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

Caracterização, classificação e análise de redes complexas

Francisco Aparecido Rodrigues

Tese apresentada ao Instituto de Física de São Carlos (USP), como requisito necessário para obtenção do título de Doutor.

Orientador: *Prof. Dr. Luciano da Fontoura Costa*

SÃO CARLOS
2007

Agradecimentos

São muitas as pessoas que contribuíram direta ou indiretamente para que este trabalho fosse realizado e é praticamente impossível citar todas elas. Primeiramente gostaria de agradecer ao professor Luciano da F. Costa, pela amizade, companhia, paciência e pela grande competência na orientação do trabalho, além do constante otimismo oferecido. Também agradeço ao professor Gonzalo Travieso, cujo auxílio e colaboração foram muito importantes em diversas pesquisas.

Agradeço à minha esposa, Camila, por todo o apoio, auxílio e revisão deste texto. Seu incentivo foi muito importante, principalmente no final do doutoramento, através de sugestões e discussões que ajudaram a enriquecer a tese.

Muitas pessoas me auxiliaram não em termos acadêmicos, mas nos afazeres e convivências do dia-a-dia. Agradeço principalmente minha mãe, Arsênia, e meus irmãos Fabrício, Gabriela, Meire, Elisa, Fátima e João, por oferecerem muita compreensão, otimismo e esperança inacabáveis.

Agradeço a todas as pessoas que de uma forma ou de outra contribuíram para o meu trabalho. Agradeço ainda ao Paulino R. Villas Boas, pelo companheirismo e assistência dispensada desde os tempos de graduação.

Finalmente, agradeço a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), pelo apoio financeiro, que foi essencial para o desenvolvimento de toda a pesquisa realizada durante o meu doutoramento.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

A teoria das redes complexas tem sido utilizada na análise, modelagem e caracterização de sistemas naturais e artificiais, tais como a Internet, interações protéicas, a sociedade, a teia mundial e diferentes ecossistemas. Basicamente, a caracterização destes sistemas é realizada através de poucas medidas de redes, apesar delas existirem em grande número. Geralmente, apenas a distribuição das conexões, o coeficiente de aglomeração médio e o menor caminho médio são considerados na caracterização de redes reais. Como os modelos são desenvolvidos em termos destas medidas, a modelagem se torna muitas vezes incompleta, pois muitos desses modelos são considerados precisos quando reproduzem apenas um conjunto reduzido de propriedades estruturais de redes reais. A obtenção de descrições mais detalhadas e conseqüentemente, modelos mais precisos, pode ser alcançada com a utilização de diversas medidas de redes. A análise de um conjunto amplo de medidas pode ser realizada através de métodos de estatística multivariada e reconhecimento de padrões, conforme apresentamos neste trabalho. Considerando a análise das variáveis canônicas e decisão Bayesiana, caracterizamos diversas redes reais, principalmente a Internet, e encontramos os modelos mais adequados para reproduzir a estrutura de redes reais. Além disso, empregamos medidas de redes no estudo da relação conectividade-letalidade em redes de interação de proteínas e domínios do *S. cerevisiae* e mostramos que essa relação é melhor definida entre os domínios. Finalmente, utilizamos a teoria das redes

complexas na análise e caracterização das relações entre os produtores de vinhos da região de Bordeaux (França). Os resultados obtidos sugerem que as técnicas de cultivo e propriedades da produção de vinho são fortemente influenciadas pelo território, o que reforça o conceito de *terroir*.

Abstract

Complex networks theory has been used for analysis, modeling and characterization of natural and artificial systems, such as the Internet, protein interaction, society, the World Wide Web and ecosystems. Basically, such systems are characterized in terms of complex networks measurements. Despite the large number of network measurements that has been developed, generally only the degree distribution, the average clustering coefficient and the average shortest path are considered for network characterization and modeling. Since the network models are generally quantified in terms of these measurements, the modeling is frequently incomplete. In order to obtain more detailed characterization of networks, and consequently more accurate models, several networks measurements must be taken into account. The analysis of a large set of measurements can be performed through multivariate statistical analysis and pattern recognition methods, as presented in this work. We characterize and classify various real networks considering canonical variable analysis and bayesian decision theory. We determine the accuracy of the most well-known networks models considering several real networks, such as the Internet. Besides, we apply the complex networks theory to study the connectivity-lethality relation for proteins interaction network and the protein domain network of the yeast *S. cerevisiae* and show that this relation is better defined between the domains. Finally, we characterize and analyze the similarities of wine producers of Bordeaux region (France) through complex

networks theory. Our results suggest that the cultivation techniques and production properties are strongly influenced by the territory, which stress the concept of *terroir*.

Capítulo 1

Introdução

A teoria das redes complexas engloba conceitos de teoria dos grafos, mecânica estatística, sistemas não-lineares e sistemas complexos, que são aplicados na modelagem, análise e simulação de sistemas naturais e artificiais formados por partes discretas que interagem. Devido a sua generalidade e caráter multidisciplinar, essa teoria é utilizada nas mais diversas áreas de pesquisa, como física, química, matemática, biologia, psicologia, medicina, computação, sociologia, lingüística, engenharia, telecomunicações e astronomia [Bar03]. No caso da biologia, a modelagem de interações entre os componentes celulares utilizando a teoria das rede complexas ajudou no desenvolvimento das pesquisas em biologia de sistemas [Kit02, Kit04], que, basicamente, está interessada no estudo das relações entre genes [SOMMA02, DdFC05], proteínas [dFCRT06, JMBO01] e metabólicos [JTA+00]. Em muitos casos, o estudo das relações entre tais componentes é mais importante do que a análise de cada um deles individualmente, o que reforça a importância da utilização da teoria das redes complexas em microbiologia. Por exemplo, conforme sugerido por Volgestein, Lane e Levine, estudar as conexões do gene p53, que é um supressor de tumores, é mais importante do que estudá-lo individualmente [VLL00].

Como aplicação em computação, a teoria das redes complexas oferece suporte para o estudo da Internet (formada por roteadores conectados por fibras ópticas ou sistemas autônomos conectados por tabelas BGP) [FFF99], da Teia Mundial (formada por portais, *blogs* e *sites* pessoais e educacionais que se referenciam) [AJB99] e de aspectos de engenharia de *software* [VS07]. No caso da Internet, fatores dinâmicos, como a propagação de falhas ou vírus entre roteadores, podem ser simulados [BLM+06]. Além disso, o desenvolvimento de modelos que representem a evolução da Internet são fundamentais para se fazer previsões quanto à sua estrutura e determinar o efeito de falhas estáticas [AJB00] ou dinâmicas (efeito cascata [Mot04]). Modelos que reproduzam as principais características da Internet também são importantes para o desenvolvimento de protocolos de roteamento e planejamento de tráfego [YJB02].

A sociedade e os ecossistemas são outros sistemas complexos que naturalmente são modelados pela teoria das redes complexas. No primeiro caso, a estrutura das relações existentes entre pessoas [LEA+01] ou grupos de indivíduos de uma determinada especialidade, como artistas [GD03], cientistas [BJR+02, New01] ou mesmo esportistas [OC04], pode ser modelada por redes. Através da caracterização da estrutura de redes sociais pode-se inferir as relações entre indivíduos. Por exemplo, Newman mostrou que as colaborações entre cientistas são distribuídas de forma não homogênea, ficando concentradas entre poucos pesquisadores [New01], sendo a rede formada por conjuntos de cientistas que colaboram muito entre si e pouco com os demais. Além da caracterização, fenômenos dinâmicos, como a propagação de opiniões e controle de epidemias também podem ser simulados em redes sociais [BLM+06]. Já no caso dos ecossistemas, as cadeias alimentares são representadas por redes formadas por espécies conectadas de acordo com relações de predatismo. Nestas redes, fenômeno dinâmicos podem ser simulados, a fim de quantificar os efeitos causados por desastres ambientais

ou extinção de espécies [KFM+03].

De acordo com estes exemplos, podemos notar que a teoria das redes complexas oferece suporte para caracterização, análise e modelagem dos mais variados sistemas complexos. Na verdade, qualquer sistema natural ou artificial formado por muitas partes que interagem pode ser representado por redes. Entretanto, apesar do grande sucesso obtido por tal teoria, ainda há um vasto campo de pesquisa a ser explorado devido à algumas limitações existentes, como a falta de medidas e métodos para analisar, caracterizar e classificar redes reais. Como nenhuma medida é suficiente para caracterizar completamente uma rede, um conjunto formado por diversas medidas deve ser considerado. No entanto, como geralmente tal conjunto se restringe a duas ou três medidas, a descrição de redes reais e daquelas geradas por modelos acaba sendo incompleta. Assim, muitos dos modelos que têm sido desenvolvidos reproduzem apenas algumas propriedades particulares de redes reais, como a distribuição livre de escala e a pequena distância média entre os vértices. Logo, tais modelos são bastante simplificados e não reproduzem grande parte das características topológicas e dinâmicas das redes consideradas. Portanto, para a obtenção de modelos mais precisos é fundamental a consideração de um conjunto amplo de medidas não redundantes, o que pode ser alcançado com a utilização de técnicas de reconhecimento de padrões e mineração de dados, conforme apresentamos neste trabalho.

1.1 Objetivos

Apesar de existir dezenas de medidas de redes complexas [dFCRTB07], apenas algumas delas são consideradas na análise e caracterização de redes. Com isso, muitos dos modelos de redes desenvolvidos reproduzem apenas algumas propriedades estruturais de redes reais, o que os torna incompletos. Para se ob-

ter caracterizações mais completas da estrutura de redes, e conseqüentemente modelagens mais precisas, é necessária a utilização de um conjunto amplo de medidas não correlacionadas, que descrevam as propriedades topológicas mais importantes. Um dos objetivos deste trabalho é oferecer uma metodologia de caracterização, análise e classificação de redes complexas baseada em métodos de estatística multivariada e reconhecimento de padrões, que solucione as limitações na caracterização e classificação de redes.

Outras finalidades deste trabalho se concentram na análise e caracterização de três sistemas complexos.

1. Internet no nível dos sistemas autônomos: A evolução da Internet é guiada por diversos fatores, incluindo ligação preferencial, distância geográfica entre roteadores e crescimento, com a adição de novos roteadores e conexões. Uma das finalidades de nossa pesquisa é o desenvolvimento de um modelo baseado nestes fatores e sua comparação com outros modelos de redes.

2. Redes de interações de proteínas e de domínios protéicos do *Sacharomices cerevisiae*: Proteínas letais (essenciais) são aquelas que quando inativas, causam a morte ou esterilidade de organismos. Como estas proteínas são fundamentais, verificou-se que há uma correlação entre letalidade e conectividade em redes de interação de proteínas, ou seja, há uma tendência das proteínas mais conectadas serem letais [JMBO01]. Como os domínios protéicos são as unidades funcionais das proteínas, é esperado que exista uma correlação entre conectividade e letalidade no nível dos domínios. Um dos objetivos deste trabalho é testar essa hipótese e comparar as correlações obtidas pela análise das redes de interação de proteínas e domínios.

3. Rede dos produtores de vinhos da região de Bordeaux: Os *chateaux* ao redor da região de Bordeaux foram divididos em distritos de forma a as-

sociar as características dos vinhos aos respectivos territórios de produção. No entanto, apesar dessa suposta associação existir há séculos, nenhuma investigação já foi realizada de forma a analisar quantitativamente essa hipótese. Este trabalho tem a finalidade de analisar de forma quantitativa como as características de produção e cultivo de vinho são influenciadas pela região de produção.

1.2 Contribuições

Dentre as contribuições principais deste trabalho, podemos citar:

- **Revisão de medidas de redes**: como as medidas de redes complexas são descritas em centenas de artigos diferentes, realizamos uma revisão na literatura e apresentamos neste trabalho suas principais características.

- **Metodologia de classificação**: geralmente os modelos de rede são considerados precisos quando reproduzem um número reduzido de propriedades estruturais de redes reais, como a distribuição das conexões e o caminho característico. No entanto, essa aproximação é bastante limitada e para se obter modelagens mais completas deve-se considerar um conjunto amplo de medidas na caracterização dos modelos. Para realizar essa tarefa, mostramos como a análise das variáveis canônicas e decisão Bayesiana podem ser utilizadas na determinação da completeza de modelos de rede. Nossos resultados mostram que muitos dos modelos que são considerados adequados para representar redes reais não reproduzem propriedades estruturais importantes, como a assortatividade e o alto coeficiente de aglomeração. A metodologia que propomos neste trabalho permite uma caracterização mais completa de modelos de redes e a visualização, em duas ou três dimensões,

da separação entre eles. Nesse caso, é possível determinar de forma precisa qual modelo é mais adequado para modelar uma dada rede real.

- **Modelagem da Internet**: desenvolvemos um modelo de Internet e o comparamos com alguns modelos de redes complexas. Mostramos que esse modelo é bastante preciso e concluímos que a evolução da Internet é guiada por três fatores limitantes: (i) ligação preferencial, (ii) distância geográfica entre os roteadores e (iii) crescimento, com a adição de novos sistemas autônomos e ligações.

- **Essencialidade em redes de interação de domínios protéicos**: comparamos a correlação entre essencialidade e conectividade em redes de interação de proteínas e domínios protéicos. Nossos resultados sugerem que essa correlação é mais bem definida no nível dos domínios, o que mostra a importância destas estruturas na definição das funções das proteínas.

- **Rede dos produtores de vinhos**: mostramos como a representação por redes complexas pode ser utilizada como um método de classificação não-supervisionado. Aplicamos essa metodologia no caso da rede dos produtores de vinho de Bordeaux e mostramos como os atributos de produção e cultivo do vinho são influenciados pelo território. Nesse caso, nossos resultados mostraram que os chateaux pertencentes ao mesmo território tentem a compartilhar técnicas de cultivo e produção semelhantes. Também analisamos quais atributos mais influenciam na qualidade dos vinhos e determinamos os *chateaux* protótipos de cada região, o que permite determinar os vinhos com as características típicas de cada distrito.

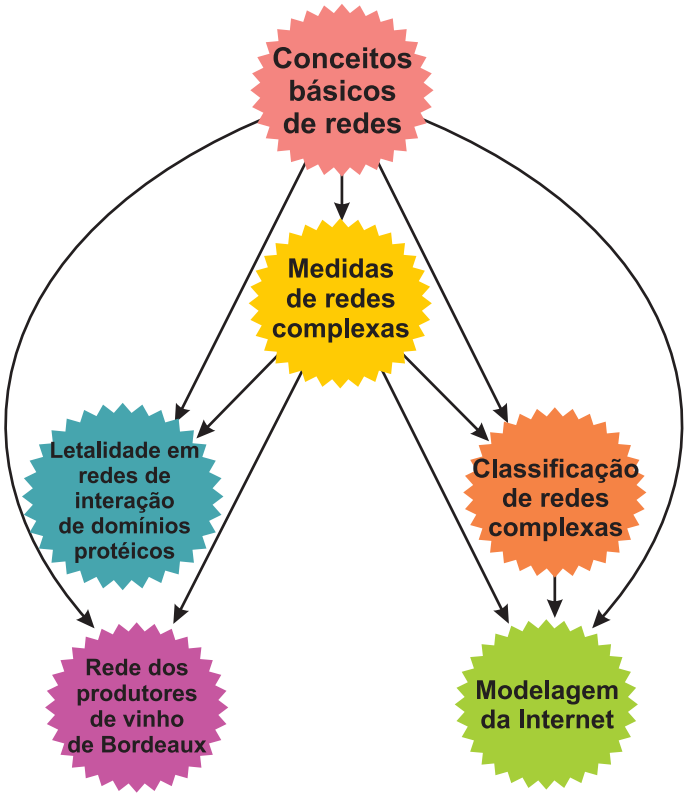


Figura 1.1: Relação entre os capítulos que constituem a tese.

1.3 Descrição dos capítulos

Nos próximos capítulos, introduzimos diversas medidas de redes complexas e métodos de análise multivariada reconhecimento de padrões, que são utilizados na caracterização, análise e classificação de diversas redes reais. Iniciamos com uma introdução sobre conceitos básicos de redes complexas, onde apresentamos um breve histórico sobre o desenvolvimento desta área de pesquisa, os conceitos fundamentais e alguns modelos de redes. A seguir, discutimos as principais medidas topológicas desenvolvidas para quantificar propriedades estruturais de redes complexas. No capítulo posterior, mostramos como a classificação de redes pode ser realizada utilizando análise multivariada e decisão Bayesiana. Nesse caso, consi-

deramos diversos modelos e medidas para definirmos o espaço de classificação e determinamos os modelos que melhor reproduzem a estrutura de várias redes reais. Um resultado interessante é que muitas das redes reais que possuem distribuição de conexões do tipo livre de escala foram classificadas como redes geográficas, cuja distribuição de conexões é do tipo Poisson. Destarte, a análise por reconhecimento de padrões que sugerimos exigirá uma revisão dos modelos existentes e uma nova perspectiva para o desenvolvimento de modelos mais completos. Finalmente, realizamos diversas aplicações nos capítulos seguintes considerando três redes reais: (i) a Internet, (ii) a rede de interação de proteínas do *S. cerevisiae* e (iii) a rede dos produtores de vinhos da região de Bordeaux (França). Propomos um modelo de crescimento da Internet e o comparamos com diversos outros utilizando análise das variáveis canônicas e decisão Bayesiana. Mostramos que dentre os modelos de redes considerados, o modelo que desenvolvemos é o mais preciso para representar a estrutura e evolução da Internet. A análise das redes de interações de proteínas e domínios nos leva a conclusão de que a correlação entre letalidade e conectividade é mais definida no nível dos domínios protéicos do que no nível das proteínas. No caso da produção de vinhos, modelamos as semelhanças entre os produtores como redes e analisamos a influência do território nas técnicas de cultivo e propriedades da produção. Comparando a classificação obtida pela teoria das redes complexas com a divisão territorial, mostramos que os atributos relativos à produção dos vinhos são fortemente influenciados pela região a que os *chateaux* pertencem. Além disso, determinamos quais fatores mais influenciam na qualidade dos vinhos e encontramos os *chateaux* protótipos de cada região. No último capítulo, discutimos as conclusões e as possibilidades de trabalhos futuros. A Figura 1.1 apresenta a estrutura de ligação entre os capítulos que constituem a tese.

Capítulo 2

Conceitos Básicos de redes complexas

Sistemas complexos são formados por muitos elementos capazes de interagir entre si e com o meio ambiente. Propriedades destes sistemas incluem (i) *emergência*: a complexidade do todo é maior do que a complexidade da soma das partes, (ii) *auto-organização*: o sistema se organiza sem um comando externo e (iii) *universalidade*: sistemas pertencentes à mesma classe possuem propriedades semelhantes [BY92]. Como esses sistemas são formados por partes discretas que se conectam, a modelagem por redes é realizada de forma natural [AO04]. Por exemplo, o cérebro é formado por células nervosas conectadas por axônios [WS98, dFCS05, dFCS06b, dFCKH07]; a sociedade é formada por pessoas ligadas por laços de amizade [WF94]; as cadeias alimentares são formadas por animais e plantas conectados por relações de predatismo [MHH98]; as conexões entre aeroportos são definidas por rotas aéreas [GA04, GMT$^+$05]; as malhas rodoviárias são formadas por rodovias que ligam cidades [GN06]; a Internet é composta por roteadores ligados por fibras-ópticas [FFF99, YJB02, AJB00]; a Teia Mundial é definida por documentos conectados por *hyperlinks* [AJB99, BAJ00, AH00]; a

linguagem é resultante da interação de palavras que se relacionam por similaridade e função [dFC04b, ANJdFC07] e os cientistas estão conectados de acordo com os trabalhos de colaboração [BJR+02, New01]. Aplicações da teoria das redes complexas na modelagem de sistemas naturais e artificiais são encontradas em diversos artigos de revisão [Str01, AB02, New03b, AO04] e livros [Bar03, Wat03, SG03, Cal07].

2.1 Introdução histórica

A teoria das redes complexas nasceu da aplicação de medidas desenvolvidas pela teoria dos grafos e conceitos provenientes da mecânica estatística, física não-linear e sistemas complexos. A teoria dos grafos, particularmente, começou com o trabalho de Leonhard Euler para resolver o famoso problema das *Sete Pontes de Königsberg* (Prússia no século XVIII, atual Kaliningrado, Rússia), onde haviam duas grandes ilhas que, juntas, formavam um complexo que continha sete pontes. Discutia-se nas ruas da cidade a possibilidade de alguém atravessar todas as pontes sem repetir nenhuma. A chance de tal façanha havia se tornado uma lenda popular quando Leonhard Euler, em 1736, provou que não existia tal caminho. Ele modelou o problema das sete pontes como um grafo, transformando caminhos em arestas e suas intersecções em vértices (ver a Figura 2.1), criando, possivelmente, o primeiro grafo da história [Bol98, Die00]. Desde então, os avanços obtidos por tal teoria se restringiram, em grande parte, a grafos estáticos e com relativamente poucas aplicações práticas [NBW06]. Dentre tais aplicações, a sociologia usou a teoria dos grafos, a partir da década de 50, para analisar redes formadas por indivíduos ligados por algum tipo de interação social (e.g. amizade e trabalho) [WF94]. Esta aplicação revelou que a teoria dos grafos pode ser usada como uma ferramenta prática para analisar dados empíricos. Além disso,

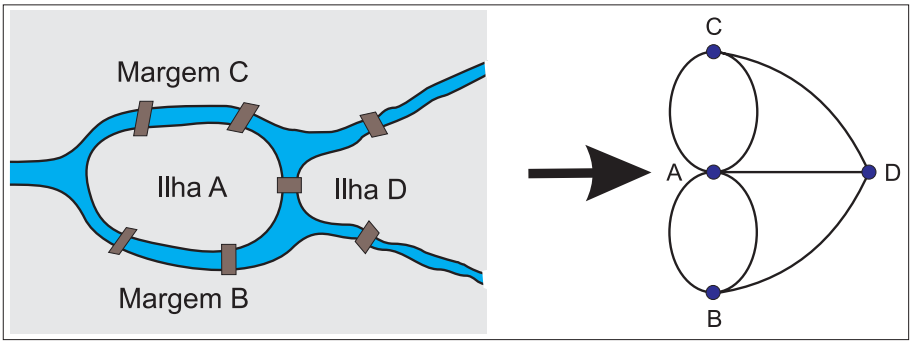


Figura 2.1: A configuração das pontes antes de 1875, com a ilha de Kneiphof (A), a área de terra (D) entre os dois braços do rio Pregel, e as duas porções de terra que circundam a ilha (C e B). Euler transformou essa configuração em um grafo e provou que não é possível alguém atravessar todas as pontes passando apenas uma vez por cada uma delas.

nesta mesma época, alguns pesquisadores começaram a usar grafos como meio de simulação de processos dinâmicos, como a propagação de doenças [NBW06], opiniões [TdFC06, RC05] e caminhadas aleatórias (*random walks*) [dFCT07].

Em 1967, uma importante propriedade presente nas redes sociais foi descoberta quando Stanley Milgram, um pesquisador de sociologia em Harvard, Estados Unidos, interessado na estrutura da sociedade americana, descobriu que a distância média entre duas pessoas quaisquer nos Estados Unidos é próxima de seis. O experimento que determinou tal resultado foi realizado com o envio de centenas de cartas a pessoas residentes em Wichita (Kansas) e Omaha (Nebraska). Estas pessoas foram escolhidas de forma aleatória e na carta era perguntado se elas conheciam a esposa de um aluno de graduação que residia em Sharon (Massachusetts) ou se conheciam um corretor de fundos públicos em Boston. Caso elas conhecessem, as cartas deveriam ser enviadas aos respectivos destinatários. Caso contrário, as pessoas deveriam colocar seus dados e enviar as cartas a outras pessoas que supostamente os conheciam. As missivas deveriam passar pelas mãos de

diversos indivíduos até chegarem ao seu destino e, dessa forma, Milgram poderia saber a rota pela qual elas teriam passado. Das 160 cartas enviadas, 42 chegaram ao destino e, com isso, Milgram determinou o caminho médio que separava duas pessoas quaisquer nos Estados Unidos. Tal distância foi determinada como sendo 5,5 e arredondada para seis. A partir desta descoberta, surgiu o famoso termo *seis graus de separação* [Mil67]. Segundo o referido princípio, a distância média entre um esquimó no Alaska e uma pessoa residente em São Paulo deve ser em torno de seis. Esse fenômeno é conhecido como efeito de mundo pequeno (*small world*). Hoje, acredita-se que a distância média que separa duas pessoas em qualquer região do planeta seja menor do que a distância encontrada por Milgram [Bar03].

O efeito *small world* é devido ao fato da distância de separação crescer mais lentamente do que o tamanho da rede. Ou seja, se cada vértice está ligado em média a k vizinhos, o número de vértices entre um dado vértice i e outro j localizado a uma distância ℓ de i, é igual a k^ℓ. Como k^ℓ não deve ser maior do que N, temos $\ell \leq \log N / \log k$. Assim, a origem do efeito *small world* é que a distância cresce com o logaritmo da rede e decresce com o logaritmo da média do número de conexões. Hoje, sabemos que a maioria das redes complexas apresentam esta propriedade, como a Internet ($\ell \approx 10$), a Teia Mundial ($\ell \approx 11$), as espécies em cadeias alimentares ($\ell \approx 2$) e as moléculas presentes nas células ($\ell \approx 3$) [AB02, New03b, Bar03].

Apesar dos avanços obtidos em sociologia, a falta de poder computacional e bases de dados de redes reais fizeram com que as investigações em redes complexas se reduzissem à análise de redes formadas por algumas dezenas ou no máximo centenas de vértices. Além disso, o estudo de redes sociais limitava-se a caracterizar a estrutura das ligações, como no trabalho de Milgram, e não em modelar as interações sociais ou investigar as propriedades dinâmicas dessas redes, como

sua evolução. Apenas a partir do final da década de 90, com o advento da Internet e com o aumento do poder computacional, a teoria das redes complexas foi estabelecida, pois bases de dados foram disponibilizadas e os computadores tornaram-se capazes de processar um grande volume de informações. Embora existam semelhanças, a teoria das redes complexas difere da teoria dos grafos em três aspectos básicos: (i) ela está relacionada com a modelagem de redes reais, por meio de análise de dados empíricos; (ii) as redes estudadas não são estáticas, mas evoluem no tempo, modificando sua estrutura; (iii) as redes, muitas vezes, não são consideradas apenas objetos topológicos, mas constituem estruturas onde processos dinâmicos (como a propagação de doenças e opiniões) podem ser simulados. Além de princípios da teoria dos grafos, a teoria das redes complexas apresenta conceitos de física estatística, sistemas não lineares, fractais e autômatos celulares, dentre outras áreas. Esse caráter multidisciplinar permite a teoria das redes complexas cobrir aplicações desde biologia até sociologia, sendo a computação responsável pelas ferramentas utilizadas na modelagem, simulação e tratamento das bases de dados.

2.2 Conceitos básicos

Fundamentalmente, as redes complexas são descritas por um conjunto de vértices (nós) que são ligados por arestas (conexões, ligações ou *links*) devido a algum tipo de interação [New03b]. Essas redes podem ser estáticas, quando não há variação no número de vértices, arestas ou mesmo na configuração das ligações; ou dinâmicas, sendo que, neste caso, é possível modelar o seu crescimento pela análise da variação de sua estrutura no tempo. Embora as redes reais sejam dinâmicas, elas podem ser analisadas como estáticas dentro de um intervalo de tempo em que as variações são inexistentes ou pouco importantes. Matema-

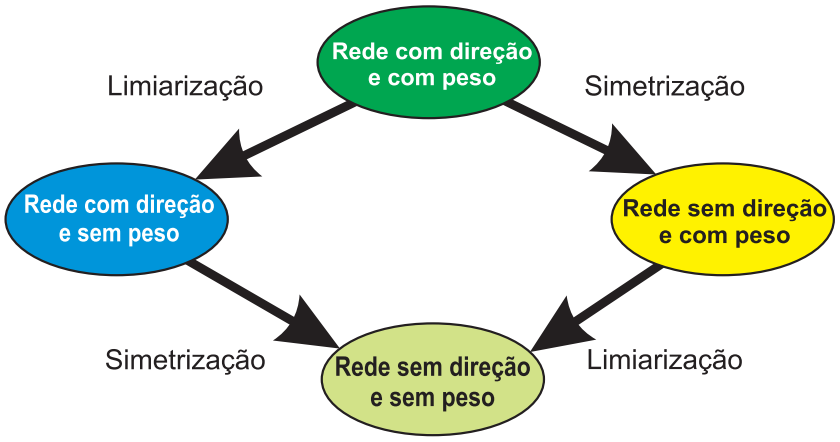


Figura 2.2: Representação esquemática das relações entre os tipos de redes. Figura adaptada de [dFCRTB07].

ticamente, uma rede $R = (\mathcal{N}, \mathcal{E})$ é formada por um conjunto de N vértices, $\mathcal{N} = \{n_1, n_2, \ldots, n_N\}$, e um conjunto de M arestas, $\mathcal{E} = \{e_1, e_2, \ldots, e_M\}$. As conexões nessas redes podem ser dirigidas, quando o sentido da ligação importa, ou não-dirigidas. Se as ligações possuem intensidade, como no caso da largura de banda em fibras ópticas que ligam roteadores, a cada aresta é associado um peso. Neste caso, a rede deve apresentar informações adicionais sobre os pesos, isto é, além de ser formada pelos conjuntos $\mathcal{N}, \mathcal{E}$, a rede possui ainda o conjunto $\mathcal{W} = \{w_1, w_2, \ldots, w_M\}$, que representa o peso das ligações, sendo a rede representada por $R = (\mathcal{N}, \mathcal{E}, \mathcal{W})$. Deste modo, o tipo de rede mais geral é aquela cujas ligações são dirigidas e possuem uma intensidade associada [dFCRTB07]. A partir deste tipo de rede mais geral, é possível obter as demais configurações, através de operações de limiarização, para obtenção de redes sem peso, e simetrização, para a obtenção de redes não-dirigidas. A Figura 2.2 apresenta um diagrama com estas operações. A limiarização é realizada retirando-se arestas cujo peso seja menor do que um limiar definido e associando peso unitário às arestas remanescentes. Já a simetrização transforma as ligações dirigidas em não-dirigidas.

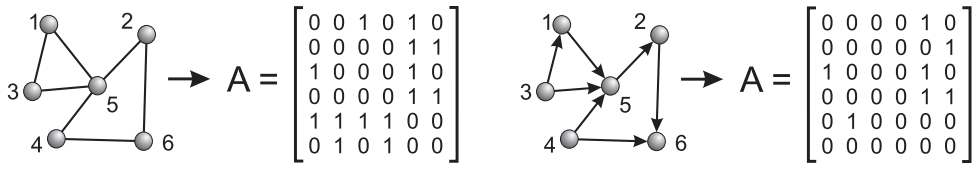
Em termos computacionais, as redes podem ser armazenadas através de listas ou matrizes de adjacência, por exemplo. No caso da lista, apenas os pares de vértices (i, j) que possuem ligações são armazenados. Já no caso da matriz de adjacência A, se dois vértices i e j estão ligados, a entrada a_{ij} na matriz será igual a 1 e igual a 0, caso contrário. A Figura 2.3 mostra um exemplo de mapeamento de uma rede não-dirigida e de uma dirigida em matrizes de adjacência. Quando as conexões na rede possuem peso, a lista tem um terceiro elemento relacionado à intensidade das ligações, (i, j, w_{ij}). Além disso, ao invés da matriz de adjacência é utilizada uma matriz de pesos W, que armazena os pesos das ligações entre os vértices, dados pelas entradas w_{ij} na matriz. Cada estrutura de armazenamento tem suas vantagens e desvantagens. O uso das listas permite maior economia de memória (quando as redes são esparsas) do que o uso das matrizes de adjacência, embora o acesso às ligações seja mais complexo, porque são necessárias buscas na lista. Neste trabalho discutiremos as medidas em termos do armazenamento usando matrizes de adjacência. Portanto, as operações de simetrização e limiarização, apresentadas na Figura 2.2, podem ser expressas da seguinte forma:

- *Limiarização*: $A = \delta_T(W)$,
- *Simetrização*: $\psi(A) = \delta_1(A + A^T)$,

onde $A = \delta_T(W)$ associa $a_{ij} = 1$ se $w_{ij} > T$ ou $a_{ij} = 0$, caso contrário. A matriz A^T é a transposta de A.

Uma medida básica para caracterização da estrutura de redes é dada pela média do número de conexões entre os vértices, denominada *conectividade média*, $\langle k \rangle$. A *conectividade* de um dado vértice i para uma rede não-dirigida é dada por

$$k_i = \sum_{j=1}^{N} A_{ij}, \tag{2.1}$$



(a) **(b)**

Figura 2.3: As redes complexas podem ser representadas por matrizes de adjacência. Em (a) temos uma rede não-dirigida e em (b) uma rede dirigida. No caso (a), os elementos a_{ij} da matriz são iguais a 1 se há uma ligação entre os vértices i e j e iguais a zero, caso contrário. Já no caso (b), os elementos da matriz a_{ij} são iguais a 1 se existe uma conexão dirigida do vértice i para o vértice j.

e a conectividade média,

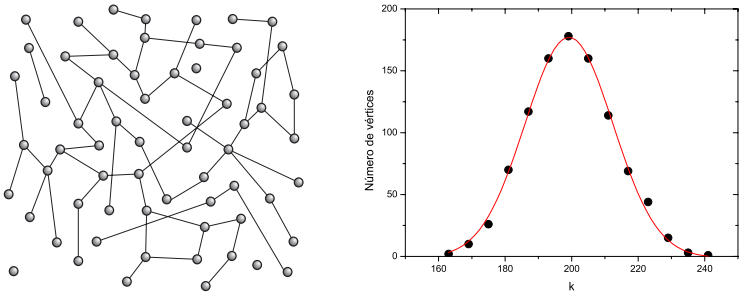
$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i. \tag{2.2}$$

O *grau* é definido como uma generalização da conectividade, pois leva em conta multi-conexões entre dois vértices [New03b]. Em nosso trabalho, como não consideramos redes cujos pares distintos de vértices possuem mais de uma ligação entre si, adotaremos grau e conectividade como sinônimos. Através da conectividade e da sua distribuição é possível, por exemplo, caracterizar diversos tipos de redes, assim como determinar se a configuração das conexões de uma dada rede é definida de forma aleatória ou se possui alguma lei de formação. A análise das conexões foi fundamental nas primeiras investigações a respeito das redes complexas, e estimulou pesquisas futuras, como o desenvolvimento de modelos para reproduzir a estrutura de redes reais.

2.3 Modelos de redes complexas

Em 1959, dois matemáticos húngaros, Paul Erdős e Alfred Rényi, consideraram os grafos como objetos estocáticos, ao invés de analisá-los de forma puramente determinística, como fazia até então a matemática discreta e a sociologia. Destarte, eles sugeriram um modelo de rede baseado em ligações aleatórias, que ficou conhecido como *grafos aleatórios de Erdős e Rényi* [ER59, ER60, ER61]. O referido grafo é construído iniciando-se com um conjunto de N vértices totalmente desconectados e a cada passo dois vértices são escolhidos aleatoriamente e conectados com uma probabilidade fixa p, sendo cada par de vértices considerado apenas uma vez. Assim sendo, todas as ligações possuem a mesma probabilidade de ocorrerem, ou seja, a rede gerada tem uma estrutura altamente homogênea. Na Figura 2.4(a) é mostrado um exemplo de rede aleatória. A distribuição da conectividade para essas redes, quando N é grande e a conectividade média é mantida constante, tende à distribuição de Poisson (ver Figura 2.4(b) e Tabela 3.1). Além disso, o caminho mínimo médio é pequeno nessas redes, caindo com o logaritmo do tamanho da rede, $\ell \sim \ln N / \ln\langle k\rangle$, sendo $\langle k\rangle = 2M/N = p(N-1)$ o número médio de conexões na rede e M o número de arestas.

