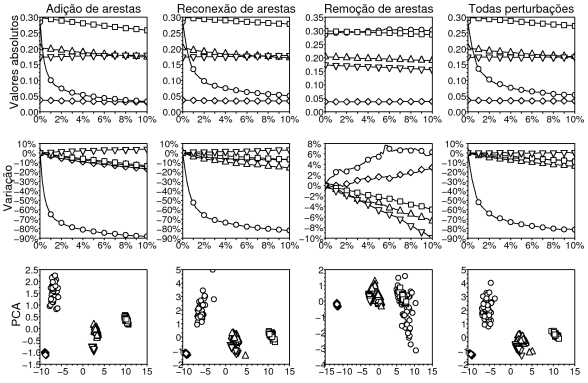
(e) Razão de divergência média de nível 2



(f) Caminho mínimo médio



(g) Grau de intermediação médio



(h) Ponto de dominância central

Figura 26: Perturbações das medidas acima para as redes reais: email ($\Diamond$), aeroportos dos Estados Unidos ($\triangle$), sistema de transmissão elétrica ($\bigcirc$), rede neural ($\square$), e interação de proteínas ($\bigtriangledown$). Para as duas primeiras linhas, o eixo x é a porcentagem de arestas adicionadas, removidas ou reconectadas.

7 Conclusões e perspectivas

Muitas das redes estudadas têm problemas de amostragem como já foram identificados em várias análises realizadas em redes reais, tais como Internet [21], WWW [19], redes de interação de proteínas [20] e redes sociais [22]. Problemas na amostragem incluem: avaliação errada do coeficiente α da lei de potência $P(k) \sim k^{-\alpha}$ se a rede analisada for livre de escala [120], redes aleatórias amostradas são identificadas como redes livres de escala [32], vários falsos positivos nas redes de interação de proteínas [23, 24], falta de dados nas redes sociais [22] e propriedades da rede WWW que dependem do *web crawler* utilizado [19], entre outros. Como se pode notar, muitos problemas podem acontecer com a amostragem de redes complexas. Neste trabalho, foi identificado mais um: a classificação de redes amostradas pode não corresponder com a da rede original. Em outras palavras, utilizar amostras pequenas de sistemas complexos pode levar a resultados e conclusões incorretas. Além disso, modelos que se baseiam em redes amostradas podem apresentar propriedades muito diferentes das que a rede completa teria.

Neste trabalho, foram ainda definidas duas estruturas em redes complexas: as árvores de borda e as cadeias de vértices. A primeira está relacionada com a borda da rede (vértices de grau unitário) e a segunda, com seqüência de vértices de grau 2. A análise dessas estruturas revelou que a ocorrência de árvores de borda e de cadeias de vértices grandes é alta em redes reais que provavelmente tiveram problemas de amostragem, portanto existe uma certa relação entre amostragem e a ocorrência dessas estruturas. Para verificar tal suposição, redes amostradas de dois modelos: aleatório de Erdős e Rényi e livre de escala de Barabási e Albert foram obtidas através de caminhada aleatória e de forma gradual, e a ocorrência de cadeias de vértices foi analisada. Quando a amostras foram pequenas, a presença de alças e caudas grandes foi elevada. A medida que as amostras foram ficando maiores, a quantidade dessas estruturas foi

diminuindo, até praticamente desaparecerem (nesta fase, a rede original foi quase totalmente obtida). As estruturas que restaram foram aquelas que já existiam na rede original, como no modelo aleatório de Erdős e Rényi. Apesar de aparecer cadeias longas de vértices em redes parcialmente amostradas, não é correto afirmar que a presença de tais estruturas implica em redes amostradas. Entretanto, se forem encontradas, é necessário fazer uma análise mais cuidadosa da rede estudada. Um exemplo de rede que provavelmente teve problemas de amostragem é a tão estudada WWW, que apresentou muitas árvores de borda, caudas e alças grandes. Vale lembrar que mais de 50% dos vértics da rede WWW pertencem às árvores de borda. Em outros casos, porém, a presença dessas estruturas não é sinal de incompleteza como revelou a análise local das árvores de borda em redes reais cujos vértices são rotulados. No caso da rede de interação de proteínas, foi descoberto que as raízes das árvores de borda são proteínas com funções mais gerais e as folhas, proteínas com funções mais específicas. Resultados semelhantes foram observados para as redes Wordnet e Roget. A análise da rede de aeroportos dos Estados Unidos revelou que a maioria das raízes das árvores de borda são aeroportos internacionais enquanto que as folhas são aeroportos regionais ou pequenos. Desta forma, o estudo das árvores da borda de redes reais pode ajudar a encontrar a importância dos vértices que estão na borda da rede.

E para finalizar este trabalho, uma análise da sensibilidade de várias propriedades frente a perturbações da conectividade das redes foi realizada. O objetivo foi verificar quais propriedades discriminam as redes analisadas e variam menos quando arestas são adicionadas, removidas ou reconectadas, sendo, portanto, as mais adequadas para caracterizar redes com problemas de amostragem. As melhores propriedades foram (na seqüência de melhor para pior): razão de divergência de nível 2 (propriedade hierárquica), caminho mínimo médio, ponto de intermediação médio e coeficiente de aglomeração hierárquico médio de nível 2. Além disso, foi verificado que não adicionar conexões duvidosas é menos crítico para alterar as propriedades da rede original do que adicioná-las.

Este trabalho apresentou, portanto, estruturas comuns a redes com problemas de amostragem; como a amostragem influencia na classificação de redes complexas; e, por último, que propriedades devem ser utilizadas quando há problemas na amostragem da rede que se queira analisar.

Há muitas perspectivas de trabalhos futuros envolvendo o estudo da amostragem em redes complexas. Algumas incluem: procurar outras estruturas relacionadas às amostragens de redes complexas, avaliar o efeito da amostragem na classificação de outras redes reais, realizar outros tipos de perturbações nas redes e avaliar o efeito da amostragem em processos dinâmicos.

Referências

[1] EICK, S. G. *Arc map showing the world-wide internet traffic*. Disponível em: <http://www.cybergeography.org/atlas/geographic.html>. Acesso em: 20/04/2008.

[2] COSTA, L. da F. et al. Characterization of complex networks: a survey of measurements. *Advances in Physics*, v. 56, n. 1, p. 167 – 242, 2007.

[3] BOAS, P. R. V. et al. Border trees of complex networks. *Journal of Physics A: Mathematical and Theoretical*, v. 41, n. 22, p. 224005, 2008.

[4] BOAS, P. R. V. et al. Chain motifs: the tails and handles of complex networks. *Physical Review E*, v. 77, n. 2, p. 26106, 2008.

[5] BOAS, P. V. et al. Sensitivity of complex networks measurements. ArXiv:0804.1104. 2008. Disponível em: <http://arxiv.org/abs/0804.1104>. Acesso em: 20/05/2008.

[6] ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. *Reviews of Modern Physics*, v. 74, n. 1, p. 48–98, 2002.

[7] NEWMAN, M. E. J. The structure and function of complex networks. *SIAM Review*, v. 45, n. 2, p. 167–256, 2003.

[8] DOROGOVTSEV, S. N.; MENDES, J. F. F. Evolution of networks. *Advances in Physics*, v. 51, n. 4, p. 1079–1187, 2002.

[9] BOCCALETTI, S. et al. Complex networks: structure and dynamics. *Physics Reports*, v. 424, n. 4-5, p. 175–308, 2006.

[10] COSTA, L. da F. et al. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. Arxiv preprint arXiv:0711.3199. 2007. Disponível em: <http://arxiv.org/abs/0711.3199>. Acesso em: 10/05/2008.