Erdős e Rényi estavam interessados apenas na riqueza matemática das redes aleatórias e não em aplicações práticas. Eles apenas mencionaram em seu artigo de 1959 que a evolução dos grafos poderia ser considerada como um modelo muito simplificado de certas redes de comunicação, como estradas e ferrovias [Bar03]. Logo, os trabalhos publicados por estes estudiosos consideram apenas propriedades matemáticas dos grafos e não aplicações práticas. Mesmo assim, o modelo de Erdős e Rényi suscitou questões relacionadas à estrutura das conexões em redes reais. Uma delas argüia se estas conexões poderiam ser representadas pelo modelo aleatório. Entretanto essa resposta só foi obtida no final da década de 90, quando novas bases de dados surgiram e o poder computacional



(a) **(b)**

Figura 2.4: (a) Um exemplo de um grafo aleatório de Erdős e Rényi. (b) A distribuição da conectividade para uma rede com 10.000 vértices, usando uma probabilidade $p = 0,2$. Cada ponto no gráfico é a média sobre 10 redes. Ilustração adaptada de [dFCRTB07].

aumentou. O primeiro passo nessa descoberta foi dado em 1998, quando Duncan Watts e Steven Strogatz, pesquisadores das Universidade de Columbia e Cornell respectivamente, observaram que em algumas redes reais, tais como a rede de neurônios do *Caernohabditis elegans* e a rede de distribuição de energia dos Estados Unidos; a presença de *loops* (caminhos fechados) de ordem três é muito maior do que nas redes aleatórias com mesmo número de vértices e arestas [WS98]. Esse foi o primeiro indício de que as redes reais não são completamente aleatórias, mas possuem uma determinada lei de formação. Baseados nesta descoberta, Watts e Strogatz sugeriram um modelo alternativo aos grafos aleatórios, chamado modelo *small world de Watts-Strogatz* (em analogia ao fenômeno descoberto por Stanley Milgram), que apresenta o efeito *small world* e a presença de um grande número de *loops* de ordem três. Neste modelo, eles assumem que as redes presentes na natureza não são completamente regulares e nem mesmo aleatórias, todavia se situam entre esses dois extremos.

Para a obtenção do modelo *small world*, inicia-se com uma rede regular for-

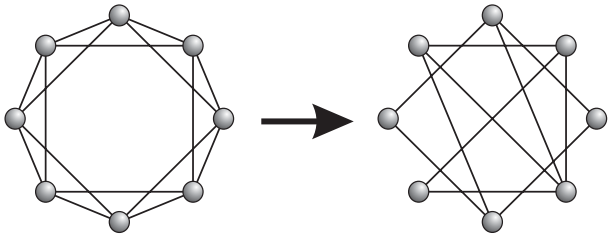
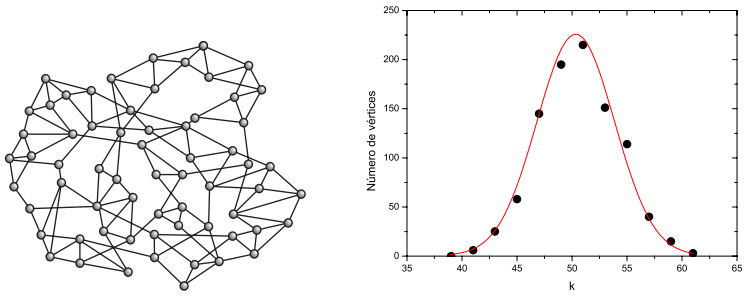


Figura 2.5: As redes *small world* de Watts e Strogatz são construídas a partir de uma rede regular, religando as arestas com probabilidade p.

mada por N vértices ligados a κ vizinhos mais próximos em cada direção, totalizando 2κ conexões iniciais, sendo $N \gg \kappa \gg \log(N) \gg 1$. A seguir, cada aresta é aleatoriamente reconectada com uma probabilidade fixa p, que introduz o caráter aleatório à rede. Quando $p = 0$ a rede é completamente regular, apresentando alta quantidade de *loops* e caminho médio alto, e quando $p = 1$, a rede é aleatória, apresentando baixa quantidade de *loops* de ordem três, mas pequeno caminho médio. Portanto, tal modelo se situa entre a completa regularidade e a aleatoriedade. A emergência do regime *small world* ocorre para $p > 0{,}01$, quando o menor caminho médio converge para o valor encontrado nos grafos aleatórios e a ocorrência de ciclos de ordem três permanece da ordem das redes regulares. Na Figura 2.5 é apresentado o mecanismo de construção do modelo *small world*. A Figura 2.6(a) ilustra um exemplo de rede gerada por tal modelo e a Figura 2.6(b) apresenta a distribuição da conectividade para uma rede *small world* formada por 10.000 vértices.

O trabalho de Watts e Strogatz formalizou um antigo problema de sociologia proposto por Mark Granovetter em 1973 [Gra73], que estudou as relações entre pessoas na sociedade e mostrou que as ligações mais fracas de amizade são extremamente importantes. Granovetter demonstrou que os laços afetivos entre familiares e entre amigos íntimos não oferecem uma diversidade de conhecimento tão grande como as relações entre pessoas conhecidas e amigos distantes. Tal



(a) **(b)**

Figura 2.6: (a) Um exemplo de uma rede *small world* formada por 64 vértices. Note a presença de um elevado número de *loops* de ordem três. (b) A distribuição da conectividade para uma rede *small world* formada por 1.000 vértices, $\kappa = 25$ e $p = 0{,}3$. Figura adaptada de [dFCRTB07].

conclusão foi obtida quando Granovetter entrevistou dezenas de trabalhadores e perguntou a estes quem os tinha ajudado a encontrar um emprego. Na maioria dos casos (27,8 % dos casos), a informação sobre tal emprego vinha de conhecidos, ao invés de amigos íntimos (16,7 % dos casos). Isto ocorre porque os amigos íntimos tendem a compartilhar as mesmas informações [Gra95]. As implicações da descoberta de Granovetter vão desde a sociologia até a economia, *marketing* e política [Gra85, Gra95]. No modelo de Watts e Strogatz, as ligações mais fracas sugeridas por Granovetter são aquelas estabelecidas pela reconexão de arestas. Estas ligações têm um papel fundamental na conexão entre os membros de grupos fechados (formado por familiares e amigos íntimos) e o mundo externo. É através delas que as redes sociais se tornam *small world*. O trabalho de Watts e Strogatz trouxe o fenômeno *small world* da sociologia para as comunidades formadas por físicos e matemáticos, inspirando novas investigações em redes complexas.

Para explicar a topologia da Internet, em 1998, Waxman [Wax88] propôs um modelo geográfico cujas ligações são estabelecidas de acordo com a distância

espacial entre os vértices. Neste caso, N vértices são distribuídos aleatoriamente em um espaço bidimensional e conectados com uma probabilidade que decai com a distância Euclidiana D entre eles (por exemplo, $P(i \rightarrow j) \sim \alpha e^{-\lambda D_{ij}}$). Esse modelo, embora incorporasse um dos fatores que são importantes para definir as conexões entre os roteadores na Internet (a distância geográfica), gera distribuição das conexões similar à encontrada no modelo aleatório de Erdős e Rényi.

Em 1999, dois pesquisadores da Universidade de Notre Dame, Estados Unidos, Albert-László Barabási e Reka Albert [BA99], decidiram verificar se o fenômeno *small world* estava presente na Teia Mundial (*World Wide Web*). Através do uso de um *web crawler*, eles mapearam a topologia das conexões entre as páginas da Teia Mundial e descobriram que além dela apresentar o fenômeno *small world* ($\ell \approx 11$), a distribuição de conexões não é aleatória, mas do tipo *livre de escala* (*scale-free*), que é da forma $P(k) \sim k^{-\gamma}$ [AJB99]. Naquele mesmo ano, o mesmo tipo de distribuição já havia sido encontrada na Internet pelos irmãos Faloutsos [FFF99], embora tal trabalho não tivesse despertado forte impacto. A partir dessas descobertas, verificou-se que diversas redes apresentam distribuição de conexões do tipo livre de escala e, portanto, o universo aleatório de Erdős e Rényi tende a não estar presente na natureza. Assim, o trabalho de Watts e Strogatz sugeriu a primeira limitação do modelo de Erdős e Rényi, que foi a ausência de ciclos de ordem três, embora tenha mantido o caráter aleatório. Já o de Barabási e Albert descartou a aleatoriedade e mostrou que há leis que regem a estrutura das redes reais.

A distribuição livre de escala é um tipo de distribuição de probabilidades que reflete invariância de escala. Leis de potência também são conhecidas como Lei de Zipf ou Distribuição de Pareto. Tais distribuições são idênticas, embora alguns autores tenham gerado confusão ao diferenciá-las [New05]. A distribuição de Pareto foi proposta por Vilfredo Pareto [Par42] no início do século 20, que de-

monstrou que certos fenômenos em economia, assim como na física, podem ser modelados matematicamente. Deste modo, a economia não é governada por simples aleatoriedade, mas possui leis que regem o seu comportamento. Quantidades geradas aleatoriamente possuem uma escala típica, sendo descritas por curvas características definidas por uma média e um desvio padrão (ver Figura 2.4(b)). Por exemplo, a distribuição de riquezas, do tamanho das cidades, dos preços de livros, dos diâmetros das crateras lunares, da intensidade dos terremotos, do número de conexões por roteadores e o número de citações por artigo; não possuem uma média e desvio característicos, sendo invariantes por escala [New05]. Pareto havia observado que em muitos fenômenos, 80% das conseqüências advém de 20% das causas (regra 80/20), o que gera uma lei de potência, que é uma curva continuamente decrescente sem um pico característico descrita por um único expoente (ver Figura 2.7(b)). Diferentemente da uniformidade, leis de potência sugerem que muitos eventos pequenos podem coexistir com poucos eventos grandes.

Motivados pela descoberta da estrutura da Teia Mundial, pesquisadores verificaram que a falta de uniformidade na estrutura das redes complexas é um fenômeno universal [Bar03], sendo observado, por exemplo, nas redes de colaboração entre cientistas [BJR+02], nas redes de interações de proteínas [JMBO01] e nas redes metabólicas [JTA+00]. Inspirados por essas descobertas, Babrabási e Albert propuseram um modelo de crescimento, que gera redes livres de escala [BA99], que é baseado em dois passos:

1. *Crescimento*: Iniciando-se com um pequeno número de vértices N_0, a cada passo é adicionado um vértice com m $(m \leq N_0)$ arestas que se conectam com vértices já presentes na rede.

2. *Ligação preferencial*: O novo vértice, que vai ser adicionado à rede, tende a se conectar com os vértices mais conectados, ou seja, a probabilidade de um

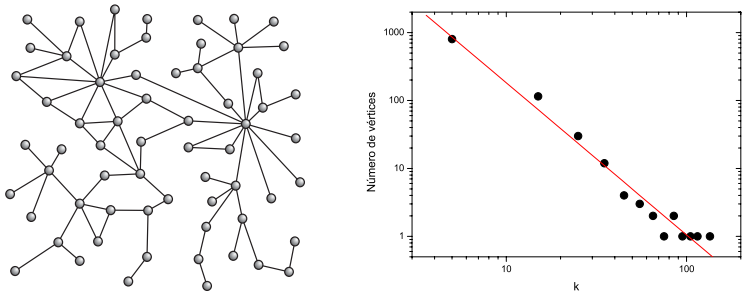
vértice j, presente na rede, ser escolhido é proporcional a sua conectividade,

$$\mathcal{P}_{i\to j}(n+1) = \frac{k_j(n)}{\sum_{u=-N_0+1}^{n} k_u(n)}, \tag{2.3}$$

onde n é o tempo e o número de vértices adicionados à rede.

Interessante notar que estes dois mecanismos de construção das redes livres de escala não estão presentes no modelo aleatório de Erdős e Rényi e no modelo *small world* de Watts e Strogatz, pois nestes não há adição de novos vértices e as conexões são estabelecidas de forma homogênea, havendo uma conectividade característica (conectividade média) na rede. Por outro lado, nas redes geradas pelo modelo livre de escala, os vértices mais conectados tendem a receber mais conexões — paradigma conhecido como *"o rico fica mais rico"*. Estas redes são formadas por um reduzido número de vértices altamente conectados, denominados *hubs*, e por uma grande quantidade de vértices pouco conectados, o que define a distribuição livre de escala. Na Figura 2.7 é mostrada uma rede gerada pelo modelo de Barabási e Albert e a distribuição da conectividade para uma rede composta por 10.000 vértices.

O modelo de Barabási e Albert possui várias similaridades com o modelos desenvolvido por Price [dSP76], em 1976, para explicar a distribuição das conexões em redes de citação, que o mesmo autor havia encontrado uma década antes [dSP65]. Entretanto, no modelo de Price a probabilidade de que um novo artigo i cite um anterior j é proporcional a $k_j^{in} + 1$, onde k_j^{in} é o número de vezes que artigo j já foi citado. O modelo de Price, por sua vez, é uma reformulação do modelo de Simon [Sim55], que foi desenvolvido em 1955 para explicar a lei de potência que aparece em vários dados empíricos. Portanto, as distribuições livre de escala têm sido observadas desde a década de 50, mas sua popularidade cresceu com os trabalhos de Barabási e Albert no final da década de 90, pois foi apenas a partir dessa época que diversas bases de dados dos mais diversos sistemas complexos tornaram-se disponíveis. Deste modo, o desenvolvimento da



(a) **(b)**

Figura 2.7: (a) Exemplo de uma rede gerada pelo modelo livre de escala de Barabási e Albert. (b) Distribuição das conexões para uma rede livre de escala formada por 10.000 vértices considerando $m = 5$. A distribuição das conexões segue uma lei de potência, diferentemente das redes apresentadas nas Figuras 2.4 e 2.6. Cada ponto é uma média sobre 10 redes. Figura adaptada de [dFCRTB07].

Internet e a disponibilidade de informações que possibilitassem a construção de redes biológicas, sociais, tecnológicas e de informação, foram fundamentais para o desenvolvimento da teoria das redes complexas.

A Tabela 8 apresenta algumas redes reais com seus respectivos número de vértices, conectividade média, menor caminho médio, expoente da distribuição das conexões e as referências onde estas redes são descritas. A Figura 2.8 apresenta a representação visual da Internet obtida em 2005 pelo *The Opte Project* (`http://www.opte.org`). A distribuição das conexões nesta rede é do tipo livre de escala.

Além do modelo de Barabási e Albert, muitos outros foram propostos para aperfeiçoá-lo, de forma a permitir maior flexibilidade no valor de γ, que é igual a três quando $N \to \infty$ no modelo de Barabási e Albert [DM03], e um coeficiente de aglomeração próximo do encontrado em redes reais. Amaral *et al.* [ASBS00]

Rede complexa	N	$\langle k \rangle$	ℓ	γ	Referências
Atores	449.913	113, 43	3, 48	2, 3	[WS98, ASBS00]
Chamadas telefônicas	47×10^6	3, 16	–	2, 1	[ACL01]
Mensagens de e-mail	59.912	1, 44	4, 95	1, 5 – 2, 0	[EMB02]
Contatos sexuais	2.810	–	–	3, 2	[LEA03, LEA+01]
WWW (nd.edu)	269.504	5, 55	11, 27	2, 1 – 2, 4	[AJB99, BAJ00]
WWW (Altavista)	203.549.046	10, 46	16, 18	2, 1 – 2, 7	[BKM+00]
Rede de citações	783.339	8, 57	–	3.0	[Red98]
Co-ocorrência de palavras	460.902	70, 13	–	2, 7	[Dor01, CS01]
Internet	10.697	5, 98	3, 31	2, 5	[CCGJ02, FFF99]
Pacotes de *softwares*	1.439	1, 20	2, 42	1, 6 – 1, 4	[New03a]
Circuitos eletrônicos	24, 097	4, 34	11, 05	3, 0	[CJS01]
Redes metabólicas	765	9, 64	2, 56	2, 2	[JTA+00]
Interações protéicas	2.115	2, 12	6, 80	2, 4	[JMBO01]

Tabela 2.1: Exemplos de redes complexas livre de escala.

generalizaram o modelo de Barabási e Albert demonstrando que o mecanismo de ligação preferencial pode ser limitado por três fatores básicos: (i) *idade dos vértices*: alguns vértices tendem a não receber mais conexões a partir de um certo tempo, mas continuam contribuindo para a formação da estrutura da rede; (ii) *custo da adição de novas arestas e capacidade limitada*: em algumas redes, os vértices podem não receber ligações devido ao alto custo que tais conexões podem gerar ou devido à sua capacidade limitada de gerenciar novos *links*; (iii) *filtragem de informações*: as conexões podem se limitar aos vértices mais semelhantes [MBESA02]. No primeiro caso, temos como exemplo a rede de colaboração de atores, que possuem um tempo limitado de atividade, não recebendo mais links quando deixam de atuar em filmes. No segundo, temos as redes de aeroportos, que podem deixar de receber conexões aéreas por não possuírem os recursos para gerenciá-las. Finalmente, no terceiro caso, temos a Teia Mundial, onde as conexões são limitadas por assuntos. Neste caso, por exemplo, há uma probabilidade baixa de uma página relacionada a futebol se conectar a outra cujo assunto seja religião.

Assim, a ligação entre os vértices, ao invés de ser regulada pela Equação 2.3, segue a seguinte lei,

$$\mathcal{P}_{i\to j}(n+1) = \frac{k_j(n) f(k_j(n), n, j, \ldots)}{\sum_{u=-m_0+1}^{n} k_u(n) f(k_u(n), n, u, \ldots)}, \tag{2.4}$$

onde $f(k_j(n), n, j, \ldots)$ é uma função de custo, que pode depender da conectividade do vértice j, de sua idade e de outros fatores limitantes. A partir destes resultados, Amaral *et al.* sugeriram que existem três classes de redes *small world*: (i) livre de escala, (ii) larga-escala ou livre de escala truncada, que possui uma distribuição livre de escala com um limite (*cutoff*) exponencial e (iii) escala simples, caracterizada por uma distribuição com um decaimento rápido, como exponencial ou gaussiano [ASBS00].

Para permitir maior variação no coeficiente da lei de potência γ, Dorogovtsev *et al.* [DMS00] propuseram um modelo com ligação preferencial da forma

$$\mathcal{P}_{i\to j}(n+1) = \frac{k_j(n) + k_0}{\sum_{u=-m_0+1}^{n} (k_u(n) + k_0)}, \tag{2.5}$$

com $-m < k_0 < \infty$. Neste caso, $\gamma = 3 + k_0/m$, ou seja γ pode variar de 2 até ∞. Quando $k_0 = 0$, obtém-se o modelo de Barabási e Albert.

Outra generalização do modelo livre de escala foi proposta por Krapivsky *et al.* [KRL00]. Neste caso, os vértices são conectados por uma probabilidade de conexão não-linear, da forma,

$$\mathcal{P}_{i\to j}(n+1) = \frac{k_j(n)^\alpha}{\sum_{u=-m_0+1}^{n} k_u(n)^\alpha}. \tag{2.6}$$

Se $\alpha = 1$, temos o modelo de Barabási e Albert. Quando $\alpha < 1$, a rede gerada tem distribuição exponencial e quando $\alpha > 1$ uma parte da rede se conecta a quase todas outras partes. Deste modo, a linearidade na probabilidade de ligação ($\alpha = 1$) é um fator determinante da estrutura livre de escala.

Na ligação preferencial, há uma tendência dos vértices mais antigos serem os *hubs* da rede. Entretanto, Bianconi e Barabási verificaram que em alguns casos

reais, vértices mais novos podem atrair um grande número de conexões e se tornam os mais conectados. Por exemplo, na Teia Mundial, o *site* de busca *Google* se tornou o maior *hub* em apenas dois anos de existência, atraindo grande parte das conexões de outros *sites* [Bar03]. Deste modo, Bianconi e Barabási sugeriram um modelo de ligação preferencial baseado na "popularidade" dos vértices, onde cada vértice é diferenciado por seu coeficiente de adaptação (*fitness*), η_j, escolhido a partir de uma distribuição η [BB01]. A ligação preferencial, neste caso, segue uma probabilidade da forma

$$\mathcal{P}_{i\to j}(n+1) = \frac{k_j(n)\eta_j}{\sum_{u=-m_0+1}^{n} k_u(n)\eta_u}. \tag{2.7}$$

Neste modelo, pode-se ainda associar níveis de energia aos vértices e mapeá-los como um condensado de Bose-Einstein [BB01]. Apesar de simples, tal modelo proporcionou uma modelagem mais precisa de diversos sistemas complexos, como a Teia Mundial e a rede de colaboração entre atores.

Além dos modelos discutidos anteriormente, muitos outros foram sugeridos de modo a permitirem uma modelagem mais precisa de diversas redes reais. Para maiores detalhes sobre tais modelos, recomendamos a revisão apresentada na Referência [BLM+06]. Embora a maioria dos modelos de redes sejam capazes de reproduzir a lei de potência na distribuição das conexões e outras propriedades estruturais, como o caminho médio pequeno e a alta probabilidade de ocorrência de *loops* de ordem três, eles falham na previsão de estruturas importantes, como as comunidades e determinados tipos de subgrafos [MSOI+02]. Conseqüentemente, as leis de formação da maioria das redes reais ainda não foram identificadas [Bar03]. Além disso, redes com estrutura completamente distinta podem apresentar a mesma distribuição de conexões [ADLW05]. Para a construção de modelos mais precisos é necessário analisar diversas medidas topológicas e tentar inferir os mecanismos dinâmicos que determinam a estrutura das redes. No próximo capítulo, discutiremos as principais medidas de redes complexas e a sua

utilidade na caracterização de redes reais.

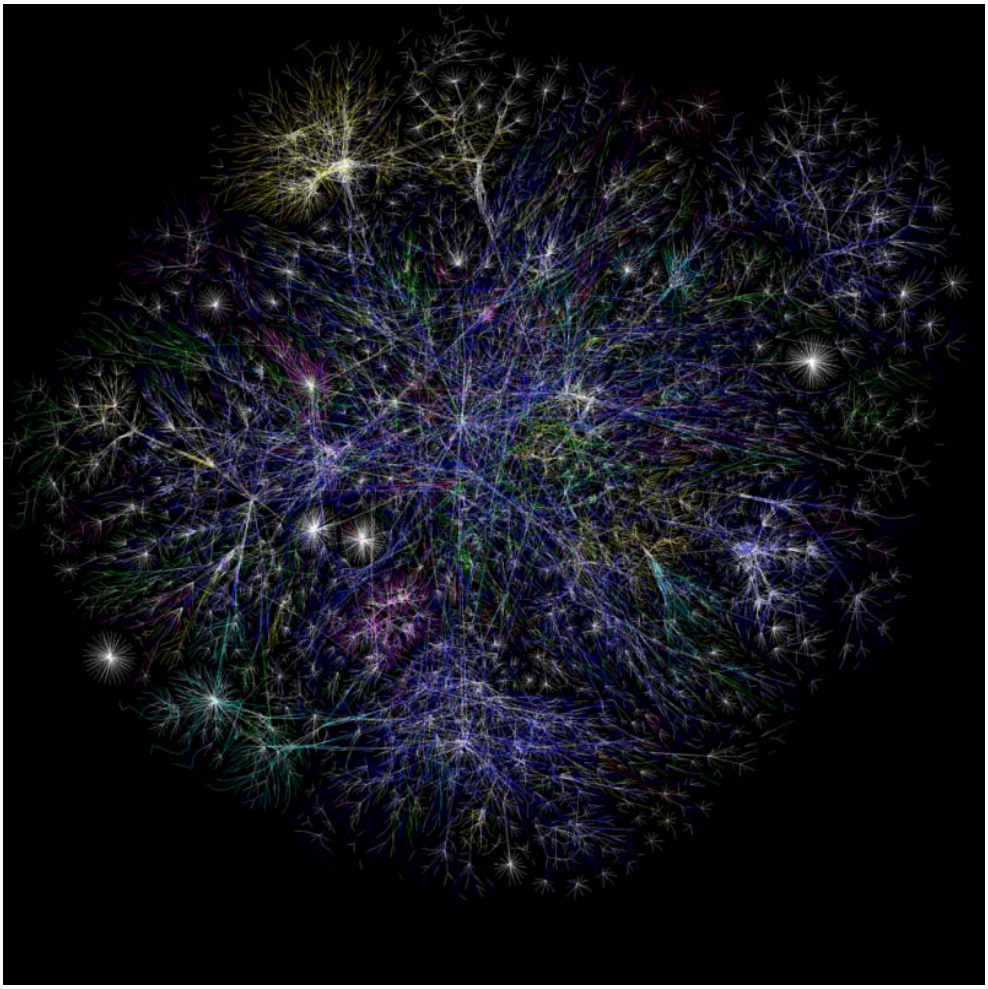


Figura 2.8: Representação visual da Internet obtida em 15 de janeiro de 2005 pelo *The Opte Project* (`http://www.opte.org`). As cores indicam os seguintes domínios: (i) net, ca, us (azul), (ii) com, org (verde), (iii) mil, gov, edu (vermelho), (iv) jp, cn, tw, au (amarelo), (v) de, uk, it, pl, fr (rosa escuro), (vi) br, kr, nl (azul claro) e (vii) desconhecido (branco).

Capítulo 3

Medidas para análise, caracterização e classificação de redes complexas

A caracterização da estrutura de redes que representam sistemas complexos é fundamental em diversos ramos da ciência. Por exemplo, para curar muitas das doenças, é necessária a compreensão da estrutura das ligações entre componentes celulares (e.g. genes, proteínas e metabólicos) em sua integridade e como as interações entre eles ocorrem, pois o conhecimento destes componentes individualmente não é suficiente para se entender a dinâmica celular [OB02]. Além disso, as células têm uma complexa rede reguladora que controla desde o metabolismo até a morte celular e a caracterização da estrutura desta rede pode oferecer relevantes informações, como a importância de cada gene e de que forma as mensagens são trocadas dentro das células [BBO05a, BBO05b]. Em epidemiologia, o conhecimento da estrutura das ligações entre as pessoas, como os laços de amizade, é fundamental para se entender como as epidemias se espalham e determinar meios eficazes de imunização [GAM89, BnPS02, New02b]. De

forma semelhante, caracterizar a topologia da Internet é importante na análise da propagação de vírus entre computadores e na determinação de como se pode bloquear a infestação [BFNW04]. Como podemos notar através destes poucos exemplos, a caracterização de redes complexas é fundamental não só para a descrição da estrutura das ligações entre seus elementos, mas também para a compreensão dos mecanismo evolutivos que governam seu crescimento.

Redes complexas podem apresentar diferentes topologias, dependendo dos mecanismos que determinam sua evolução. Por exemplo, as redes sociais apresentam arquitetura muito diferente das redes tecnológicas, como a Internet, ou biológicas, como as redes de interação de proteínas [NP03]. Para se quantificar a estrutura das ligações, diversas medidas têm sido desenvolvidas. Através de medidas, redes podem ser analisadas, caracterizadas, classificadas e modeladas [dFCRTB07].

A análise e caracterização de redes é importante no estudo das relações entre forma e função em redes complexas. Por exemplo, se uma rede apresenta topologia do tipo livre de escala, é esperado que sua estrutura seja altamente tolerante à falhas. Ademais, a caracterização da estrutura das ligações de redes é importante na análise de simulações de processos dinâmicos, como fluxos de informações [TRT06, TRRdFC06], falhas [AJB00] e autômatos celulares [RC05]. Neste caso, conhecendo-se a estrutura das redes, é possível determinar quais aspectos topológicos influenciam na simulação de processos dinâmicos.

A classificação de redes pode ser realizada agrupando-se em uma mesma categoria as redes cujas medidas estruturais forneçam valores estatisticamente semelhantes. Com isso, é possível a construção de uma taxonomia de redes complexas. Neste caso, a escolha das medidas é fundamental, uma vez que a utilização de um conjunto reduzido de medidas, ou mesmo a consideração de medidas que forneçam resultados redundantes, pode ocasionar classificações equi-

vocadas. Ademais, também é possível determinar qual modelo melhor representa uma dada rede real, dentre um conjunto de modelos considerados. Deste modo, as medidas permitem determinar a compleza de modelos, isto é, quanto mais características topológicas um dado modelo é capaz de reproduzir, mais completo ele deve ser.

De acordo com as aplicações aqui descritas, vemos que para realizar a caracterização, comparação, classificação e modelagem de redes complexas, é fundamental o conhecimento das medidas e sua utilidade. A seguir, são descritas as principais medidas de redes complexas que serão utilizadas nos capítulos subseqüentes.

3.1 Medidas relacionadas à conectividade

Uma medida simples associada a um vértice i é a *conectividade*, k_i, cujo valor é obtido pela Equação 2.1 (ver Figura 3.1). A partir dela é possível calcular a respectiva medida global, chamada *conectividade média* da rede, $\langle k \rangle$, pela Equação 2.2. Para os modelos: (i) grafos aleatórios de Erdős e Rényi, $\langle k \rangle = pN$, (ii) *small-world* de Watts e Strogatz, $\langle k \rangle = 2\kappa$, e (iii) livre de escala de Barabási e Albert, $\langle k \rangle = 2m$. Se as conexões na rede possuem intensidade, como a largura de banda associada aos roteadores que formam a Internet, é utilizada a *força* do vértice, s_i, ao invés da conectividade, que é calculada por,

$$s_i = \sum_{j=1}^{N} w_{ij}. \tag{3.1}$$

A respectiva medida global é a *força média* da rede,

$$\langle s \rangle = \frac{1}{N} \sum_{i=1}^{N} s_i. \tag{3.2}$$

Estas medidas, apesar de simples, podem ser usadas na identificação de *hubs* (vértices altamente conectados) e para quantificar a densidade de conexões [AJB00].

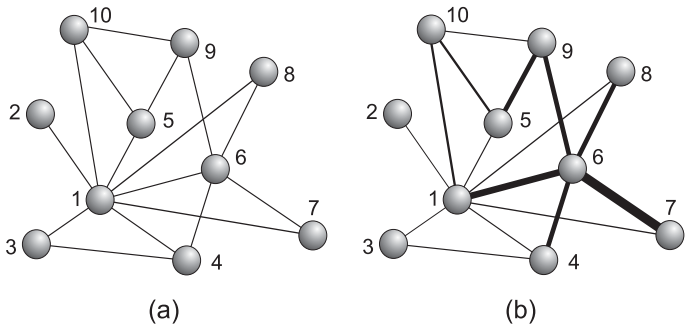


Figura 3.1: Exemplo de duas redes formadas por 10 vértices e 15 arestas. Enquanto que na rede sem peso (a), o vértice 1 é um *hub* porque concentrar grande parte das conexões ($k_1 = 8$), na rede (b) o vértice 6 é um *hub* por possuir ligações com maior intensidade (a intensidade da ligação é representada pela largura da aresta).

Os *hubs* têm importância fundamental na formação da estrutura das redes complexas, já que a sua remoção pode causar a fragmentação da rede, resultando em componentes não conectados[1]. Esse fenômeno está relacionado às falhas aleatórias e ataques em redes, sendo que, dependendo da aplicação, pode-se desejar redes robustas ou vulneráveis. No caso da robustez, um exemplo é a Internet, que deve manter o tráfego apesar da interrupção de alguns roteadores [AJB00]. Já no caso da vulnerabilidade, um exemplo é a sociedade, pois no caso de propagação de epidemias, é desejável que a parte infectada seja isolada do restante da rede quando alguns indivíduos morrem ou são colocados em quarentena (vértices são removidos). Com isso, há o bloqueio da transmissão da doença para o restante da rede [BnPS02].

A *distribuição das conexões*, $P(k)$, representa a probabilidade de um vértice escolhido aleatoriamente ter conectividade k (ver Figuras 2.4 e 2.7, que apresen-

[1]Componentes não conectados são formados por vértices sem conexões ou grupos de vértices conectados entre si e não conectados ao restante da rede.

tam as distribuições de probabilidades para uma rede aleatória de Erdős e Rényi e livre de escala de Barabási e Albert, respectivamente). Da mesma forma é definida a *distribuição da força*, $P(s)$, sendo que ao invés da conectividade, são consideradas as forças dos vértices [BBPSV04]. Tais medidas são importantes, por exemplo, para determinar se as conexões entre os vértices são uniformes, servindo como uma medida básica de classificação, já que se uma rede possui uma distribuição de Poisson, essa pode ser associada ao modelo de Erdős e Rényi. Uma maneira de determinar o quanto uma distribuição se aproxima da lei de potência, sem levar em conta a inclinação da distribuição, pode ser realizada através da correlação (coeficiente de Pearson) entre a probabilidade $P(k)$ e a respectiva conectividade k na escala logarítmica, chamada medida de *"retidão" da distribuição*, st [dFCRTB07]. Tal correlação é obtida da seguinte forma,

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

onde x_i e y_i são valores do i-ésimo elemento dos vetores x e y, de tamanho n, e $\langle x \rangle$ and $\langle y \rangle$ representa a respectiva média de x e y. Para calcular st, basta substituir x pelo logaritmo da conectividade k e y pelo logaritmo probabilidade $P(k)$. O valor de r_{xy} pode variar entre -1 e 1. O caso $r = -1$ indica a presença de uma rede com distribuição livre de escala (o gráfico é uma reta decrescente), pois o coeficiente de Pearson tem valor mínimo quando duas variáveis são completamente anti-correlacionadas.

A análise das conexões pode também ser utilizada para determinar as correlações entre os vértices, i.e., se os *hubs* têm uma tendência a se ligarem. Neste caso, se a probabilidade de dois vértices i e j se conectarem independe da conectividade de cada um deles, a rede é dita não-correlacionada. A grande maioria das redes reais apresentam correlação entre os vértices [New03b] e portanto, os modelos de redes devem reproduzir essa característica. Como os modelos

aleatórios de Erdős e Rényi e livre de escala de Barabási e Albert geram redes não-correlacionadas [New02a], eles não são apropriados para representar a maioria das redes reais. Em algumas redes, a correlação entre os vértices pode ser fundamental no estudo de processos dinâmicos, como em análises de resistência à falhas e propagação de epidemias [BLM+06]. Uma maneira de determinar a correlação entre as conectividades é por meio da medida de *assortatividade*, que pode ser determinada pelo *coeficiente de Pearson* considerando-se as conectividades em ambos os lados de uma aresta [New02a],

$$r = \frac{\frac{1}{M}\sum_{j>i} k_i k_j a_{ij} - \left[\frac{1}{M}\sum_{j>i}\frac{1}{2}(k_i + k_j)a_{ij}\right]^2}{\frac{1}{M}\sum_{j>i}\frac{1}{2}(k_i^2 + k_j^2)a_{ij} - \left[\frac{1}{M}\sum_{j>i}\frac{1}{2}(k_i + k_j)a_{ij}\right]^2}. \tag{3.3}$$

Caso $r > 0$, os vértices de conectividades similares tendem a se ligar e a rede é chamada assortativa. Se $r < 0$, vértices mais conectados tendem a se ligar com os menos conectados e a rede é denominada dissassortativa. Entretanto, se $r = 0$, não há qualquer correlação entre as conectividades, ou seja, a rede é dita não-correlacionada.