[11] BOLLOBÁS, B. *Modern graph theory*. New York: Springer-Verlag, 1998. (Graduate Texts in Mathematics).

[12] WATTS, D. J.; STROGATZ, S. H. Collective dynamics of small-world networks. *Nature*, v. 393, n. 6684, p. 440–442, 1998.

[13] BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. *Science*, v. 286, n. 5439, p. 509–512, 1999.

[14] GIRVAN, M.; NEWMAN, M. E. J. Community structure in social and biological networks. *Proceedings of the National Academy of Science USA*, v. 99, n. 12, p. 7821–7826, 2002.

[15] STUMPF, M. P. H.; WIUF, C.; MAY, R. M. Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proceedings of the National Academy of Science USA*, v. 102, n. 12, p. 4221–4224, 2005.

[16] PASTOR-SATORRAS, R.; VESPIGNANI, A. Epidemic spreading in scale-free networks. *Physical Review Letters*, v. 86, n. 14, p. 3200–3203, 2001.

[17] COSTA, L. da F.; TRAVIESO, G. Exploring complex networks through random walks. *Physical Review E*, v. 75, n. 1, p. 16102, 2007.

[18] TADIĆ, B.; RODGERS, G. J.; THURNER, S. Transport on complex networks: flow, jamming and optimization. *International Journal of Bifurcation and Chaos*, v. 17, n. 7, p. 2363–2385, 2007.

[19] SERRANO, M. Á. et al. Decoding the structure of the www: a comparative analysis of web crawls. *ACM Transactions on the Web*, v. 1, n. 2, 2007.

[20] HAN, J. D. et al. Effect of sampling on topology predictions of protein-protein interaction networks. *Nature Biotechnology*, v. 23, n. 7, p. 839–844, 2005.

[21] LAKHINA, A. et al. Sampling biases in ip topology measurements. In: *IEEE INFOCOM 2003: TWENTY-SECOND ANNUAL JOINT CONFERENCE OF THE IEEE COMPUTER AND COMMUNICATIONS SOCIETIES*. [S.l.]: IEEE, 2003. v. 1, p. 332–341.

[22] KOSSINETS, G. Effects of missing data in social networks. *Social Networks*, v. 28, n. 3, p. 247–268, 2006.

[23] MROWKA, R.; PATZAK, A.; HERZEL, H. Is there a bias in proteome research? *Genome Research*, v. 11, n. 12, p. 1971–1973, 2001.

[24] SAITO, R. et al. Interaction generality, a measurement to assess the reliability of a protein–protein interaction. *Nucleic Acids Research*, v. 30, n. 5, p. 1163–1168, 2002.

[25] SPRINZAK, E.; SATTATH, S.; MARGALIT, H. How reliable are experimental protein–protein interaction data. *Journal of Molecular Biology*, v. 327, n. 5, p. 919–23, 2003.

[26] JEONG, H. et al. Lethality and centrality in protein networks. *Nature*, v. 411, n. 6833, p. 41–42, 2001.

[27] KHANIN, R.; WIT, E. How scale-free are biological networks. *Journal of Computational Biology*, v. 13, n. 3, p. 810–818, 2006.

[28] KROGAN, N. J. et al. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, v. 440, n. 7084, p. 637–643, 2006.

[29] ALBERT, R.; JEONG, H.; BARABÁSI, A.-L. Diameter of the world wide web. *Nature*, v. 401, n. 6749, p. 130–131, 1999.

[30] BECCHETTI, L. et al. A comparison of sampling techniques for web graph characterization. In: *ACM LINKKDD'06: WORKSHOP ON LINK ANALYSIS*. Philadelphia, PA: ACM, 2006.

[31] LEGUAY, J. et al. Describing and simulating internet routes. *Computer Networks*, v. 51, n. 8, p. 2067–2085, 2007.

[32] CLAUSET, A.; MOORE, C. Accuracy and scaling phenomena in internet mapping. *Physical Review Letters*, v. 94, n. 1, p. 18701, 2005.

[33] AMARAL, L. A. N.; OTTINO, J. M. Complex networks. *European Physical Journal B – condensed matter*, v. 38, n. 2, p. 147–162, 2004.

[34] WASSERMAN, S.; FAUST, K. *Social network analysis*. Cambridge: Cambridge University Press, 1994.

[35] SCOTT, J. P. *Social network analysis: a handbook*. London: Sage, 2000.

[36] BARABÁSI, A.-L. et al. On the topology of the scientific collaboration networks. *Physica A*, v. 311, p. 590–614, 2002.

[37] NEWMAN, M. E. J. From the cover: The structure of scientific collaboration networks. *Proceedings of the National Academy of Science USA*, v. 98, n. 2, p. 404–409, 2001.

[38] FALOUTSOS, M.; FALOUTSOS, P.; FALOUTSOS, C. On power-law relationships of the internet topology. In: *ACM SIGCOMM '99: PROCEEDINGS OF THE CONFERENCE ON APPLICATIONS, TECHNOLOGIES, ARCHITECTURES, AND PROTOCOLS FOR COMPUTER COMMUNICATION*. Cambridge: ACM, 1999. p. 251–262.

[39] YOOK, S.; JEONG, H.; BARABÁSI, A. Modeling the internet's large-scale topology. *Proceedings of the National Academy of Sciences*, v. 99, n. 21, p. 13382–13386, 2002.

[40] BARABÁSI, A.-L.; ALBERT, R.; JEONG, H. Scale-free characteristics of random networks: the topology of the world wide web. *Physica A: statistical mechanics and its applications*, v. 281, n. 1-4, p. 69–77, 2000.

[41] ADAMIC, L. A.; HUBERMAN, B. A. Power-law distribution of the world wide web. *Science*, v. 287, n. 5461, p. 2115, 2000.

[42] SPORNS, O. Network analysis, complexity and brain function. *Complexity*, v. 8, n. 1, p. 56–60, 2002.

[43] COSTA, L. da F.; SPORNS, O. Hierarchical features of large-scale cortical connectivity. *The European Physical Journal B - condensed matter and complex systems*, v. 48, n. 4, p. 567–573, 2005.

[44] COSTA, L. da F.; KAISER, M.; HILGETAG, C. C. Predicting the connectivity of primate cortical networks from topological and spatial node properties. *BMC Systems Biology*, v. 1, p. 16, 2007.

[45] MCCANN, K. S.; HASTINGS, A.; HUXEL, G. R. Weak trophic interactions and the balance of nature. *Nature*, v. 395, n. 6704, p. 794–798, 1998.

[46] DROSSEL, B.; MCKANE, A. Modelling food webs. In: S. Bornholdt and H. G. Schuster. Weinheim: Wiley-VCH, 2003.

[47] GUIMERÀ, R.; AMARAL, L. A. N. Modeling the world-wide airport network. *European Physical Journal B – condensed matter*, v. 38, n. 2, p. 381–385, 2004.

[48] GUIMERÀ, R. et al. The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Science USA*, v. 102, n. 22, p. 7794–7799, 2005.

[49] COSTA, L. da F. What's in a name? *International Journal of Modern Physics C*, v. 15, n. 1, p. 371–379, 2004.

[50] AMANCIO, D. R. et al. Complex networks analysis of manual and machine translations. *International Journal of Modern Physics C*, v. 19, n. 4, p. 1–16, 2008.

[51] NEWMAN, M. E. J.; BARÁBASI, A.-L.; WATTS, D. J. (Ed.). *The structure and dynamics of networks*. [S.l.]: Princeton University Press, 2006.

[52] ERDŐS, P.; RÉNYI, A. On random graphs. *Publicationes Mathematicae*, v. 6, n. 290, p. 290–297, 1959.