3.2 Medidas relacionadas a ciclos

As redes reais apresentam uma alta ocorrência de *loops* de ordem três (subgrafos formados por três vértices totalmente conectados). Em redes sociais, essa propriedade, chamada aglomeração (*clustering*) ou transitividade, indica a probabilidade de dois amigos quaisquer A e B terem um amigo C em comum. Para medir a fração de tais subgrafos é usada a medida chamada *coeficiente de aglomeração* (*clustering coefficient*), que mede a razão entre o número de arestas entre os vizinhos de um dado vértice i, denotado por e_i, e o número máximo possível de arestas entre esses vizinhos, que é dado por $k_i(k_i - 1)/2$. Na Figura 3.2 são apresentados três exemplos de configurações que geram diferentes coeficientes de

aglomeração. Em termos da matriz de adjacência, o coeficiente de aglomeração é calculado por [WS98],

$$cc_i = \frac{2e_i}{k_i(k_i-1)} = \frac{\sum_{j=1}^{N}\sum_{m=1}^{N} a_{ij}a_{jm}a_{mi}}{k_i(k_i-1)}. \tag{3.4}$$

De maneira similar, uma possível definição do coeficiente de aglomeração para uma rede com peso é a seguinte [BBPSV04],

$$cc_i^w = \frac{1}{s_i(k_i-1)}\sum_{j>m}\frac{w_{ij}+w_{im}}{2}a_{ij}a_{im}a_{jm}. \tag{3.5}$$

Como medida de caracterização global da rede, é calculada a média do coeficiente de aglomeração entre todos os vértices da rede,

$$\langle cc\rangle = \frac{1}{N}\sum_{i=1}^{N} cc_i, \tag{3.6}$$

e para redes com peso,

$$\langle cc^w\rangle = \frac{1}{N}\sum_{i=1}^{N} cc_i^w. \tag{3.7}$$

A Tabela 3.1 apresenta formas analíticas do coeficiente de aglomeração para os modelos de Erdős e Rényi, de Watts e Strogatz e de Barabási e Albert.

O coeficiente de aglomeração pode ainda ser expresso como função da conectividade dos vértices. Neste caso,

$$cc(k) = \frac{\sum_i cc_i\delta_{k_ik}}{\sum_i \delta_{k_ik}}, \tag{3.8}$$

onde δ_{ij} é a função delta de Kronecker ($\delta_{ij} = 1$ se $i = j$ ou zero, caso contrário). Para algumas redes reais, Ravasz e Barabási mostraram que essa função tem a forma $cc(k) \sim k^{-\omega}$, onde ω é chamado *expoente hierárquico*, já que o comportamento de $cc(k)$ é associado com a estrutura hierárquica da rede [RB03]. A dependência entre cc e k está associada às correlações na rede, onde os vértices mais conectados tendem a se ligarem com os menos conectados, conforme observado por Soffer e Vázquez [SV05].

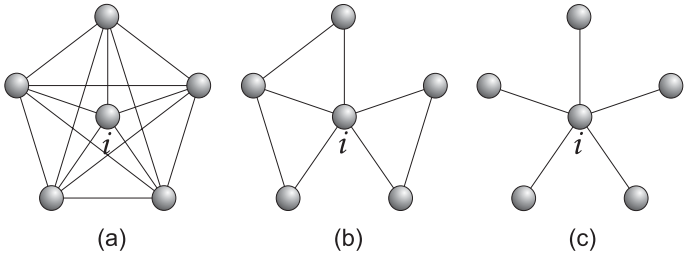


Figura 3.2: Ilustração esquemática de três situações onde o coeficiente de aglomeração tem diferentes valores. Em (a) é apresentado um exemplo de clique, onde todos os vértices estão conectados entre si. Neste caso, $cc_i = 1$. Na figura (b), $cc_i = 3/10$. Já em (c) $cc_i = 0$, pois seus vizinhos não possuem conexões entre si.

Além da distribuição livre de escala e da ocorrência de ciclos de ordem três, outras propriedades importantes podem ser quantificadas em redes complexas. Dependendo da rede, ciclos de ordem superior podem ser fundamentais em processos dinâmicos, como transporte e robustez [RKBbA05]. O número de ciclos pode ser computado através da matriz de adjacência. Neste caso, o número de ciclos de ordem três, quatro e cinco são dados por:

$$N_3 = \frac{1}{6}\sum_i (A^3)_{ii}, \tag{3.9}$$

$$N_4 = \frac{1}{8}\left[\sum_i (A^4)_{ii} - 2\sum_i (A^2)_{ii}(A^2)_{ii} + \sum_i (A^2)_{ii}\right], \tag{3.10}$$

$$N_5 = \frac{1}{10}\left[\sum_i (A^5)_{ii} - 5\sum_i (A^2)_{ii}(A^3)_{ii} + 5\sum_i (A^3)_{ii}\right]. \tag{3.11}$$

Um fator interessante relacionado a estes ciclos observado por Bianconi *et al.*, é que as distribuições estatísticas de ciclos de ordens três, quatro e cinco permanecem constantes durante a evolução da Internet [BCC05]. Muitos outros trabalhos foram elaborados de forma a estudar a distribuição de ciclos em redes livre de escala [BM06b].

Outras medidas relacionadas a ciclos são o *coeficiente cíclico* [KK05] e o coeficiente *rich-club* [ZM04a, CFSV06]. O *coeficiente cíclico local* é dado por,

$$\Theta_i = \frac{2}{k_i(k_i-1)} \sum_{k>j} \frac{1}{S_{ijk}} a_{ij} a_{ik}, \tag{3.12}$$

onde S_{ijk} representa o tamanho do menor ciclo que passa entre os vértices i, j e k. Se não há conexões entre estes vértices, $S_{ijk} = \infty$. O *coeficiente cíclico* é dado pela média sobre todos os vértices,

$$\Theta = \frac{1}{N} \sum_i \Theta_i. \tag{3.13}$$

O coeficiente *rich-club* mede a tendência dos *hubs* se conectarem, formando comunidades. Esse efeito é largamente observado em diversas redes reais, como nas redes de colaboração, onde os cientistas mais conectados tendem a formar grupos de colaboração e publicar artigos em conjunto [CFSV06]. Para medir essa relação entre os *hubs*, Zhou e Mondragon [ZM04a] definiram o coeficiente *rich-club* para uma dada conectividade k, da seguinte forma,

$$\phi(k) = \frac{1}{|\mathcal{R}(k)|(|\mathcal{R}(k)|-1)} \sum_{i,j \in \mathcal{R}(k)} a_{ij}, \tag{3.14}$$

onde $\mathcal{R}(k) = \{v \in \mathcal{N}(G) | k_v > k\}$ representa o conjunto de vértices da rede G cujas conectividades são maiores do que k. De forma análoga, Colizza *et al.* definiram este coeficiente para redes com peso,

$$\phi^w(s) = \frac{\sum_{i,j \in \mathcal{R}^w(s)} w_{ij}}{\sum_{i \in \mathcal{R}^w(s)} s_i}, \tag{3.15}$$

onde $\mathcal{R}^w(s) = \{v \in \mathcal{N}(G) | s_v > s\}$ representa o conjunto de vértices com força maior do que s.

3.3 Medidas relacionadas à distância

A distância entre os vértices é um fator importante que está relacionado ao transporte e comunicação em redes. Por exemplo, a chance de perda de pacotes tro-

cados entre computadores na Internet pode aumentar com a distância entre eles, já que quanto maior a distância, maior o número de roteadores no caminho e, conseqüentemente, maior a probabilidade de ocorrer falhas ou congestionamento. Caso ocorram falhas nos principais roteadores da Internet, o diâmetro da rede aumenta e há atraso na transmissão ou perda de pacotes [Tan02], o que prejudica o tráfego de informações. Portanto, as medidas relacionadas à distância são fundamentais no estudo da estrutura e dinâmica de redes reais.

O *comprimento do caminho* que conecta dois vértices i e j é dado pelo número de arestas ao longo deste caminho. O *comprimento do menor caminho* entre dois vértices i e j, d_{ij}, é dado pela extensão de todos os caminhos que conectam estes vértices cujos comprimentos são mínimos [WS98]. Sua determinação é importante para caracterização da estrutura interna das redes e na investigação de efeitos dinâmicos relativos ao transporte e à comunicação [BLM+06]. Os menores caminhos entre todos os vértices em uma rede podem ser representados através de uma matriz de distâncias D, cujos elementos d_{ij} expressam o valor do menor caminho entre os vértices i e j. O valor $d_{max} = \max_{i,j} d_{ij}$ é chamado *diâmetro* da rede. A média entre os valores na matriz D exprime o *caminho característico da rede* (*menor caminho médio*), sendo calculada por

$$\ell = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij}. \tag{3.16}$$

Caso i e j não pertençam a um mesmo componente conectado, $d_{ij} = \infty$. A Tabela 3.1 apresenta as formas analíticas de ℓ para os modelos de Erdős e Rényi, de Watts e Strogatz e de Barabási e Albert.

Um problema na definição de ℓ é que se há vértices desconectados seu valor diverge. Uma possibilidade para evitar tal problema foi proposta por Latora e Machiori [LM01], que introduziram uma medida chamada *eficiência global*, cujo

Tabela 3.1: Resultados analíticos de algumas medidas básicas para os modelos de Erdős-Rényi,Watts-Strogatz e Barabási-Albert. Tabela extraída de [dFCRTB07].

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

$\star$ Nas redes de Watts-Strogatz, os valores 2κ representam o número de vizinhos de cada vértice na rede regular (na Figura 2.5, $\kappa = 4$).

$*$ A função $f(u) =$ constante se $u \ll 1$ ou $f(u) = \ln(u)/u$ se $u \gg 1$.

cálculo é realizado da seguinte forma,

$$E = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{ij}}. \tag{3.17}$$

Essa medida é um indicador da capacidade de tráfego de uma rede. O recíproco da eficiência global é a *média harmônica do menor caminho médio*,

$$h = \frac{1}{E}, \tag{3.18}$$

que evita o problema da divergência que ocorre com a Equação 3.16. Portanto, a média harmônica do menor caminho médio é uma medida mais apropriada para redes que possuam componentes desconectados.

A medida de eficiência pode ainda ser empregada para determinar quais vértices são os mais importantes na obtenção do melhor transporte na rede. Na maioria dos casos, os *hubs* correspondem a estes vértices. No entanto, mesmo quando os vértices têm aproximadamente a mesma conectividade, pode haver uma hierarquia de importância entre eles. Por exemplo, no caso de uma árvore binária, os vértices mais próximos da raiz são os mais fundamentais, pois quando eles são retirados ocorre a quebra da rede em diferentes componentes conectados. Na Figura 3.3(a), os vértices A, B e C ocupam a hierarquia mais alta.

Se associarmos a performance de uma rede com sua eficiência global, ou seja, que uma maior eficiência resulta em uma maior performance; a *vulnerabilidade* de um vértice i é determinada pela queda na performance da rede quando esse vértice e suas respectivas conexões são removidos [GKS04]. Assim, a vulnerabilidade de um vértice i é calculada da seguinte forma,

$$V_i = \frac{E - E_i}{E}, \tag{3.19}$$

onde E_i é a eficiência calculada após a remoção do vértice i. O vértice com maior valor de V_i ocupa a posição mais alta na hierarquia da rede. A *vulnerabilidade da*

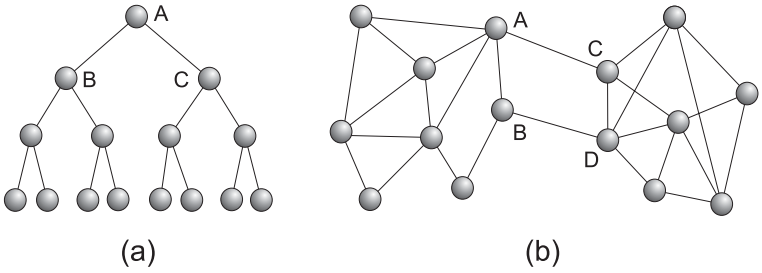


Figura 3.3: (a) Numa árvore binária, os vértices que ocupam a posição mais alta na hierarquia são os mais vulneráveis (A, B e C), pois sua remoção causa a ruptura da rede. (b) Os vértices com maior grau de intermediação (*betweenness centrality*) são aqueles que estão entre comunidades (A, B, C e D), pois participam da maioria dos menores caminhos da rede.

rede é expressa pelo máximo valor da vulnerabilidade obtido considerando todos os vértices,

$$V = \max_i V_i. \tag{3.20}$$

Ainda com relação ao transporte na rede, alguns vértices ou arestas recebem um tráfego mais intenso do que outros. Tais elementos representam os chamados "gargalos" e estão situados entre muitos dos menores caminhos. Quando removidos, podem ocorrer rupturas na estrutura da rede, surgindo componentes não conectados, que são formados por vértices densamente conectados entre si, mas não conectados com o restante da rede. Para medir o tráfego que passa em um dado vértice (ou aresta), é usada a medida chamada *grau de intermediação* (*betweenness centrality*) [Fre77], que mede o quanto um vértice ou aresta está no caminho entre outros vértices, e é calculada da seguinte forma,

$$B_u = \sum_{ij} \frac{\sigma(i,u,j)}{\sigma(i,j)}, \tag{3.21}$$

onde $\sigma(i,u,j)$ é o número de menores caminhos entre os vértices i e j que passam pelo vértice (ou aresta) u e $\sigma(i,j)$ é o número total de menores caminhos entre i e

j. A soma é feita sobre todos os pares distintos i, j de vértices. Na Figura 3.3(b), os vértices A, B, C e D são os que possuem maior grau de intermediação. A média do grau de intermediação pode ser utilizada como uma medida de caracterização global da rede,

$$\langle B \rangle = \frac{1}{N} \sum_i B_i. \tag{3.22}$$

A partir do *grau de intermediação* pode-se ainda obter uma medida global chamada *dominância do ponto central*, que é calculada pela seguinte equação,

$$c_D = \frac{1}{N-1} \sum_i (B_{\max} - B_i), \tag{3.23}$$

onde $B_{\max}$ é o maior valor da *betweenness* na rede [Fre77, Fre79]. A dominância do ponto central será 0 para uma rede completamente conectada (ver Figura 3.2(a)) e será 1 para uma rede tipo estrela (ver Figura 3.2(c)), na qual um vértice central está incluso em todos os caminhos da rede. Outras medidas de centralidade são discutidas na Referência [KLP+05].

3.4 Subgrafos

Um grafo g é um *subgrafo* do grafo G se $\mathcal{N}(g) \subseteq \mathcal{N}(G)$ e $\mathcal{E}(g) \subseteq \mathcal{E}(G)$, com as arestas em $\mathcal{E}(g)$ se estendendo sobre vértices em $\mathcal{N}(g)$. Há muitas maneiras de definir subgrafos conforme apresentado na Figura 3.4. Importantes tipos de subgrafos são os ciclos de ordens três, quatro ou superiores, conforme discutido na Seção 3.2. Além destes, há os subgrafos chamados *k-núcleos* (*k-cores*), que são obtidos através da decomposição da rede. Neste caso, remove-se todos os vértices com conectividade menor que k e suas respectivas conexões [DVGM06, GDM06]. Depois de cada remoção, o número de conexões de alguns dos vértices que foram mantidos na rede pode ser menor do que k, já que muitas arestas são perdidas quando os vértices são extraídos. O processo de decomposição é repetido até que

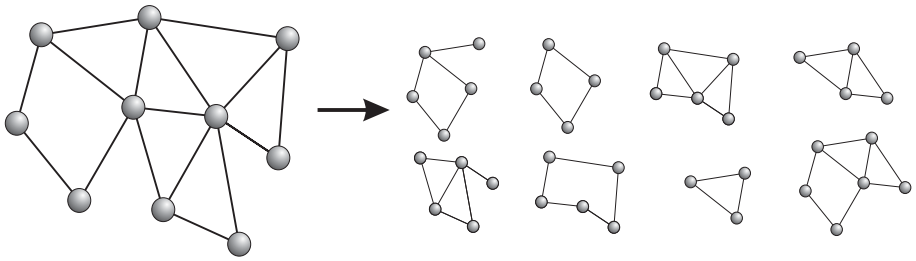


Figura 3.4: Exemplo de uma rede e alguns dos seus possíveis subgrafos.

não haja mais vértices a serem retirados. O subgrafo resultante é o *k-núcleo* (*k-core*) da rede. Uma aplicação importante de tal metodologia é na visualização de redes [AHDV05]. Além disso, os k-núcleos são utilizados no estudo da topologia de diversas redes reais, como a Internet [CHK+06] e as redes de interações de proteínas. Neste último caso, os k-núcleos podem ser considerados na previsão da função de proteínas [AUANK+03] e na análise de aspectos de letalidade [WA05].

Em rede reais, alguns tipos de subgrafos aparecem mais freqüentemente do que nas redes aleatórias equivalentes[2]. Tais subgrafos são denominados *motivos* (*motifs*) e estão intimamente relacionados à estrutura e evolução de redes complexas [MSOI+02]. Os motivos podem ser considerados em redes dirigidas e não-dirigidas. Na Figura 3.5 são apresentados os principais tipos de motivos encontrados em redes reais.

Para quantificar a ocorrência de um dado motivo i em uma rede, a medida chamada *Z-score* é calculada por,

$$Z_i = \frac{N_i^{(\text{real})} - \langle N_i^{(\text{rand})} \rangle}{\sigma_i^{(\text{rand})}}, \tag{3.24}$$

onde $N_i^{(\text{real})}$ é o número de vezes que o motivo i aparece na rede real, $\langle N_i^{(\text{rand})} \rangle$ e $\sigma_i^{(\text{rand})}$ são respectivamente a média e o desvio padrão do número de ocorrências

[2]Redes aleatórias equivalentes são construídas pelo modelo de configuração [BC78] ou métodos de reconexão de arestas [MKI+03]. Estas redes possuem número de vértices, arestas e distribuição de conexões iguais aos da rede original.

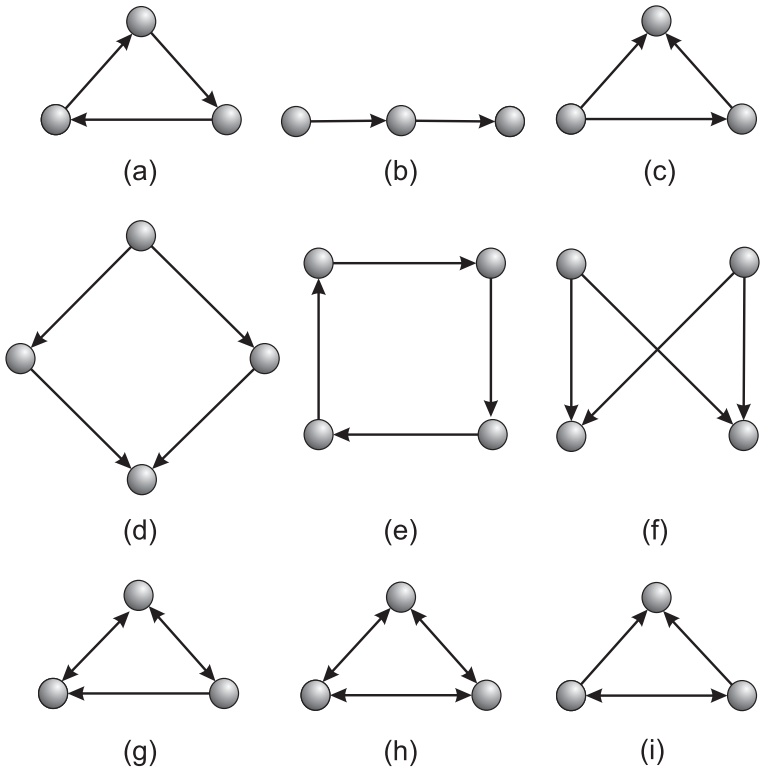


Figura 3.5: Exemplos de motivos: (a) *three-vertex feedback*, (b) *three chain*, (c) *feed-forward loop*, (d) *bi-parallel*, (e) *four-vertex feedback*, (f) *bi-fan*, (g) *feedback with two mutual dyads*, (h) *fully connected triad* e (i) *uplinked mutual dyad*. Mantivemos os nomes em inglês conforme encontrados na literatura. Figura adaptada de [dFCRTB07].

do motivo no grupo de redes aleatórias. Deste modo, Z_i quantifica a ocorrência do motivo i na rede. Um subgrafo é considerado um motivo se a probabilidade dele aparecer na rede randomizada é menor do que $P = 0{,}01$ [MSOI+02].

Redes reais podem ser classificadas em famílias de acordo com os motivos que possuem [MIK+04]. Para isso, é associado a cada rede um vetor SP, chamado *perfil de significância* (*significance profile*), cujo elemento SP_i contém o *Z-score*

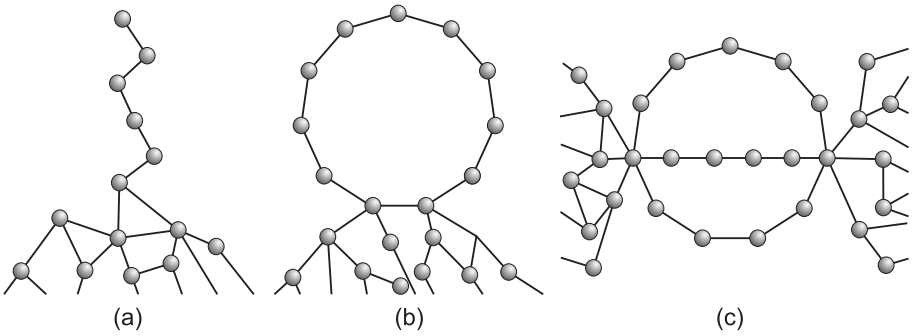


Figura 3.6: Tipos de cordões (*chains*): (a) um cauda (*tail*), (b) uma corrente (*handle*) e (c) uma corrente de ordem três (*3-handle*).

do motivo i e é calculado da seguinte forma,

$$\mathrm{SP}_i = \frac{Z_i}{\sum_j Z_j^2}. \tag{3.25}$$

A partir da análise comparativa de diversas redes, Milo *et al.* observaram que os motivos presentes em cadeias alimentares são diferentes dos encontrados em redes genéticas do *E. coli* e *S. cerevisiae* ou daqueles encontrados na Teia Mundial [MSOI+02]. Motivos semelhantes foram encontrados em redes que realizam processamento de informações, como entre os neurônios na rede neural do *C. elegans* e entre as biomoléculas dentro de uma célula. Assim sendo, os motivos podem definir classes de redes complexas. Além disso, a ocorrência de motivos também foi analisada em redes biológicas [SOMMA02, DBBO04, YLSK+04, BL04]. Neste caso, verificou-se que diferentes motivos são conservados ao longo da evolução de diversos organismos [WOB03]. Em outro tipo de análise, Albert e Albert [AA04] utilizaram a ocorrência de motivos na previsão de funções de proteínas.

Um outro tipo de motivo que encontramos em redes complexas são denominados "cordões" (*chains*): um cordão C é definido por um conjunto de vértices $V_C = \{i_1, i_2, \ldots, i_{k-1}, i_k, i_{k+1}, \ldots, i_{N_C}\}$ conectados de maneira seqüencial onde os vértices internos i_k, com $k = 2, \ldots, N_C - 1$ têm apenas dois vizinhos, i_{k-1} e

i_{k+1}. Os vértices das extremidades, i_1 e i_{N_C}, podem ser conectados com o restante da rede [VBRTdFC07]. O tamanho do cordão é dado pelo número de conexões entre os vértices em V_C, que é igual a $N_C - 1$. No caso em que os cordões têm uma das extremidades desconectada do restante da rede, a rede é chamada "*cauda*" (*tail*), ou seja, i_1 ou i_{N_C} tem conectividade igual a um. Caso as duas extremidades estejam conectadas com a rede, o cordão é denominado "*corrente*" (*handles*). Quando n cordões compartilham da mesma extremidade, o motivo é denominado "*corrente de ordem* n" (n-*handle*). A Figura 3.6 apresenta alguns tipos de cordões, correntes e caudas. Na Referência [VBRTdFC07] são analisadas a ocorrência de cordões em diversas redes reais e em redes geradas por modelos. Um resultado interessante é que dentre os modelos considerados, aleatório de ErdErdős e Rényi, *small world* de Watts e Strogatz e livre de escala de Barabási e Albert, apenas o segundo apresenta a presença de caudas e correntes quando a conectividade média da rede é igual a dois. No caso das redes reais, os cordões aparecem predominantemente na Teia Mundial, seguido pelas redes de livros e pela rede de distribuição de energia elétrica do EUA.

3.5 Medidas hierárquicas

Redes complexas podem ser analisadas e caracterizadas via medidas hierárquicas, as quais são definidas a uma distância d de um subgrafo g. Algumas dessas medidas provém de estudos de morfologia matemática em grafos [Vin89, HNTV90], enquanto outras têm sido desenvolvidas para estudos de redes complexas [dFC04a, dFCS06a, dFCdR06, dFCA07].

A maioria das medidas tradicionais de redes podem ser generalizadas hierarquicamente. Por exemplo, o *grau hierárquico* de um subgrafo g é definido pelo número de arestas entre g e o restante da rede, sem considerar as arestas que ligam

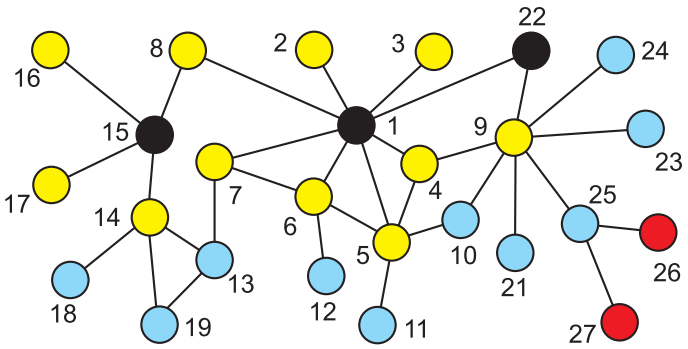


Figura 3.7: O subgrafo de interesse é definido pelos vértices de cor preta, $g = \{1, 15, 22\}$, cujos graus hierárquicos $k_0(g) = 12$, $k_1(g) = 12$ e $k_2(g) = 2$. O primeiro grau hierárquico é dado pelo número de arestas entre os vértices amarelos e azuis. Já o segundo grau hierárquico é dado pelo número de arestas entre os vértices azuis e vermelhos.

os vértices internos de g. A Figura 3.7 apresenta um exemplo da determinação de grau hierárquico. Note que o grau hierárquico $k_0(g)$ é igual a k_i, quando g é formado apenas pelo vértice i.

Outra medida que pode ser definida hierarquicamente é o coeficiente de aglomeração, chamado *coeficiente de aglomeração hierárquico*. Neste caso, é necessário definir o conceito de *anel-rs*, que é dado pelo conjunto de vértices localizados entre as distâncias $r - 1$ e s. O coeficiente de aglomeração do subgrafo g, $cc_{rs}(g)$ é computado pela razão entre o número de arestas no respectivo anel-rs, dada por $n_{rs}(g)$, e o número de arestas possíveis [dFCS06a], ou seja,

$$cc_{rs}(g) = \frac{2n_{rs}(g)}{|\mathcal{N}(R_{rs}(g))|(|\mathcal{N}(R_{rs}(g))| - 1)}, \tag{3.26}$$

onde $|\mathcal{N}(R_{rs}(g))|$ expressa o número de vértices no anel-rs. Essa equação é uma generalização da Equação 3.4. O coeficiente de aglomeração hierárquico de $c_{dd}(i)$ considera os vértices distantes $d - 1$ e d arestas de i.

A *razão de convergência* de nível d, $cv_d(g)$, é dada pelo quociente entre o grau

hierárquico à distância $d-1$ e o número de vértices no anel à distância d, ou seja,

$$cv_d(g) = \frac{k_{d-1}(g)}{|\mathcal{N}(R_d(g))|}. \tag{3.27}$$

Essa medida pode ser entendida como o número médio de arestas recebidas por cada vértice no nível hierárquico d, provenientes do nível inferior. O recíproco da razão de convergência é a razão de divergência, $dv_d(g)$, que é calculada por,

$$dv_d(g) = \frac{|\mathcal{N}(R_d(g))|}{k_{d-1}(g)}. \tag{3.28}$$

Outras medidas hierárquicas podem ser encontradas na Referência [dFCS06a], onde também são analisados modelos de redes e redes reais em termos de tais medidas.

3.6 Identificação de comunidades

A maioria das redes complexas possui estrutura modular, isto é, as conexões são densamente distribuídas entre vértices que pertençam a um mesmo grupo e esparsamente entre os vértices de grupos distintos. Os módulos são formados por vértices que possuem alguma relação de similaridade, como ocorre na Teia Mundial, onde as páginas que correspondem a tópicos semelhantes tendem a ser mais densamente conectadas entre si do que com o restante da rede [FLGC02]. A estrutura modular é comum na maioria das redes reais, como nas redes biológicas (e.g. redes metabólicas, genéticas e de interação de proteínas) [Bar03, HHLM99, GA05], nas redes tecnológicas [GA04], nas redes sociais [GD03, GN02] e nas redes de informação [FLGC02, HKKS04]. A presença de *hubs* e organização modular são propriedades comuns presentes nas redes complexas, principalmente em redes biológicas, e estão relacionadas a adaptação das redes com relação à robustez e multitarefa [HHLM99]. Determinar como a necessidade da multitarefa afeta a estrutura e dinâmica de redes complexas é uma questão muito investigada

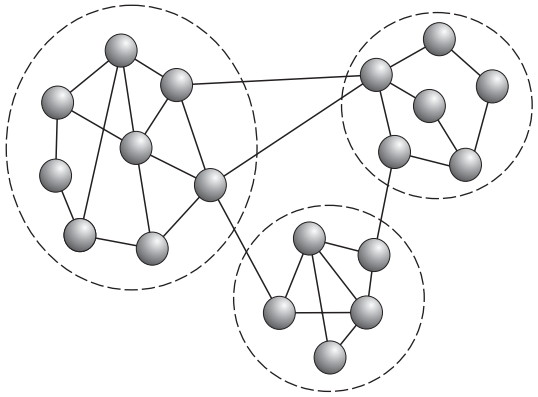


Figura 3.8: Exemplo de uma rede com estrutura modular. As comunidades são indicadas pelas linhas tracejadas.

atualmente [Kit02]. Para se obter conclusões a esse respeito, é necessário primeiramente identificar os módulos organizacionais das redes [GA05], denominados *comunidades* [DDADG05], que são definidas por vértices altamente conectados entre si e pouco conectados com o restante da rede.

A identificação das comunidades tem ainda várias aplicações práticas, como no caso da Teia Mundial, onde páginas relacionadas ao mesmo assunto são altamente conectadas. Neste caso, a identificação das comunidades pode ajudar na busca por informações [FLGC02], já que os algoritmos de busca podem se basear nesta divisão.

Um fator negativo quanto ao processo de identificação das comunidades é que essa tarefa constitui um problema *NP-completo* [DDADG05] de difícil solução, pois, geralmente não se tem idéia de quantas comunidades formam a rede. Além disso, em muitos casos, comunidades podem ser definidas hierarquicamente, quando se tem comunidades dentro de outras comunidades [RB03]. Apesar dessas dificuldades, muitos métodos têm sido propostos para identificar tais estruturas, mas o seu uso depende dos resultados desejados e das limitações existentes quanto ao poder computacional disponível, já que o tempo de execução dos algoritmos

atualmente existentes varia de $O(N^2 \log(N))$ a $O(N^4)$ [DDADG05], sendo N o tamanho da rede.

Para se determinar a qualidade de uma divisão particular, é utilizada a medida de *modularidade*, Q, que foi proposta por Newman e Girvan [New04a]. Para uma dada divisão de uma rede em c comunidades, constrói-se uma matriz E, $c \times c$, cujos elementos ao longo da diagonal principal, e_{ii}, fornecem a fração das conexões entre os vértices na mesma comunidade e os elementos e_{ij}, $j \neq i$, representam a fração de conexões entre as comunidades i e j. A modularidade Q é calculada da seguinte forma,

$$Q = \sum_i [e_{ii} - (\sum_j e_{ij})^2] = \mathrm{Tr}E - ||E^2||. \tag{3.29}$$

$Q = 1$ identifica uma rede formada por módulos desconectados. Valores altos de Q determinam redes com estrutura modular definida.

Os métodos de identificação de comunidades podem ser agrupados de acordo com a metodologia adotada. Conforme discutido na revisão de medidas de redes complexas apresentada na Referência [dFCRTB07], os métodos podem ser classificados em: (i) *espectrais*, que são baseados na análise dos autovetores de matrizes derivadas da matriz de adjacência [New06a]; (ii) *divisivos*, cuja metodologia é baseada na remoção das conexões entre as comunidades de forma interativa até a obtenção da maior modularidade possível [GN02, RCC+04]; (iii) *aglomerativos*, que são baseados no princípio de que vértices pertencentes às mesmas comunidades possuem similaridades [New04c, CNM04]; (iv) *maximização da modularidade*, que tenta obter a melhor divisão quando o maior valor da modularidade é obtido [DA05]; e (v) *métodos locais*, que determinam comunidades localmente, sem considerar informações globais da rede [Cla05, BB05].

Uma metodologia bastante popular usada na identificação de comunidades é baseada no grau de intermediação (*betweenness centrality*) das conexões da

rede [GN02]. Neste caso, são removidos os vértices de maior grau de intermediação, que representam os gargalos na rede, ou seja, aqueles que estão entre as comunidades. A cada passo é calculado o grau de intermediação de cada aresta e é removida aquela que corresponda ao maior valor. Posteriormente, o processo é repetido até que a rede seja completamente dividida em vértices isolados. A divisão que ofereça o valor mais alto da modularidade é então adotada e as respectivas comunidades são obtidas. A maior limitação desse método é o tempo de computação, que é $O(M^2N)$.

O método aglomerativo mais largamente utilizado foi proposto inicialmente por Newman [New04c] e posteriormente otimizado por Clauset *et al.* [CNM04]. Neste caso, as comunidades são aglomeradas de forma a obter o maior valor possível da modularidade. Inicia-se com uma rede completamente desconectada, onde cada vértice é considerado individualmente como uma comunidade, e a cada passo, escolhe-se as duas comunidades i e j cuja aglomeração forneça o maior acréscimo (ou menor decréscimo) no valor da modularidade. A variação obtida pela aglomeração das comunidades i e j é calculada por,

$$\Delta Q_{ij} = 2\left(e_{ij} - \frac{\sum_j e_{ij} \sum_i e_{ij}}{2M}\right). \tag{3.30}$$

A aglomeração que corresponda ao maior valor da modularidade é então adotada e obtém-se as comunidades.

Danon *et al.* [DDGA06] observaram que o valor de ΔQ_{ij} calculado conforme a Equação (3.30) tem uma limitação quando o tamanho das comunidades não é homogêneo. Para eliminar tal efeito, o valor de ΔQ_{ij} deve ser normalizado pelo número de conexões dentro da comunidade i,

$$\Delta \widehat{Q}_{ij} = \frac{\Delta Q_{ij}}{a_i}. \tag{3.31}$$

A principal vantagem na utilização deste método está relacionado ao tempo de computação ($O(N \log^2 N)$), o que fornece a melhor divisão da rede em um menor

tempo de processamento.