[53] ERDŐS, P.; RÉNYI, A. On the evolution of random graphs. *Publicationes Mathematicae*, v. 5, p. 17–61, 1960.

[54] ERDŐS, P.; RÉNYI, A. On the strenght of connectedness of a random graph. *Acta Mathematica Scientia Hungary*, v. 12, p. 261–267, 1961.

[55] MILGRAN, S. The small world problem. *Psychology Today*, v. 1, n. 1, p. 60–67, 1967.

[56] BARABÁSI, A.-L. *Linked: how everything is connected to everything else and what it means for business, science, and everyday life*. New York: Plume, 2003.

[57] NEWMAN, M. E. J. Scientific collaboration networks: I. network construction and fundamental results. *Physical Review E*, v. 64, n. 1, p. 016131, 2001.

[58] NEWMAN, M. E. J. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical Review E*, v. 64, n. 1, p. 16132, 2001.

[59] NEWMAN, M. E. J. *Mark Newman's network data*. Disponível em: <http://www-personal.umich.edu/∼mejn/netdata>. Acesso em: 25/04/2008.

[60] EBEL, H.; MIELSCH, L.; BORNHOLDT, S. Scale-free topology of e-mail networks. *Physical Review E*, v. 66, n. 3, p. 35103, 2002.

[61] GUIMERÀ, R. et al. Self-similar community structure in a network of human interactions. *Physical Review E*, v. 68, n. 6, p. 65103, 2003.

[62] ROGET, P. M.; ROBERT, A. *Roget's thesaurus of English words and phrases*. [S.l.]: Longman Harlow, Essex, 1982.

[63] BATAGELJ, V.; MRVAR, A. *Pajek datasets*. 2006. Disponível em: <http://vlado.fmf.uni-lj.si/pub/networks/data>. Acesso em: 20/05/2008.

[64] BARABÁSI, A.-L. *Center for Complex Network Research*. Disponível em: <http://www.nd.edu/∼networks/resources.htm>.Acesso em: 13/04/2008.

[65] ANTIQUEIRA, L. et al. Strong correlations between text quality and complex networks features. *Physica A: Statistical Mechanics and its Applications*, v. 373, p. 811–820, 2007.

[66] ANTIQUEIRA, L. et al. Some issues on complex networks for author characterization. In: *PROCEEDINGS OF THE INTERNATIONAL JOINT CONFERENCE IBERAMIA/SBIA/SBRN 2006 – WORKSHOP IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY*. Ribeirão Preto: [s.n.], 2006.

[67] WHITE, J. G. et al. The structure of the nervous system of the nematode *Caenorhabditis Elegans*. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, v. 314, n. 1165, p. 1–340, 1986.

[68] SHEN-ORR, S. S. et al. Network motifs in the transcriptional regulation network of *Escherichia Coli*. *Nature Genetics*, v. 31, n. 1, p. 64–68, 2002.

[69] RAPOPORT, A. Nets with distance bias. *Bulletin of Mathematical Biophysics*, v. 13, n. 2, p. 85–91, 1951.

[70] RAPOPORT, A. Spread of information through a population with socio-structural bias: I. assumption of transitivity. *Bulletin of Mathematical Biophysics*, v. 15, n. 4, p. 523–533, 1953.

[71] RAPOPORT, A. Contribution to the theory of random and biased nets. *Bulletin of Mathematical Biophysics*, v. 19, n. 4, p. 257–277, 1957.

[72] NEWMAN, M. E. J. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, v. 46, n. 5, p. 323–351, 2005.

[73] GASTNER, M. T.; NEWMAN, M. E. J. The spatial structure of networks. *The European Physical Journal B*, v. 49, n. 2, p. 247, 2006.

[74] WAXMAN, B. M. Routing of multipoint connections. *IEEE Journal on Selected Areas in Communications*, v. 6, n. 9, p. 1617–1622, 1988.

[75] MOLLOY, M.; REED, B. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, v. 6, n. 213, p. 161–179, 1995.

[76] MOLLOY, M.; REED, B. The size of the giant component of a random graph with a given degree sequence. *Combinatorics, Probability and Computing*, v. 7, n. 03, p. 295–305, 1998.

[77] AMARAL, L. A. N. et al. Classes of small-world networks. *Proceedings of the National Academy of Science*, v. 97, n. 21, p. 11149 – 11152, 2000.

[78] BAR, S.; GONEN, M.; WOOL, A. A geographic directed preferential internet topology model. *Computer Networks: The International Journal of Computer and Telecommunications Networking*, v. 51, n. 14, p. 4174–4188, 2007.

[79] JIN, C.; CHEN, Q.; JAMIN, S. Inet: Internet topology generator. Michigan, 2000. (Technical Report CSE-TR-433-00).

[80] WINICK, J.; JAMIN, S. Inet-3.0: Internet topology generator. Michigan, 2002. (Technical Report CSE-TR-456-02).

[81] COSTA, L. da F.; Cesar Jr., R. M. *Shape analysis and classification: theory and practice*. [S.l.]: CRC Press, 2001.

[82] BARTHÉLEMY, M. et al. Characterization and modeling of weighted networks. *Physica A*, v. 346, p. 34–43, 2005.

[83] EDWARDS., A. L. *An introduction to linear regression and correlation*. San Francisco: W. H. Freeman and Co, 1993.

[84] NEWMAN, M. E. J. Assortative mixing in networks. *Physical Review Letters*, v. 89, n. 20, p. 208701, 2002.

[85] LATORA, V.; MARCHIORI, M. Efficient behavior of small-world networks. *Physics Review Letters*, v. 87, n. 19, p. 198701, 2001.

[86] FREEMAN, L. C. A set of measures of centrality based on betweenness. *Sociometry*, v. 40, n. 1, p. 35–41, 1977.

[87] RAVASZ, E.; BARABÁSI, A.-L. Hierarchical organization in complex networks. *Physical Review E*, v. 67, n. 2, p. 026112, 2003.

[88] BIANCONI, G.; CAPOCCI, A. Number of loops of size h in growing scale-free networks. *Physical Review Letters*, v. 90, n. 7, p. 078701, 2003.

[89] KIM, H. J.; KIM, J. M. Cyclic topology in complex network. *Physical Review E*, v. 72, p. 036109, 2005.

[90] ZHOU, S.; MONDRAGON, R. J. The rich-club phenomenon in the internet topology. *IEEE Communications Letters*, v. 8, n. 3, p. 180–182, 2004.

[91] ARENAS, A. et al. Community analysis in social networks. *European Physical Journal B*, v. 38, p. 373–380, 2004.

[92] GLEISER, P. M.; DANON, L. Community structure in jazz. *Advances in complex systems*, v. 6, n. 4, p. 565–573, 2003.

[93] GUIMERÀ, R.; AMARAL, L. A. N. Functional cartography of complex metabolic networks. *Nature*, v. 433, n. 7028, p. 895–900, 2005.

[94] RADICCHI, F. et al. Defining and identifying communities in networks. *Proceedings of the National Academy of Science USA*, v. 101, n. 9, p. 2658–2663, 2004.

[95] REICHARDT, J.; BORNHOLDT, S. Statistical mechanics of community detection. *Physical Review E*, v. 74, n. 1, p. 16110, 2006.

[96] NEWMAN, M. E. J.; GIRVAN, M. Finding and evaluating community structure in networks. *Physical Review E*, v. 69, p. 026113, 2004.

[97] NEWMAN, M. E. J. Detecting community structure in networks. *The European Physical Journal B*, v. 38, p. 321–330, 2004.

[98] DANON, L. et al. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, v. 9, p. P09008, Setembro 2005.