Os métodos locais são comumente empregados quando se quer obter informações de uma comunidade especificamente, sem considerar a rede toda [Cla05, BB05]. A sua utilização somente é viável em redes cujos métodos globais não podem ser utilizados por falta de poder computacional, como no caso da Teia Mundial no domínio Altavista, que é formada por $203.549.046$ vértices [BKM+00]. A precisão fornecida por estes métodos é inferior àquela obtida quando os métodos globais são utilizados [DDADG05].

Dentre todos os métodos desenvolvidos até agora, o método espectral proposto por Newman, baseado na análise dos autovalores e autovetores da *matriz de modularidade*, fornece resultados mais precisos [New06b]. Para cada sub-rede g, os elementos da matriz de modularidade $B^{(g)}$ são calculados por,

$$b_{ij}^{(g)} = a_{ij} - \frac{k_i k_j}{2M} - \delta_{ij} \sum_{u \in \mathcal{N}(g)} \left[a_{iu} - \frac{k_i k_u}{2M} \right], \tag{3.32}$$

para vértices i e j em g. A partir dessa matriz, são calculados seus autovalores e autovetores. Então, é determinado o autovalor mais positivo e o respectivo autovetor. De acordo com os sinais dos elementos deste vetor, a rede é dividida em duas partes. A seguir, o processo é repetido recursivamente, até que nenhuma divisão seja possível. As sub-redes indivisíveis correspondem às comunidades[New06b].

Uma descrição detalhada dos métodos de detecção de comunidades pode ser encontrada nas revisões apresentadas nas Referências [dFCRTB07, DDADG05, New04b, BLM+06]. Quanto à escolha do método a ser utilizado, depende de dois fatores básicos: precisão e tempo de processamento. Caso a precisão seja o fator mais importante no problema considerado, o método espectral de Newman [New06b] é o mais indicado. Caso contrário, quando é necessário se obter uma divisão menos precisa, mas que seja realizada no menor tempo possível, o método de Clauset *et al.* [CNM04] é o mais indicado. O nível de precisão deste

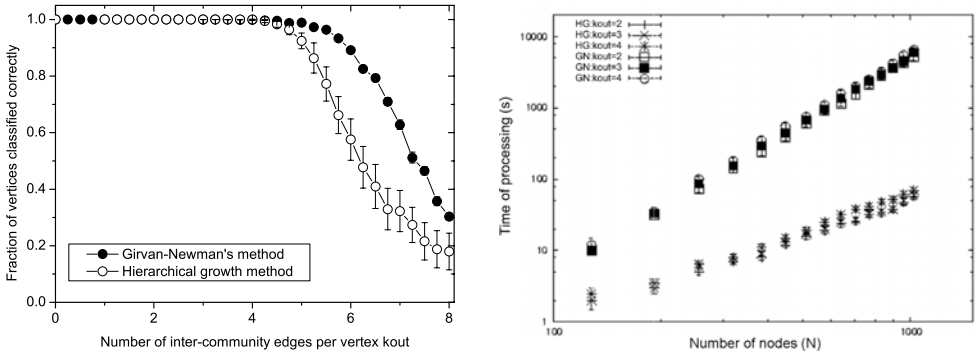
último método apenas é inferior ao método espectral de Newman [New06b] e no baseado na otimização extrema da modularidade de Duch e Arenas [DA05]. Logo, sua utilização é a mais recomendada na maioria dos casos. Quando a necessidade exige um tempo de processamento ainda menor, o método que propomos na Referência [RTC07][3] é o mais adequado, pois apesar de não fornecer uma divisão tão precisa quanto os métodos descritos anteriormente, esse método oferece o menor tempo de processamento. A Figura 3.9 apresenta a precisão e o tempo de processamento de tal método, quando comparado com o método desenvolvido por Girvan e Newman [GN02]. Conforme notamos, apesar do método baseado no crescimento hierárquico não ser tão preciso como muitos já propostos, ele é bastante rápido na determinação das comunidades.

3.7 Redes com diferentes tipos de vértices

Algumas redes reais são formadas por vértices que podem ser classificados diferentemente, como, por exemplo, em redes sociais cujos vértices podem ser divididos por tipo de raça, sexo, religião ou nível de escolaridade. Neste caso, Newman propôs uma medida chamada assortatividade, que quantifica a tendência de vértices do mesmo tipo se conectarem. Se E é uma matriz cujos elementos e_{st} representam o número de ligações entre vértices do tipo s e t. Sua normalização é dada por

$$\hat{E} = \frac{E}{\|E\|}. \tag{3.33}$$

[3]O método de crescimento hierárquico é baseado na análise sucessiva de vizinhos na rede, alcançados via crescimento hierárquico a partir de um vértice inicial, e na definição de comunidades como sendo subgrafos cujo número de conexões internas é maior do que externas.



(a) **(b)**

Figura 3.9: (a) Fração de vértices corretamente classificados em termos do grau entre comunidades (k_{out}) para uma rede com 128 vértices divididos em quatro comunidades, considerando $k_{in} + k_{out} = 16$. Quanto maior k_{out}, mais difícil é a separação entre as comunidades. (b) Enquanto o tempo de processamento do método desenvolvido por Girvan e Newman cresce da forma $O(N^{3,0\pm0,1})$ com o tamanho da rede, o tempo para o método baseado no crescimento hierárquico cresce da forma $O(N^{1,6\pm0,1})$. Figura extraída de [RTC07].

O *coeficiente de assortatividade* é calculado por [New02a],

$$\eta = \frac{\mathrm{Tr}\,\hat{E} - \|\hat{E}^2\|}{1 - \|\hat{E}^2\|}. \tag{3.34}$$

Quando $\eta = 1$ a rede é perfeitamente assortativa, isto é, os vértices se conectam apenas a outros vértices do mesmo tipo. Se $\eta = 0$, há uma conexão aleatória entre os vértices de diferentes tipos. Note que essa medida é parecida com a assortatividade, descrita na Seção 3.1, onde os tipos dos vértices são determinados pelo grau dos vértices. Outra medida semelhante ao coeficiente de assortatividade é o coeficiente *rich-club*, descrito na Seção 3.2, que mede a tendência dos *hubs* se conectarem.

3.8 Outras medidas

Outras medidas de redes complexas, como a dimensão fractal, medidas relacionadas à entropia, medidas de centralidade, medidas espectrais e medidas de complexidade, bem como maiores detalhes sobre as medidas aqui descritas, são apresentadas na Referência [dFCRTB07].

3.9 Como escolher as medidas para caracterizar as redes complexas?

Apesar de existir um grande número de medidas para caracterização das redes, a sua escolha depende da aplicação que se deseja estudar. Geralmente, deve-se escolher as medidas que quantifiquem as principais propriedades das redes que serão analisadas. Neste caso, é fundamental o conhecimento do problema a ser investigado e das medidas de redes. Por exemplo, para se avaliar como uma dada rede está estruturada, pode-se utilizar métodos de detecção de comunidades e a medida de modularidade.

Se o objetivo é a classificação de redes complexas ou a escolha do modelo que melhor reproduz a arquitetura de uma dada rede real, deve-se escolher um conjunto de medidas que descreva as propriedades estruturais das redes de forma mais completa possível, pois caso contrário, pode-se obter classificações equivocadas. Por exemplo, as evolução das redes de interações de proteínas poderiam ser representadas pelo modelo de Barabási e Albert se fossem consideradas apenas o menor caminho médio e a distribuição das conexões como medidas de caracterização. Entretanto, sabe-se que o mecanismo de crescimento das rede de interações de proteínas é completamente distinto do utilizado no modelo de Barabási e Albert — as redes de interações de proteínas são geradas por dois processos básicos,

duplicação e mutação [VFMV03b, SPSSK02], que não envolvem ligação preferencial. Além disso, deve-se evitar medidas que resultem em redundâncias, já que estas podem incrementar o erro da classificação [DHS01]. Na Tabela 3.2 são apresentadas as correlações entre as medidas, calculadas pela Equação 3.3, para os grafos aleatórios de Erdős e Rényi (ER), modelo livre de escala de Barabási e Albert (BA) e o modelo geográfico de Waxman (GN). Cada uma das redes geradas são formadas por $N = 1.000$ vértices e grau médio $\langle k \rangle = 4$. Conforme podemos notar, os valores mais altos de correlação foram observados para o modelo BA. Além disso, a correlação obtida para os modelos é diferente daquela obtida quando todos os modelos são levados em conta conjuntamente. Deste modo, a análise das correlações não é trivialmente determinada, pois a intensidade das correlações depende especificamente do tipo dos modelos considerados.

Nos próximos capítulos, apresentamos a utilização de medidas na caracterização, classificação e análise de redes complexas. Discutimos como as medidas são escolhidas em cada análise e apresentamos a utilização de medidas na classificação de redes, com o uso de métodos estatísticos multivariados.

Tabela 3.2: Correlações entre as medidas obtidas considerando-se os modelos BA, ER e GN e todos os modelos conjuntamente. Os valores foram estimados de 1.000 realizações de cada modelos. Cada rede é formada por $N = 1.000$ e possui grau médio $\langle k \rangle = 4$. Tabela extraída de [dFCRTB07].

		st	r	$\langle cc \rangle$	ℓ	c_D	$\langle k_2 \rangle$	$\langle cc_{22} \rangle$
r	BA	–0,22						
	ER	–0,01						
	GN	–0,13						
	Todos	0,71						
$\langle cc \rangle$	BA	0,06	–0,29					
	ER	–0,01	0,07					
	GN	0,04	–0,00					
	Todos	0,31	0,82					
ℓ	BA	–0,01	0,38	–0,63				
	ER	–0,06	0,04	–0,08				
	GN	–0,10	0,02	0,03				
	Todos	0,69	0,96	0,88				
c_D	BA	–0,09	0,23	0,39	–0,58			
	ER	–0,61	0,10	0,03	0,07			
	GN	–0,05	–0,02	0,03	0,23			
	Todos	–0,87	–0,44	0,02	–0,41			
$\langle k_2 \rangle$	BA	0,01	–0,30	0,63	–0,99	0,60		
	ER	0,04	0,03	0,08	–0,90	–0,06		
	GN	0,08	0,28	–0,02	–0,65	–0,13		
	Todos	–0,96	–0,80	–0,43	–0,79	0,85		
$\langle cc_{22} \rangle$	BA	0,02	0,02	0,58	–0,74	0,59	0,76	
	ER	–0,03	0,04	0,45	–0,16	0,02	0,19	
	GN	–0,00	0,09	0,59	0,18	0,07	–0,11	
	Todos	0,37	0,86	0,99	0,91	–0,05	–0,49	
$\langle dv_3 \rangle$	BA	0,01	0,26	–0,57	0,91	–0,52	–0,94	–0,69
	ER	0,03	–0,10	–0,01	–0,25	–0,01	–0,16	–0,04
	GN	–0,02	–0,28	–0,09	–0,03	–0,00	–0,50	–0,21
	Todos	–0,14	–0,74	–0,97	–0,79	–0,18	0,27	–0,96

Capítulo 4

Classificação de redes complexas

Em um contexto geral, classificar significa associar classes ou categorias a elementos de acordo com suas propriedades [dFCJ01]. As classes podem ser divididas hierarquicamente, onde uma classe geral pode ser fragmentada em classes mais específicas. Tomemos como exemplo os seres vivos, cuja classificação científica mais aceita internacionalmente constitui 5 grandes reinos, quais sejam: Monera, Protistas, Fungi, Metafita ou Plantae e Metazoa ou Animália. Estes reinos podem ser divididos em filos, que podem ainda ser decompostos em classes, que são divididas em ordens, que são separadas em famílias, que são subdividas em gêneros e que, finalmente, são separados em espécies. Se desejarmos caracterizar os seres humanos, podemos seguir essa hierarquia: o homem pertence ao reino Meatazoa, ao filo Chordata, à classe Mammalia, à ordem dos Primatas, à família Hominidae, ao gênero Homo e à espécie *Homo sapiens*. Um dos objetivos da classificação é evitar redundância na descrição dos objetos, onde um dado elemento pode ser descrito pelas características básicas da classe à qual pertence. Por exemplo, quando nos referimos a alguma das 235 espécies que constituem a

família dos primatas, ficam subentendidas suas características básicas: (i) visão mais utilizada do que o olfato, (ii) visão binocular, (iii) membros do corpo e mãos adaptados para pendurar-se, saltar e balançar nas árvores, (iv) habilidade para pegar e manipular objetos pequenos, usando dedos com unhas em lugar de garras, (v) cérebro grande em relação ao tamanho corporal e (vi) vida social complexa. Embora a classificação auxilie na descrição e caracterização de objetos, tal processo pode ser extremamente complexo quando ocorre, por exemplo, sobreposição de características de diferentes classes [dFCJ01]. Por exemplo, além dos primatas, a maioria das aves também utilizam mais a visão do que o olfato. Além disso, os felinos também possuem visão binocular. Deste modo, a consideração de um número limitado de características pode fornecer classificações equivocadas, pois elas podem estar associadas a mais de uma classe. Em muitos casos, a classificação pode não só ser difícil, mas até mesmo impossível [dFCJ01].

Do ponto de vista das redes complexas, elas podem ser classificadas de acordo com propriedades estruturais e dinâmicas. No primeiro caso, são calculadas determinadas medidas para um dado conjunto de redes, que são armazenadas em vetores de características. Estes vetores são associados a cada rede conforme ilustrado na Figura 4.1. Posteriormente, métodos de classificação são aplicados sobre tais vetores de forma a agrupar as redes por semelhanças (ver Figura 4.2). O método de classificação utilizado pode ser essencialmente de dois tipos: (i) *supervisionado*, quando se conhece *a priori* as classes das redes, que podem ser definidas por modelos de redes complexas, por exemplo; e (ii) *não-supervisionado*, quando não se tem qualquer informação sobre as classes [DHS01]. No primeiro caso, a classificação envolve dois processos básicos: (i) aprendizado: corresponde ao estágio em que os métodos e critérios são treinados em classes conhecidas; e (ii) reconhecimento: quando o sistema que foi treinado é utilizado para classificar redes desconhecidas. No segundo caso, uma das possibilidades é determinar

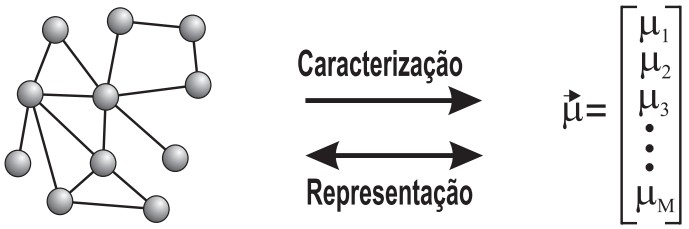


Figura 4.1: Mapeamento de uma rede complexa em um vetor de características. Quando é possível obter a rede original a partir de suas medidas, diz-se que esse mapeamento é uma representação.

classificações que maximizem a similaridade entre as redes pertencentes à mesma categoria e minimizem a semelhança entre redes em classes distintas. Tal metodologia também é chamada de *clustering*. Assim sendo, as classes são criadas de acordo com as características das redes, sem qualquer conhecimento anterior das categorias reais. A classificação não-supervisionada pode ser utilizada para se criar uma taxonomia de redes complexas. Neste caso, é esperado que as redes pertencentes à mesma categoria (sociais, biológicas, de informação e tecnológicas) compartilhem as mesmas propriedades básicas.

A classificação supervisionada de redes complexas pode ser realizada considerando-se diversos modelos e medidas para a construção do espaço de classificação. Redes reais são então projetadas sobre esse espaço e associadas ao modelo que gera redes cujas topologias mais se assemelhem à da rede real [DHS01, dFCJ01]. Essa forma de classificação permite, por exemplo, determinar o modelo que melhor reproduz a estrutura de uma dada rede real. Esse tipo de análise é fundamental no estudo das redes complexas, pois auxilia no aperfeiçoamento dos modelos e, conseqüentemente, no entendimento dos processos dinâmicos que geram redes reais. Entretanto, apesar de simples, a classificação supervisionada de redes complexas é limitada por três questões fundamentais: Quais os modelos mais adequados para representar a rede real em questão? Quais as medidas que devem ser

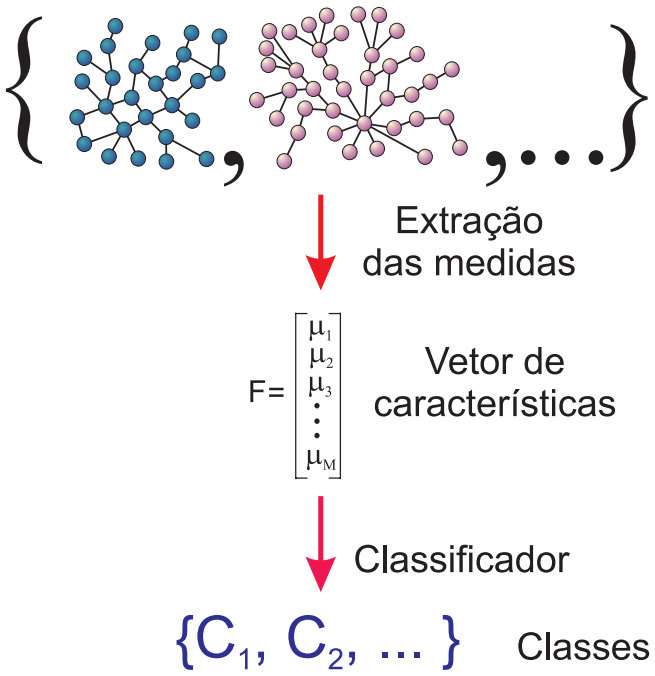


Figura 4.2: Processo de classificação: as medidas são extraídas de um conjunto de redes, que serão utilizadas por um classificador, definindo as classes a que as redes pertencem.

utilizadas para se obter uma classificação precisa? Qual método de classificação deve ser utilizado?

De maneira geral, alguns dos modelos de redes complexas podem ser classificados conforme a Figura 4.3. Note que além de existir uma hierarquia de classificação, alguns modelos situam-se na intersecção de diversas classes. Por exemplo, o modelo *small-world* de Watts-Strogatz pode apresentar conexões regulares ou completamente aleatórias, dependendo do valor da probabilidade de reconexão adotado. Além disso, dependendo desta probabilidade, a rede também pode apresentar estrutura modular. O mesmo ocorre com o modelo de configuração, que gera redes com distribuições de conexões pré-definidas [MR95], podendo gerar redes do tipo Erdős-Rényi, livres de escala ou qualquer outro tipo cuja

distribuição das conexões seja previamente conhecida. Outro exemplo são os modelos livres de escala, pois conforme discutido na Seção (2.3), os modelos de Krapivsky *et al.* [KRL00] e Dorogovtsev *et al.* [DMS00] são generalizações do modelo de Barabási-Albert[1]. Destarte, estes parâmetros devem ser ajustados de acordo com as redes que se deseja analisar, pois dependendo dos parâmetros empregados, os modelos podem gerar redes com estruturas mais ou menos semelhantes à redes reais, o que pode influenciar na classificação.

Com relação às medidas de redes complexas, algumas são específicas a determinadas características topológicas e não são suficientes para oferecer uma representação precisa de redes complexas [dFCRTB07]. Além disso, algumas medidas ainda podem ser correlacionadas, o que resulta em redundâncias nos resultados. Infelizmente, o problema da escolha de quais medidas utilizar na caracterização, classificação e comparação de redes, não pode ser resolvido por qualquer método matemático. Na prática, a escolha das medidas para realizarem estas tarefas refletem interesses e aplicações específicas [dFCRTB07].

4.1 Análise das variáveis canônicas

Uma maneira de classificar redes utilizando-se métodos supervisionados é realizada considerando-se a análise das variáveis canônicas e classificação Bayesiana, conforme apresentamos na Referência [dFCRTB07]. O interesse da análise das variáveis canônicas é verificar a relação da magnitude de diferenças entre grupos especificados *a priori* relativa àquela dentro dos grupos, achando combinações lineares das variáveis para cada um deles que são maximamente correlacionadas entre si, através das projeções que otimizem a separação das categorias [MdR99,

[1] Os modelos de Krapivsky *et al.* e Dorogovtsev *et al.* geram o modelo de Barabási-Albert quando consideram como parâmetros $\alpha = 1$ e $k_0 = 0$, respectivamente (para maiores detalhes, ver Seção (2.3)).

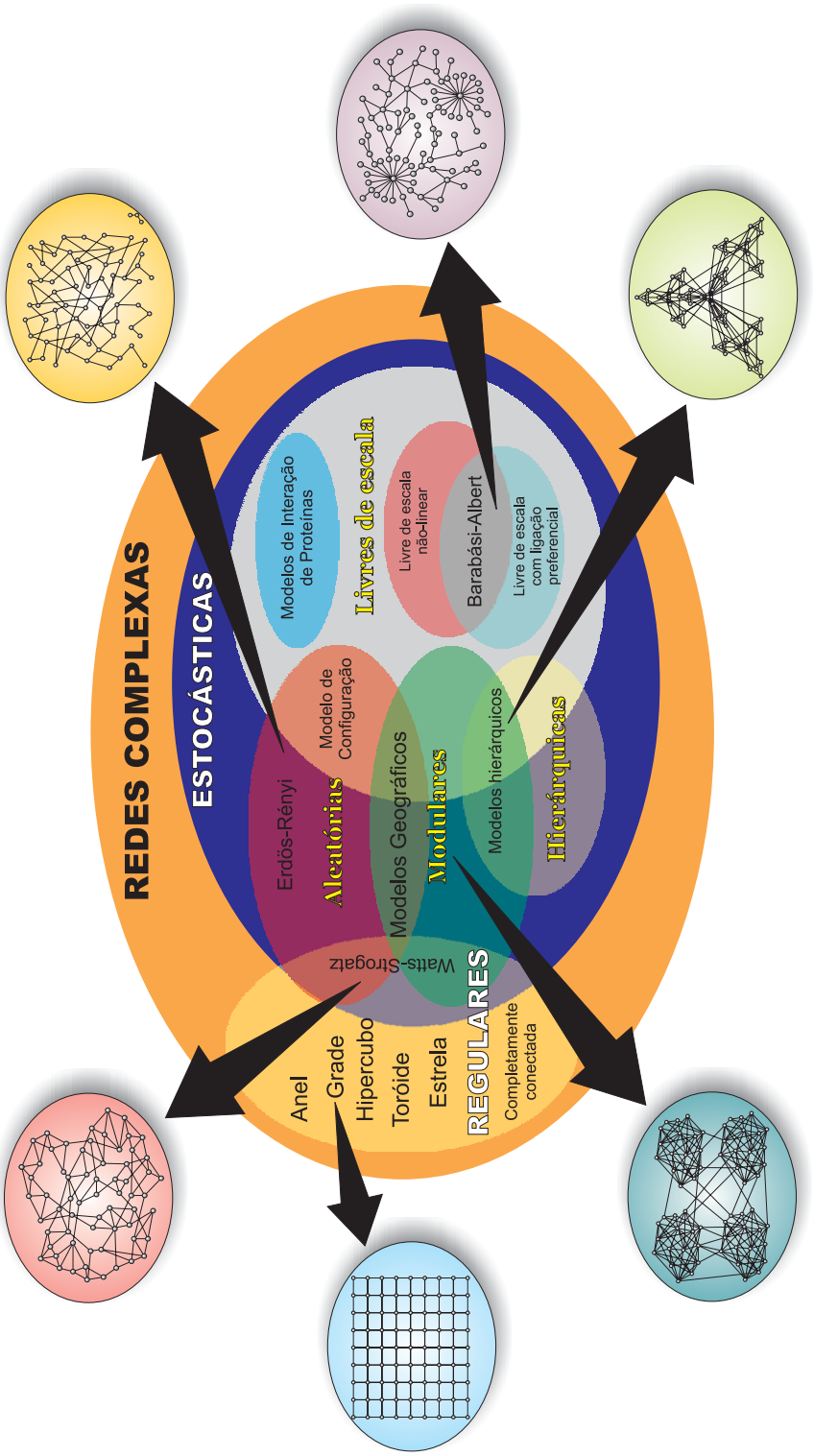


Figura 4.3: Uma possível classificação das redes complexas considerando os modelos mais utilizados. Note que várias redes possuem características de mais de uma classe.

And58, DHS01]. Em outras palavras, a análise das variáveis canônicas permite projetar um espaço de n medidas em duas ou três dimensões que maximize a separação entre as classes definidas *a priori*. Para realizar a análise das variáveis canônicas, é necessária a construção de uma matriz que quantifique a variação dentro dos grupos e outra que quantifique a variação entre tais grupos. Portanto, considerando-se R redes divididas em c classes, cada uma com N_i redes, $i = 1, 2, \ldots, c$, sendo cada rede ξ representada por seu respectivo vetor de medidas $\vec{x_\xi} = (x_1, x_2, \ldots, x_p)^T$; definimos a matriz de dispersão total (S) que expressa a dispersão completa das medidas, da seguinte forma,

$$S = \sum_{\xi=1}^{R} \left(\vec{x_\xi} - \langle\vec{x}\rangle\right)\left(\vec{x_\xi} - \langle\vec{x}\rangle\right)^T. \tag{4.1}$$

Já a matriz de dispersão para cada classe, é dada por,

$$S_i = \sum_{\xi\in\, N_i} \left(\vec{x_\xi} - \langle\vec{x}\rangle_i\right)\left(\vec{x_\xi} - \langle\vec{x}\rangle_i\right)^T, \tag{4.2}$$

onde $\langle\vec{x}\rangle_i$ é a média do vetor de medidas da classe i. A matriz de soma dos quadrados dentro das classes é definida pela expressão,

$$S_{\text{dentro}} = \sum_{i=1}^{c} S_i. \tag{4.3}$$

Finalmente, a matriz de soma dos quadrados entre as classes é definida da seguinte forma,

$$S_{\text{entre}} = \sum_{i=1}^{c} N_i \left(\langle\vec{x}\rangle_i - \langle\vec{x}\rangle\right)\left(\langle\vec{x}\rangle_i - \langle\vec{x}\rangle\right)^T. \tag{4.4}$$

Pode ser verificado que,

$$S = S_{\text{dentro}} + S_{\text{entre}}. \tag{4.5}$$

Os autovalores extraídos da matriz $S^{-1}_{\text{dentro}} S_{\text{entre}}$ são interpretados como a quantidade de variação associada a cada autovetor ou eixo de maior variação. Assim, para uma projeção em duas dimensões, escolhem-se os dois autovetores ($\vec{e_a}$ e $\vec{e_b}$)

associados aos dois maiores autovalores (λ_a e λ_b) e faz-se o produto escalar entre a matriz com o atributos originais e tais autovetores. Cada um destes produtos reflete os dados projetados no espaço definido pelos eixos canônicos.

4.2 Decisão Bayesiana

A partir da projeção canônica, é feita a classificação Bayesiana. O principal problema considerado por tal classificação é associar uma dada rede ξ a uma das classes c (modelos de rede, em nosso caso) minimizando o erro de associação em outras classes. Em princípio, é considerado que as probabilidades P_i que corresponde à probabilidade de uma rede pertencer à classe i, bem como a função densidade de probabilidade condicional $p(\vec{x_\xi}|i)$ são conhecidas ou podem ser estimadas. A probabilidade P_i pode ser estimada da respectiva freqüência. Quando a função densidade de probabilidade é conhecida e apenas seus parâmetros (por exemplo, média e desvio padrão) têm que ser determinados, tal estimação é chamada *paramétrica*. Por outro lado, quando tal função é desconhecida, é necessário deduzir o seu tipo, sendo tal classificação chamada *não-paramétrica*. A estimativa de tal função, neste último caso, pode ser realizada pelo método de Parzen, o qual representa cada ponto (rede) como uma função, chamada *núcleo de Parzen* [DHS01], no espaço definido pela análise das variáveis canônicas. Em nossa classificação, utilizamos essa aproxiamação. O critério de decisão pode ser expresso como: se $f(\vec{x}_\xi|i)P(i) = \mathrm{Max}_{b=1,c}\{f(\vec{x}_\xi|b)P(b)\}$ então a rede ξ pertence à classe i [dFCJ01]. Na Figura 4.4 são mostradas duas aproximações obtidas por estimação paramétrica e não-paramétrica para diversas redes derivadas de três modelos básicos: redes geográficas (de Waxman), redes *small-world* (de Watts e Strogatz) e redes aleatórias (de Erdős e Rényi). Neste caso, foram consideradas apenas duas medidas, o menor caminho médio (ℓ) e a medida de assortatividade

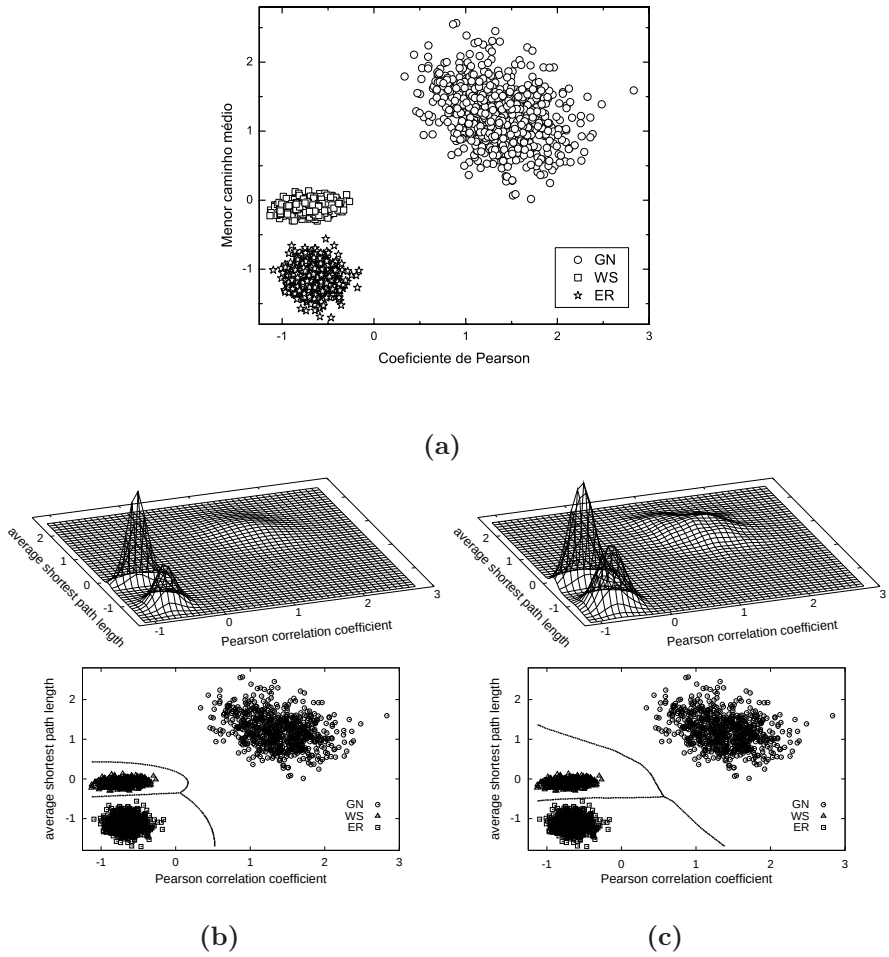


Figura 4.4: (a) Espaço definido pela assortatividade e pelo menor caminho médio para redes derivadas de três modelos básicos: redes geográficas (de Waxman), redes *small-world* (de Watts e Strogatz) e redes randômicas (de Erdős e Rényi). A seguir, temos as respectivas funções gaussianas e as regiões de decisão considerando estimação paramétrica (b) e não-paramétrica (c). Os parâmetros das redes são: $N = 250, \langle k \rangle = 20$, sendo realizadas 1.000 simulações para cada modelo. A probabilidade de reconexão no modelo *small-world* é 0.4. Note que os valores das medidas consideradas foram normalizados. Figura extraída de [dFCRTB07].

(r). Cada ponto nesse espaço de classificação representa uma rede gerada por um dos modelos.

4.3 Classificação

O método de classificação das redes complexas foi dividido em duas fases distintas: (i) *treinamento* e (ii) *classificação*. Na fase de treinamento, são escolhidas as medidas e modelos de redes para a criação do espaço e das regiões de classificação. Em nossa análise, utilizamos 9 diferentes medidas de redes, (i) o menor caminho médio (ℓ), (ii) a "retidão" da distribuição (st), (iii) o grau médio ($\langle k \rangle$), (iv) o coeficiente de aglomeração médio ($\langle cc \rangle$), (v) o grau hierárquico médio de nível 2 ($\langle k_2 \rangle$), (vi) o coeficiente de aglomeração hierárquico de nível 2 ($\langle cc_{22} \rangle$), (vii) a razão de divergência de nível 3 (dv_3), (viii) a medida de assortatividade (r), (ix) e a dominância do ponto central (c_D). Todas estas medidas são discutidas no Capítulo 3. Utilizamos a seguinte combinação de medidas para avaliar seu efeito na classificação,

i. $\{\ell, st\}$,

ii. $\{\langle k \rangle, \langle cc \rangle, \ell\}$,

iii. $\{\langle k_2 \rangle, \langle cc_{22} \rangle, \langle dv_3 \rangle\}$,

iv. $\{st, r, c_D\}$,

v. $\{\langle k \rangle, \langle cc \rangle, \ell, st, r, c_D\}$,

vi. $\{\langle k \rangle, \langle cc \rangle, \ell, \langle k_2 \rangle, \langle cc_{22} \rangle, \langle dv_3 \rangle\}$,

vii. $\{st, r, c_D, \langle k_2 \rangle, \langle cc_{22} \rangle, \langle dv_3 \rangle\}$,

viii. Todas as medidas.

O conjunto de medida (i) e (ii) são os mais largamente utilizados para caracterizar redes complexas. Já o conjunto (iii) é composto pelas medidas hierárquicas. No caso (iv) são adotadas medidas de centralidade e assortatividade. Em (v) são combinadas as medidas de (ii) com (v), de forma a obter classificações mais detalhadas. A combinação entre as medidas de (ii) e (iii) é considerada em (vi). Em (vii) há uma combinação entre as medidas dos conjuntos (iii) e (vi). Finalmente, em (viii) são consideradas todas as medidas.

Utilizamos três modelos de redes complexas para definir o espaço de classificação: (i) Erdős-Rényi (ER), (ii) Barabási-Albert (BA) e (iii) geográfico de Waxman (GN). Para cada modelo, foram geradas redes com o mesmo número de vértices e conectividade média próximos dos encontrados nas redes reais. Para essas realizações, calculamos as nove medidas descritas anteriormente, definindo os vetores de características de cada rede. Finalmente, aplicamos análise das variáveis canônicas e classificação Bayesiana nestes vetores a fim de finalizarmos a fase de treinamento da classificação supervionada. As medidas utilizadas neste processo foram consideradas conforme enumerado anteriormente. Na segunda fase da classificação, projetamos as redes reais sobre o espaço bidimensional definido pela análise das variáveis canônicas, obtendo a respectiva classificação em um dos modelos considerados. As redes reais consideradas na classificação são as seguintes:

Rede de transporte aéreo americana (USATN): Essa rede é composta por 332 aeroportos e foi construída segundo dados sobre conexões aéreas obtidos em 1997 [BM06a]. Conforme discutido por Guimerá *et al.* [GA04, GMT+05], é esperado que esse tipo de rede apresente distribuição das conexões tipo livre de escala.

Rede de interação de proteínas do* S. cerevisiae *(PPIN): A rede de interação é formada por 1.922 proteínas conectadas de acordo com interações físicas. A base

que utilizamos é disponibilizada pelo *Center for Complex Network Research* (Universidade de Notre Dame). Conforme observado por Jeong *et al.*, as interações entre as proteínas seguem uma distribuição livre de escala [JMBO01].

Internet (AS): A base de dados da Internet no nível dos sistemas autônomos é formada por um conjunto de roteadores e redes sob a mesma administração, como universidades, empresas privadas ou provedores de acesso à Internet. As conexões são determinadas por ligações físicas entre tais elementos. Os dados utilizados em nossa análise foram coletados do *National Laboratory of Applied Network Research* (`http://www.nlanr.net`), obtidos em fevereiro de 1998. A rede, formada por 3.522 vértices e 6.324 conexões, tem distribuição de conexões do tipo livre de escala [PSVV01].

Rede de transcrição genética do E. coli (TRNE): Neste tipo de rede, os vértices representam *operons*[2]. As conexões são dirigidas de um *operon* i a outro j, que é regulado pelo fator de transcrição codificado por i. Esse tipo de rede basicamente controla a expressão genética e a distribuição das conexões entre os *operons* segue uma lei de potência [SOMMA02]. A base de dados que utilizamos é uma versão não-direcionada utilizada por Shen-Orr *et al.* [SOMMA02], formada por 577 *operons* e 424 conexões.

Rede de Delaunay (DLN): Essa rede é do tipo geográfica e é obtida distribuindo-se os vértices uniformemente em um quadrado de largura unitária e conectando-os seguindo a triangulação de Delaunay [SKM96]. Segundo essa técnica, os vértices se conectam apenas aos seus vizinhos imediatos geograficamente. Em nossa análise, utilizamos redes formadas por 251 vértices conectados por 700 arestas.

[2]Os operons são definidos como unidades transcricionais nas quais existem genes contíguos (geralmente relacionados a uma única via metabólica) sob o controle de um promotor e um operador, que regulam sua transcrição. Esse tipo de organização gênica é somente vista em procariontes.

Os resultados obtidos pela classificação supervisionada destas redes reais são apresentados na Tabela (4.1). Apenas no caso (i) não foi necessária a utilização de análise das variáveis canônicas, porque somente duas medidas são consideradas. As classificações contrárias ao modelo esperado são assinaladas em negrito e aquelas que correspondem ao modelo esperado, mas cujos graus diferem do observado na rede original, em itálico. As classes indicadas por $*$ representam classificações completamente distintas da esperada, cujo modelo e grau não correspondem ao esperado. Conforme podemos notar, a melhor compatibilidade entre a classificação esperada e a obtida ocorreu para as redes de Delaunay (DN), que foi classificada como geográfica em todos os casos.

A rede de transporte aérea americana (USTAN) pode ser modelada pelos modelos que incorporam a ligação preferencial, pois esse tipo de rede apresenta distribuição das conexões do tipo livre de escala, e fatores geográficos, já que as ligações são limitadas por território e os aeroportos menores tendem a se conectarem com os maiores mais próximos [GA04, GMT$^+$05]. Deste modo, era esperado que a classificação desta rede se situasse entre os modelos BA e GN. Entretanto, apenas para o conjunto de medidas (iii) a classificação obtida correspondeu à esperada. Para as demais configurações, na maioria dos casos a classificação foi completamente distinta da esperada. A Figura 4.5(a) apresenta a classificação da rede de transporte aéreo considerando todas as nove diferentes medidas. A distância entre a posição da rede real e daquelas geradas pelos modelos no espaço de classificação sugere que nenhum dos modelos considerados é adequado para representar a rede de transporte aéreo americana.

No caso da Internet (AS), a classificação ocorreu de duas formas distintas: ou a classificação correspondeu ao modelo esperado ou ao grau da rede original. O modelo esperado é determinado pela caracterização usual que inclui a análise da distribuição das conexões. Logo, era esperado que redes livre de escala fossem

associadas ao modelo de Barabási e Albert. Em nenhum dos casos, a classificação foi completamente igual à esperada. Em metade dos casos, a rede foi classificada como correspondendo ao modelo BA e na outra metade, ao modelo GN. Semelhante ao discutido para a rede de transporte aéreo, vale lembrar que a ligação dos sistemas autônomos que formam a Internet também é influenciada por fatores geográficos, ou seja, as ligações físicas entre os roteadores estão limitadas pela distância entre eles. Logo, a classificação destas redes como geográficas não foi completamente equivocada. Mesmo assim, novamente observamos que nenhum dos modelos adotados é adequado para modelar a Internet. Para se obter uma modelagem mais precisa, é necessário considerar a ligação preferencial e a distância geográfica entre os roteadores, conforme discutimos no próximo capítulo.

De maneira similar à Internet, a rede de transcrição genética não foi classificada conforme o esperado em nenhum dos casos. A classificação mais próxima da esperada foi obtida considerando-se o conjunto (i) de medidas (ℓ e st). Nos demais casos, a rede foi associada a outros dois modelos (ER e GN). Essa variação, sugere que nenhum dos modelos considerados é adequado para representar a rede de transcrição genética. A Figura 4.5(b) apresenta a classificação da rede de transcrição genética considerando todas as nove diferentes medidas e, neste caso, a rede é classificada como geográfica.

Finalmente, a rede de interação de proteínas também não foi classificada conforme o esperado em nenhum dos casos. Embora essa rede inicialmente pudesse ser associada ao modelo BA, por apresentar distribuição de conexões do tipo livre de escala, foi classificada como geográfica na maioria dos casos. Isto aconteceu porque as redes de interações de proteínas apresentam certos atributos que podem ocorrer nas redes geográficas, como o alto coeficiente de aglomeração. Desta forma, a distribuição das conexões pode não ser o fator principal que caracterize a estrutura das redes complexas, pois redes com estruturas completamente

distintas podem apresentar a mesma distribuição de conexões [ADLW05]. As redes de interação de proteínas, segundo Vázquez *et al.* [VFMV03b] e Solé *et al.* [SPSSK02], são geradas por dois processos básicos, duplicação e mutação. No primeiro estágio, uma dada proteína é duplicada e herda as ligações da proteína original. No segundo processo, a proteína duplicada sofre uma mutação e perde parte das conexões. Desta maneira, o coeficiente de aglomeração médio é alto nessas redes, como acontece com o modelo geográfico. As redes BA, por outro lado, apresentam baixo coeficiente de aglomeração médio. Além disso, as redes de interação de proteínas têm algumas propriedades que não são encontradas nos modelos considerados, como a estrutura modular. Outra diferença é que os modelos BA, GN e ER não geram redes com qualquer tipo de assortatividade ($r \sim 0$), enquanto as redes de interações de proteínas são do tipo disassortativas ($r = -0{,}156$), ou seja, os *hubs* tendem a se ligarem aos vértices menos conectados [New03b, MS02]. Logo, é evidente que nenhum dos modelos considerados são adequados a representarem redes de interações de proteínas. Nas Figuras 4.5(c) e 4.5(d) são apresentadas classificações obtidas para essa rede considerando todas as medidas e excluindo-se as medidas hierárquicas, respectivamente.

Como primeira conclusão fundamental que podemos tirar de nossa análise é que a consideração de um conjunto reduzido de medidas pode não proporcionar uma classificação precisa de redes reais. Devido à alta complexidade e aos diferentes fatores que geram a estrutura e dinâmica destas redes, apenas a consideração de um conjunto amplo de medidas pode fornecer resultados precisos, pois mais propriedades estruturais são quantificadas. Porém, as medidas devem ser escolhidas de forma a não fornecerem resultados redundantes. Além disso, esse método mostra que um modelo será preciso se, independente da configuração de medidas utilizada, a rede real será associada a ele. Se a classificação depende

Tabela 4.1: A classificação das redes reais segundo as 8 combinações de medidas enumeradas no texto e os três modelos básicos considerados (ER, BA e GN). As classes em negrito indicam classificações diferente da esperada e as em itálico, indicam classificações cujos modelos possuem conectividade média diferente da rede original. As classes indicadas por $*$ representam classificações completamente distintas da esperada. As redes reais correspondem à rede de transporte aéreo americana (USATN), à Internet (AS), à rede de transcrição genética do *E. coli* (TRNE), à rede de interação de proteínas do *S. cerevisiae* (PPIN) e à uma dada rede de Delaunay (DLN). Resultados obtidos de [dFCRTB07].

Rede complexa analisada	**Classificação esperada**	**Redes identificadas pelas seguintes combinações de medidas:**							
		(i)	**(ii)**	**(iii)**	**(iv)**	**(v)**	**(vi)**	**(vii)**	**(viii)**
USATN	BA/GN	*BA*	*GN**	GN	*BA*	*BA**	*BA**	*GN**	BA*
$\langle k \rangle = 12.8$	$\langle k \rangle = 12.8$	$\langle k \rangle = 10.0$	$\langle k \rangle = 10.0$	$\langle k \rangle = 12.8$	$\langle k \rangle = 10.0$	$\langle k \rangle = 10.0$	$\langle k \rangle = 10.0$	$\langle k \rangle = 14.0$	$\langle k \rangle = 12.0$
AS	BA	*BA*	**GN**	*BA*	BA	**GN**	**GN**	*BA*	**GN**
$\langle k \rangle = 3.59$	$\langle k \rangle = 3.59$	$\langle k \rangle = 6.0$	$\langle k \rangle = 3.59$	$\langle k \rangle = 6.0$	$\langle k \rangle = 4.0$	$\langle k \rangle = 3.59$	$\langle k \rangle = 3.59$	$\langle k \rangle = 6.0$	$\langle k \rangle = 3.59$
TRNE	BA	BA	**GN**	***GN***	*BA*	**ER**	**ER**	**GN**	**ER**
$\langle k \rangle = 2.45$	$\langle k \rangle = 2.45$	$\langle k \rangle = 2.0$	$\langle k \rangle = 2.45$	$\langle k \rangle = 4.0$	$\langle k \rangle = 4.0$	$\langle k \rangle = 2.45$	$\langle k \rangle = 2.45$	$\langle k \rangle = 2.45$	$\langle k \rangle = 2.45$
PPIN	BA	***ER***	**GN**	***GN***	***ER***	**GN**	**GN**	***ER***	**GN**
$\langle k \rangle = 3.03$	$\langle k \rangle = 3.03$	$\langle k \rangle = 2.0$	$\langle k \rangle = 3.03$	$\langle k \rangle = 2.0$	$\langle k \rangle = 2.0$	$\langle k \rangle = 3.03$	$\langle k \rangle = 3.03$	$\langle k \rangle = 2.0$	$\langle k \rangle = 3.03$
DLN	GN	*GN*	*GN*	GN	*GN*	*GN*	GN	GN	GN
$\langle k \rangle = 6.0$	$\langle k \rangle = 6.0$	$\langle k \rangle = 4.0$	$\langle k \rangle = 4.0$	$\langle k \rangle = 6.0$	$\langle k \rangle = 4.0$	$\langle k \rangle = 4.0$	$\langle k \rangle = 6.0$	$\langle k \rangle = 6.0$	$\langle k \rangle = 6.0$

do conjunto de medidas, é um indicativo de que os modelos considerados não são precisos o suficiente e esse é um indício que eles precisam ser aperfeiçoados. Essa maneira de avaliar as redes com o uso de estatística multivariada tende a fornecer resultados muito mais completos do que os utilizados atualmente, que consideram apenas duas ou três medidas de redes na avaliação. Logo, alguns modelos para representar redes reais que atualmente são considerados adequados, podem necessitar de aperfeiçoamento se forem submetidos à análise multivariada.

A identificação do modelo que melhor representa uma dada rede real é fundamental para se entender sua evolução, o que permite fazer previsões e análises do comportamento dinâmico quando a rede é sujeita a perturbações, como a remoção de vértices ou arestas. Por exemplo, um modelo preciso da Internet permite avaliar a evolução da rede após um ataque dirigido ou falha em roteadores. Além disso, caso nenhum dos modelos considerados reproduzam a topologia da rede real de forma completa, pode-se identificar as estruturas que não estão sendo bem reproduzidas com o uso de análise das variáveis canônicas — a análise dos autovalores e autovetores permite determinar as medidas que mais contribuem na discriminação. Como isso, os modelos podem ser aperfeiçoados para uma melhor modelagem da rede real.

4.4 Classificação hierárquica

A classificação hierárquica das redes pode ser obtida considerando-se métodos aglomerativos (e.g. [DHS01, dFCJ01, dFCS06a]). Neste caso, as classes são diversificadas de um nível mais geral para um mais particular. Deste modo, o nível de semelhança entre as classes depende do nível hierárquico observado. Redes que pertencem à mesma classe em um nível pode não pertencer em outro. A Figura 4.6 apresenta o dendograma de classificação para a situação descrita na Fi-

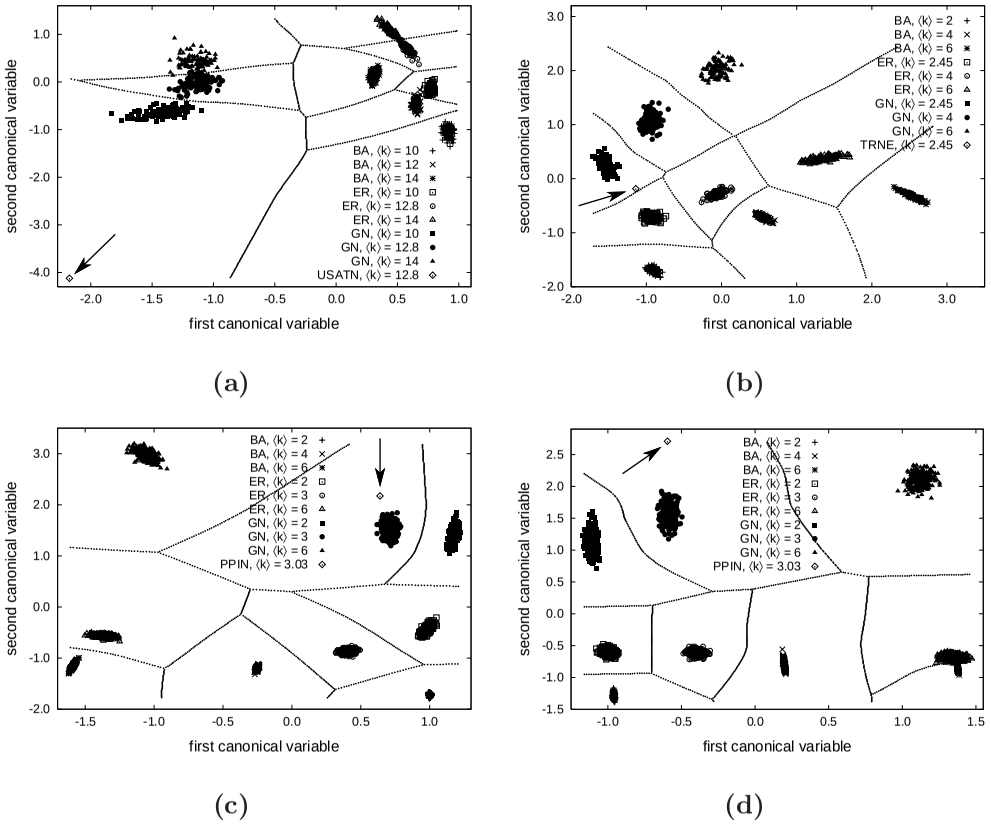


Figura 4.5: Exemplos de classificação via análise das variáveis canônicas e decisão Bayesiana. As redes reais consideradas são: (a) rede de aeroportos dos Estados Unidos (USATN); (b) rede de transcrição genética do *E. coli* (TRNE); e a rede de interações de proteínas do *S. cerevisiae* (PPIN) considerando todas as medidas (c) e excluindo as medidas hierárquicas (d). As setas representam a classificação das redes reais no espaço das medidas. Figura extraída de [dFCRTB07].

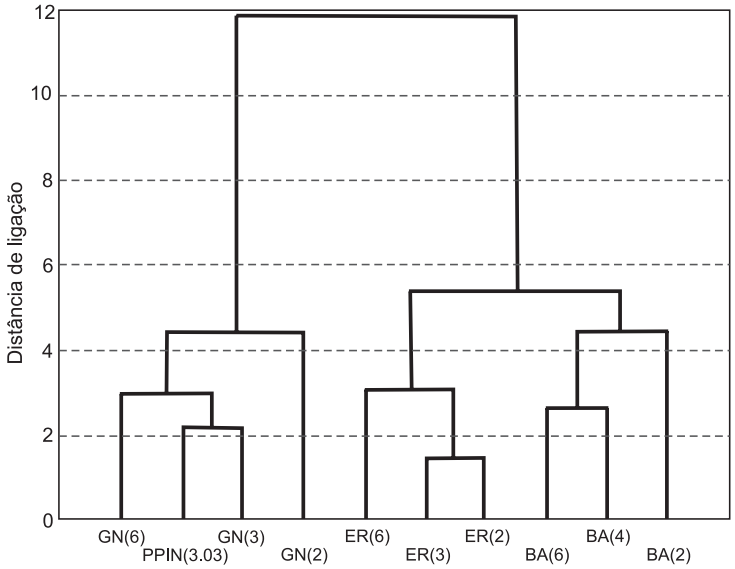


Figura 4.6: Dendograma obtido para a rede de interação de proteínas considerando o seguinte conjunto de medidas: $\{\ell, st, \langle k\rangle, \langle cc\rangle, r, c_D\}$. Note que enquanto os modelos BA, ER e GN resultam em ramos bem separados, a rede de interação de proteínas foi incluída entre as redes GN. Figura extraída de [dFCRTB07].

gura 4.5(c) utilizando o método aglomerativo de Ward [WJ63], que forma grupos de dados buscando minimizar a soma das diferenças entre os elementos de cada grupo e o valor médio do grupo, minimizando o desvio padrão entre os dados de cada grupo formado. A distância de ligação entre as classes é medida ao longo do eixo y, indicando os pontos onde as classes são agrupadas. A rede de interação de proteínas foi incluída no grupo das redes geográficas (GN), que é diferente das redes ER e BA, que por sua vez foram incluídas no mesmo ramo, quando a distância de ligação é maior do que cinco.

4.5 Conclusão

A utilização de análise de variáveis canônicas e decisão Bayesiana, torna possível a consideração de um número elevado de medidas para a caracterização de redes, além de permitir a determinação de qual modelo representa uma dada rede real de forma mais precisa. Essa metodologia auxilia também na determinação de falhas em modelos, já que grande parte deles são considerados adequados para modelar redes reais porque reproduzem apenas algumas propriedades estruturais dessas redes. Destarte, a técnica que apresentamos pode ajudar no desenvolvimento de modelos mais precisos através da análise das medidas que mais contribuem para a discriminação entre modelos de redes. Por exemplo, se o coeficiente de aglomeração médio mais contribui para a separação, isso implica que os modelos considerados devem ser aperfeiçoados de forma a melhor reproduzir essa característica topológica. Conforme notamos em nossa análise de redes reais livre de escala, o modelo de Barabási e Albert não se revelou o mais adequado, embora esse modelo gere redes com distribuição da conectividade do tipo livre de escala, como nas redes reais. Isto se deve ao fato desse modelo não gerar redes com alto coeficiente de aglomeração médio. Como o modelo geográfico de Waxman, por outro lado, gera redes com alto número de ciclos de ordem três, ele se mostrou mais adequado na maioria dos casos. Portanto, para que o modelo de Barabási e Albert reproduza a estrutura da maioria das redes reais livre de escala, ele deve ser aperfeiçoado pela introdução de aspectos dinâmicos de crescimento que permitam-no gerar redes com um maior coeficiente de aglomeração médio. Assim, vemos que a metodologia descrita neste capítulo é fundamental para uma análise mais profunda da estrutura de redes complexas e pode auxiliar no desenvolvimento de modelos mais precisos.

No próximo capítulo, aplicamos a metodologia de classificação aqui descrita na análise de modelos de Internet. Já nos capítulos subseqüentes, apresentamos

uma aplicação do uso da teoria das redes complexas na caracterização de estruturas presentes em redes de interações de proteínas e analisamos a rede dos produtores de vinho da região de Bordeaux (França).

Capítulo 5

Classificação e modelagem da Internet

A Internet surgiu a partir de iniciativa da *Advanced Research Project Agency* (ARPA) que criou a ARPANet com fins militares em meio à Guerra Fria, no final década de 60. A ARPANet tinha o objetivo de conectar as bases militares e os departamentos de pesquisa do governo americano, criando a primeira infraestrutura global de comunicações e os respectivos protocolos. Inicialmente, as conexões eram estabelecidas de forma a oferecerem comunicação entre o leste e o oeste dos Estados Unidos, mas a partir da década de 80, quando a Guerra Fria foi enfraquecida, uma nova utilidade para a ARPANet surgiu: iterligar laboratórios e universidades nos EUA e, mais tarde, em outros países. Foi nessa época que surgiu o nome *Internet*, quando a ARPANet entregou à *National Science Foundation* (NSF) a responsabilidade de manter e aumentar o *backbone*. No final da década de 80, Tim Berners-Lee, conjuntamente com sua equipe do *European Organization for Nuclear Research* (CERN), em Genebra, desenvolveu um sistema de hipertexto que deveria funcionar em redes de computadores. Em 1991, estes pesquisadores criaram a Teia Mundial (*World Wide Web*), quando as informações

disponíveis eram basicamente no formato texto. Entretanto, em 1992, Marc Andressen, do *National Center for Supercomputer Applications* (NCSA), criou o primeiro navegador para Internet, chamado *Mosaic*, que era capaz de interpretar gráficos e realizar navegações através de *links*, como ocorre atualmente. Além disso, neste mesmo ano, Al Gore, que era senador do governo americano, passou a falar sobre a *Information Highway*, ou Superestrada da Informação, o que levou ao crescimento da popularidade da Internet, surgindo grande interesse comercial. A parir daí, o crescimento da Internet foi acelerado.

Hoje em dia, o DARPA (*Defense Advanced Research Projects Agency*), denominação atual do ARPA, investe milhões de dólares em pesquisas para mapear a Internet. O projeto mais visível é o *Cooperative Association for Internet Data Analysis* (CAIDA), cujo objetivo é monitorar grande parte das características da Internet, desde o tráfego até a topologia. O mapeamento da Internet é fundamental, pois sem um mapa é impossível desenvolver novas ferramentas e serviços que ofereçam uma rápida e confiável estrutura de comunicação. A topologia da Internet pode ser analisada considerando-se as conexões entre roteadores ou entre sistemas autônomos [YJB02], sendo que, neste último caso, cada vértice na rede representa uma entidade administrativa, como universidades ou empresas privadas.

Em 1999, Michalis Faloutsos, da Universidade da California-Riverside, Petros Faloutsos, da Universidade de Toronto, e Christos Faloutsos, da Universidade Carnegie Mellon, analisando dados coletados pelo *National Laboratory for Applied Network Research* (NLANR) entre novembro de 1997 e dezembro de 1998, descobriram que as conexões entre os sistemas sistemas autônomos que formam a Internet é distribuída segundo uma lei de potência [FFF99]. A partir desta descoberta, ocorreu um grande interesse da comunidade científica para caracterizar a topologia dessa rede [PSVV01, VPSV02, ZM04a, BCC05] e desenvolver modelos pre-

cisos para reproduzir sua estrutura e evolução [YJB02, CCGJ02, RT04, ZM04b, CHK+06]. Além disso, diversas investigações relacionadas a perturbações, como ataques e falhas [AJB00, CEbAH01], ou simulações de processos dinâmicos, como a propagação de vírus [BFNW04], têm sido desenvolvidas.

Embora a Internet seja uma criação humana, ela possui todas as características de sistemas complexos evolutivos, sendo mais similar a uma célula do que a um circuito de computador. A Internet é um exemplo típico de sistema complexo pois suas conexões não são estabelecidas através de comando externo, mas de acordo com necessidades locais — desde corporações até universidades adicionam roteadores e conexões sem a permissão de qualquer autoridade central. No entanto, apesar das conexões serem criadas de forma descentralizadas, elas não são estabelecidas de forma aleatória, mas segundo diversos mecanismos como os relacionados a ligação preferencial, posição geográfica dos roteadores e recursos disponíveis para estabelecer tais conexões. Enquanto poucos roteadores concentram grande parte das conexões, muitos outros possuem apenas um único *link*. O fato da estrutura da Internet seguir uma lei de potência a torna altamente tolerante a falhas aleatórias, mas vulnerável a ataques direcionados [CEbAH01].

Modelos que reproduzam as principais características da Internet são importantes para o desenvolvimento de protocolos de roteamento e no planejamento de tráfego [PSV04]. Apesar da estrutura das conexões da Internet mudar constantemente, suas propriedades globais, como a conectividade média, o menor caminho médio e coeficiente de aglomeração se mantém praticamente constante, o que sugere que as propriedades topológicas da Internet alcançaram um estado estacionário [PSVV01, VPSV02]. Logo, os modelos para representar a evolução da Internet devem levar em conta estes aspectos. Na Referência [RBTdFC07], comparamos diferentes modelos de redes complexas na modelagem da Internet e verificamos que nenhum dos modelos considerados é preciso o suficiente para

reproduzir a maioria das propriedades topológicas da Internet.

Para se obter modelos precisos, é necessário analisar como as conexões são estabelecidas entre os roteadores. Quando uma universidade, empresa privada ou um provedor de Internet vai estabelecer um novo *link*, o principal parâmetro considerado é o custo de comunicação. Porém, às vezes, conexões um pouco mais longas do que a mais próxima possível pode resultar em maior largura de banda e, portanto, em uma melhor relação custo-benefício. A ligação preferencial é resultante do fato de que os roteadores que fornecem mais largura de banda tendem a receber mais conexões. Além disso, roteadores estão localizados onde há maior população e mais demanda por tráfego [YJB02], o que mostra a importância da sua localização geográfica. Como estes fatores são fundamentais na definição da estrutura da Internet, para se obter modelos precisos é necessário levá-los em conta. Portanto, os modelos de Internet devem ser baseados no crescimento do número de roteadores e conexões, na ligação preferencial, na distância entre os roteadores e na distribuição geográfica dos roteadores ou sistemas autônomos.

O modelo que sugerimos neste trabalho basicamente considera o crescimento, a ligação preferencial e a distância geográfica entre os sistemas autônomos. A rede é construída da seguinte forma:

i. A cada passo um novo sistema autônomo i é adicionado à rede num posição geográfica (x, y) escolhida aleatoriamente dentro de uma caixa $L \times L$.

ii. São escolhidos os m vizinhos geográficos mais próximos de i. Dentre eles, são escolhidos os r com maior conectividade e cada um deles é conectado a i com uma probabilidade α.

iii. São escolhidos aleatoriamente q sistemas autônomos, já presentes na rede, que são conectados aos s sistemas autônomos de maior conectividade dentro de uma distância $L/4$, onde cada ligação é estabelecida com uma proba-

bilidade β.

iv. O processo termina quando N sistemas autônomos tiverem sido adicionados à rede.

As conexões são limitadas pela conectividade máximo da rede real, como ocorre no modelo livre de escala limitado de Amaral *et al.* [ASBS00].

Cada passo do modelos é justificado pela análise de como são estabelecidas as conexões entre sistemas autônomos na Internet. O crescimento da rede se realizada através da adição de novos roteadores e conexões. Quando se vai estabelecer uma conexão, primeiramente analisa-se quais os sistemas autônomos mais próximos. Dentre eles, a tendência é a escolha daquele que ofereça maior largura de banda. No entanto, nem sempre é possível estabelecer a conexão, devido a fatores como a limitação do número de conexões do sistema autônomo escolhido. Estas características são representadas pelos parâmetros m, r e α que consideramos em nosso modelo. As conexões que são estabelecidas entre os sistema autônomo já existentes na rede também ocorrem constantemente [PSVV01]. Neste caso, como alguns sistemas autônomos podem estar distribuídos por mais de uma região, como ocorre com universidades que possuem institutos espalhados por diversas cidades, conexões de maior alcance geográfico são estabelecidas. Novamente, fatores limitantes, como os recursos para estabelecer a conexão ou limitações dos sistemas autônomos, podem bloquear as ligações. Neste caso, os parâmetros q, s e β são justificados.

Para verificar a adequação do nosso modelo, utilizamos uma base de dados coletada pela Universidade de Oregon (*University of Oregon Route Views Project*) através da análise de tabelas de roteamento BGP (*Border Gateway Protocol*). O BGP é um protocolo de roteamento para ser usado entre múltiplos sistemas autônomos baseado no protocolo TCP/IP. O BGP constrói um gráfico dos sistemas autônomos, usando as informações trocadas pelos “vizinhos BGP”

(*BGP neighbors*), que são compostas dos números de identificação dos sistemas autônomos. Baseados nestas informações, os sistemas autônomos conseguem trocar informações e determinar o melhor caminho para as redes que formam a Internet. Essas tabelas são então processadas e disponibilizadas pelo *National Laboratory of Applied Network Research* (NLANR). Utilizamos a rede obtida em dois de abril de 1998, que é constituída de 3.522 roteadores e 6.324 conexões.

Em nossas simulações, encontramos os parâmetros que melhor representam a estrutura da Internet: $m = 40$, $r = 3$, $\alpha = 0{,}15$, $q = 2$, $s = 2$ e $\beta = 0{,}4$. O tamanho da caixa, L, pode ser definido por um valor arbitrário. Consideramos 17 medidas de redes, diferentemente do capítulo anterior, quando utilizamos nove medidas. As medidas são: (i) a conectividade média ($\langle k \rangle$), (ii) a conectividade máxima (k_{max}), (iii) o coeficiente de aglomeração médio (cc), (iv) a conectividade médio entre os vizinhos (k_{nn}), (v) o menor caminho médio (ℓ), (vi) a medida de assortatividade (r), (vii) a dominância do ponto central ($\mathrm{c_D}$), (viii) o grau de intermediação médio ($\langle B \rangle$), (ix) a medida de retidão da distribuição (st), (x) o grau médio hierárquico de nível dois ($\langle k_2 \rangle$), (xi) o coeficiente de aglomeração hierárquico de nível dois (cc_{22}), (xii) a razão de convergência de nível dois (cv_2), (xiii) a razão de divergência de nível dois (dv_2), (xiv) o grau médio hierárquico de nível três ($\langle k_3 \rangle$), (xv) o coeficiente de aglomeração hierárquico de nível três (cc_{33}), (xvi) a razão de convergência de nível três (cv_3) e (xvii) a razão de divergência de nível três (dv_3). Todas estas medidas estão descritas no Capítulo 3.

Compararmos o modelo que sugerimos com nove modelos de redes complexas: (i) grafos aleatórios de Erdős e Rényi (ER) [ER59, ER60, ER61], (ii) *small world* de Watts e Strogatz (WS) [WS98], (iii) livre de escala de Barabási e Albert (BA) [BA99], (iv) o modelo geográfico de Waxman (GN) [Wax88], (v) o modelo livre de escala limitado de Amaral *et al.* (LSF) [ASBS00], (vi) o modelo livre de escala com ligação preferencial de Dorogovtsev *et al.* (DMS) [DMS00],

(vii) o modelo livre de escala não linear de Krapivsky *et al.* (KP) [KRL00], (viii) o modelo preferencial dirigido de Bar *et al.* (GdTang) [BGWB05] e (ix) o gerador de topologia da Internet inet (Inet) [JCJ00, WJ02]. Os modelos (i)-(vii) estão descritos no Capítulo 2. Os modelos (viii) e (ix) foram desenvolvidos especificamente para reproduzirem a estrutura da Internet. Os modelos BA, LSF, DMS, KP, GdTang e Inet geram redes livre de escala como ocorre na Internet. Já os modelos ER, WS e GN foram considerados em nossa análise porque refletem outras importantes propriedades estruturais como o alto coeficiente de aglomeração (WS e GN) e o caminho médio pequeno (WS e ER). Para cada modelo, geramos 50 redes com o mesmo número de vértices que a rede original e com conectividade média semelhante. Conforme podemos notar na Tabela 5.1, o modelo que desenvolvemos reproduz de forma precisa as principais propriedades estruturais das redes complexas. Apenas o menor caminho médio ℓ e a conectividade média entre os vizinhos k_{nn} não corresponderam ao valor encontrado na rede real, embora tenham ficado bastante próximos.

Para obtermos o espaço de classificação, utilizamos o seguinte conjunto de medidas:

i. $\{\langle k\rangle, k_{max}, \ell\}$,

ii. $\{k_{max}, cc, k_{nn}, r, c_D\}$,

iii. $\{\langle k\rangle, k_{max}, cc, \ell, r, \langle k_3\rangle, cc_{33}\}$,

iv. $\{cc, k_{nn}, \ell, c_D, \langle k_2\rangle, cc_{22}, \langle k_3\rangle, cc_{33}\}$,

v. $\{\langle k\rangle, k_{max}, cc, k_{nn}, \ell, r, c_D, st, \langle k_2\rangle, cc_{22}, dv_2\}$,

vi. $\{\langle k\rangle, k_{max}, cc, k_{nn}, \ell, r, \mathrm{c_D}, \langle B\rangle, st, \langle k_2\rangle, cc_{22}, cv_2, \langle k_3\rangle, cc_{33}, cv_3\}$

Esta combinações foram escolhidas a partir de combinações das medidas mais comuns com medidas de centralidade e hierárquicas. Os resultados obtidos, con-

Tabela 5.1: Comparação entre as medidas não hierárquicas da rede real e daquelas geradas pelo modelo que propomos.

Medida	Internet	Modelo
$\langle k \rangle$	3,56	3,59 $\pm$ 0,03
k_{max}	742	742
cc	0,19	0,18 $\pm$ 0,01
k_{nn}	177,7	156,7 $\pm$ 3,7
ℓ	3,4	4,2 $\pm$ 0,1
r	-0,21	-0,19 $\pm$ 0,01
c_D	0,33	0,36 $\pm$ 0,01
$\langle B \rangle$	0,0008	0,0006 $\pm$ 0,0001

siderando a metodologia de classificação descrita no Capítulo 4, são apresentados na Figura 5.1. Conforme podemos observar, em todos os casos a rede real foi associada ao modelo que desenvolvemos, o que mostra que nosso modelo é mais preciso do que os geradores de topologia da Internet. Deste modo, os fatores dinâmicos que consideramos para a construção desse modelo são determinantes da estrutura da Internet. Interessante notar que mesmo quando todas as 17 medidas são consideradas, a rede real ainda é associada ao modelo que desenvolvemos. Deste modo, esse modelo se mostrou bastante preciso, reproduzindo todas as propriedades estruturais quantificadas pelas medidas.

O modelo que apresentamos constitui um importante passo no entendimento da evolução da Internet. Os fatores utilizados na sua construção, como a ligação preferencial, a distância geográfica entre os roteadores e a adição constante de conexões entre roteadores antigos, se mostraram fundamentais para uma modelagem precisa da Internet. Além disso, a aplicação da metodologia de classificação

descrita no Capítulo 4 nos mostra que a consideração de um conjunto amplo de medidas é importante para uma melhor classificação e determinação da precisão de modelos de redes complexas. Nos capítulos posteriores apresentamos mais duas aplicações da teoria das redes complexas na caracterização e análise de sistemas complexos.

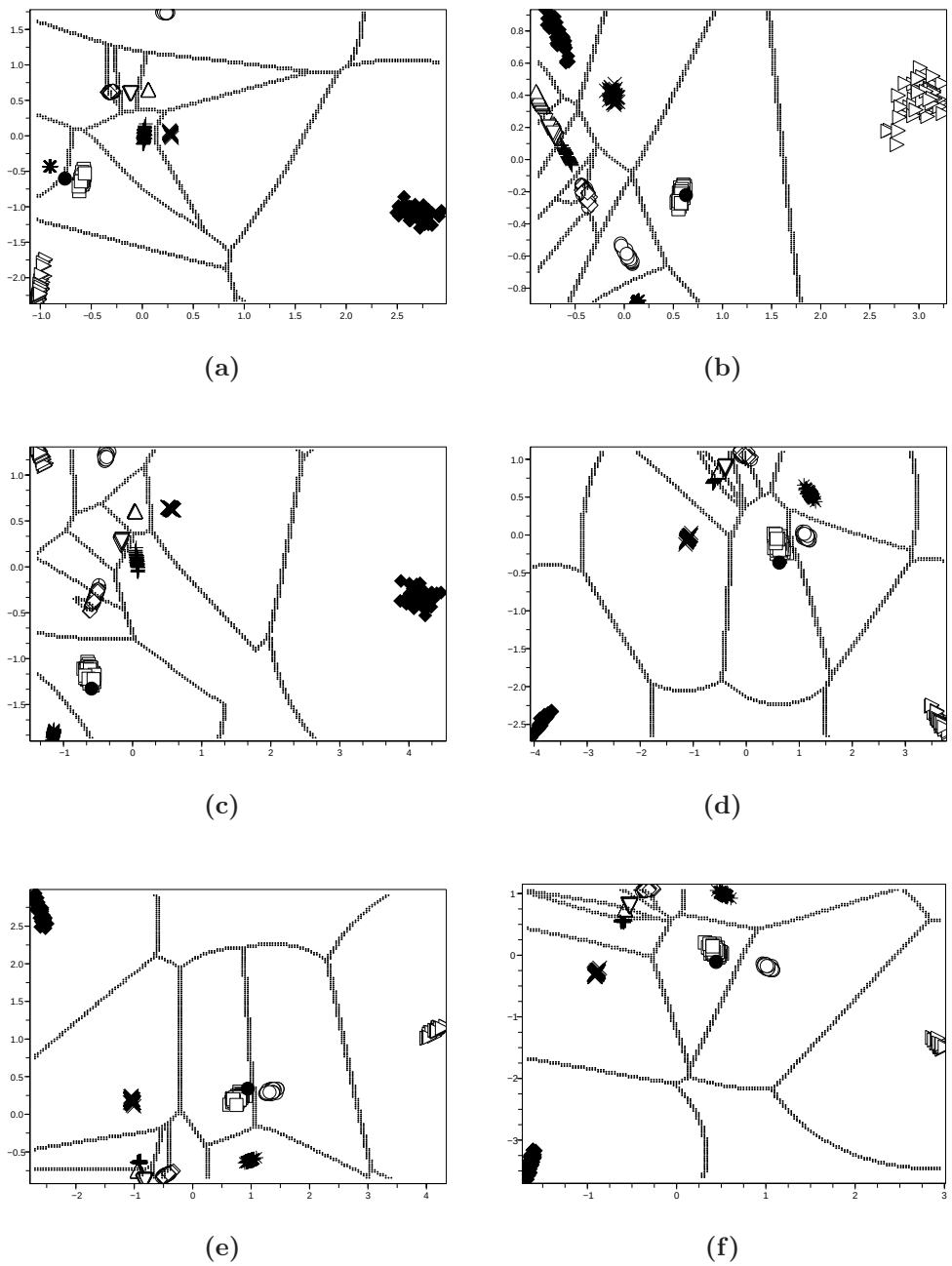


Figura 5.1: Classificações obtidas considerando os seguintes conjuntos de medidas descritos no texto: (a) (i), (b) (ii), (c) (iii), (d) (iv), (e) (v) e (f) (vi). Os modelo são representados por: $+$ ER, $\times$ WS, $\oplus$ BA, $\blacklozenge$ GN, $\lozenge$ LSF, $\triangle$ DMS, $\triangledown$ KP com $\alpha = 0,5$, $\triangleright$ KP com $\alpha = 1,5$, $\circ$ GdTang, $*$ Inet e $\square$ o modelo que desenvolvemos. A rede real é indicada por $\bullet$.

Capítulo 6

O aspecto da letalidade em redes de interações de domínios protéicos

A biologia tem usado o reducionismo ao longo dos séculos para entender os componentes celulares associando funções a elementos específicos, como proteínas e genes. Entretanto, apesar do enorme sucesso obtido, grande parte das funções celulares não podem ser descritas pela análise individual dos seus componentes [HHLM99]. Deste modo, é fundamental estudar a estrutura e dinâmica das interações entre os constituintes celulares para se entender o complexo processo que gera a vida [RG03, HHLM99]. A atividade celular pode fundamentalmente ser dividida em três níveis básicos, (i) *metabólico*, que é determinado pelas (ii) *interações protéicas*, cuja produção é controlada pela (iii) pelas *regulações gênicas* [BO04]. Assim, para se compreender os processos biológicos que originam a vida, é indispensável analisar como a energia é obtida pela célula através de processos bioquímicos entre os componentes metabólicos (produtos, substratos e enzimas); como as proteínas participam em diversos processos, como na

formação de complexos protéicos; e como as informações são transmitidas de um fator de transcrição para o gene que tal transcrição regula. Como estes processos biológicos são determinados por componentes discretos (moléculas, genes ou proteínas) que interagem, é natural modelá-los por redes [BO04].

Redes de interações de proteínas podem ser obtidas representando cada proteína como um vértice na rede e cada possível interação como uma aresta não-direcionada, ligando os respectivos vértices. Geralmente, as conexões não possuem peso, embora em alguns casos um índice de credibilidade possa ser associado às interações [Kro]. Essencialmente, o estudo das redes de interações de proteínas pode ser dividido em quatro áreas básicas: (i) caracterização da estrutura das conexões [JMBO01, RSDR+01, MS02, WOB03, GBB+03, RB03, Alb05], (ii) predição e caracterização das funções das proteínas [M+99, EMXY00, MH00, VFMV03a, AA04], (iii) modelagem e simulação da evolução das interações entre proteínas [Wag01, SPSSK02, VFMV03b, Wag03, PSSS03, QLWL03, BLW04], e (iv) caracterização e modelagem das redes de interações entre domínios protéicos [Wuc01, Wuc02, BIH04]. O trabalho que descrevemos neste capítulo está relacionado aos ítens (i) e (iv). Deste modo, caracterizamos a estrutura das redes de interações de proteínas e domínios protéicos, utilizando medidas de redes (ver Capítulo 3), e analisamos a relação entre letalidade e conectividade nestas redes.

6.1 Domínios protéicos

As proteínas são compostas de unidades estruturais denominadas domínios [JC85, PR02]. O conceito de domínios protéico foi proposto em 1973 por Wetlaufer, após a realização de estudos cristalográficos de raios-X, que definiu domínios como sendo unidades estáveis que compõem a estrutura das proteínas e que podem dobrar-se de forma independente [Phi66]. Outras definições posteriores definem

os domínios como estruturas compactas [Ric81], relacionadas à função, evolução e dobramento das proteínas [Bor91, Wet73]. Na natureza, diversos domínios se unem para formar proteínas multi-funcionais [Cho92], onde cada domínio pode realizar uma função independentemente ou em associação com seus vizinhos [GH02]. Assim sendo, grande parte da funcionalidade das proteínas é definida pela presença de domínios protéicos específicos, que são formados de porções de aminoácidos que definem a estrutura primária das proteínas [NZT03]. Como as interações físicas das proteínas envolvem interações físicas entre domínios, o estudo da ligação entre tais estruturas pode ser particularmente útil para validar, explicar ou mesmo predizer interações de proteínas. A Figura 6.1 apresenta a enzima *piruvato kinase*, que participa do processo de glicólise[1] e é formada por três domínios, um domínio regulador β, um domínio α/β de ligação de substratos e um domínio α/β de ligação de nucleotídeos [GH02].

6.2 Interações de proteínas

Foi pela utilização de dois domínios com funções específicas, o domínio de ligação ao DNA (BD) e o domínio de ativação da transcrição (AD), que os estudos relativos à rede de interação de proteínas se difundiram a partir do final da década de 80. Stanley Fields, um pesquisador da Universidade de Washington, propôs em 1989 uma técnica semi-automática, chamada análise dupla híbrida (*two-hybrid assay*), para detectar interações entre proteínas de uma forma relativamente simples e rápida [FS89]. Naquela época, a necessidade de um método com estas características era fundamental, pois na maioria dos organismos há milhões de

[1] A glicólise é a seqüência metabólica de várias reações enzimáticas, na qual a glicose é oxidada produzindo duas moléculas de ácido pirúvico e dois equivalentes reduzidos de NAD+, que ao serem introduzidos na cadeia respiratória, produzem duas moléculas de ATP.

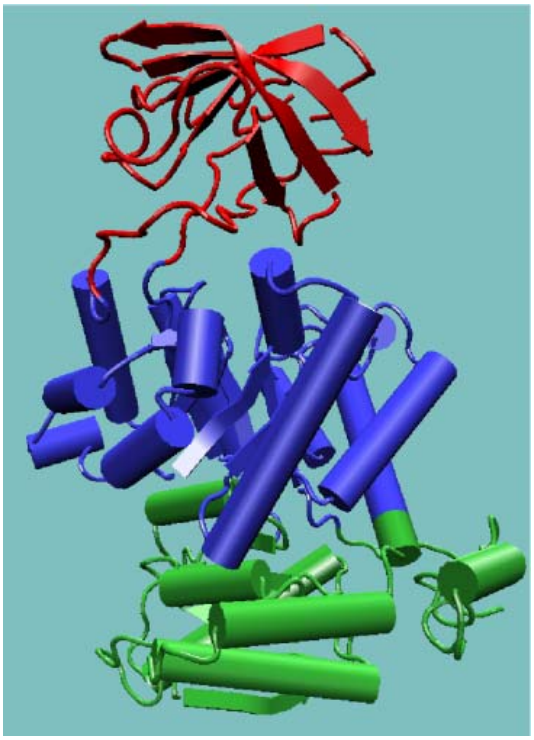


Figura 6.1: A enzima *piruvato kinase* contém três domínios: (i) um domínio regulador β, (ii) um domínio α/β de ligação de substratos e (iii) um domínio α/β de ligação de nucleotídeos [GH02]

interações possíveis. Por exemplo, no caso do *Saccharomyces cerevisiae*, cujo genoma foi anunciado em 24 de abril de 1996, que possui 6.300 genes que codificam aproximadamente o mesmo número de proteínas, a determinação das interações exige o exame de 6.300×6.300 pares, que totaliza quase 40 milhões de interações possíveis. Deste modo, se fossem utilizadas técnicas padrões de biologia molecular, poderia se levar décadas e centenas de pessoas para se determinar o interactoma dos organismos mais simples. Embora a técnica dupla híbrida ofereça um alto número de falso positivos, que são as interações que ocorrem no experimento, mas nunca no organismo vivo, os mapas que ela tem determinado geram oportunidades únicas de estudo das interações entre proteínas de forma a

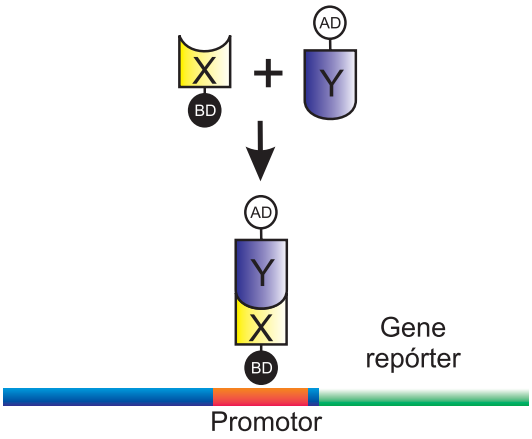


Figura 6.2: Na técnica dupla híbrida, o domínio de ligação ao DNA se liga a uma seqüência promotora específica, que se situa no início de um gene repórter. Este, por sua vez, interage com o domínio de ativação da transcrição (AD), o qual atrai os componentes críticos do complexo de iniciação de transcrição. O gene que codifica a proteína de interesse X é fusionado ao gene que codifica o domínio de ligação ao DNA (BD), enquanto uma biblioteca de cDNA, a qual codifica várias proteínas potencialmente interativas a serem testadas, Y, é fundida ao gene que codifica o domínio de ativação de transcrição (AD). Quando ocorre uma interação entre a proteína de interesse, X, e uma proteína da biblioteca, Y, o fator de transcrição é reconstituído e os genes repórteres que estão sob seu controle são ativados.

determinar a estrutura de sua conexões [Bar03]. A Figura 6.2 ilustra a técnica dupla híbrida. Além da análise dupla híbrida, outros métodos de detecção de interação de proteínas têm sido largamente utilizados atualmente, como técnicas de co-imunoprecipitação [EIK99] e *microarray* [MS96]. A partir das bases geradas por esses métodos, iniciou-se a análise da estrutura das conexões que formam a rede de interação de proteínas.

A caracterização da estrutura das redes de interações de proteínas é funda-

mental para modelar o surgimento das interações, prever funções protéicas, ou mesmo analisar o comportamento de processo dinâmicos, como falhas e transporte. Até o final da década de 90, acreditava-se que as interações entre as proteínas eram completamente aleatórias, sem qualquer organização. Entretanto, Jeong *et al.* [JMBO01], analisando a rede de interações do fungo *Saccharomyces cerevisiae*, descobriu que as ligações não ocorrem uniformemente, mas segundo ligação preferencial, o que resulta em uma distribuição do tipo lei de potência com um corte exponencial, dada por $\mathrm{P(k)} \approx (\mathrm{k}+\mathrm{k_0})^{\gamma}\mathrm{e}^{-(\mathrm{k}+\mathrm{k_0})/\mathrm{k_c}}$, onde k_c é a conectividade de corte. Deste modo, a estrutura das redes de interações de proteínas é altamente heterogênea, formada por poucas proteínas muito conectadas (*hubs*) e muitas proteínas pouco conectadas. Em estudos posteriores, a mesma estrutura foi encontrada nas redes de interações de proteínas da bactéria *Helicobacter pylori* [RSDR+01] e do inseto *Drosophila melanogaster* [GBB+03]. Tais descobertas sugerem que a distribuição livre de escala é uma característica presente em todos os organismos e provavelmente essa é a configuração das ligações que fornece melhor adaptação às pressões exigidas pela seleção natural. Além de apresentar distribuição das conexões livre de escala, as redes de interações de proteínas ainda são *small-world* ($\ell \approx 7$), apresentam uma alta quantidade de ciclos de ordem três [RB03] e estrutura modular [RG03, Alb05].

A presença de *hubs* nas redes livre de escala permite que elas sejam altamente tolerantes a falhas [AJB00], que ocorrem quando vértices são retirados aleatoriamente. Como os *hubs* constituem uma pequena fração do número total de vértices, a probabilidade do vértice que sofreu a falha ser um *hub* é pequena. Assim, como a arquitetura não-homogênea das redes livres de escala torna-as altamente robustas a falhas, simples organismos continuam a crescer e se reproduzir apesar de viverem às vezes em ambientes hostis [Kit04]. Entretanto, quando há ataques, ou seja, os *hubs* são removidos, estas redes entra em colapso rapidamente [AJB00], se

decompondo em componentes não conectados. Esta fragilidade a ataques direcionados é um preço que as redes livres de escala pagam pela sua robustez [Bar03].

As proteínas letais (essenciais) são as mais importantes para a manutenção da vida e reprodução dos organismos. Devido à essa importância, Jeong *et al.* investigaram se haveria alguma relação entre a conectividade das proteínas e sua letalidade, isto é, se há uma tendência para que as proteínas letais sejam as mais conectadas, uma vez que os *hubs* são fundamentais para garantir a robustez das redes. Embora eles tenham notado uma correlação entre a letalidade e número de conexões, ela não é bem definida. Assim sendo, como os domínios são as unidades funcionais das proteínas, decidimos analisar a relação conectividade-letalidade em redes de interação de domínios.

6.3 Redes de interação de domínios protéicos

As redes de interações de domínios podem ser obtidas via complexos protéicos, seqüências de Rosetta Stone e usando as redes de interações de proteínas [Wuc01, Wuc02, NZT03]. Em nossas investigações, apresentada em detalhes na Referência [dFCRT06], nós construímos a rede de interações de domínios considerando a rede de interações de proteínas do *Saccharomyces cerevisiae* obtida por Sprinzak *et al.* [SSM03], que utilizaram bases de dados não redundantes. A rede de interação de proteínas é formada por 4.135 proteínas e 8.695 interações. Utilizando a base de dados do Pfam [BCD+04], a qual é composta por uma larga coleção de múltiplas seqüências e alinhamentos obtidos por modelos de Markov escondidos (HMM), foram identificados 1.424 domínios nas proteínas que formam a rede. Alguns desses domínios são altamente abundantes, podendo aparecer em mais de uma proteína. Além disso, muitas proteínas podem ser formadas por mais de um domínio. A Figura 6.3 apresenta um exemplo da distribuição de

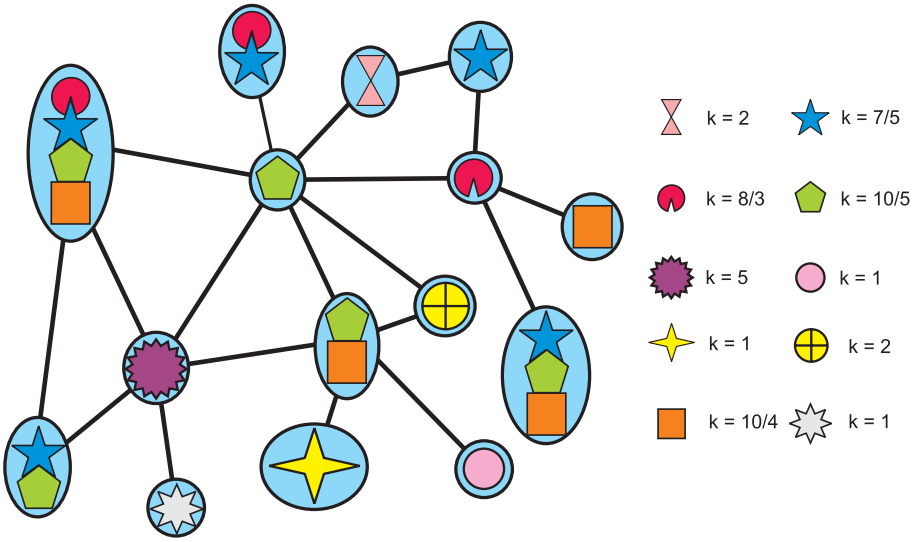


Figura 6.3: Exemplo de uma rede de interação de domínios. As proteínas são representadas pelas elipses e os domínios que as compõem são indicados por figuras geométricas circunscritas. A conectividade k de um domínio é dada pelo número de conexões entre as proteínas que são formadas por este domínio e o restante da rede (excetuando-se as conexões entre as proteínas que possuem o mesmo domínio) dividido pelo número de proteínas onde tal domínio aparece.

domínios em uma rede de interação de proteínas.

Para cada domínio, é definido um sugrafo formado pelas proteínas que contém o respectivo domínio. Nesse caso, a *conectividade de um dado domínio* é definida hierarquicamente como sendo igual número de conexões entre as proteínas que formam tal subrede e o restante da rede, dividido pelo número de proteínas presentes em tal subrede, de forma a não favorecer os domínios mais abundantes, que tendem a ter maior número de conexões. As conexões internas ao subgrafo não são consideradas no cálculo da conectividade. Note a definição de conectividade dos domínios é diferente do grau hierárquico, apresentada na Seção 3.5.

Assim como no caso das proteínas, a letalidade de domínios pode ser definida. Entretanto, como não há dados experimentais ou qualquer consenso sobre a letalidade de domínios, nós usamos duas hipóteses para definir tal letalidade da seguinte forma:

- I. *Letalidade de domínios em um sentido fraco*: um domínio é letal se ele aparece em proteínas letais.
- II. *Letalidade de domínios em um sentido forte*: um domínio é letal se ele aparece apenas em proteínas letais formadas por um único domínio.

A primeira definição é considerada fraca porque um domínio definido como letal pode aparecer em proteínas viáveis (não-letais) e letais simultaneamente. Entretanto, tal suposição é ainda potencialmente interessante porque domínios que possuem funções similares, têm maior probabilidade de ocorrerem conjuntamente em proteínas do que separados [NZT03, VTPL05], o que sugere que proteínas letais devem ser uniformemente formadas por domínios letais. A segunda hipótese, por outro lado, é considerada forte porque se um domínio aparece apenas em proteínas letais formadas por um único domínio, ele deve ser responsável pela essencialidade de tais proteínas. Quando adotamos a primeira hipótese, toda a rede de interação de proteínas é considerada. No segundo caso, por outro lado, a rede de interação é formada apenas pelas proteínas que possuam um único domínio. A Figura 6.4 apresenta um exemplo de domínios não-letais, letais em um sentido fraco e letais em um sentido forte.

É importante notar que as duas situações de letalidade descritas anteriormente são apenas hipóteses que devem ser confirmadas por resultados experimentais. Em outras palavras, eventual identificação de alta correlação entre conectividade e letalidade, deve ser entendida como suporte para tal suposição, devido à regra *centralidade-letalidade*, que reflete a importância de *hubs* na organização da ar-

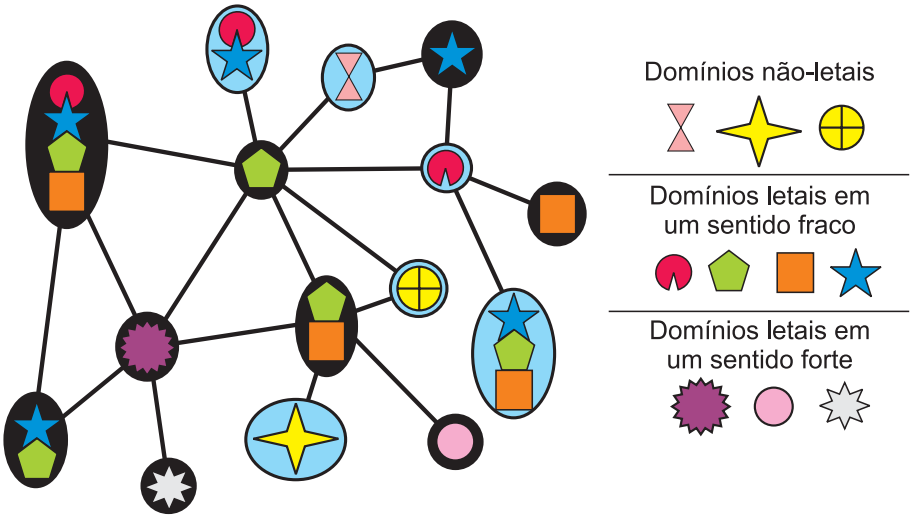


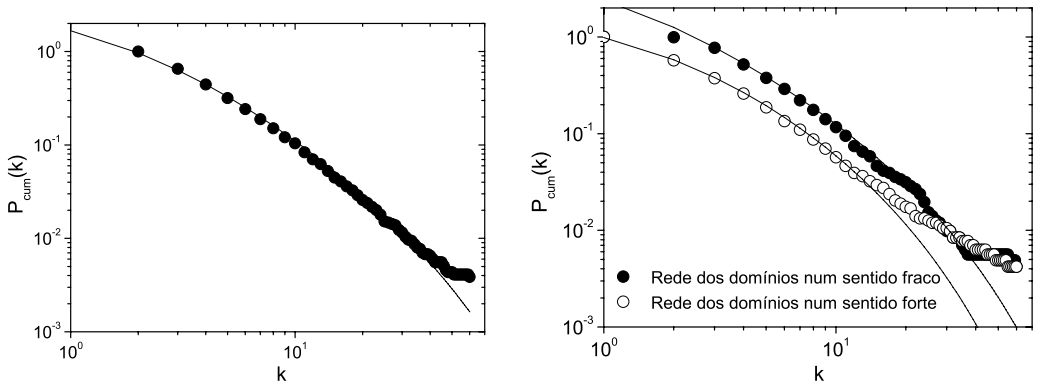
Figura 6.4: Exemplo de uma rede formada por domínios não-letais, letais em um sentido fraco e letais em um sentido forte. As proteínas letais são representadas pelas elipses pretas e as não letais por elipses azuis. Veja o texto para maiores detalhes.

quitetura e dinâmica de redes [HZ06].

6.4 Resultados e discussões

A Figura 6.5 apresenta os histogramas cumulativos das conectividades de proteínas e domínios, cujas letalidades foram definidas num sentido fraco e forte. Todas essas distribuições seguem uma lei de potência com um limite (*cutoff*) exponencial, sendo descritas por $\mathrm{P_{cum}(k)} \approx (\mathrm{k}+\mathrm{k_0})^{\gamma}\mathrm{e}^{-(\mathrm{k}+\mathrm{k_0})/\mathrm{k_c}}$ [JMBO01]. O valor de k_c e γ obtidos a partir das distribuições cumulativas são apresentados na Tabela 6.1.

A Figura 6.6 mostra as relações entre conectividade e letalidade verificadas em redes de proteínas e domínios. A letalidade das proteínas foi determinada usando-se a base de dados do *Munich Information Center for Protein Sequences* (MIPS) [MFG+02] e o número de proteínas e domínios letais, N_L, é mos-



(a) **(b)**

Figura 6.5: As distribuições cumulativas das (a) redes de proteínas e (b) domínios, usando ambas hipóteses (I) e (II), seguem uma lei de potência com um corte exponencial $\mathrm{P_{cum}(k)} \approx (\mathrm{k}+\mathrm{k_0})^{\gamma}\mathrm{e}^{-(\mathrm{k}+\mathrm{k_0})/\mathrm{k_c}}$, representadas pelas linhas contínuas.

Tabela 6.1: Valores estatísticos obtidos para as redes de proteínas e domínios. N é número de vértices, $\langle k \rangle$ é a conectividade média, k_c a conectividade de *cutoff*, γ o expoente da lei de potência, r o coeficiente de Pearson e ρ o coeficiente de Spearman.

Rede de interação	N	N_L	$\langle k \rangle$	k_c	γ	r	ρ
Proteínas	4.135	795	2,10	38	2,8	0,12	0,10
Domínios em um sentido fraco	1.424	499	2,24	22	2,5	0,61	0,61
Domínios em um sentido forte	818	243	1,27	15	2,6	0,73	0,77

trado na Tabela 6.1. A abscissa da Figura 6.6 representa a conectividade k das proteínas (ou dos domínios), cujo limite (*cutoff*) (ver Figura 6.5) são apresentados na mesma tabela. Já a ordenada representa a fração de proteínas (ou de domínios) que tem conectividade k, ou seja, este valor varia de zero a um, dependendo do número de proteínas ou domínios que tenham conectividade k. A fim de determinarmos a correlação entre a fração de proteínas (ou domínios) letais e a respectiva conectividade, nós calculamos os coeficientes de Pearson [Edw76], r, e de Spearman [LD98, FW93], ρ, que é um coeficiente não paramétrico usado no caso de relações não lineares. O coeficiente de Pearson é calculado pela Equação 3.3. Já o de Spearman é calculado por

$$\rho_{xy} = 1 - 6 \sum \frac{d_{xy}^2}{N(N^2-1)}, \tag{6.1}$$

sendo a soma realizada considerando todos os N elementos dos vetores e d_{xy} é igual à diferença entre os níveis (*ranks*) dos valores correspondentes dos elementos dos vetores x e y. Os níveis são atribuídos associando valor 1 aos maiores valores dos vetores x e y, 2 aos próximos maiores e assim sucessivamente. O coeficiente de Spearman corresponde ao coeficiente de Pearson em níveis [Edw76]. Os valores de r e ρ são mostrados na Tabela 6.1, onde podemos ver que a correlação entre letalidade e conectividade é maior para os domínios, usando as hipóteses (I) e (II), do para as proteínas. Para verificarmos a significância estatística de tais resultados, utilizamos o teste de Fisher para comparação de coeficientes de correlação, p [Fis21]. A correlação entre os coeficientes obtidos para proteínas e domínios letais no sentido fraco, resulta em $p \leq 0,035$ para r e $p \leq 0,001$ para ρ. Já a comparação com domínios letais no sentido forte, resulta em $p \leq 0,015$ para r e $p \leq 0,001$ para ρ. Como r e ρ são menor do que $0,05$, isso indica que existe associação entre a conectividade e a letalidade dos domínios tanto num sentido forte com fraco.

Para compararmos as redes de domínios com uma versão aleatória, redis-

tribuímos as proteínas na rede como rótulos, isto é, mudamos as posições das proteínas mantendo a mesma estrutura da rede. Assim, a conectvidade dos domínios também tende a mudar, já que as proteínas redistribuídas carregam consigo seus respectivos domínios. Repetindo a análise anterior para redes de proteínas e domínios, verificamos pelos resultados apresentados na Tabela 6.1 e na Figura 6.6(d), que a correlação encontrada nas redes de domínios não é um efeito casual. Os resultados apresentados são uma média tomada sobre 100 versões aleatórias.

Como os domínios representam a unidade evolutiva básica que forma as proteínas, não é novidade concluir que os domínios têm um papel fundamental na definição da letalidade de proteínas [VTPL05]. Desta forma, os resultados que obtivemos indicam que as interações entre proteínas, bem como sua letalidade, devem ser definidas no nível dos domínios.

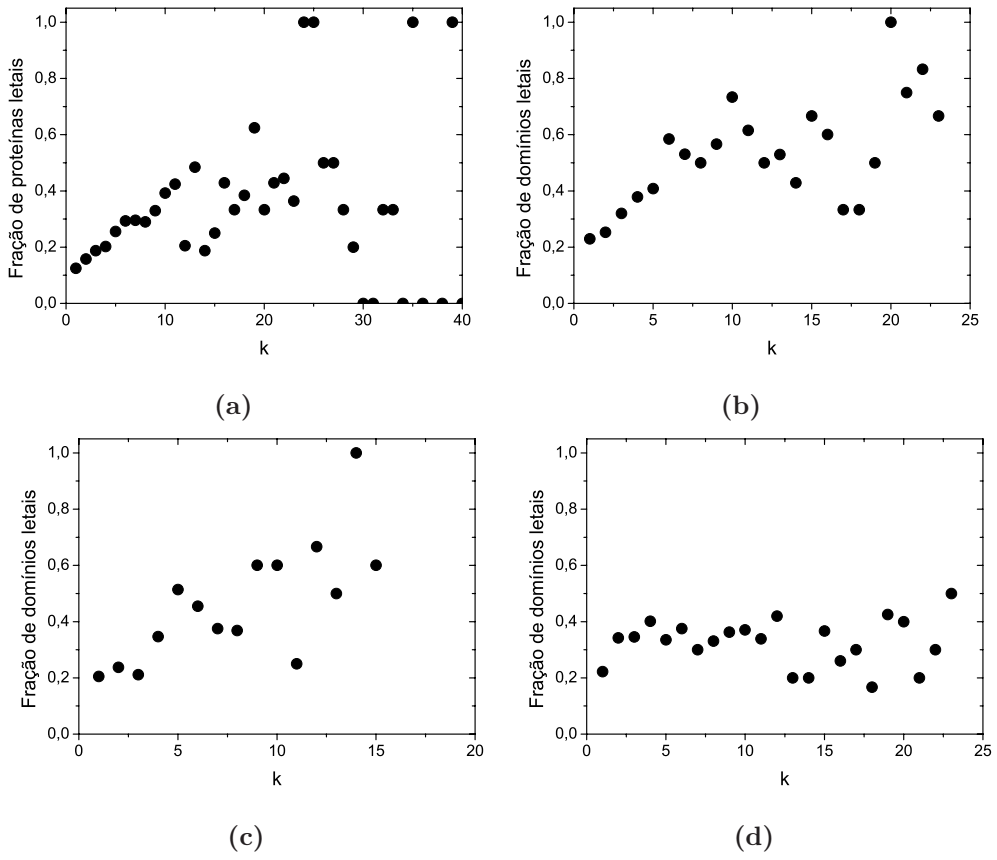


Figura 6.6: Fração de proteínas e domínios letais com uma conectividade particular para as redes (a) de proteínas, (b) domínios letais no sentido fraco, (c) domínios letais no sentido forte e (d) versão aleatória da rede de interação de domínios. A fração de proteínas letais com uma conectividade k é dada pela razão entre o número de proteínas letais que apresentam esta conectividade e o número total de proteínas com a mesma conectividade k.

Capítulo 7

A rede complexa dos produtores de vinhos de Bordeaux

A qualidade dos vinhos é fortemente influenciada por fatores geológicos, climáticos e por técnicas de cultivo e produção [JD00, Par03]. Como *chateaux* pertencentes à mesma região devem produzir vinhos de qualidade e características similares, as regiões de produção em Bordeaux foram delimitadas de modo a haver uma relação entre o território em que o vinho é cultivado e suas propriedades. No entanto, a classificação das regiões de produção tem sido feita de forma subjetiva, já que a qualidade dos vinhos é determinada através de análise olfativa e gustativa, sem qualquer metodologia técnica ou científica [Jac00]. A falta de um exame quantitativo entre os produtores de vinhos gerou diversas classificações

dos *chateaux*, conforme as ocorridas em Bordeaux em 1855[1], em 1959 [2] e 1985 [3]. Uma maneira de se obter uma classificação quantitativa, que considere aspectos relativos à produção e cultivo, pode ser realizada considerando-se a teoria das redes complexas, modelando as similaridades dos produtores por redes e utilizando as medidas e métodos descritos no Capítulo 3, conforme discutimos nas próximas seções.

7.1 Introdução histórica sobre a produção de vinhos em Bordeaux

Acredita-se que a produção de vinho tenha se iniciado há mais de 6.000 anos, já que achados arqueológicos indicam o cultivo da uva neste período [Phi01]. Registros históricos deixados pelos egípcios, como pinturas e documentos datados de 3.000 a 1.000 a.C., indicam que o vinho era utilizado em celebrações e rituais [Phi01]. Com o avanço do comércio no Mediterrâneo por volta de 2.500 a.C., os vinhos egípcios começaram a ser exportados pelos fenícios para a África Central e para a Ásia. Na Grécia, a partir de 2.000 a.C. uvas começaram a ser cultivadas ao longo do Mediterrâneo e o comércio de vinhos se tornou uma atividade importante nesta região. Diferentemente do Egito, na Grécia o vinho era apreciado por todas as classes sociais e, deste modo, ele se popularizou e seu comércio se tornou intenso [McG03]. A partir de 1.000 a.C., os gregos expandiram as plantações para a Itália, Sicília, chegando até a península Ibérica e Marselha, na França. Com a fundação de Roma em 753 a.C. e a posterior expansão do Império

[1] A classificação de 1855 foi instituída por Napoleão III, que considerou apenas as regiões *Médoc* e *Barsac e Sauternes* [Phi01].

[2] A classificação de 1959 considera os vinhos da região do Graves [Phi01].

[3] A classificação de 1985 considera os vinhos de *Saint Emilion* que não haviam sido classificados anteriormente.

Romano, os vinhedos se expandiram pela Europa, chegando à Grã-Bretanha, à Germânia e à Gália (França) [McG03].

Na França, particularmente, a produção de vinhos iniciou-se por volta do século III a.C. na região de *Saint Emilion* (Bordeaux), quando os vinhos eram produzidos para os soldados do exército romano e a produção era pouco exportada. Entretanto, a partir do século XII, a popularidade dos vinhos de Bordeaux aumentou devido à exportação de vinhos para a Inglaterra. Deste modo, novos *chateaux* surgiram para compensar a demanda. Em 1725, de forma a classificar os vinhos de acordo com o território de produção, a região de Bordeaux foi delimitada em distritos. Essa classificação foi feita levando-se em conta o conceito de *terroir*, que é definido como um conjunto de vinhedos que compartilham de um mesmo tipo de solo, condições climáticas, espécies de uvas e técnicas de produção, que contribuem na determinação das propriedades do vinho [JD00]. Deste modo, acredita-se que as propriedades da produção e cultivo de vinhos, e conseqüentemente a qualidade, devem estar fortemente associada ao território. A Figura 7.1 representa a distribuição geográfica dos *chateaux* ao redor da cidade de Bordeaux.

7.2 Fatores que influenciam na qualidade

A fim de realizarmos um estudo quantitativo sobre os produtores de vinho na região de Bordeaux, coletamos informações fornecidas por Robert Parker [Par03] e pelo *Wine Journal* [4] sobre 571 *chateaux* pertencentes às regiões *Saint Julien*, *Pauillac*, *Pomerol*, *Saint Stèphe*, *Margaux*, *Saint Emilion*, *Barsac e Sauternes*, e *Pessac-Léognan e Graves*. *Saint Emilion* e *Saint Julien* têm o maior e menor número de *chateaux*, respectivamente. Todas as regiões produzem principal-

[4] Wine Journal: `http://www.wine-journal.com`.

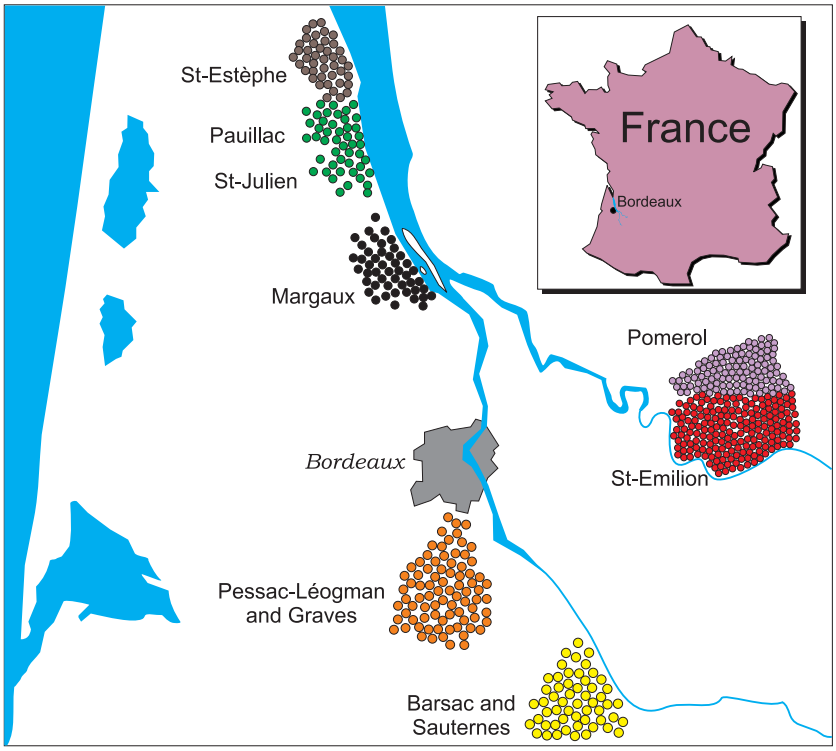


Figura 7.1: Divisão geográfica da produção de vinho ao redor da cidade de Bordeaux. Os 571 *chateaux* são distribuídos em oito regiões de produção.

mente vinhos tintos, exceto *Barsac e Sauternes*, que são especializadas em vinhos brancos. A cada *chateau* é associado um vetor de características que contém informações sobre a produção de vinho tais como:

i. **Castas das uvas**:as casta de uvas cultivadas na região de Bordeaux são *cabernet sauvignon*, *merlot*, *cabernet franc*, *petit verdot*, *semillon* e *sauvignon*. A casta *cabernet sauvignon* é a mais cultivada, seguida pela *merlot*, *semillon* e *sauvignon*, respectivamente.

ii. **Área de cultivo**: está relacionada ao tamanho dos *chateaux* e é medida em hectares quadrados.

iii. **Densidade da plantação**: indica o número de vinhas plantadas por hectare.

iv. **Idade média das vinhas**: as vinhas não são plantadas todos os anos e acredita-se que a sua idade influencia na qualidade dos vinhos [Jac00].

v. **Produção por hectare**: quantidade de hectolitros de sumo de uva produzida por hectare.

vi. **Filtragem**: esta técnica é utilizada para remover pedaços de frutas, possíveis bactérias e leveduras depois da fermentação ou antes do vinho ser engarrafado.

vii. ***Finning***: é uma técnica utilizada para remover partículas suspensas que escurecem o vinho. Um agente é introduzido no vinho, como pó de argila ou claras de ovos, que se liga aos elementos suspensos a fim de torná-los pesados o suficiente para que se precipitem e sejam removidos.

viii. **Quantidade de garrafas produzidas**: indica a produção anual dos *chateaux*.

ix. **Tempo de fermentação**: corresponde ao número de meses que o vinho fermenta em barris de carvalho.

Enquanto as propriedades (i)-(v) são relacionadas ao cultivo, as (vi)-(ix) correspondem à produção. A Tabela 7.2 apresenta os valores do coeficiente de correlação de Pearson obtidos considerando-se pares de atributos. Conforme podemos notar, algumas correlações já eram esperadas, como entre a área de plantação e o número de garrafas produzidas ($r = 0{,}84$) e entre a produção por hectare e a densidade de plantação ($r = 0{,}21$). A análise das correlações sugere ainda que enquanto as uvas da casta *cabernet sauvignon* tendem a ser cultivadas nos maiores *chateaux* ($r = 0{,}42$), as da casta *merlot* nos menores ($r = -0{,}43$).

Tabela 7.1: Correlação entre os atributos. A: área de cultivo, B: idade das vinhas, C: densidade da plantação, D: produção por hectare, E: tempo de fermentação, F: *finning*, G: filtragem, H: quantidade de garrafas produzidas, I: porcentagem de *cabernet sauvignon*, J: porcentagem de *merlot*, K: porcentagem de *cabernet franc*, L: porcentagem de *petit verdot*

	B	C	D	E	F	G	H	I	J	K	L
A	-0,08	0,37	0,04	0,11	0,15	0,06	0,84	0,42	-0,43	-0,23	0,16
B		-0,01	-0,26	0,13	-0,11	-0,15	-0,12	-0,22	0,04	0,17	-0,02
C			0,21	0,10	0,13	-0,10	0,41	0,63	-0,50	-0,30	0,32
D				-0,36	0,27	-0,05	0,31	0,47	0,22	0,07	0,20
E					-0,12	-0,05	-0,05	-0,07	-0,15	-0,05	-0,02
F						0,15	0,17	0,20	-0,17	-0,08	0,12
G							0,02	-0,02	-0,16	-0,17	-0,09
H								0,53	-0,30	-0,18	0,22
I									-0,48	-0,42	0,30
J										0,25	-0,17
K											-0,11

A análise das correlações pode ainda ser usada para determinar quais propriedades mais interferem na qualidade dos vinhos. Robert Parker fornece uma classificação para 153 vinhos produzidos em 2003, sendo definidos como *First Growth* os melhores vinhos e *Fifth Growth*, os piores. Deste modo, associamos nota 5 aos melhores vinhos e nota 1 aos piores. Calculando a correlação entre os atributos descritos anteriormente e tal nível de qualidade, nós obtivemos os resultados apresentando na Tabela 7.2. Conforme observamos, idade média das vinhas ($r = 0{,}287$ e $\rho = 0{,}327$) e o tempo de fermentação ($r = 0{,}278$ e $\rho = 0{,}310$) são os fatores mais correlacionados com a qualidade do vinho. Logo, *chateaux* que possuem vinhas mais antigas e que fermentam os vinhos por mais tempo em barris, tendem a produzir melhores vinhos. Interessante notar também que alguns fatores pioram a qualidade do vinho, como a produção por hectare ($r = -0{,}162$ e $\rho = -0{,}127$) e a filtragem ($r = -0{,}151$ e $\rho = -0{,}057$). Logo, *chateaux* que extraem maior quantidade de sumo por hectare e cujos vinhos são filtrados, produzem vinhos piores. Além disso, é interessante observar que muitas outras características não possuem qualquer correlação com a qualidade do vinho, o que sugere que elas não têm qualquer influência na determinação da qualidade. Além disso, o fato das correlações encontradas serem relativamente baixas sugere que outros fatores que não foram considerados em nossa análise podem influenciar na qualidade do vinho, como os climáticos (temperatura, umidade do ar) ou geológicos (tipo de solo). Deste modo, a consideração de novos atributos, bem como uma investigação levando em conta um maior número de *chateaux* pode ajudar na descoberta de quais fatores determinam a qualidade do vinho, o que terá forte impacto na indústria vinícola mundial. A análise das correlações é altamente robusta e simples, podendo ser largamente utilizada na investigação não só da qualidade dos vinhos, mas de outros produtos agrícolas.

Tabela 7.2: Correlações entre os atributos e a qualidade dos vinhos medidas pelo coeficientes de Pearson (r) e Spearman (ρ).

Attributo	r	ρ
Área de produção	0,118	0,094
Idade média das vinhas	0,287	0,327
Densidade da plantação	0,076	0,079
Produção por hectare	-0,162	-0,127
Tempo de fermentação	0,278	0,310
Finning	-0,079	-0,078
Filtragem	-0,151	-0,057
Produção	0,062	0,040
% de Cabernet Sauvignon	0,013	0,072
% de Merlot	-0,011	-0,024
% de Cabernet Franc	0,040	-0,031
% de Petit Verdot	-0,100	0,006

7.3 A classificação dos vinhos de Bordeaux por redes complexas

A rede complexa dos produtores de vinhos pode ser obtida representado cada *chateau* como um vértice na rede e conectando-os de acordo com as suas similaridades. A Figura 7.2 apresenta uma representação esquemática da construção da rede dos produtores de vinhos. A cada *chateau* i é associado um vetor x_i cujos n elementos representam características de cultivo e produção de vinhos, $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,n})^T$. Cada elemento m do vetor x_i é normalizado da se-

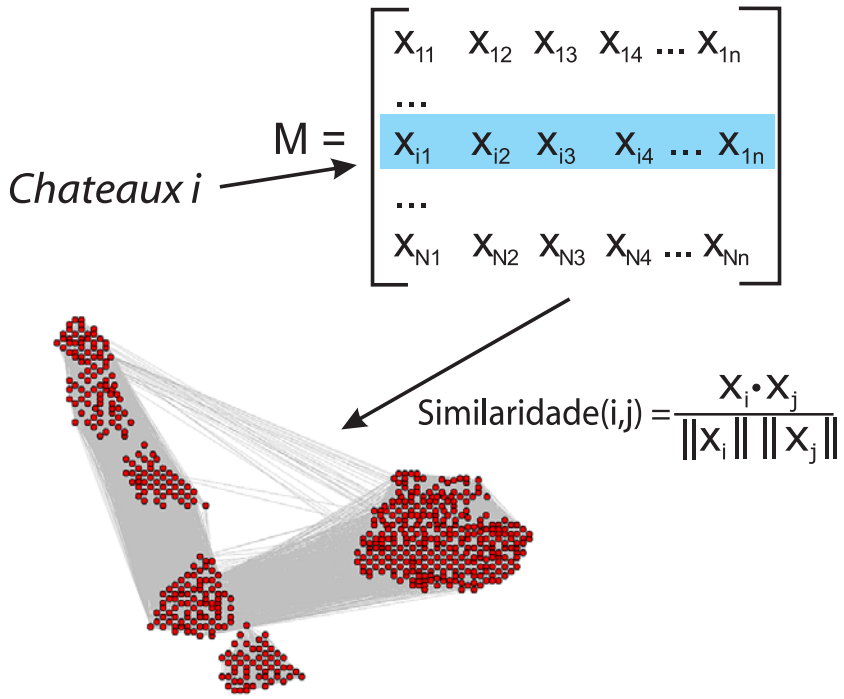


Figura 7.2: Representação esquemática da construção da rede dos produtores de vinho da região de Bordeaux. A cada *chateaux* é associado um vetor de propriedades (cultivo e produção). Cada linha da matriz de atributos representa um *chateaux* e cada coluna uma característica.

guinte forma,

$$y_i(m) = \frac{x_i(m) - \langle x_i \rangle}{\sigma_{x_i}}, \tag{7.1}$$

onde $\langle x_i \rangle$ e σ_{x_i} são a média e o desvio padrão do vetor de propriedades x_i, respectivamente. O vetor resultante dessa normalização, y_i é caracterizado por $\langle y_i \rangle = 0$ e $\sigma_{y_i} = 1$.

Os atributos considerados para compor os vetores são aqueles descritos anteriormente, ou seja, (i) castas das uvas, (ii) área de cultivo, (iii) densidade da plantação, (iv) idade média das vinhas, (v) produção por hectare, (vi) se o vinho é filtrado, (vii) se é utilizado *finning*, (viii) quantidade de garrafas produzidas

anualmente e (ix) tempo de fermentação. As castas das uvas, particularmente, são subdivididas em seis tipos, (i) *cabernet sauvignon*, (ii) *merlot*, (iii) *cabernet franc*, (iv) *petit verdot*, (v) *semillon* e (vi) *sauvignon*. A cada tipo de uva é associado um valor que corresponde à sua porcentagem na composição do vinho. Deste modo, os elementos do vetor que representam tais porcentagens não necessitam ser normalizados pela Equação 7.1.

As conexões na rede dos produtores são determinadas pelas similaridades entre os *chateaux* que pode ser quantificada pelo cosseno do ângulo entre os vetores de atributos. Para dois *chateaux* i e j, a similaridade é calculada da seguinte forma,

$$\text{similaridade}(i,j) = \frac{x_i \cdot x_j}{||x_i||\,||x_j||}, \tag{7.2}$$

onde $x_i \cdot x_j$ representa o produto interno entre os vetores x_i e x_j, e $||x_i||$ o módulo do vetor x_i. Assim, se dois *chateaux* compartilham exatamente os mesmo atributos, a similaridade é máxima (o ângulo entre eles é igual a 0^0 graus). Por outro lado, se dois *chateaux* têm características completamente distintas, a similaridade é mínima (ângulo de 90^0 graus). A Tabela 7.3 apresenta as medidas básicas da rede dos produtores considerando um valor de limiarização $T = 0{,}3$ para calcular as medidas sem peso. Conforme notamos, os valores do grau médio e do coeficiente de aglomeração médio são maiores do que os valores observados na maioria das redes reais (ver Tabela 2.3). Este fato reflete a alta similaridade entre os *chateaux*, o que resulta em uma rede altamente conectada e com estrutura modular, o que sugere que esta rede pode ser modelada pelo modelo *small world* de Watts-Strogatz (ver Seção 2.3). Neste caso, as redes geradas por este modelo considerando uma probabilidade de reconexão $p \approx 0{,}05$, formadas por 571 vértices conectados inicialmente a $\kappa \approx 25$ vizinhos diretos, apresentam grau médio $\langle k \rangle = 50$, coeficiente de aglomeração médio $cc \sim 0{,}63$ e o menor caminho médio $\ell \sim 2{,}3$, que correspondem aproximadamente aos valores apresentados na Tabela 7.3.

Tabela 7.3: Propriedades estatísticas da rede dos produtores de vinhos.

Número de vértices	571
Número de conexões*	14.485
$\langle k \rangle$*	50,7
$\langle s \rangle$	32
ℓ*	2,8
$\langle cc \rangle$*	0,63
$\langle cc^w \rangle$	0,64
Diâmetro	5

*Considerando um valor de limiarização T = 0,3.

Como a rede é construída através das similaridades entre os produtores, os *chateaux* mais semelhantes são ligados por conexões mais fortes, formando comunidades nas redes [GN02]. A comparação entre a divisão topológica e a localização geográfica dos *chateaux* permite determinar o quanto o território influencia nas técnicas de cultivo e produção e, conseqüentemente, na qualidade do vinho. Deste modo, aplicamos o método de detecção de comunidades baseado na maximização da modularidade desenvolvido por Newman [New04c], e posteriormente otimizado por Clauset *et al.* (ver Seção 3.6), e obtivemos a separação topológica dos produtores, conforme apresentado na Figura 7.3. Nesta figura, observamos que há uma forte correspondência entre a localização e as características dos *chateaux*. *Chateaux* que compartilham de técnicas semelhantes de produção e cultivo tendem a estar localizados proximamente. Além disso, vemos que algumas regiões são incluídas dentro de uma mesma comunidade, como ocorre por exemplo com as regiões do *Pomerol* e *Saint Emilion*, além de *Saint Julien*, *Saint-Estèphe*, *Pauillac* e *Margaux*. Esse resultado indica que embora os *chateaux* localizados nestas regiões sejam incluídos em diferentes distritos, eles compartilham das mesmas

propriedades relacionadas ao cultivo e produção dos vinhos. Interessante notar ainda que na região *Pessac-Léognan e Graves*, por estar localizada entre *Saint Emilion* e *Margaux*, parte dos seus *chateaux* se assemelham aos da primeira região e parte aos da segunda. Assim, a influência do território é um fator predominante na produção dos vinhos em Bordeaux.

Este método de separação dos *chateaux* de acordo com similaridades pode ser considerado como sendo um tipo de classificação não supervisionada, pois não utilizamos a classificação dos *chateaux* definida pelas regiões a que eles pertencem. Caso esta categorização fosse considerada, poderíamos utilizar análise das variáveis canônicas, mas a pré-definição da classificação forçaria a divisão entre as classes, pois, conforme discutimos no Seção 4.1, a análise das variáveis canônicas otimiza a separação entre as categorias. Logo, para termos uma classificação que não levasse em conta informações sobre a divisão pré-definida, utilizamos a modelagem por redes e obtivemos as comunidades.

A variação entre as propriedades dos *chateaux* pode ser observada na rede considerando-se o menor caminho entre os *chateaux* mais diferentes. Os *chateaux* que possuem as propriedades mais comuns e as mais específicas dentre todos os da rede podem ser determinados pela distância de Mahalanobis [DHS01], que é diferente da distância Euclidiana, pois leva em conta as correlações dos conjuntos de dados e é invariante por escala, isto é, não depende da escala das medidas. A distância de Mahalanobis para um dado *chateaux* i é calculada da seguinte forma:

$$D_M(i) = \sqrt{(x_i - \mu)^T M_{cov}^{-1}(x - \mu)}, \tag{7.3}$$

onde x_i é o vetor que representa o conjunto de atributos do *chateaux* i, μ é a média dos atributos entre todos os *chateaux* e M_{cov} é a matriz de covariância obtida a partir da matriz de atributos. No caso em que pretendemos determinar os *chateaux* de cada região, μ e M_{cov} são calculadas considerando os apenas os atributos dos *chateaux* pertencentes ao mesmo distrito. Em nossa análise, *Arsac* (*Margaux*) é o

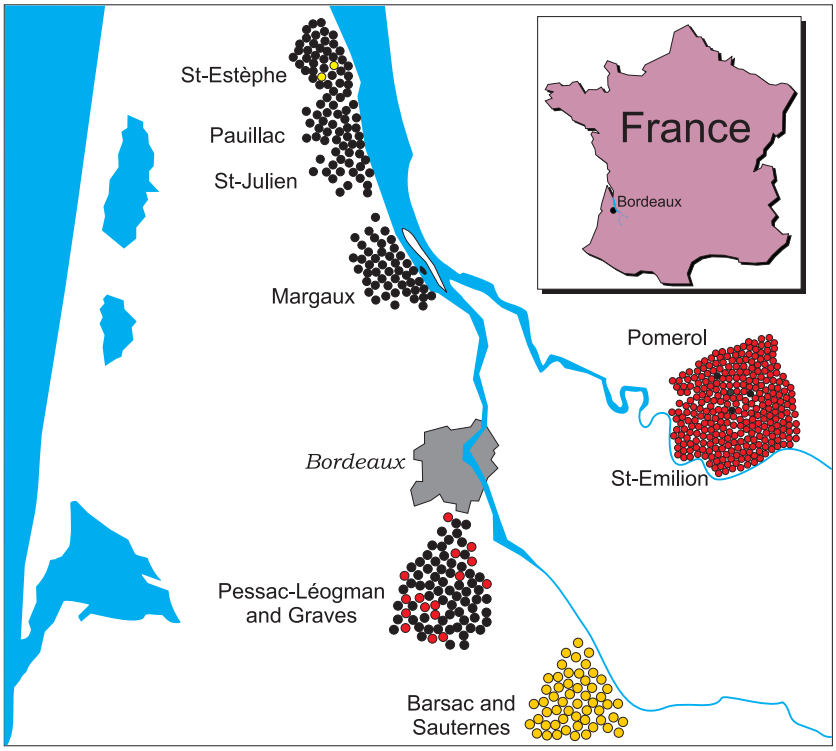


Figura 7.3: As cores dos *chateaux* representam a classificação obtida por meio da aplicação de detecção de comunidades na rede complexa dos produtores de vinho. A região geográfica está fora de escala para melhor visualização.

chateau protótipo, ou seja, o que possui as propriedades mais comuns; e *Laplagnotte Bellevue* (*Saint Emilion*) é o *chateau* mais diferente. A Figura 7.4 mostra a variação das propriedades dos *chateaux* presentes no menor caminho entre *Arsac* e *Laplagnotte Bellevue*. Interessante notar que a variação das propriedades é influenciada pelo território, ou seja, as propriedades de cultivo e produção mudam gradualmente ao longo do território.

A distância de Mahalanobis permite ainda a identificação dos *chateaux* protótipos de cada território. Em nossa análise, identificamos *Hortevie* (*Saint Julien*), *Batailley* (*Pauillac*), *Vray Croix de Gay* (*Pomerol*), *Saint Estèphe* (*Saint Stephe*),

Labégorce Zédé (*Margaux*), *Balestard la Tonnele* (*Saint Emilion*), *De France* (*Pessac-Léogman e Graves*) e *De Rolland* (*Barsac e Sauternes*). As propriedades relacionadas ao cultivo e as técnicas de produção de vinhos destes *chateaux* são as mais típicas em cada uma de suas regiões. Como tais fatores influenciam na qualidade do vinho, os vinhos produzidos pelos *chateaux* protótipos devem ser típicos de cada região.

7.4 Conclusão

Embora há muito tempo, acredita-se que as propriedades dos vinhos eram determinadas por fatores típicos da região que eles eram produzidos, como a casta de uva cultivada, fatores climáticos e geológicos (conceito de *terroir*), muitas pesquisa sugerem que escolhas tecnológicas podem ser mais importantes na determinação das propriedades e qualidade do vinho do que os fatores ambientais [GG01]. A modelagem das semelhanças entre os produtores de vinho da região de Bordeaux permitiu determinar como os atributos relacionados ao cultivo e produção são influenciados pelo território. Verificamos que os *chateaux* pertencentes aos mesmos distritos tendem a utilizar semelhantes técnicas de cultivo e produção. A metodologia empregada, com a ligação dos produtores de acordo com as semelhanças e a determinação das comunidades, se mostrou um método alternativo de classificação não-supervisionada. Deste modo, esta técnica pode ser empregada a qualquer sistema cujos constituintes possuem vetores de atributos associados.

Outro resultado importante foi obtido com a análise das correlações entre a qualidade dos vinhos e os respectivos atributos. Neste caso, observamos que os vinhos que por mais tempo são fermentados em barris de carvalho, ou cujas vinhas são mais antigas, tendem a terem um grau de qualidade melhor (segundo a

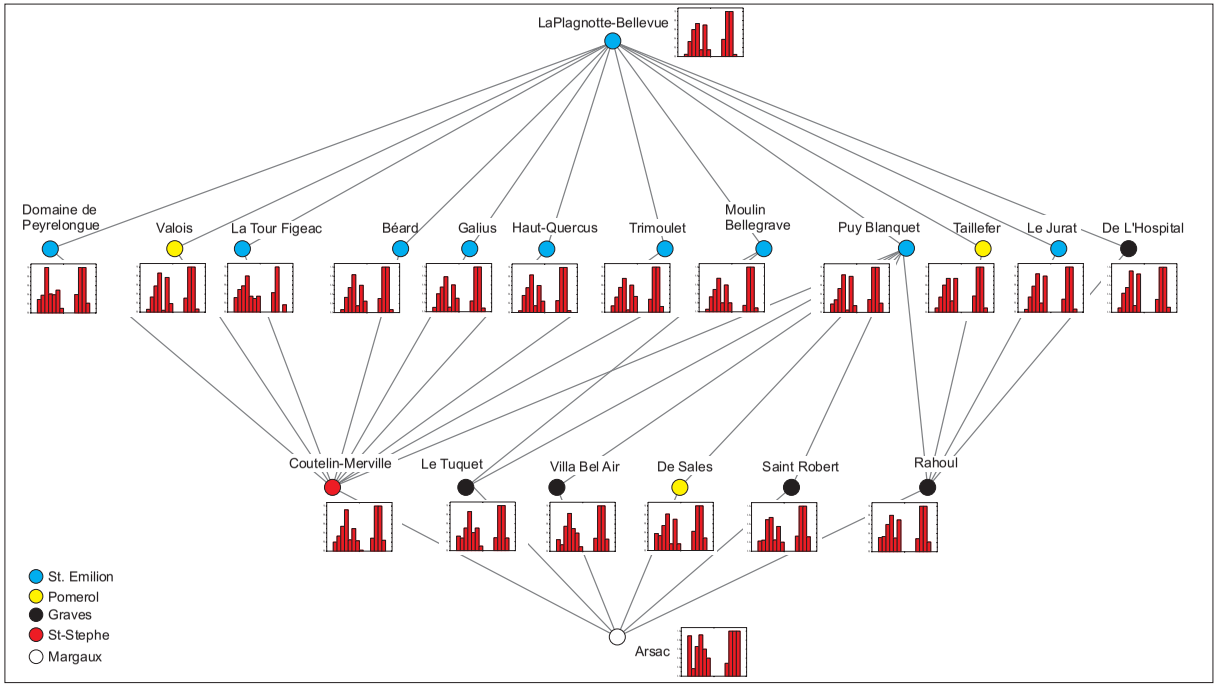


Figura 7.4: O *chateau* protótipo e o mais diferente encontrado na rede são *Arsac* e *Laplagnotte Bellevue*, respectivamente. Os *chateaux* que se situam no menor caminho entre tais *chateaux* possuem propriedades que mudam gradualmente de acordo com a distância entre eles. As diferenças são expressas pelos histrogramas dos atributos dispostos na seguinte seqüência: área de cultivo, idade média das vinhas, densidade da plantação, produção por hectare, tipo de uva (porcentagens de *cabernet sauvignon*, *merlot*, *cabernet franc*, *petit verdot*, *semillon* e *sauvignon*), tempo de fermentação, se o vinho é filtrado, se o *finning* é utilizado e o número de garrafas produzidas.

classificação de Robert Parker). O baixo valor das correlações sugerem que outros atributos que não foram considerados podem ser mais importantes, embora os resultados que obtivemos devam ser levados em conta, já que tais fatores podem estar associados a outros atributos que não foram analisados.

Finalmente, determinamos os *chateaux* protótipo de cada região, que devem produzir o vinho típico de cada distrito. Além disso, analisamos os menores caminhos na rede entre os *chateaux* mais diferentes, ou seja, o com maior e menor distância de Mahalanobis. Neste caso, pudemos verificar como os atributos variam na rede. As investigações descritas neste capítulo podem ser largamente estendidas para outras regiões de produção, sendo possível a consideração de outros atributos. Na verdade, nossa análise pode ser empregada a qualquer sistema formado por elementos que possuem atributos associados, como no caso da classificação de imagens ou mesmo outros produtos agrícolas.

Capítulo 8

Conclusões e trabalhos futuros

A metodologia descrita neste trabalho abre novas perspectivas de pesquisa na área de redes complexas. A utilização de métodos estatísticos de classificação proporciona uma classificação completa dos modelos de redes com a utilização de um conjunto amplo de medidas. Como, geralmente são utilizadas apenas a distribuição das conexões, o grau médio, o coeficiente de aglomeração e o menor caminho médio para caracterizar redes reais, muitos dos modelos existentes são incompletos. A distribuição das conexões não é uma boa medida de classificação, pois, conforme mostrado por Alderson *et al.*, redes com estruturas completamente distintas podem apresentar a mesma distribuição de conexões [ADLW05]. Para a obtenção de uma classificação precisa, é necessário considerar o maior número de medidas possível de forma a quantificar um maior número de propriedades estruturais. Neste caso, é importante a utilização de medidas não redundantes, pois redundâncias podem contribuir para um aumento do erro na classificação [DHS01].

Conforme mostramos, muitos modelos que geram redes livre de escala falham na representação precisa da estrutura de redes reais. Apesar da distribuição das conexões seguir a lei de potência na maioria das redes reais, a estrutura destas redes é diferente daquelas geradas por modelos. Por exemplo, embora a rede de

interação de proteínas pudesse ser caracterizada pelo modelo de Barabàsi e Albert, por apresentar distribuição de conexões do tipo livre de escala, ela foi classificada como geográfica quando consideramos um conjunto amplo de medidas. Tal fato ocorreu porque as redes de interações de proteínas apresentam certos atributos que são típicos em redes geográficas, como o alto coeficiente de aglomeração. Desta forma, a distribuição das conexões pode não ser o fator principal que caracteriza a estrutura das redes complexas. Portanto, a utilização de métodos estatísticos de classificação poderá exigir uma revisão da maioria dos modelos de redes.

A metodologia de classificação, que utiliza análise das variáveis canônicas e decisão Bayesiana, foi aplicada na modelagem da Internet. Modelos que reproduzam as principais características da Internet são importantes para o desenvolvimento de protocolos de roteamento e no planejamento do tráfego. Através do modelos que desenvolvemos, mostramos que a evolução da Internet é guiada principalmente pela ligação preferencial, pela distância geográfica entre os roteadores e pela adição constante de conexões entre roteadores antigos. O modelo que apresentamos se mostrou mais preciso do que todos os modelos de rede considerados.

As medidas de redes complexas foram utilizadas na caracterização e análise da rede de interação de proteínas e domínios protéicos do *S. cerevisiae*. Mostramos que a distribuição das conexões, tanto na rede de interação de proteínas como na de domínios, segue uma lei de potência com um limite (*cutoff*) exponencial. A fim de compararmos a relação conectividade-letalidade nestas duas redes, sugerimos duas definições de essencialidade para os domínios protéicos em um sentido fraco (quando domínios estão presentes em qualquer proteína letal) e forte (quando domínios estão presentes apenas em proteínas letais formadas por um único domínio). A partir destas definições, mostramos que a correlação entre conectividade e letalidade é mais definida para as redes dos domínios do que das proteínas. Tal conclusão mostra a importância dos domínios na definição das

funções das proteínas.

Também caracterizamos e analisamos a rede dos produtores de vinho da região de Bordeaux. Quando modelamos as semelhanças entre os produtores de vinho por meio de redes complexas e aplicamos um método de identificação de comunidades, observamos que há uma alta correspondência entre a separação obtida e a localização geográficas dos *chateaux*. De fato, verificamos que os produtores mais próximos geograficamente tendem a compartilhar técnicas semelhantes de cultivo e produção de vinho. Esta conclusão comprova a antiga crença de que as propriedades do vinho são fortemente definidas pelo território, o que está relacionado ao conceito de *terroir*. Além disso, através da análise das correlações entre as propriedades de produção/cultivo e a qualidade dos vinhos, observamos que os *chateaux* que possuem vinhas mais antigas e que fermentam os vinhos por mais tempo em barris, tendem a produzir vinhos de melhor qualidade. Entretanto, os valores baixos para as correlações sugerem que outras propriedades que não foram consideradas em nossa análise podem influenciar na qualidade do vinho, como os fatores climáticos (temperatura, umidade do ar) ou geológicos (tipo de solo).

Todos os resultados obtidos mostram a utilidade da teoria das redes complexas na análise, caracterização e classificação de sistemas complexos. Novas perspectivas de pesquisa são possíveis com a extensão do trabalho aqui apresentado. Entre elas, podemos sugerir (i) a utilização de técnicas de mineração de dados para analisar a precisão de modelos ou mesmo construir uma taxonomia das redes complexas, onde redes semelhantes poderão ser agrupadas em uma mesma classe; (ii) análise da estabilidade das medidas na classificação; (iii) o aperfeiçoamento do modelos de Internet, comparando com outros modelos existentes e utilizando técnicas de otimização, de forma a estudar a influência dos parâmetros do modelo ou mesmo reduzir seu número; (iv) analise das proteínas considerando as

funções específicas de cada uma, como metabolismo e síntese de ATP; e, por fim, (v) a continuação da análise dos produtores de vinho considerando outros atributos ou mesmo outras regiões de produção. Nesse caso, pode-se considerar regressão linear multivariada de forma a determinar a qualidade de vinhos desconhecidos [dFC07a, dFC07b] pela comparação de atributos e níveis de qualidade de um conjunto de vinhos pré-definidos com os de um vinho que se deseja analisar.

Todas estas investigações podem ser desenvolvidas considerando novas bases de dados e os métodos estatísticos e medidas de redes aqui descritos.

Glossário

Aresta: conexão que liga dois vértices na rede.

Assortatividade: refere-se à preferência que os nós da rede têm de se ligarem com outros que são similares ou diferentes de alguma maneira. Esta semelhança, no caso das redes complexas, pode ser dada pela conectividade dos vértices. Deste modo, uma rede é dita assortativa se os vértices com conectividade similares tendem a se ligarem.

Caminho: A seqüência de vértices $(i, k_1), (k_1, k_2), \ldots, (k_{m-1}, j)$ que liga dois vértices i e j através de m arestas, é chamada caminho entre i e j.

Clique: conjunto de vértices totalmente conectados.

Componente conectado: conjunto de vértices que possui um caminho entre quaisquer dois nós.

Comunidade: conjunto de vértices altamente conectados entre si, mas pouco conectados com o restante da rede.

Conectividade: número de conexões de um vértice.

Diâmetro: comprimento do maior caminho na rede.

Distância: igual ao comprimento do menor que liga dois vértices i e j.

Distância geodésica: menor caminho entre dois nós. O caminho geodésico não é necessariamente único.

Distribuição lei de potência: é qualquer distribuição que tem a forma geral: $p(x) = Cx^{-\alpha}$, onde a constante α é chamada expoente da lei de potência e C é ajustada de forma que a função $p(x)$ tenha soma igual a um.

Efeito mundo pequeno (*Small world*): está presente nas redes cuja distância geodésica média entre os pares de vértices escala logaritmicamente (ou mais lentamente) com o tamanho da rede, para um grau médio fixo.

Grafo: é um objeto matemático constituído por um conjunto de vértices (nós) que são ligados por arestas (conexões, ligações ou *links*).

***Hub*:** vértice altamente conectado.

Ligação preferencial: propriedade das redes (particularmente as livre de escala) onde a probabilidade dos vértices adquirem novas conexões é proporcional a sua conectividade.

Matriz de adjacência: matriz que armazena as ligações da rede. As entradas a_{ij} na matriz serão igual a um se existir uma conexão entre i e j ou iguais a zero, caso contrário.

Modularidade: propriedade de algo ter uma organização modular. Os módulos são formados por componentes que possuem alguma relação de similaridade. A modularidade é uma medida utilizada para determinar a qualidade de uma particular divisão de uma rede em comunidades.

Motivos: subgrafos que aparecem mais freqüentemente em redes reais do que nas redes aleatórias equivalentes.

Rede complexa: rede que possui certas propriedades topológicas não-triviais que não ocorrem em redes simples (grades ou grafos aleatórios). Tais propriedades não-triviais incluem distribuição livre de escala, alto coeficiente de aglomeração médio, assortatividade ou disassortatividade entre os vértices, estrutura de comunidades em muitas escalas, estrutura fractal e hierárquica.

Redes com peso: rede cujas arestas possuem um valor associado.

Rede dirigida: redes cujas arestas possuem direção. Neste caso, o sentido da ligação tem importância fundamental.

Vértice: unidade fundamental que constitui as redes. Também denominado nó.

Vulnerabilidade: propriedade das redes serem vulneráveis a ataques dirigidos. No caso das redes livre de escala, o ataque resulta na divisão da rede em componentes desconectados.

Lista de símbolos e abreviações

- $R = (\mathcal{N}, \mathcal{E})$: rede formada por um conjunto de N vértices, $\mathcal{N} = \{n_1, n_2, \ldots, n_N\}$, e um conjunto de M arestas, $\mathcal{E} = \{e_1, e_2, \ldots, e_M\}$.
- A: matriz de adjacência.
- a_{ij}: elemento da matriz de adjacência que é igual a um se existe uma conexão entre i e j e igual zero, caso contrário.
- W: matriz de pesos.
- w_{ij}: elemento da matriz de pesos que representa o peso da ligação entre os vértices i e j.
- k_i: conectividade do vértice i.
- s_i: força do vértice i.
- $P(k)$: distribuição de probabilidade das conexões.
- $\langle k \rangle$: conectividade média.
- $\langle s \rangle$: força média.
- p: probabilidade de ligação usada no modelo de Erdős e Rényi e de reconexão de arestas no modelo de Watts e Strogatz.

- ℓ: menor caminho médio.
- κ: número de vizinhos iniciais considerados no modelo de Watts e Strogatz.
- m: número de arestas adicionadas a cada passo no modelo de Barabási e Albert.
- α: expoente da ligação preferencial não linear no modelo de Krapivisky *et al.*
- γ: expoente da lei de potência.
- ER: modelo de Erdős e Rényi.
- WS: modelo de Watts e Strogatz.
- BA: modelo de Barabási e Albert.
- GN: modelo geográfico de Waxman.

Lista de modelos

Grafos aleatórios de Erdős e Rényi: iniciando-se com um conjunto de N vértices totalmente desconectados, a cada passo dois vértices são escolhidos aleatoriamente e conectados com uma probabilidade fixa p, sendo cada par de vértices considerado apenas uma vez.

Modelo *small world* de Watts e Strogatz: inicia-se com uma rede regular formada de N vértices ligados a κ vizinhos mais próximos em cada direção. A seguir, cada aresta é aleatoriamente reconectada com uma probabilidade fixa p.

Modelos de configuração: modelo aleatório construído a partir de uma distribuição de conexões pré-definida $P(k)$.

Modelo livre de escala de Barabási e Albert: baseado em dois princípios básicos: (i) crescimento: a cada passo um novo vértice i é adicionado com m arestas; e (ii) ligação preferencial: a probabilidade de um vértice j, na rede, ser escolhido é proporcional a sua conectividade,

$$\mathcal{P}_{i\to j}(n+1) = \frac{k_j(n)}{\sum_{u=-N_0+1}^{n} k_u(n)}.$$

Este modelo gera redes com distribuição livre de escala.

Modelo livre de escala de Dorogovtev *et al.*: baseado nas mesma propriedades do modelo de Barabási e Albert, mas com ligação preferencial da

forma

$$\mathcal{P}_{i\to j}(n+1) = \frac{k_j(n)+k_0}{\sum_{u=-m_0+1}^{n}(k_u(n)+k_0)},$$

com $-m < k_0 < \infty$.

Modelo livre de escala limitado de Amaral *et al.*: baseado no modelo de Barabási e Albert, mas o mecanismo de ligação preferencial pode ser limitado por três fatores básicos: (i) idade dos vértices, (ii) custo da adição de novas arestas e capacidade limitada e (iii) filtragem de informações. Deste modo, a conectividade dos vértices é limitada.

Modelo livre de escala não-linear: os vértices que são adicionados à rede são conectados com uma probabilidade de conexão não-linear da forma,

$$\mathcal{P}_{i\to j}(n+1) = \frac{k_j(n)^\alpha}{\sum_{u=-m_0+1}^{n} k_u(n)^\alpha}.$$

Quando $\alpha = 1$, temos o modelo de Barabási e Albert.

Tabela de medidas

Medida	**Símbolo**	**Equação**
Conectividade	k_i	2.1
Conectividade média	$\langle k \rangle$	2.2
Força	s_i	3.1
Força média	$\langle s \rangle$	3.2
Medida de retidão da distribuição	st	3.3
Assortatividade	r	3.3
Coeficiente de aglomeração	cc_i	3.4
Coeficiente de aglomeração para redes com peso	cc_i^w	3.5
Coeficiente de aglomeração médio	$\langle cc \rangle$	3.6
Coeficiente de aglomeração médio para redes com peso	$\langle cc^w \rangle$	3.7
Número de ciclos de ordem três	N_3	3.9
Número de ciclos de ordem quatro	N_4	3.10
Número de ciclos de ordem cinco	N_5	3.11
Coeficiente cíclico	Θ	3.13
Coeficiente *rich-club*	$\phi(k)$	3.14
Menor caminho médio	ℓ	3.16
Eficiência global	E	3.17
Média harmônica do menor caminho médio	h	3.18
Vulnerabilidade	V	3.20
Grau de intermediação	B_u	3.21
Grau de intermediação médio	$\langle B \rangle$	3.22
Dominância do ponto central	c_D	3.23
Coeficiente de aglomeração hierárquico	cc_{rs}	3.26
Razão de convergência	cv_d	3.27
Razão de divergência	dv_d	3.28
Modularidade	Q	3.29
Coeficiente de assortatividade	η	3.34

Referências Bibliográficas

[AA04] I. Albert and R. Albert. Conserved network motifs allow protein–protein interaction prediction. *Bioinformatics,* 20(18):3346–3352, 2004.

[AB02] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:48–98, 2002.

[ACL01] W. Aiello, F. R. K. Chung, and L. Lu. Random evolution in massive graphs. *IEEE Symposium on Foundations of Computer Science*, pages 510–519, 2001.

[ADLW05] D. Alderson, J. C. Doyle, L. Li, and W. Willinger. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Mathematics*, 2(4):431–523, 2005.

[AH00] L. A. Adamic and B. A. Huberman. Power-law distribution of the World Wide Web. *Science*, (287), 2000.

[AHDV05] J. I. Alvarez-Hamelin, L. Dall'Asta, and A. Vespignani. k-core decomposition: a tool for the visualization of large scale networks. cs.NI/0504107, 2005.

[AJB99] R. Albert, H. Jeong, and A.-L. Barabási. Diameter of the World Wide Web. *Nature*, 401:130–131, 1999.

[AJB00] R. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406:378–382, 2000.

[Alb05] R. Albert. Scale-free networks in cell biology. *Journal of Cell Science*, 118(Pt 21):4947–57, 2005.

[And58] T.W. Anderson. *An introduction to multivariate statistical analysis*. Wiley New York, 1958.

[ANJdFC07] L. Antiqueira, M. G. V. Nunes, O. N. Oliveira Jr, and L. da F. Costa. Strong correlations between text quality and complex networks features. *Physica A*, 373:811–820, 2007.

[AO04] L. A. N. Amaral and J. M. Ottino. Complex networks. *The European Physical Journal B*, 38:147–162, 2004.

[ASBS00] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences of the United States of America*, 97(21):11149–11152, 2000.

[AUANK+03] M. Altaf-Ul-Amin, K. Nishikata, T. Koma, T. Miyasato, Y. Shinbo, M. Arifuzzaman, C. Wada, M. Maeda, T. Oshima, H. Mori, and S. Kanaya. Prediction of protein functions based on k-cores of protein-protein interaction networks and amino acid sequences. *Genome Informatics*, 14:498–499, 2003.

[BA99] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–12, 1999.

[BAJ00] A.-L. Barabási, R. Albert, and H. Jeong. Scale-free characteristics of random networks: The topology of the World Wide Web. *Physica A*, 281:69–77, 2000.

[Bar03] A.-L. Barabási. *Linked: How Everything is Connected to Everything Else and what it Means for Business, Science, and Everyday Life*. Plume, 2003.

[BB01] G. Bianconi and A.-L. Barabási. Bose-Einstein condensation in complex networks. *Physical Review Letters*, 86:5632–5635, 2001.

[BB05] J. P. Bagrow and E. M. Bollt. A local method for detecting communities. *Physical Review E*, 72(046108), 2005.

[BBO05a] G. Balazsi, A.-L. Barabási, and Z. N. Oltvai. Functional organization of transcriptional-regulatory networks. *FEBS Journal*, 272:103–103, 2005.

[BBO05b] G. Balazsi, A.-L. Barabási, and Z. N. Oltvai. Topological units of environmental signal processing in the transcriptional regulatory network of *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, 102(22):7841–7846, 2005.

[BBPSV04] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752, 2004.

[BC78] E. A. Bender and E. R. Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory*, Ser. A 24:296–307, 1978.

[BCC05] G. Bianconi, G. Caldarelli, and A. Capocci. Loops structure of the Internet at the autonomous system level. *Physical Review E*, 71(6):066116, 2005.

[BCD+04] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. R. Eddy. The Pfam protein families database. *Nucleic Acids Research*, (32):D138–D141, 2004.

[BFNW04] J. Balthrop, S. Forrest, M. E. J. Newman, and M. M. Williamson. Technological networks and the spread of computer viruses. *Science*, 304(5670):527–529, 2004.

[BGWB05] S. Bar, M. Gonen, A. Wool, and S. Bar. A geographic directed preferential Internet topology model. *Modeling, Analysis, and Simulation of Computer and Telecommunication Systems. 13th IEEE International Symposium on*, pages 325–328, 2005.

[BIH04] D. Betel, R. Isserlin, and C. W. V. Hogue. Analysis of domain correlations in yeast protein complexes. *Bioinformatics*, 20 Suppl 1:I55–I62, 2004.

[BJR+02] A.-L. Barabási, H. Jeong, R. Ravasz, Z. Néda, T. Vicsek, and A. Schubert. On the topology of the scientific collaboration networks. *Physica A*, 311:590–614, 2002.

[BKM+00] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the Web. *Computer Networks*, 33(1–6):309–320, 2000.

[BL04] J. Berg and M. Lässig. Local graph alignment and motif search in biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2004. In press.

[BLM+06] S. Boccaletti, V. Latora, Y. Moreno, M. Chaves, and D.-U. Hwang. Complex networks: structure and dynamics. *Physics Reports*, 424:175–308, 2006.

[BLW04] J. Berg, M. Lassig, and A. Wagner. Structure and evolution of protein interaction networks: A statistical model for link dynamics and gene duplications. *BMC Evolutionary Biology*, 4:51, 2004.

[BM06a] V. Batagelj and A. Mrvar. Pajek datasets, 2006. `http://vlado.fmf.uni-lj.si/pub/networks/data`.

[BM06b] G. Bianconi and M. Marsili. Effect of degree correlations on the loop structure of scale-free networks. *Physical Review E*, 73(6):066127, 2006.

[BnPS02] M. Boguñá and R. Pastor-Satorras. Epidemic spreading in correlated complex networks. *Physical Review E*, 66:047104, 2002.

[BO04] A.-L. Barabási and Z. N. Oltvai. Network biology: understanding the cells functional organization. *Nature*, 5:101–113, 2004.

[Bol98] B. Bollobás. *Modern Graph Theory*. Graduate Texts in Mathematics, Springer-Verlag, New York, 1998.

[Bor91] P. Bork. Shuffled domains in extracellular proteins. *FEBS Letters*, 286(1–2):47–54, 1991.

[BY92] Y. Bar-Yam. *Dynamics of Complex Systems*. Perseus Books, 1992.

[Cal07] G. Caldarelli. *Scale-Free Networks: Complex Webs in Nature and Technology*. Oxford University Press, 2007.

[CCGJ02] Q. Chen, H. Chang, R. Govindan, and S. Jamin. The origin of power laws in Internet topologies revisited. *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies.*, 2, 2002.

[CEbAH01] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin. Breakdown of the Internet under intentional attack. *Physical Review Letters*, 86(16):3682–3685, 2001.

[CFSV06] V. Colizza, A. Flammini, M. A. Serrano, and A. Vespignani. Detecting rich-club ordering in complex networks. *Nature Physics*, 2:110–115, 2006.

[CHK+06] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir. New model of Internet topology using k-shell decomposition. cs.NI/0607080, 2006.

[Cho92] C. Chothia. One thousand families for the molecular biologist. *Nature*, 357(6379):543–544, 1992.

[CJS01] R. F. Cancho, C. Janssen, and R. V. Solé. Topology of technology graphs: Small world patterns in electronic circuits. *Physical Review E*, 64(4):46119, 2001.

[Cla05] A. Clauset. Finding local community structure in networks. *Physical Review E*, 72:026132, 2005.

[CNM04] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(066111), 2004.

[CS01] R. Ferrer Cancho and R. V. Solé. The small world of human language. *Proceedings of Royal Society of London B*, 268:2261–2266, 2001.

[DA05] J. Duch and A. Arenas. Community detection in complex networks using extremal optimization. *Physical Review E*, 72:027104, 2005.

[DBBO04] R. Dobrin, Q. K. Beg, A.-L. Barabási, and Z. N. Oltvai. Aggregation of topological motifs in *Escherichia coli transcriptional* regulatory networks. *BMC Bioinformatics*, 5(10), 2004.

[DDADG05] L. Danon, J. Duch, A. Arenas, and A. Díaz-Guilera. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, page P09008, 2005.

[DdFC05] L. Diambra and L. da F. Costa. Complex networks approach to gene expression driven phenotype imaging. *Bioinformatics*, 21(20):3846–3851, 2005.

[DDGA06] L. Danon, A. Díaz-Guilera, and A. Arenas. Effect of size heterogeneity on community identification in complex networks. *Journal of Statistical Mechanics*, page P11010, 2006.

[dFC04a] L. da F. Costa. The hierarchical backbone of complex networks. *Physical Review Letters*, 93(098702), 2004.

[dFC04b] L. da F. Costa. What's in a name? *International Journal of Modern Physics C*, 15(1):371–379, 2004.

[dFC07a] L. da F. Costa. On the separability of attractors in grandmother dynamic systems with structured connectivity. arXiv:physics/0701089v1, 2007.

[dFC07b] L. da F. Costa. Seeking for simplicity in complex networks. arXiv:physics/0702102v1, 2007.

[dFCA07] L. da F. Costa and R. F. S. Andrade. What are the best hierarchical descriptors for complex networks? arXiv:0705.4251v1, 2007.

[dFCdR06] L. da F. Costa and L. E. C. da Rocha. A generalized approach to complex networks. *The European Physical Journal B*, 50(1):237–242, 2006.

[dFCJ01] L. da F. Costa and R. M. Cesar Jr. *Shape Analysis and Classification: Theory and Practice*. CRC Press, 2001.

[dFCKH07] L. da F. Costa, M. Kaiser, and C. C. Hilgetag. Predicting the connectivity of primate cortical networks from topological and spatial node properties. *BMC Systems Biology*, 1:16, 2007.

[dFCRT06] L. da F. Costa, F. A. Rodrigues, and G. Travieso. Protein domain connectivity and essentiality. *Applied Physics Letters*, 89:174101, 2006.

[dFCRTB07] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167 – 242, 2007.

[dFCS05] L. da F. Costa and O. Sporns. Hierarchical features of large-scale cortical connectivity. *The European Physical Journal B*, 48(4):567–573, 2005.

[dFCS06a] L. da F. Costa and F. N. Silva. Hierarchical characterization of complex networks. *Journal of Statistical Physics*, 125(4):841–872, 2006.

[dFCS06b] L. da F. Costa and O. Sporns. Correlating thalamocortical connectivity and activity. *Applied Physics Letters*, 89:013903, 2006.

[dFCT07] L. da F. Costa and G. Travieso. Exploring complex networks through random walks. *Physical Review E*, 75(1):016102, 2007.

[DHS01] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, Inc., 2001.

[Die00] R. Diestel. *Graph Theory*. Springer, 2000.

[DM03] S. N. Dorogovtsev and J. F. F. Mendes. *Evolution of Networks - From Biological Nets to the Internet and WWW*. Oxford University Press, 2003.

[DMS00] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin. Structure of growing networks with preferential linking. *Physical Review Letters*, 85:4633–4636, 2000.

[Dor01] S. N. Dorogovtsev. Language as an evolving word web. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, 268(1485):2603–2606, 2001.

[dSP65] D. J. de S. Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

[dSP76] D. J. de S. Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5–6):292–306, 1976.

[DVGM06] S. N. Dorogovtsev, A. V., Goltsev, and J. F. F. Mendes. k-core organization of complex networks. *Physical Review Letters*, 96(4):40601, 2006.

[Edw76] A. L. Edwards. *An introduction to linear regression and correlation*. W. H. Freeman, 1976.

[EIK99] A. J. Enright, I. Iliopoulos, and N. C. Kyrpides. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402:86–90, 1999.

[EMB02] H. Ebel, L. I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Physical Review E*, 66(3):35103, 2002.

[EMXY00] D. Eisenberg, E.M. Marcotte, I. Xenarios, and T.O. Yeates. Protein function in the post-genomic era. *Nature*, 405:823–826, 2000.

[ER59] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[ER60] P. Erdős and A. Rényi. b. *Publ. Math. Inst. Hungar. Acad. Sci*, 5:17–61, 1960.

[ER61] P. Erdős and A. Rényi. On the strenght of connectedness of a random graph. *Acta Mathematica Scientia Hungary*, 12:261–267, 1961.

[FFF99] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the Internet topology. *Computer Communication Review*, 29(4):251–262, 1999.

[Fis21] R. A. Fisher. On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, 1(4), 1921.

[FLGC02] G. W. Flake, S. R. Lawrence, C. L. Giles, and F. M. Coetzee. Self-organization and identification of web communities. *IEEE Computer*, 35, 2002.

[Fre77] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40:35–41, 1977.

[Fre79] L. C. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1:215–239, 1979.

[FS89] S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246, 1989.

[FW93] R. J. Freund and W. J. Wilson. *Statistical Methods*. Academic Press, 1993.

[GA04] R. Guimerà and L. A. N. Amaral. Modeling the world-wide airport network. *The European Physical Journal B*, 38:381–385, 2004.

[GA05] R. Guimerà and L. A. N. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433:895–900, 2005.

[GAM89] S. Gupta, R. M. Anderson, and R. M. May. Network of sexual contacts: Implications for the pattern of spread of HIV. *AIDS*, 03:807–817, 1989.

[GBB+03] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y.L. Hao, C. E. Ooi, B. Godwin, E. Vitols, et al. A protein interaction map of *Drosophila melanogaster*. *Science*, 302(5651):1727–1736, 2003.

[GD03] P. M. Gleiser and L. Danon. Community structure in jazz. *Advances in complex systems*, 6(4), 2003.

[GDM06] A.V. Goltsev, S. N. Dorogovtsev, and J. F. F. Mendes. k-core (bootstrap) percolation on complex networks: Critical phenomena and nonlocal effects. *Physical Review E*, 73:056101, 2006.

[GG01] O. Gergaud and V. Ginsburgh. Natural endowments, production technologies and the quality of wines in bordeaux. is it possible to produce wine on paved roads?, 2001. Working Paper, presented at the 9th Annual Meeting of the Vineyard Data Quantification Society, May.

[GH02] R. A. George and J. Heringa. An analysis of protein domain linkers: their classification and role in protein folding. *Protein Engineering*, 15(11):871–879, 2002.

[GKS04] V. Gol'dshtein, G. A. Koganov, and G. I. Surdutovich. Vulnerability and hierarchy of complex networks, 2004. cond-mat/0409298.

[GMT+05] R. Guimerà, S. Mossa, A. Turtschi, Amaral, and L. A. N. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(22):7794–7799, 2005.

[GN02] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826, 2002.

[GN06] M. T. Gastner and M. E. J. Newman. The spatial structure of networks. *The European Physical Journal B*, 49(2):247–252, 2006.

[Gra73] M. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[Gra85] M. Granovetter. Economic action and social structure: The problem of embeddedness. *American Journal of Sociology*, 91(3):481–510, 1985.

[Gra95] M. Granovetter. *Getting a Job: A Study of Contacts and Careers*. Chicago University Press, 1995.

[HHLM99] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402(6761):C47–C52, 1999.

[HKKS04] J. Hopcroft, O. Khan, B. Kulis, and B. Selman. Tracking evolving communities in large linked networks. *Proceedings of the*

National Academy of Sciences of the United States of America, 101:5249–5253, 2004.

[HNTV90] H. Heijmans, P. Nacken, A. Toet, and L. Vincent. Graph morphology. *Journal of Visual Communication and Image Representation*, 3(1):24–38, 1990.

[HZ06] X. He and J. Zhang. Why do hubs tend to be essential in protein networks? *PLOS Genetics*, 2(6):826–834, 2006.

[Jac00] R. S. Jackson. *Wine Science: Principles, Practice, Perception*. Academic Press, 2000.

[JC85] J. Janin and C. Chothia. Domains in proteins: definitions, location, and structural principles. *Methods in Enzymology*, 115:420–30, 1985.

[JCJ00] C. Jin, Q. Chen, and S. Jamin. Inet: Internet topology generator. *University of Michigan Technical Report CSE-TR-433-00*, 2000.

[JD00] G. V. Jones and R. E. Davis. Climate influences on grapevine phenology, grape composition, and wine production and quality for Bordeaux, France. *American Journal of Enology and Viticulture*, 51(3):249–261, 2000.

[JMBO01] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

[JTA+00] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–4, 2000.

[KFM+03] A. Krause, K. A. Frank, D. M. Mason, R. E. Ulanowicz, and W. M. Taylor. Compartments exposed in food-web structure. *Nature*, 426:282–285, 2003.

[Kit02] H. Kitano. Systems biology: A brief overview. *Science*, 295, 2002.

[Kit04] H. Kitano. Biological robustness. *Nature Reviews Genetics*, 5(11):826–837, 2004.

[KK05] H. J. Kim and J. M. Kim. Cyclic topology in complex network. *Physical Review E*, 72:036109, 2005.

[KLP+05] D. Koschützki, K. A. Lehmann, L. Peeters, S. Richter, D. Tenfelde-Podehl, and O. Zlotowski. Centrality indices. *Lecture Notes in Computer Science*, 3418, 2005.

[KRL00] P. L. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random netwoks. *Physical Review Letters*, 85(4629), 2000.

[Kro] Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*, author= Krogan, N. J. and Cagney, G. and Yu, H. and Zhong, G. and Guo, X. and Ignatchenko, A. and Li, J. and Pu, S. and Datta, N. and Tikuisis, A. P. and others, journal= Nature, volume= 440, pages= 637–643, year= 2006.

[LD98] E. L. Lehmann and H. J. M. D'Abrera. *Nonparametrics: Statistical Methods Based on Ranks*. Prentice-Hall, 1998.

[LEA+01] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Aaberg. The web of human sexual contacts. *Nature*, 411(6840):907–908, 2001.

[LEA03] F. Liljeros, C.R. Edling, and L. A. N. Amaral. Sexual networks: implications for the transmission of sexually transmitted infections. *Microbes and Infection*, 5(2):189–196, 2003.

[LM01] V. Latora and M. Marchiori. Efficient behavior of small-world networks. *Physics Review Letters*, 87(19):198701, 2001.

[M+99] E.M. Marcotte et al. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753, 1999.

[MBESA02] S. Mossa, M. Barthélémy, H. Eugene Stanley, and L. A. N. Amaral. Truncation of power law behavior in ”scale-free´´ network models due to information filtering. *Physical Review Letters*, 88(13):138701, 2002.

[McG03] P. E. McGovern. *Ancient Wine: The Search for the Origins of Viniculture*. Princeton University Press, 2003.

[MdR99] L.R. Monteiro and S.F. dos Reis. *Princípios de morfometria geométrica*. Holos, 1999.

[MFG+02] H. W. Mewes, D. Frishman, U. Guldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgesnter, M. Munsterkotter, S. Rudd, and B. Weil. Mips: A database for genomes and protein sequences. *Nucleic Acids Res.*, 30:31–34, 2002.

[MH00] M. L. Mayer and P. Hieter. Protein networks built by association. *Nature Biotechnology*, 18:1242–1243, 2000.

[MHH98] K. S. McCann, A. Hastings, and G. R. Huxel. Weak trophic interactions and the balance of nature. *Nature*, 395:794–798, 1998.

[MIK+04] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, 2004.

[Mil67] S. Milgran. The small world problem. *Psychology Today*, 1(1):60–67, 1967.

[MKI+03] R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Newman, and U. Alon. On the uniform generation of random graphs with prescribed degree sequences, 2003. cond-mat/0312028.

[Mot04] A. E. Motter. Cascade control and defense in complex networks. *Physical Review Letters*, 93(9):98701, 2004.

[MR95] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, 6:161–179, 1995.

[MS96] G. MacBeath and S. L. Schreiber. Printing proteins as microarrays for high-throughput function determination. *Nature*, 382:607, 1996.

[MS02] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296:910–913, 2002.

[MSOI+02] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298:824–827, 2002.

[NBW06] M. E. J. Newman, A.-L. Barábasi, and D. J. Watts, editors. *The Structure and Dynamics of Networks*. Princeton University Press, 2006.

[New01] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2):404–409, 2001.

[New02a] M. E. J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20):–, 2002.

[New02b] M. E. J. Newman. Spread of epidemic disease on networks. *Physical Review E*, 66(1):–, 2002.

[New03a] M. E. J. Newman. Mixing patterns in networks. *Physical Review E*, 67(2):–, 2003.

[New03b] M. E. J. Newman. Structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[New04a] M. E. J. Newman. Analysis of weighted networks. *Physical Review E*, 70:056131, 2004.

[New04b] M. E. J. Newman. Detecting community structure in networks. *The European Physical Journal B*, 38:321–330, 2004.

[New04c] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6):–, 2004.

[New05] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351, 2005.

[New06a] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):36104, 2006.

[New06b] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23):8577–8582, 2006.

[NP03] M. E. J. Newman and J. Park. Why social networks are different from other types of networks. *Physical Review E*, 68(3), 2003.

[NZT03] S.K. Ng, Z. Zhang, and S. H. Tan. Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, 19(8):923–9, 2003.

[OB02] Z. N. Oltvai and A.-L. Barabási. Life's complexity pyramid. *Science*, 298(5594):763–764, 2002.

[OC04] R. N. Onody and P. A. de Castro. Complex network study of brazilian soccer players. *Physical Review E*, 70:037103, 2004.

[Par42] V. Pareto. *The Mind and Society: Trattato Di Sociologia Generale*. Harcourt, Brace and Co, 1942.

[Par03] R. M. Parker. *Bordeaux : A Consumer's Guide to the World's Finest Wines*. Simon and Schuster Adult Publishing Group, 2003.

[Phi66] D. C. Phillips. The three-dimensional structure of an enzyme molecule. *Scientific American*, 215(5):78–90, 1966.

[Phi01] R. Phillips. *A Short History of Wine*. Ecco Press, 2001.

[PR02] C. P. Ponting and R. R. Russell. The natural history of protein domains. *Annual Review of Biophysics and Biomolecular Structure*, 31:45–71, 2002.

[PSSS03] R. Pastor-Satorras, E. Smith, and R. V. Solè. Evolving protein interaction networks through gene duplication. *Journal of Theoretical Biology*, 222(2):199–210, 2003.

[PSV04] R. Pastor-Satorras and A. Vespignani. *Evolution and structure of the Internet*. Cambridge Univeristy Press New York, 2004.

[PSVV01] R. Pastor-Satorras, A. Vázquez, and A. Vespignani. Dynamical and correlation properties of the Internet. *Physical Review Letters*, 87(25):258701, 2001.

[QLWL03] H. Qin, H. H. S. Lu, W. B. Wu, and W. H. Li. Evolution of the yeast protein interaction network. *Proceedings of the National Academy of Sciences of the United States of America*, 100(22):12820–12824, 2003.

[RB03] E. Ravasz and A.-L. Barabási. Hierarchical organization in complex networks. *Physical Review E*, 67:026112, 2003.

[RBTdFC07] F. A. Rodrigues, P. R. Villas Boas, G. Travieso, and L. da F. Costa. Seeking the best Internet model. arXiv:0706.3225v2, 2007.

[RC05] F. A. Rodrigues and L. da F. Costa. Surviving opinions in Sznajd models on complex networks. *International Journal of Modern Physics C*, 16(11), 2005.

[RCC+04] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(09):2658–2663, 2004.

[Red98] S. Redner. How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B*, 4(2):131–134, 1998.

[RG03] A. W. Rives and T. Galitski. Modular organization of cellular networks. *Proceedings of the National Academy of Sciences of the United States of America*, 100(3):1128–1133, 2003.

[Ric81] J. S. Richardson. The anatomy and taxonomy of protein structure. *Advances in Protein Chemistry*, 34:167–339, 1981.

[RKBbA05] H. D. Rozenfeld, J. E. Kirk, E. M. Bollt, and D. ben Avraham. Statistics of cycles: how loopy is your network? *Journal of Physics A: Mathematical and General*, 38:4589–4595, 2005.

[RSDR+01] J. C. Rain, L. Selig, H. De Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schaechter, et al. The protein-protein interaction map of *Helicobacter pylori*. *Nature*, 409(6817):211–215, 2001.

[RT04] V. Rosato and F. Tiriticco. Growth mechanisms of the AS-level Internet network. *Europhysics Letters*, 66(4):471–477, 2004.

[RTC07] F. A. Rodrigues, G. Travieso, and L. da F. Costa. Fast community identification by hierarchical growth. *International Journal of Modern Physics C*, 18(6), 2007.

[SG03] S. Bornholdt Schuster and H. G., editors. *Handbook of Graphs and Networks: From the Genome to the Internet*. Wiley-VCH, 2003.

[Sim55] H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42(3–4):425–440, 1955.

[SKM96] D. Stoyan, W. S. Kendall, and J. Mecke. *Stochastic Geometry and Its Applications*. John Wiley and Sons, 1996.

[SOMMA02] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of *Escherichia Coli*. *Nature Genetics*, 31:64–68, 2002.

[SPSSK02] R. V. Solè, R. Pastor-Satorras, E. Smith, and T. B. Kepler. A model of large-scale proteome evolution. *Advances in Complex Systems*, 5(1):43–54, 2002.

[SSM03] E. Sprinzak, S. Sattath, and H. Margalit. How reliable are experimental protein-protein interaction data? *J. Mol. Biol.*, 327(5):919–923, 2003.

[Str01] S. H. Strogatz. Exploring complex networks. *Nature*, (410):268–276, 2001.

[SV05] S. N. Soffer and A. Vázquez. Clustering coefficient without degree correlations biases. *Physical Review E*, 71:057101, 2005.

[Tan02] A. S. Tanenbaum. *Computer Networks*. Prentice Hall, 2002.

[TdFC06] G. Travieso and L. da F. Costa. Spread of opinions and proportional voting. *Physical Review E*, 74(3), 2006.

[TRRdFC06] G. Travieso, F. A. Rodrigues, C. A. Rugiero, and L. da F. Costa. Complex network modeling and simulation of distributed systems processing. In *II TIDIA Workshop*, São Paulo - Brazil, 2006.

[TRT06] B. Tadic, G. B. Rodgers, and S. Thurner. Transport on complex networks: Flow, jamming and optimization. arXiv:physics/0606166, 2006.

[VBRTdFC07] P. R. Villas-Boas, F. A. Rodrigues, G. Travieso, and L. da F. Costa. Chain motifs: The tails and handles of complex networks. arXiv:0706.2365v1, 2007.

[VFMV03a] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani. Global protein function prediction from protein-protein interaction networks. *Nature Biotechnology*, 21(6):697–700, 2003.

[VFMV03b] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani. Modeling of protein interaction networks. *Complexus*, 1(1):38–44, 2003.

[Vin89] L. Vincent. Graphs and mathematical morphology. *Signal Processing*, 16(4):365–388, 1989.

[VLL00] B. Vogelstein, D. Lane, and A. J. Levine. Surfing the p53 network. *Nature*, 408:307–310, 2000.

[VPSV02] A. Vázquez, R. Pastor-Satorras, and A. Vespignani. Large-scale topological and dynamical properties of the Internet. *Physical Review E*, 65(6):66130, 2002.

[VS07] S. Valverde and R.V. Solè. Hierarchical small worlds in software architecture. *Dynamics of Continuous Discrete and Impulsive Systems: Series B; Applications and Algorithms*, 2007.

[VTPL05] C. Vogel, S. A. Teichmann, and J. Pereira-Leal. The relationship between domain duplication and recombination. *Journal of Molecular Biology*, 346(1):355–65, 2005.

[WA05] S. Wuchty and E. Almaas. Peeling the yeast protein network. *Proteomics*, 5(2):444–449, 2005.

[Wag01] A. Wagner. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Molecular Biology and Evolution*, 18:1283–1292, 2001.

[Wag03] A. Wagner. How the global structure of protein interaction networks evolves. *Proceedings: Biological Sciences*, 270(1514):457–466, 2003.

[Wat03] D. J. Watts. *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton University Press, Princeton, New Jersey, 2003.

[Wax88] BM Waxman. Routing of multipoint connections. *IEEE Journal on Selected Areas in Communications*, 6(9):1617–1622, 1988.

[Wet73] D. B. Wetlaufer. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 70(3):697–701, 1973.

[WF94] S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge University Press, Cambridge, 1994.

[WJ63] J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[WJ02] J. Winick and S. Jamin. Inet-3.0: Internet topology generator. *University of Michigan Technical Report CSE-TR-456-02*, 2002.

[WOB03] S. Wuchty, Z. N. Oltvai, and A.-L. Barabási. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Genetics*, 35(2):176–179, 2003.

[WS98] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, 1998.

[Wuc01] S. Wuchty. Scale-free behavior in protein domain networks. *Molecular Biology and Evolution*, 18:1694–1702, 2001.

[Wuc02] S. Wuchty. Interaction and domain networks of yeast. *Proteomics*, 2(12):1715–23, 2002.

[YJB02] S. H. Yook, H. Jeong, and A.-L. Barabási. Modeling the Internet's large-scale topology. *Proceedings of the National Academy of Sciences of the United States of America*, 99:13382–13386, 2002.

[YLSK+04] E. Yegger-Lotem, S. Sattah, N. Kashtan, S.Itzkovitz, R. Milo, R. Y. Pinter, and U. Alon. Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):5934–5939, 2004.

[ZM04a] S. Zhou and R. J. Mondragon. The rich-club phenomenon in the Internet topology. *Communications Letters, IEEE,* 8(3):180–182, 2004.

[ZM04b] S. Zhou and R.J. Mondragón. Accurately modeling the Internet topology. *Physical Review E,* 70(6):66108, 2004.