[99] COSTA, L. da F. The hierarchical backbone of complex networks. *Physical Review Letters*, v. 93, p. 098702, 2004.

[100] VIANA, M. P.; COSTA, L. da F. The dynamics of hierarchical evolution of complex networks. *Arxiv preprint cond-mat/0504346*, 2005. Disponível em: <http://arxiv.org/abs/cond-mat/0504346>. Acesso em: 20/05/2008.

[101] COSTA, L. da F.; SILVA, F. Hierarchical characterization of complex networks. *Journal of Statistical Physics*, v. 125, n. 4, p. 841–872, 2006.

[102] TRAVENÇOLO, B. A. N.; COSTA, L. da F. Hierarchical spatial organization of geographical networks. *Journal of Physics A: Mathematical and Theoretical*, v. 41, n. 22, p. 224004, 2008.

[103] COSTA, L. da F.; ROCHA, L. E. C. da. A generalized approach to complex networks. *The European Physical Journal B - condensed matter*, v. 50, n. 1, p. 237–242, 2006.

[104] VINCENT, L. Graphs and mathematical morphology. *Signal Processing*, v. 16, n. 4, p. 365–388, 1989.

[105] HEIJMANS, H. et al. Graph morphology. *Journal of Visual Communication and Image Representation*, v. 3, n. 1, p. 24–38, 1990.

[106] DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern classification*. New York: John Wiley & Sons, Inc., 2001.

[107] JOHNSON, R. A.; WICHERN, D. W. *Applied multivariate statistical analysis*. 5 ed.. ed. [S.l.]: Prentice Hall, 2002.

[108] MONTEIRO, L. R.; REIS, S. F. dos. *Princípios de morfometria geométrica*. [S.l.]: Holos, Ribeirão Preto, 1999.

[109] FUKUNAGA, K. *Introduction to statistical pattern recognition*. [S.l.]: Academic Press, 1990.

[110] RODRIGUES, F. A. et al. Seeking the best internet model. Arxiv preprint arXiv:0706.3225. 2007. Disponível em: <http://arxiv.org/abs/0706.3225>. Acesso em: 03/02/2008.

[111] NOH, J. D.; RIEGER, H. Random walks on complex networks. *Physical Review Letters*, v. 92, n. 11, p. 118701, 2004.

[112] DOROGOVTSEV, S. N. et al. k-core organization of complex networks. *Physical Review Letters*, v. 96, n. 4, p. 40601, 2006.

[113] CARMI, S. et al. New model of internet topology using k-shell decomposition. Cs.NI/0607080. 2006.

[114] WUCHTY, S.; ALMAAS, E. Peeling the yeast protein network. *Proteomics*, v. 5, n. 2, p. 444–449, 2005.

[115] ALTAF-UL-AMIN, M. et al. Prediction of protein functions based on k-cores of protein-protein interaction networks and amino acid sequences. *Genome Informatics*, v. 14, p. 498–499, 2003.

[116] DIESTEL, R. *Graph theory*. [S.l.]: Springer, 2000.

[117] MILO, R. et al. On the uniform generation of random graphs with prescribed degree sequences. Cond-mat/0312028. 2003. Disponível em: <http://arxiv.org/abs/cond-mat/0312028>. Acesso em: 20/05/2008.

[118] COSTA, L. da F.; RODRIGUES, F. A.; TRAVIESO, G. Protein domain connectivity and essentiality. *Applied Physics Letters, (in press.)*, 2006.

[119] SCHWIKOWSKI, B.; UETZ, P.; FIELDS, S. A network of protein- protein interactions in yeast. *Nature Biotechnology*, v. 18, p. 1257–1261, 2000.

[120] PETERMANN, T.; RIOS, P. L. Exploration of scale-free networks. *The European Physical Journal B-Condensed Matter*, v. 38, n. 2, p. 201–204, 2004.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo