**UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS**

LUCAS ANTIQUEIRA

**Relações da estrutura de redes complexas com
as dinâmicas do passeio aleatório, de
transporte e de sincronização**

São Carlos

2011

LUCAS ANTIQUEIRA

**Relações da estrutura de redes complexas com
as dinâmicas do passeio aleatório, de
transporte e de sincronização**

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

Área de Concentração: Física Aplicada
Opção: Física Computacional

Orientador: Prof. Dr. Luciano da Fontoura Costa

VERSÃO CORRIGIDA

(versão original disponível na unidade que aloja o programa)

São Carlos

2011

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

Ficha catalográfica elaborada pelo Serviço de Biblioteca e Informação do IFSC,
com os dados informados pelo(a) autor(a)

Antiqueira, Lucas

Relações da estrutura de redes complexas com as dinâmicas do passeio aleatório, de transporte e de sincronização / Lucas Antiqueira; orientador Luciano da Fontoura Costa - versão corrigida -- São Carlos, 2011.

169 p.

Tese (Doutorado - Programa de Pós-Graduação em Física Aplicada Computacional) -- Instituto de Física de São Carlos, Universidade de São Paulo, 2011.

1. Redes complexas. 2. Grafos. 3. Estrutura. 4. Dinâmica. 5. Sistemas complexos. I. da Fontoura Costa, Luciano, orient. II. Título.

Página reservada à folha de aprovação.

*À minha querida e encantadora mãe,
Valquíria (1962-2009).*

Agradecimentos

Neste espaço, ofereço uma homenagem:

- Aos meus pais, Gilberto e Valquíria;
- Ao meu irmão, Moisés;
- À luz de todos os meus dias, Lia;
- Aos meus avôs, João e Amélia;
- Aos “cãopanheiros” Bob e Dig;
- Ao meu orientador, Prof. Luciano.

Faltam-me palavras para expressar completa e exatamente minha gratitude. Podem ter certeza disto: lembro-me exatamente da importância de cada um em minha vida e, consequentemente, no desenvolvimento desta tese. Tenho esperança de que minhas atitudes em nosso dia a dia tenham refletido meu apreço por vocês.

Ofereço também agradecimentos:

- Aos amigos do IFSC que ajudaram-me neste trabalho. Mais especificamente: João, pelo suporte em sincronização de redes; Francisco, pelo suporte em análise canônica; e Matheus, pelo cálculo da acessibilidade em redes grandes;
- Aos demais amigos do IFSC com quem tive a oportunidade de conviver durante esses quatro anos: Mauro, Bruno, Alexandre, Marquinhos, Débora, Mônica, Krissia, Renato F., Renato P., Diego, Paulino, César, Osvaldo, André e Thomas.

Que Deus abençoe todos.

Agradeço também a FAPESP pelo apoio financeiro (processo #06/61743-7).

“Não há escola que se iguale a um lar decente,
nem professores que se igualem a pais honestos e virtuosos.”
— MAHATMA GANDHI (1869-1948)

“Fly, on your way, like an eagle.
Fly as high as the sun.”
— BRUCE DICKINSON (1958-...)

Resumo

ANTIQUEIRA, Lucas. **Relações da estrutura de redes complexas com as dinâmicas do passeio aleatório, de transporte e de sincronização**. 2011. 169 p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2011.

O relacionamento entre estrutura e dinâmica em redes complexas foi considerado utilizando-se uma ampla gama de diferentes técnicas. Diversas redes reais foram estudadas em termos das correlações entre grau e atividade. A medida de atividade é definida como a proporção de visitas por vértice no regime estacionário do passeio aleatório simples. O estudo desse tipo de correlação é importante pois pode fornecer subsídios para que uma propriedade dinâmica de um vértice possa ser obtida somente analisando-se seu(s) grau(s). O conceito de acessibilidade foi abordado nesse contexto, permitindo que fossem evidenciadas diferentes correlações, em redes como a WWW, de acordo com a intensidade de acessibilidade dos vértices. Propôs-se também um novo modelo de rede baseado no crescimento do número de vértices em que novas conexões são criadas com probabilidade proporcional à atividade de cada vértice. Esse modelo pode ser entendido como uma generalização do modelo de Barabási e Albert para redes com arestas direcionadas. Utilizando-se um conjunto de diversas medidas estruturais, mostrou-se que o novo modelo apresenta, entre outros modelos tradicionais de redes, a maior compatibilidade com três redes corticais. Foi também desenvolvido um método para caracterização da distribuição de subgrafos e seus inter-relacionamentos. O principal aspecto dessa metodologia é a expansão gradual dos subgrafos, desenvolvida para que os vértices que encontram-se fora de subgrafos possam ter suas relevâncias quantificadas em termos da importância no estabelecimento das conexões entre subgrafos. Experimentos para ilustração do método foram realizados utilizando-se quatro modelos de redes e cinco redes reais, e os resultados obtidos foram relacionados aos processos dinâmicos de transporte e de espalhamento. Outro tópico aqui considerado é o dos efeitos da amostragem de redes corticais, quantificados por meio de análise multivariada e classificação, fazendo uso de um conjunto de medidas estruturais de redes. Esses efeitos também foram mensurados em termos do comportamento dinâmico das redes (sincronização e acessibilidade). Simulações dos métodos de encefalografia MEG e EEG mostraram que as redes amostradas podem apresentar características bem diferentes das da rede original, principalmente no caso de amostras pequenas. Adicionalmente, a rede integrada da bactéria *Escherichia coli* foi analisada, a qual incorpora (i) regulação de transcrição gênica, (ii) vias metabólicas e de sinalização e (iii) interações entre proteínas. *Outliers* foram identificados no relacionamento entre grau e atividade, os quais representam reguladores globais de transcrição. Além disso, verificou-se que esses *outliers* são genes altamente expressos em diferentes condições, apresentando, portanto, uma natureza global no controle de diversos outros genes da célula.

Palavras-chave: Redes complexas. Grafos. Estrutura. Dinâmica. Sistemas complexos.

Abstract

ANTIQUEIRA, Lucas. **Relationships between the structure of complex networks and the random walk, transport and synchronization dynamics**. 2011. 169 p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2011.

The relationship between structure and dynamics was addressed by employing a wide range of different approaches. First, the correlations between degree and activity were studied in various real-world networks. The activity is defined as the proportion of visits to each node in the steady-state regime of the simple random walk. This type of correlation can provide means to assess node activity only in terms of the degree. The concept of accessibility was included in this analysis, showing an intimate relationship (in networks such as the WWW) between the type of correlation and the level of accessibility observed on nodes. A new complex network model founded on growth was also proposed, with new connections being established proportionally to the current activity of each node. This model can be understood as a generalization of the Barabási-Albert model for directed networks. By using several topological measurements we showed that this new model provides, among several other traditional theoretical types of networks, the greatest compatibility with three real-world cortical networks. Additionally, we developed a novel approach considering non-overlapping subgraphs and their interrelationships and distribution through a given network. The main aspect of the methodology is a novel merging procedure developed to assess the relevance of nodes (in relation to the overall subgraph interconnectivity) lying outside subgraphs. Experiments were carried out on four types of network models and five instances of real-world networks, in order to illustrate the application of the method. Furthermore, these results were related to the properties of the transport and spreading processes. Other topic here addressed is the sampling problem in cortical networks. Effects of sampling were quantified using multivariate analysis and classifiers based on structural network measurements. Samples were also evaluated in terms of their dynamical behavior using a synchronization model and the measure of accessibility. By simulating MEG/EEG recordings it was found that sampled networks may substantially deviate from the respective original networks, mainly for small sample sizes. We also report an analysis of the integrated network of *Escherichia coli*, which incorporates (i) transcriptional regulatory interactions, (ii) metabolic/signaling feedback and (iii) protein-protein interactions. Network outliers, which represent global transcriptional regulators, were identified in the relationship between out-degree and activity. These outliers are highly and widely expressed across conditions, therefore supporting their global nature in controlling many genes in the cell.

Keywords: Complex networks. Graphs. Structure. Dynamics. Complex systems.

Lista de Figuras

Figura 1.1 - Resumo das investigações realizadas neste projeto.	29
Figura 1.2 - Principais conceitos trabalhados na tese e suas inter-relações.	37
Figura 2.1 - Exemplo de cálculo da acessibilidade.	53
Figura 2.2 - Número de pacotes de dados durante a dinâmica de transporte.	56
Figura 2.3 - Classificação de máximo a posteriori unidimensional.	71
Figura 2.4 - Classificação de máximo a posteriori bidimensional.	71
Figura 2.5 - Visão geral do método de classificação de redes.	72
Figura 3.1 - Exemplos de correlações grau-atividade em um grafo não direcionado e em um direcionado.	74
Figura 3.2 - Intensidades das correlações do tipo B como função da reciprocidade. .	75
Figura 3.3 - Intensidades das correlações do tipo A como função da reciprocidade. .	76
Figura 3.4 - Ilustração do tipo de análise realizada nas correlações do tipo A.	77
Figura 3.5 - Correlações grau de entrada vs. atividade na rede Macaco'85.	78
Figura 3.6 - Correlações grau de entrada vs. atividade na rede <i>C. elegans</i>	80
Figura 3.7 - Correlações grau de entrada vs. atividade na rede Roget.	80
Figura 3.8 - Correlações grau de entrada vs. atividade na rede Massey.	81
Figura 3.9 - Correlações grau de entrada vs. atividade na rede Flinders.	82
Figura 3.10 - Correlações grau de entrada vs. atividade na rede Nottingham.	82

Figura 3.11 - Correlações entre grau de entrada e atividade (visualização parcial) nas redes WWW.	82
Figura 3.12 - Intensidades das correlações do tipo A como função da reciprocidade após a análise borda/núcleo.	83
Figura 4.1 - Passo T ao longo do crescimento de uma rede APA com $m = 2$	86
Figura 4.2 - Distribuições de grau e atividade nos modelos APA, APA' e BA.	87
Figura 4.3 - Histogramas dos coeficientes de correlação grau-atividade nos modelos APA e APA'.	89
Figura 4.4 - Correlações grau-atividade em três redes APA.	90
Figura 4.5 - Correlações grau-atividade em três redes APA'.	91
Figura 4.6 - Comparação de redes reais com os modelos APA e APA'.	92
Figura 4.7 - Correlações grau-atividade nas redes corticais.	93
Figura 5.1 - Exemplo de um grafo e seus subgrafos.	95
Figura 5.2 - Grafo utilizado como exemplo de aplicação do método de análise de subgrafos.	98
Figura 5.3 - Operação de dilatação de um subgrafo.	100
Figura 5.4 - Valores de relevância $r(v)$ para os vértices do grafo da Figura 5.2. . . .	103
Figura 5.5 - Expansão dos subgrafos da Figura 5.2.	104
Figura 5.6 - Tamanhos de (e distâncias entre) subgrafos nos modelos de rede.	108
Figura 5.7 - Número de subgrafos e de vértices na expansão gradual realizada nos modelos de rede.	109
Figura 5.8 - Taxas críticas de geração de pacotes na dinâmica de transporte.	111
Figura 5.9 - Tempo de cobertura do passeio aleatório simples.	112

Figura 5.10 - Tamanhos de (e distâncias entre) subgrafos nas redes reais.	113
Figura 5.11 - Número de subgrafos e de vértices na expansão gradual realizada nas redes reais.	114
Figura 6.1 - Ilustração do processo de amostragem de redes.	118
Figura 6.2 - Ilustrações do modelo neuronal e do método de amostragem.	120
Figura 6.3 - Classificação das redes neuronais.	124
Figura 6.4 - Classificação das redes amostradas com decaimento linear de sinal. . . .	125
Figura 6.5 - Classificação das redes amostradas com decaimento quadrático de sinal.	126
Figura 6.6 - Distribuição de subgrafos nas redes neuronais e nas respectivas amostras.	128
Figura 6.7 - Níveis de sincronia ao longo do tempo nas redes neuronais e nas respec- tivas amostras.	130
Figura 6.8 - Histogramas de acessibilidade para $h = 2$ nas redes neuronais e nas res- pectivas amostras.	131
Figura 6.9 - Histogramas de acessibilidade para $h = 3$ nas redes neuronais e nas res- pectivas amostras.	132
Figura 7.1 - Identificação dos <i>outliers</i> de dinâmica na rede integrada <i>E. coli</i>	139
Figura 7.2 - Identificação dos <i>outliers</i> de estrutura na rede integrada <i>E. coli</i>	140
Figura 7.3 - Diagramas em caixa para a acessibilidade na rede integrada <i>E. coli</i>	141
Figura 7.4 - Motivos identificados na rede integrada <i>E. coli</i>	142
Figura 7.5 - Diagramas em caixa para a expressão gênica na <i>E. coli</i>	146

Lista de Tabelas

Tabela 2.1 - Informações básicas a respeito das redes reais utilizadas neste projeto. . .	59
Tabela 2.2 - Disponibilidade das redes reais utilizadas neste projeto.	60
Tabela 4.1 - Medidas calculadas para os modelos APA, APA' e BA.	88
Tabela 6.1 - Medidas calculadas para o modelo neuronal e para os modelos ER, WS, BA e GG.	123
Tabela 7.1 - Medidas computadas para as redes TRN, MSFN e PPIN.	137
Tabela 7.2 - Medidas obtidas para a rede integrada <i>E. coli</i> e seu maior componente conexo.	138
Tabela 7.3 - Genes da <i>E. coli</i> mais frequentes nos motivos de tamanho 3.	143
Tabela 7.4 - Genes da <i>E. coli</i> mais frequentes nos motivos de tamanho 4.	144

Lista de Abreviaturas

APA	Activity-based Preferential Attachment
BA	Barabási-Albert
BGP	Border Gateway Protocol
CPD	Central Point Dominance
EAT	Edinburgh Associative Thesaurus
EEG	Electroencephalography
ER	Erdős-Rényi
FFL	Feed-Forward Loop
FOLDOC	Free On-line Dictionary of Computing
(f)MRI	(functional) Magnetic Resonance Imaging
GG	Modelo Geográfico
MEG	Magnetoencephalography
MSFN	Metabolic and Signaling Feedback Network
ODLIS	Online Dictionary for Library and Information Science
P2P	Peer-to-Peer
PCA	Principal Components Analysis
PPIN	Protein-Protein Interaction Network
TF	Transcription Factor
TG	Target Gene

TRN	Transcriptional Regulatory Network
USFFAN	University of South Florida Free Association Norms
WS	Watts-Strogatz
WWW	World Wide Web

Lista de Símbolos

$G(V,E)$	Grafo definido pelo conjunto de vértices V interconectados pelas arestas especificadas no conjunto E
A	Matriz de adjacências de um grafo
N	Número de vértices em um grafo
E	Número de arestas em um grafo
P	Matriz de transições do passeio aleatório simples
k_i	Grau de um vértice i
k_i^{in}	Grau de entrada de um vértice i
k_i^{out}	Grau de saída de um vértice i
$k_i^{(d)}$	Grau hierárquico no nível d de um vértice i
$k_i^{(d \text{in})}$	Grau hierárquico de entrada no nível d de um vértice i
$k_i^{(d \text{out})}$	Grau hierárquico de saída no nível d de um vértice i
$k_i^{(nm)}$	Grau médio da vizinhança de um vértice i
$er_i^{(d)}$	Grau inter-anel no nível hierárquico d de um vértice i
$ar_i^{(d)}$	Grau intra-anel no nível hierárquico d de um vértice i
cc_i	Coefficiente de agrupamento de um vértice i
cc_i^{in}	Coefficiente de agrupamento de entrada de um vértice i
cc_i^{out}	Coefficiente de agrupamento de saída de um vértice i
$cc_i^{(d)}$	Coefficiente de agrupamento hierárquico no nível d de um vértice i

$cc_i^{(d in)}$	Coeficiente de agrupamento hierárquico de entrada no nível d de um vértice i
$cc_i^{(d out)}$	Coeficiente de agrupamento hierárquico de saída no nível d de um vértice i
ℓ_i	Comprimento médio dos caminhos mínimos que partem de um vértice i
ℓ	Comprimento médio dos caminhos mínimos em um grafo
r	Assortatividade de um grafo
r^{ii}	Assortatividade de entrada de um grafo
r^{oo}	Assortatividade de saída de um grafo
$B_i^{(n)}$	Centralidade de um vértice i
$B_i^{(e)}$	Centralidade de uma aresta i
E_{global}	Eficiência global de um grafo
E_{local}	Eficiência local de um grafo
CPD	Dominância do ponto central em um grafo
ρ	Reciprocidade das arestas direcionadas de um grafo
ρ_{min}	Reciprocidade mínima possível em um grafo
ρ_{corr}	Coeficiente de correlação linear de Pearson
$cv_i^{(d)}$	Taxa de convergência no nível hierárquico d de um vértice i
$dv_i^{(d)}$	Taxa de divergência no nível hierárquico d de um vértice i
π_i	Atividade de um vértice i
$Ac_i^{(h)}$	Acessibilidade do vértice i a h passos
λ_c	Taxa crítica de criação de pacotes na dinâmica de transporte
r_{sync}	Nível global de sincronia em uma rede

Sumário

1	Introdução	27
2	Material e Métodos	39
2.1	Representação e Caracterização Estrutural de Redes	39
2.2	Dinâmica do Passeio Aleatório Simples	49
2.3	Dinâmica de Acessibilidade	51
2.4	Dinâmica de Transporte	54
2.5	Dinâmica de Sincronização	55
2.6	Modelos de Redes	57
2.7	Redes Reais	58
2.8	Análise de Componentes Principais	66
2.9	Análise de Variável Canônica e Classificação	68
3	Correlações entre Grau e Atividade	73
3.1	Tipos de Correlação	73
3.2	Reciprocidade e Correlações Tipo B	75
3.3	Acessibilidade e Correlações Tipo A	76
4	Conectividade Guiada pela Atividade	85
4.1	Modelo Proposto	85

4.2	Propriedades do Modelo	87
4.3	Comparação com Redes Reais	90
5	Caracterização da Distribuição de Subgrafos	95
5.1	Método Proposto	96
5.2	Subgrafos com Alto Agrupamento e Processos Dinâmicos	105
6	Efeitos da Amostragem de Redes Corticais	117
6.1	Modelo de Rede Neuronal	119
6.2	Amostragem da Rede Neuronal	119
6.3	Efeitos Estruturais e Dinâmicos da Amostragem	121
7	Rede <i>Escherichia coli</i>: Uma Abordagem Integrada	135
7.1	Integração dos Dados	135
7.2	Relacionamento Estrutura-Dinâmica: <i>Outliers</i>	138
7.3	Funções Biológicas dos <i>Outliers</i>	145
8	Conclusões	149
	Referências	155

1 Introdução

A origem dos estudos em redes complexas pode ser traçada até as primeiras investigações em grafos realizadas por Leonhard Euler no século XVIII, quando o problema das pontes de Königsberg foi solucionado utilizando-se uma representação formada por um conjunto de vértices conectados por arestas (1), ou seja, empregando-se o conceito de grafo. Euler desejava saber se existia um caminho em Königsberg (hoje Kaliningrado, Rússia) pelo qual cada uma das sete pontes da cidade seria cruzada uma única vez – ele provou que tal caminho não existia. Desde então, a teoria dos grafos evoluiu a ponto de se tornar um importante ramo da matemática (2), tendo como exemplo fundamental de investigação a teoria dos grafos aleatórios desenvolvida por Erdős e Rényi (3). Redes, que são estruturas naturalmente representadas por grafos, surgem em uma ampla gama de circunstâncias, o que faz do ramo das redes complexas um campo de pesquisa excepcionalmente interdisciplinar (4–6). Exemplos envolvem, entre muitos outros, (i) cientistas da computação, que, por exemplo, desenvolvem estruturas de dados e algoritmos para grafos (7), (ii) sociologistas, os quais investigam a estrutura e a dinâmica das relações sociais (tome como exemplo o famoso experimento de Milgram que deu origem ao termo *seis graus de separação* (8)), (iii) biólogos, que frequentemente representam e estudam interações entre proteínas (9), genes (10) e metabólitos (11) utilizando o conceito de grafo, e (iv) físicos, que fazem uso recorrente da mecânica estatística a fim de solucionar problemas importantes em redes (4).

O interesse em redes cresceu vertiginosamente há aproximadamente uma década, quando grandes estruturas do mundo real foram analisadas e descobriu-se importantes diferenças entre essas estruturas e o grafo aleatório clássico. A partir desse momento, o termo *redes complexas* passou a ser largamente utilizado para denominar estruturas com propriedades não triviais. Primeiramente, descobriu-se que muitas redes do mundo real apresentam o chamado efeito *pequeno mundo*: as distâncias entre seus vértices são curtas (como em uma rede aleatória) e apre-

sentam alto agrupamento local (diferentemente de uma rede aleatória) (12). Exemplo de rede pequeno mundo é a rede elétrica. Outra propriedade importante está relacionada à distribuição do número de conexões de cada vértice – o chamado grau. Nesse caso, a distribuição segue uma lei de potência (da forma $P(k) \sim k^{-\gamma}$, onde k representa o grau e o expoente γ depende da rede analisada), revelando um padrão de conexões bem diferente do das redes aleatórias: alguns poucos vértices, os chamados *hubs*, possuem um alto número de conexões, enquanto a maior parte dos vértices apresenta poucas ligações (13, 14). Redes desse tipo, como a Internet e a World Wide Web (WWW), são frequentemente chamadas de redes *sem escala* (no sentido de invariância de escala). Os *hubs* têm papel importante na resistência a ataques (15), propagação de doenças (16) e navegabilidade (17).

Redes complexas são frequentemente estudadas por meio da análise de suas propriedades estruturais ou dinâmicas (5, 6). A estrutura está relacionada à maneira como as conexões organizam-se entre os vértices da rede. Por exemplo, propriedades estruturais bem conhecidas são o efeito pequeno mundo e a distribuição de graus seguindo uma lei de potência. Por outro lado, a dinâmica é baseada em um processo dependente do tempo que faz uso da estrutura da rede – por exemplo, transporte (18) e sincronização (19). Diversas investigações acerca da dinâmica já foram realizadas em redes regulares utilizadas como modelos de matéria condensada (20), onde a estrutura é ditada pelas propriedades físicas do material e é frequentemente conhecida de maneira completa. Desenvolvimentos mais recentes passaram a considerar processos dinâmicos em estruturas mais complexas. Exemplos são os processos de contágio em redes sociais e de sincronização em redes neuronais (5, 6, 20). Além disso, entender o relacionamento entre estrutura e dinâmica é de grande importância no avanço do conhecimento sobre os princípios que norteiam a organização das redes complexas (21). Diversos trabalhos na literatura demonstram uma preocupação nesse sentido (5), como o estudo da relação do fluxo do passeio aleatório com a estrutura de comunidades (22) e o uso de medidas de compressão de dados como indicadores do relacionamento estrutura-dinâmica (23). Nesta tese, o relacionamento estrutura-dinâmica é investigado utilizando-se diferentes técnicas em diferentes contextos (vide esquema na Figura 1.1), conforme detalhado ao longo desta introdução.

Figura 1.1 – Resumo das investigações realizadas neste projeto, inseridas no contexto do estudo do relacionamento da estrutura de redes complexas com diferentes tipos de processos dinâmicos.

A estrutura de cada vértice pode ser representada pelo grau, o mais simples indicador de conectividade, enquanto a dinâmica pode ser representada pelo conceito aqui identificado por *atividade*, que é a frequência estacionária de visitas a cada vértice em um passeio aleatório simples. O passeio aleatório é um modelo bastante conhecido de dinâmica em redes (20, 24–27), cujas variantes incluem o passeio preferencial e o passeio com reforço. Escolheu-se o passeio aleatório por ser um modelo geral e um dos mais importantes modelos de dinâmica em diversas disciplinas, inclusive redes complexas (20). Exemplos de investigações nessa direção incluem o estudo das propriedades do passeio aleatório em redes sem escala (28–30), pequeno-mundo (31, 32) e aleatórias (33). No passeio aleatório do tipo simples, em que é permitido visitar apenas vértices vizinhos do vértice atual com probabilidade independente do histórico do passeio, o grau é perfeitamente correlacionado com a atividade em redes não direcionadas (24, 25). No caso de redes com arestas direcionadas as correlações tendem a ser bastante intrincadas, como na WWW (34). Portanto, foi empregada nesta tese uma nova abordagem para o estudo das correlações grau-atividade em redes direcionadas, as quais são pouco trabalhadas na literatura (34). Primeiramente, o nível de reciprocidade das arestas de uma rede (35) foi utilizado como indicador da intensidade da correlação entre o grau de saída e a atividade em diversas redes reais. A reciprocidade funciona da seguinte maneira: seu nível máximo ocorre em

redes não direcionadas, enquanto menores valores de reciprocidade são encontrados em redes onde as arestas direcionadas não apresentam total simetria (ou seja, para um conjunto de arestas $i \rightarrow j$ não existe a aresta recíproca $j \rightarrow i$). Adicionalmente, o conceito de acessibilidade (36) foi empregado na análise das correlações entre grau de entrada e atividade. A acessibilidade está relacionada ao número (e respectivas probabilidades) de diferentes caminhos que conectam um vértice específico ao restante dos vértices da rede. Vértices com baixa acessibilidade tendem a se posicionar na periferia (borda) da rede, enquanto vértices com alta acessibilidade tendem a se posicionar no núcleo da rede (36). As correlações entre grau de entrada e atividade foram então analisadas separadamente na borda e no núcleo em várias redes reais, o que evidenciou a existência de dois regimes separados de correlação (exemplos dessas redes são as corticais e a WWW). Esses resultados fornecem subsídios importantes para uma melhor compreensão das intrincadas correlações que podem existir em redes direcionadas (vide Capítulo 3)*.

Outro aspecto relevante na área de redes complexas é a criação e análise de modelos de rede, que são desenvolvidos para que a estrutura e a evolução de importantes classes de redes sejam melhor compreendidas. Exemplos são o modelo de grafos aleatórios de Erdős e Rényi (ER) (3) e o modelo pequeno mundo de Watts e Strogatz (WS) (12) (mais detalhes a respeito desses modelos são fornecidos na Seção 2.6, p. 57). Outro modelo bastante conhecido é o de Barabási-Albert (BA) (13), o qual reproduz a distribuição de graus observada nas redes sem escala. O modelo BA é baseado nos mecanismos de crescimento e ligação preferencial: vértices são sequencialmente adicionados à rede, e suas conexões são criadas com outros vértices já presentes na rede dando-se preferência aos vértices que já são bem conectados – ou seja, a probabilidade de conexão é proporcional ao grau (13). Nesta pesquisa, procurou-se generalizar o modelo BA de modo que passasse a considerar (i) redes com arestas direcionadas, pois são casos extremamente comuns e pouco estudados, e (ii) um processo dinâmico que porventura ocorra na rede, ao invés de se utilizar somente uma propriedade estrutural, como o grau. Uma motivação para essa investigação reside no fato de que, em muitos casos, uma propriedade relacionada a um processo dinâmico está mais intimamente relacionada à importância de um

* Parte desse trabalho foi publicada nos anais do *2nd Workshop on Complex Networks* (CompleNet'2010) (37).

vértice do que o grau ou outra medida puramente estrutural. Por exemplo, a frequência de visitas às páginas da WWW é um indicador particularmente eficiente da relevância de cada página (por exemplo, o índice PageRank do Google (38)). Outra motivação é a de que o estudo de modelos como o proposto fornecem base para que o relacionamento estrutura-dinâmica seja estudado em maior profundidade.

Alguns modelos de rede complexa foram propostos como modificações ou generalizações do modelo BA (39). Entretanto, esses modelos baseiam-se apenas em propriedades estruturais das redes em crescimento. Outros modelos consideram de fato alguma propriedade dinâmica do sistema na formação da rede. Por exemplo, a ligação preferencial já foi combinada com a dinâmica evolucionária de jogos (dilema do prisioneiro) (40). Nesse caso, os vértices são considerados indivíduos de um grupo social, e novos indivíduos são conectados preferencialmente aos vértices que apresentam maior aptidão na dinâmica evolucionária. Dependendo dos parâmetros escolhidos, o modelo pode gerar redes homogêneas ou sem escala. Em outro trabalho, redes com grau de sincronização ótimo foram construídas por meio de *simulated annealing* (41). O grau de sincronização é alto em redes propensas a manter um estado global de sincronia entre osciladores definidos em cada um dos seus vértices. A criação de conexões nesse caso é, portanto, guiada por um parâmetro de um processo dinâmico que ocorre na rede (vide outros exemplos em (42–45)). Um modelo para a Internet baseado em agentes foi criado considerando-se o fluxo de tráfego de dados, além de outras características, como a geografia (46). Sub-redes espacialmente posicionadas, os sistemas autônomos, têm um interesse econômico em aumentar seu tráfego e assim incrementar sua receita. Um aumento na receita gera uma melhoria na infraestrutura (incremento no número de conexões), a qual leva a um aumento no tráfego. A dinâmica de interações entre os agentes leva a uma rede que reproduz diversas características da Internet (46). Outra abordagem considera a teoria Hebbiana (47), em que dois neurônios conectam-se se um neurônio repetidamente ou persistentemente participa do disparo de outro neurônio. Essa teoria já foi considerada no estudo de redes complexas (48). Contudo, modelo similar que incorpore o mecanismo de crescimento da rede não foi encontrado. Além disso, a existência de um modelo de rede direcionada com crescimento e ligação preferencial

baseada em uma propriedade de um processo dinâmico é desconhecida.

Foi portanto proposto neste projeto um novo modelo de ligação preferencial em que vértices com alta atividade (definida na Seção 2.2, p. 49) têm maior chance de receber novas conexões. Foi provado que o novo modelo é uma generalização do modelo BA para o caso de redes com arestas direcionadas. Ou seja, quando a direção das arestas é desconsiderada, o novo modelo reproduz exatamente o modelo BA. Foi mostrado também que o modelo proposto, ao integrar propriedades estruturais e dinâmicas em seu crescimento, aproxima as características das redes corticais do gato e do macaco (49, 50), estruturas que são base da atividade cerebral nesses animais. A metodologia empregada envolve a comparação dessas redes corticais com diversos outros modelos, inclusive o aqui proposto, obtendo-se de cada rede um conjunto de métricas (vetor de atributos) comumente utilizadas na caracterização de redes complexas. A partir disso, o modelo que melhor se aproxima da rede real é escolhido por meio da análise de variável canônica e da classificação por máxima probabilidade a posteriori (vide Capítulo 4)*.

Redes complexas são frequentemente analisadas na mesoescala, ou seja, em uma escala intermediária entre a dos vértices isolados e a de toda a rede. Diversos trabalhos levam em consideração a estrutura de comunidades[†] ou dos caminhos existentes entre diferentes porções da rede (53, 54). Subgrafos são grupos de vértices, em princípio, genéricos; contudo, podem representar vértices importantes em uma rede, tais como os *hubs*. Subgrafos podem também ser formados por motivos[‡] e, portanto, podem ser analisados de acordo com sua distribuição na rede. Outra possibilidade refere-se à característica de algumas redes possuírem alguma informação associada aos seus vértices como, por exemplo, tópicos na WWW ou categorias morfosintáticas em uma rede de palavras. De maneira geral, vértices que compartilham um atributo de interesse podem formar subgrafos.

Poucos trabalhos foram realizados com relação ao estudo da *distribuição* de subgrafos. Muitos trabalhos existem, de fato, com relação à *detecção* de subgrafos (de motivos e co-

* Esse trabalho foi publicado no periódico *Journal of Statistical Mechanics* (51).

† Comunidades, ou agrupamentos em grafos, são conjuntos de vértices densamente conectados entre si, enquanto as conexões entre comunidades são relativamente mais esparsas (52).

‡ Motivos (do inglês, *motifs*) são grupos de vértices, geralmente pequenos, com estrutura (e por vezes função) bem definidas (55).

munidades, e também de subgrafos cujos vértices apresentam características similares (56)). Contudo, o enfoque aqui é caracterizar as relações entre subgrafos previamente conhecidos. A posição de subgrafos em grafos moleculares já foi estimada anteriormente, inclusive em termos da distância entre subgrafos (57). Contudo, pretende-se aqui investigar (i) a proximidade entre os subgrafos e (ii) o papel dos demais vértices da rede no estabelecimento da conexão entre subgrafos. Tais informações são importantes para uma maior compreensão da organização e da dinâmica de uma rede. Por exemplo, se subgrafos formados por *hubs* estão muito próximos uns dos outros, a porção da rede contendo tais subgrafos pode ser entendida como crítica para a integridade da rede e comunicação entre seus vértices. Outras situações similares podem surgir ao se considerar outra medida na seleção de subgrafos. A motivação para esse estudo surgiu a partir da observação de que regiões distintas de diversas redes direcionadas apresentam diferentes comportamentos com relação às correlações grau-atividade (vide parágrafos prévios desta Introdução). As propriedades dessas regiões (subgrafos) podem fornecer indícios que ajudem a explicar o comportamento da relação estrutura-dinâmica. Portanto, foi desenvolvida uma nova metodologia para a caracterização da distribuição de subgrafos. Embora essa metodologia possa apenas ser aplicada a redes não direcionadas, seu potencial foi mostrado por meio do estudo de redes artificiais (por exemplo, modelo pequeno mundo) e reais (como a rede elétrica dos EUA). Exemplos de subgrafos de interesse, formados por vértices com alto coeficiente de agrupamento, foram selecionados. Adicionalmente, a análise das propriedades desses subgrafos permitiu uma caracterização do seu relacionamento com os processos dinâmicos de transporte e de espalhamento (vide Capítulo 5)*.

Diversas questões em redes complexas continuam em aberto. Um exemplo é o problema da amostragem: a rede existe, mas é difícil coletá-la de maneira precisa e completa. Dado que utilizamos dados incompletos em diversos casos, é importante saber se nossas conclusões podem ser extrapoladas para a rede completa (e inacessível). Em alguns casos, a grande maioria dos vértices ou arestas pode estar oculta. Por exemplo, a WWW é uma rede enorme que evolui rapidamente, sendo virtualmente impossível dela obter uma amostra exata em determi-

* Esse trabalho foi publicado no periódico *New Journal of Physics* (58).

nado momento, pois até o fim da coleta de dados, a rede já teria sido alterada. Muitas redes sociais são, por razões práticas, apenas esparsamente amostradas, já que em alguns casos é necessário utilizar questionários ou realizar algum tipo de consulta pessoal. É interessante também observar como o procedimento de amostragem influencia a natureza do relacionamento estrutura-dinâmica, ou seja, de que maneira esse relacionamento é afetado como resultado do processo de amostragem.

A questão da amostragem de redes complexas é de extrema importância para a área, e alguns estudos evidenciam os problemas que podem surgir ao se analisar redes amostradas. Por exemplo, distribuições de graus foram avaliadas em redes aleatórias e sem escala utilizando amostragem aleatória (59) ou inspirada na coleta de conexões da Internet (amostragem *traceroute*) (60). Inclusive, amostras do tipo *traceroute* implicam em uma probabilidade de detecção de vértices/arestas proporcional aos respectivos valores de centralidade (essa métrica é definida na Seção 2.1.13, p. 47) (61). Outros trabalhos reportam análises de redes sem escala estimadas (62, 63) e da precisão de amostras como uma função do tamanho da rede coletada (64). Em outro estudo (65), foi descrito como redes de interação de proteínas são afetadas pelo processo de amostragem. Os autores consideraram diversas estruturas diferentes como representantes da rede original a ser amostrada, tais como a rede aleatória e a sem escala, entre outras. Dessas redes, amostras aleatórias foram coletadas e tiveram suas distribuições de grau analisadas. Independentemente da distribuição da rede original, todas as amostras apresentaram uma lei de potência, demonstrando um problema sério que pode ocorrer na análise de redes de proteínas.

Neste projeto optou-se por estudar a amostragem cortical realizada por meio de sensores posicionados na superfície do crânio. Investigou-se até que ponto as redes amostradas reproduzem características relevantes da rede original, ou seja, da estrutura fisiológica que forma a base da atividade gravada pelos sensores. A rede original é a dos neurônios no sistema nervoso central, que contém aproximadamente 10^{10} – 10^{12} neurônios com 10^3 – 10^4 conexões cada. A conectividade funcional pode ser estimada por meio de técnicas como MEG/EEG (do inglês *magnetoencephalography/electroencephalography*), as quais captam padrões de sincronização em populações de neurônios com alta resolução temporal. O problema da amostragem nesse

caso é um grande desafio, pois dispositivos disponíveis atualmente contêm tipicamente apenas 20–300 sensores MEG ou eletrodos EEG. Procurou-se aqui reproduzir, por meio de simulações, esse tipo de amostragem, para então avaliar seus efeitos comparando-se métricas comumente utilizadas na área de redes complexas. O processo começa com uma rede artificial criada para representar a rede neuronal original. A partir disso, amostras de diversos tamanhos são obtidas e suas características são comparadas às da rede completa. Por meio de análise de variável canônica seguida de classificação, foi possível mensurar importantes imprecisões estruturais que podem surgir como efeito desse tipo de amostragem. Propriedades dinâmicas também foram analisadas, evidenciando que o nível de sincronização das redes originais pode ser bem reproduzido nas amostras, ao contrário da dinâmica de acessibilidade (vide Capítulo 6)*. Essa metodologia deve ainda ser generalizada para redes direcionadas, nas quais o processo de amostragem pode ser investigado de acordo com a sua influência nas correlações grau-atividade, permitindo assim verificar se (e quando) a amostragem é responsável pelo enfraquecimento das correlações observadas em diversas redes (Capítulo 3).

De fato, tópicos em biologia têm sido frequentemente abordados utilizando-se o conceito de rede (67, 68). No caso da célula, os processos guiados por vias metabólicas e de sinalização, as interações proteína-proteína e as regulações de transcrição gênica podem ser representados por grafos (69–71). Entretanto, enquanto na maioria das investigações nessa área esses processos são estudados de uma maneira isolada, uma compreensão mais completa requer uma integração entre os diferentes tipos de interações que ocorrem na célula (72). Embora desenvolvimentos tenham sido realizados nessa direção (interações proteína-proteína integradas com dados de expressão e de localização subcelular na levedura *Saccharomyces cerevisiae* (73)), a maior parte dos estudos é limitada a subsistemas particulares. Um dos processos fundamentais, mesmo em um sistema unicelular simples como o de bactérias, é o de regulação da transcrição gênica. Ultimamente, uma grande quantidade de informação a respeito desse processo tem sido gerada, o que possibilitou a representação, por meio de grafos, do conjunto completo de interações em bactérias como *Escherichia coli* e *Bacillus subtilis* (74–76). Entretanto, as redes

* Esse trabalho foi publicado no periódico *NeuroImage* (66).

de transcrição gênica já estudadas limitam-se a um conjunto de fatores de transcrição que controlam um conjunto de genes-alvo (71, 77–80). Outra limitação é o emprego de um pequeno número de medidas de redes relativamente simples (como grau e coeficiente de agrupamento) de maneira mutuamente exclusiva, a fim de se compreender as propriedades locais dos vértices (81–83). Além disso, a associação entre estrutura e dinâmica é frequentemente deixada de lado no estudo dessas redes. Entender o relacionamento estrutura-dinâmica é de fato importante, como demonstrado no caso da sincronização em redes corticais (84, 85). Por outro lado, um tópico bastante explorado é a identificação de motivos em redes biológicas, dada a modularidade inerente dos processos celulares (71, 86, 87).

Nesta tese, a regulação da transcrição gênica na bactéria *Escherichia coli* foi estudada por meio da incorporação dos diferentes aspectos discutidos no parágrafo anterior: integração de diferentes tipos de informação na rede, análise do relacionamento estrutura-dinâmica, utilização de medidas mais complexas para caracterização de redes e detecção de motivos. Foi construída uma rede que incorpora três diferentes processos celulares: (i) transcrição gênica (74), (ii) vias metabólicas e de sinalização (88) e (iii) interações proteína-proteína (83, 89). A rede de transcrição foi tomada como a rede “base” da rede integrada utilizada neste estudo, na qual foram inseridas novas conexões que representam as vias metabólicas e de sinalização e também as interações entre proteínas. O conceito de atividade foi incluído na análise, o qual simula, neste caso, a dinâmica de interação gênica em termos da frequência relativa de ativação de cada vértice. Foi possível observar que alguns genes *outliers* (ou seja, com propriedades discrepantes ou não usuais) apresentam características particulares com relação ao relacionamento estrutura-dinâmica: eles são fracamente ativados embora controlem muitos outros genes. Convencionou-se aqui denominar esses genes como *outliers* de dinâmica. A seguir, as propriedades estruturais desses *outliers* foram analisadas por meio do cálculo de diversas medidas e posterior análise de componentes principais. As medidas selecionadas representam atributos em escala local (como o grau e o coeficiente de agrupamento), global (como a centralidade) e intermediária (medidas hierárquicas) (90). Verificou-se que os *outliers* de dinâmica coincidem com os de estrutura, apresentam alta acessibilidade e têm participação importante na formação de motivos. Além

disso, foi constatado que os *outliers* são reguladores globais na rede de transcrição gênica e que também representam genes intensamente expressos em diferentes condições. Esses resultados sugerem que o método proposto para detecção de *outliers* na rede integrada pode ser utilizado na identificação de reguladores globais em outros organismos (vide Capítulo 7)*.

Figura 1.2 – Principais conceitos trabalhados na tese e suas inter-relações. Para cada inter-relacionamento, é especificado o respectivo capítulo em que ele é tratado.

A Figura 1.2 contém um diagrama que integra todos os conceitos estruturais e dinâmicos tratados nesta tese. A estrutura é representada por três itens: cálculo de *medidas estruturais*, análise da distribuição de *subgrafos* e estudo dos efeitos da *amostragem* de redes corticais. Já a dinâmica relaciona-se aos processos do *passeio aleatório* (o qual inclui as dinâmicas de atividade e de acessibilidade), de *transporte* e de *sincronização*. Grande parte das contribuições desta tese está resumida na Figura 1.2, a qual mostra em quais capítulos são abordados os diversos relacionamentos existentes entre os conceitos estruturais e dinâmicos aqui considerados.

A seguir, no Capítulo 2, os métodos e recursos utilizados neste projeto são detalhados. Os capítulos posteriores referem-se ao estudo do relacionamento estrutura-dinâmica nos diversos casos já comentados nesta introdução: análise das correlações grau-atividade em redes direcionadas utilizando os conceitos de reciprocidade e acessibilidade (Capítulo 3); definição e análise de um modelo de crescimento de rede com ligação preferencial baseada na atividade (Capí-

* Esse trabalho foi submetido ao periódico *Bioinformatics*.

tulo 4); desenvolvimento de uma metodologia para caracterização da distribuição de subgrafos e aplicação no estudo do relacionamento estrutura-dinâmica (Capítulo 5); estudo dos efeitos estruturais e dinâmicos da amostragem de redes corticais funcionais (Capítulo 6); e construção de uma rede integrada da bactéria *Escherichia coli*, com detecção de *outliers* no relacionamento grau-atividade e análise da estrutura e função desses *outliers* com relação ao processo de regulação gênica (Capítulo 7). Por fim, o Capítulo 8 contém as conclusões a respeito deste trabalho de doutorado* e indicações para trabalhos futuros.

* Durante o período de doutoramento também houve participação na elaboração de um artigo de revisão, publicado no periódico *Advances in Physics* (91), o qual trata do uso de redes complexas em diversas áreas do conhecimento. Além disso, encontra-se disponível um manuscrito (92) desenvolvido durante a participação na Complex Systems Summer School de 2010, um evento organizado pelo Santa Fe Institute.

2 Material e Métodos

2.1 Representação e Caracterização Estrutural de Redes

Uma rede é representada por um grafo $G(V, E)$, onde $V = \{v_1, v_2, \dots, v_N\}$ é o seu conjunto de N vértices (ou nós) e $E = \{e_1, e_2, \dots, e_L\}$ é o seu conjunto de L arestas (ou conexões). Uma aresta e_k é um par (v_i, v_j) que representa uma conexão entre os vértices $v_i \in V$ e $v_j \in V$. Uma rede pode ser codificada em uma matriz de adjacência A de ordem $N \times N$, cuja posição $A(i, j)$ é igual a 1 se existe uma aresta direcionada do vértice i ao j . Caso contrário, $A(i, j)$ é igual a 0. Em uma rede não direcionada, uma conexão entre i e j implica $A(i, j) = A(j, i) = 1$, ou seja, a matriz A é necessariamente simétrica. Note que aqui não consideramos autoconexões (*loops*), de modo que $A(i, i) = 0$ para todo i . O grau* de entrada de um vértice i representa o número de conexões que terminam em i , e é dado por $k_i^{\text{in}} = \sum_j A(j, i)$. O grau de saída é o número de ligações que partem de i : $k_i^{\text{out}} = \sum_j A(i, j)$. Em redes não direcionadas o grau k_i de um vértice i é dado por $k_i = k_i^{\text{in}} = k_i^{\text{out}}$. Os valores máximos de grau são denotados por k_{max} , $k_{\text{max}}^{\text{in}}$ e $k_{\text{max}}^{\text{out}}$. A conectividade global de uma rede pode ser quantificada pelo grau médio $\langle k \rangle$ em redes não direcionadas ou pelas médias $\langle k^{\text{in}} \rangle$ ou $\langle k^{\text{out}} \rangle$ no caso direcionado – note que a média do grau de entrada é sempre igual à media do grau de saída.

Nas subseções seguintes, diversas métricas de redes complexas utilizadas neste projeto são definidas (93), bem como outros conceitos relevantes. As implementações foram feitas nas linguagens C e Scilab, sendo que C foi utilizada nos casos com grandes demandas de eficiência (como no cálculo da centralidade em redes grandes).

* Com frequência muito menor, grau é também chamado de conectividade por alguns autores (6). Grau é o termo derivado da teoria dos grafos.

2.1.1 Coeficiente de Agrupamento

Se dois vértices q e j estão ambos conectados a i , existe uma probabilidade de que q e j estejam também conectados entre si. Essa probabilidade pode ser quantificada pelo coeficiente de agrupamento cc_i , o qual reflete a densidade de conexões existentes entre os vizinhos de um vértice i . Considere o conjunto η_i dos vizinhos de i e o número de arestas ε_i que conectam os vizinhos de i entre si em uma rede não direcionada:

$$\eta_i = \{j \mid A(i,j) = A(j,i) = 1, i \neq j\} , \quad (2.1)$$

$$\varepsilon_i = \sum_{u,v \in \eta_i, u \neq v} A(u,v) / 2 . \quad (2.2)$$

O coeficiente de agrupamento de um vértice i é então definido como (12):

$$cc_i = \frac{2\varepsilon_i}{|\eta_i|(|\eta_i| - 1)} , \quad (2.3)$$

onde $|\eta_i|$ é a cardinalidade de η_i , ou seja, é o número de vizinhos de i . No caso especial em que o coeficiente médio da rede $\langle cc \rangle$ é zero, a rede é na verdade uma árvore, e se $cc_i = 1$ os vizinhos de i formam um clique, ou seja, formam um subgrafo completo.

Em redes direcionadas é possível aplicar as medidas definidas para redes não direcionadas desconsiderando-se as direções de suas arestas. Portanto, o coeficiente de agrupamento pode ser calculado em redes direcionadas de acordo com as explicações dadas acima (essa abordagem foi utilizada no Capítulo 4). Outra abordagem diferencia os vizinhos do vértice i entre os vizinhos de saída η_i^{out} e os de entrada η_i^{in} :

$$\eta_i^{\text{out}} = \{j \mid A(i,j) = 1, i \neq j\} , \quad (2.4)$$

$$\eta_i^{\text{in}} = \{j \mid A(j,i) = 1, i \neq j\} . \quad (2.5)$$

O número de arestas entre esses vizinhos é dado, respectivamente, por:

$$\varepsilon_i^{\text{out}} = \sum_{u,v \in \eta_i^{\text{out}}, u \neq v} A(u,v) , \quad (2.6)$$

$$\epsilon_i^{\text{in}} = \sum_{u,v \in \eta_i^{\text{in}}, u \neq v} A(u,v) . \quad (2.7)$$

Portanto, os coeficientes de agrupamento de saída e de entrada são obtidos, respectivamente, por:

$$cc_i^{\text{out}} = \frac{\epsilon_i^{\text{out}}}{|\eta_i^{\text{out}}| (|\eta_i^{\text{out}}| - 1)} , \quad (2.8)$$

$$cc_i^{\text{in}} = \frac{\epsilon_i^{\text{in}}}{|\eta_i^{\text{in}}| (|\eta_i^{\text{in}}| - 1)} . \quad (2.9)$$

Esses coeficientes de agrupamento que diferenciam o sentido das arestas em redes direcionadas foram utilizados no Capítulo 7.

2.1.2 Grau da Vizinhaça

O grau médio $k_i^{(nm)}$ da vizinhaça de um determinado vértice i pode ser calculada considerando-se a média dos graus k_j de todos os vizinhos j de i (16). Essa medida complementa o coeficiente de agrupamento ao considerar todas as conexões existentes na vizinhaça de um vértice, e não apenas as arestas que conectam os vizinhos entre si. O grau da vizinhaça para a rede toda é simplesmente tomado como a média entre todos os valores individuais, e é denotado por $\langle k^{(nm)} \rangle$.

2.1.3 Assortatividade

A assortatividade é outra medida relacionada à conectividade entre vizinhos, e mensura a correlação entre os graus dos vértices que compartilham uma conexão. Em outras palavras, o coeficiente de assortatividade indica se vértices de grau similar tendem a se conectar entre si (rede assortativa), ou se vértices com alto grau tendem a se conectar com vértices de baixo grau (rede não assortativa) (94). Esse coeficiente pode ser computado ao se aplicar o coeficiente de correlação de Pearson nos graus que aparecem nas extremidades de cada aresta:

$$r = \frac{\frac{1}{L} \sum_{j>i} k_i k_j A(i,j) - \left[\frac{1}{L} \sum_{j>i} \frac{1}{2} (k_i + k_j) A(i,j) \right]^2}{\frac{1}{L} \sum_{j>i} \frac{1}{2} (k_i^2 + k_j^2) A(i,j) - \left[\frac{1}{L} \sum_{j>i} \frac{1}{2} (k_i + k_j) A(i,j) \right]^2} , \quad (2.10)$$

onde L é o número total de arestas na rede. No caso $r \rightarrow 1$, a rede é assortativa, e se $r \rightarrow -1$, a rede é não assortativa. Quando $r \rightarrow 0$, não há correlação significativa entre os graus dos vértices. No caso de redes direcionadas, considerou-se aqui as correlações entre os graus de entrada (r^{ii} , *indegree-indegree*) e entre os graus de saída (r^{oo} , *outdegree-outdegree*).

2.1.4 Grau Hierárquico

A estrutura de uma rede pode ser quantificada ao longo de progressivos níveis mesoscópicos. Uma maneira de se realizar uma análise em escala intermediária é utilizando-se medidas hierárquicas (ou concêntricas), as quais são calculadas levando-se em consideração vizinhanças sucessivas em torno de cada vértice (95–97). Tais medidas hierárquicas são particularmente eficientes no fornecimento de informações adicionais a respeito da estrutura de uma rede, pois elas consideram um número maior de vizinhanças do que as medidas tradicionais. Um conceito essencial é o do anel $R_i^{(d)}$, o qual é formado por todos os vértices distantes $d - 1$ arestas do vértice de referência i . $R_i^{(d)}$ é também conhecido como nível hierárquico d do vértice i . O grau hierárquico no nível d , denotado por $k_i^{(d)}$, é então definido como o número de arestas que conectam os anéis $R_i^{(d)}$ e $R_i^{(d+1)}$ – note que o grau tradicional k_i é equivalente ao grau hierárquico $k_i^{(1)}$. Considerando a rede toda, o grau hierárquico no nível d é apenas a média de todos os graus hierárquicos da rede: $\langle k^{(d)} \rangle$.

No caso de redes direcionadas, pode-se considerar que as arestas não têm direção, e assim obter os níveis e graus hierárquicos como explicado acima (essa abordagem foi utilizada no Capítulo 4). Outra opção é considerar que os anéis sejam definidos pela direção das arestas: $R_i^{(d|out)}$ é então formado por todos os vértices j tal que o menor caminho que ligue i a j (agora desconsiderando-se o sentido inverso) tenha comprimento $d - 1$; e, analogamente, $R_i^{(d|in)}$ contém os vértices j tal que o menor caminho que ligue j a i tenha comprimento $d - 1$. O grau de saída hierárquico $k_i^{(d|out)}$ no nível d é então definido como o número de conexões que partem do nível $R_i^{(d|out)}$ para o nível $R_i^{(d+1|out)}$; e o grau de entrada hierárquico $k_i^{(d|in)}$ no nível d é, portanto, igual ao número de arestas que originam-se no nível $R_i^{(d+1|in)}$ e terminam no nível $R_i^{(d|in)}$. Essa

abordagem foi utilizada no Capítulo 7. Note que o grau de saída tradicional k_i^{out} é equivalente a $k_i^{(1|\text{out})}$, e o grau de entrada tradicional k_i^{in} é equivalente a $k_i^{(1|\text{in})}$.

2.1.5 Coeficiente de Agrupamento Hierárquico

Essa medida é uma generalização do coeficiente de agrupamento tradicional por meio do emprego de vizinhanças mais distantes. Esse coeficiente é dado pelo número de arestas $\varepsilon_i^{(d)}$ no respectivo anel $R_i^{(d)}$ dividido pelo número total de possíveis arestas que podem existir entre os vértices do referido anel (96). Formalmente, tem-se:

$$cc_i^{(d)} = \frac{2\varepsilon_i^{(d)}}{|R_i^{(d)}|(|R_i^{(d)}| - 1)}, \quad (2.11)$$

onde $|R_i^{(d)}|$ denota o número de vértices presentes no anel $R_i^{(d)}$. O coeficiente tradicional cc_i equivale ao hierárquico de nível 2, $cc_i^{(2)}$. Analogamente às outras métricas explicadas nesta seção, a média $\langle cc^{(d)} \rangle$ é tomada como o coeficiente de agrupamento hierárquico de toda a rede.

Em redes direcionadas, o coeficiente de agrupamento hierárquico pode ser computado simplesmente desconsiderando-se a direcionabilidade das arestas, conforme já discutido para outras medidas desta seção. Dessa maneira, a Equação 2.11 pode ser aplicada diretamente a redes direcionadas (abordagem utilizada no Capítulo 4). Outro método (utilizado no Capítulo 7) envolve a discriminação entre níveis hierárquicos de entrada e de saída. Nesse caso, tem-se, respectivamente:

$$cc_i^{(d|\text{out})} = \frac{\varepsilon_i^{(d|\text{out})}}{|R_i^{(d|\text{out})}|(|R_i^{(d|\text{out})}| - 1)}, \quad (2.12)$$

$$cc_i^{(d|\text{in})} = \frac{\varepsilon_i^{(d|\text{in})}}{|R_i^{(d|\text{in})}|(|R_i^{(d|\text{in})}| - 1)}, \quad (2.13)$$

onde $\varepsilon_i^{(d|\text{out})}$ e $\varepsilon_i^{(d|\text{in})}$ fornecem o número de arestas nos níveis hierárquicos $R_i^{(d|\text{out})}$ e $R_i^{(d|\text{in})}$, respectivamente, e $|R_i^{(d|\text{out})}|$ e $|R_i^{(d|\text{in})}|$ denotam o número de vértices nos níveis $R_i^{(d|\text{out})}$ e $R_i^{(d|\text{in})}$, respectivamente. Note que os coeficientes de agrupamento cc_i^{out} e cc_i^{in} , definidos na Seção 2.1.1,

são equivalentes aos coeficientes hierárquicos $cc_i^{(2|out)}$ e $cc_i^{(2|in)}$, respectivamente.

2.1.6 Taxa de Divergência

A taxa de divergência à distância d do vértice i corresponde à razão entre o número de vértices presentes no anel $R_i^{(d)}$ e o grau hierárquico de i no nível $d - 1$ (34):

$$dv_i^{(d)} = \frac{n_i^{(d)}}{k_i^{(d-1)}}. \quad (2.14)$$

Essa medida quantifica a conectividade internível da seguinte maneira: $dv_i^{(d)}$ aproxima-se de zero quando o número de arestas entrando no nível d (vindas de $d - 1$) é muito maior que o número de vértices nesse nível. A média $\langle dv^{(d)} \rangle$ fornece a taxa de divergência para toda a rede.

2.1.7 Taxa de Convergência

Essa medida é o inverso da taxa de divergência:

$$cv_i^{(d)} = \frac{1}{dv_i^{(d)}} = \frac{k_i^{(d-1)}}{n_i^{(d)}}. \quad (2.15)$$

Portanto, ela aproxima-se de zero quando o número de vértices no nível d é muito maior que o número de arestas vindas do nível anterior.

2.1.8 Grau Inter-anel

Outra maneira de se mensurar a conectividade entre níveis é utilizando-se o grau inter-anel, que é igual ao número médio de conexões do anel $R_i^{(d)}$ que o conectam ao anel $R_i^{(d+1)}$ (96),

$$er_i^{(d)} = \frac{k_i^{(d)}}{n_i^{(d)}}. \quad (2.16)$$

Em outras palavras, essa medida pode ser utilizada para verificar o número de arestas médio (por vértice) dentro do nível d que conecta os níveis d e $d + 1$. O grau inter-anel de uma rede é dado pela média $\langle er^{(d)} \rangle$ considerando-se todos os valores individuais de $er_i^{(d)}$.

2.1.9 Grau Intra-anel

Como no coeficiente de agrupamento hierárquico, o grau intra-anel considera arestas que conectam vértices dentro do mesmo nível hierárquico. Entretanto, o grau intra-anel é um índice diferente pois toma o número de arestas médio entre os vértices presentes no anel $R_i^{(d)}$ (96),

$$ar_i^{(d)} = \frac{2e_i^{(d)}}{n_i^{(d)}}. \quad (2.17)$$

Novamente, o grau intra-anel de uma rede é dado pela média $\langle ar^{(d)} \rangle$.

2.1.10 Comprimento do Caminho Mínimo

A proximidade entre vértices é usualmente quantificada em termos de caminhos mínimos. Um caminho p_{ij} que se estende do vértice i ao j é denotado pela sequência de vértices vizinhos:

$$p_{ij} = (v_1, v_2, \dots, v_l, v_{l+1}), \quad (2.18)$$

onde $A(v_i, v_{i+1}) = 1$, $v_1 = i$, $v_{l+1} = j$, e o comprimento do caminho é $|p_{ij}| = l$ (98). Note que o comprimento de um caminho é o número de arestas contidas ao longo de sua extensão. O comprimento do caminho *mínimo* entre dois vértices i e j é, portanto, dado por:

$$d_{ij} = \min_{i,j} \{|p_{ij}|\}, \quad (2.19)$$

que é a quantidade mínima de passos necessários para que se possa alcançar j partindo-se de i (ou vice-versa, no caso de redes não direcionadas). A média ℓ (12) dos comprimentos dos caminhos mínimos entre todos os pares de vértices de uma rede é calculada da seguinte

maneira:

$$\ell = \frac{\sum_{i=1}^N \sum_{j=1}^N d_{ij}}{N(N-1)}. \quad (2.20)$$

E a média dos comprimentos de todos os caminhos mínimos que partem de um vértice i em direção ao restante da rede é igual a:

$$\ell_i = \frac{\sum_{j=1}^N d_{ij}}{N-1}. \quad (2.21)$$

2.1.11 Eficiência Global

A eficiência na comunicação ϵ_{ij} entre dois vértices i e j pode ser interpretada como sendo inversamente proporcional ao comprimento do caminho mínimo d_{ij} que conecta esses dois vértices (99). Quando tal caminho não existe, ou seja, quando $d_{ij} \rightarrow \infty$, tem-se que $\epsilon_{ij} = 0$. Nesse caso, nenhuma informação pode ser trocada entre i e j . Vértices imediatamente vizinhos apresentam $\epsilon_{ij} = 1$, a maior eficiência possível. Seguindo essa ideia, a eficiência global de um grafo G pode ser definida como:

$$E_{global} = E(G) = \frac{\sum_{i \neq j} \epsilon_{ij}}{N(N-1)} = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{ij}}. \quad (2.22)$$

Note que a eficiência foi tomada como a média entre todas as $N(N-1)$ arestas possíveis, isto é, excluindo-se *loops*. Note que essa medida é naturalmente interpretada no caso de redes direcionadas.

2.1.12 Eficiência Local

Adicionalmente à eficiência global, uma medida de eficiência local também foi utilizada neste projeto (99). Nesse caso, a eficiência é calculada para um subgrafo G_i que contém os vizinhos (e as respectivas arestas) de um dado vértice i . Quando considera-se todos os vértices de uma

rede e seus respectivos subgrafos, a eficiência local pode ser calculada da seguinte maneira:

$$E_{local} = \frac{1}{N} \sum_{i=1}^N E(G_i). \quad (2.23)$$

Essa métrica quantifica o nível de tolerância a falhas de uma determinada rede, pois indica a eficiência de comunicação entre os vizinhos de i caso o vértice i seja removido.

2.1.13 Centralidade

Outra medida relacionada à distância é a centralidade (nesse caso, utilizamos a *betweenness centrality*), a qual pode ser calculada tanto para vértices quanto para arestas (100). A centralidade, em redes direcionadas ou não, é definida como:

$$B_i = \sum_{j \neq k} \frac{\sigma(j,i,k)}{\sigma(j,k)}, \quad (2.24)$$

onde $\sigma(j,i,k)$ é o número de caminhos mínimos *distintos* que conectam os vértices j e k necessariamente passando pelo vértice (ou aresta) i , e $\sigma(j,k)$ é o número total de caminhos mínimos que conectam j e k . Quando i é um vértice, B_i é aqui denotada por $B_i^{(n)}$; quando i é uma aresta, B_i é identificada por $B_i^{(e)}$. As médias $\langle B^{(n)} \rangle$ e $\langle B^{(e)} \rangle$ são aqui utilizadas como medidas de centralidade para toda a rede.

2.1.14 Dominância do Ponto Central

A centralidade de vértices dá origem a uma outra medida, chamada dominância do ponto central (100). Ela pode ser calculada da seguinte maneira:

$$CPD = \frac{1}{N-1} \sum_{i=1}^N (B_{\max}^{(n)} - B_i^{(n)}), \quad (2.25)$$

onde $B_{\max}^{(n)}$ é o máximo valor de centralidade da rede. Portanto, CPD (do inglês *Central Point Dominance*) é útil para se verificar o quão distante da máxima centralidade estão os demais

valores de centralidade da rede. CPD é igual a zero em grafos completos (ou seja, com todas as arestas possíveis), já que todos os vértices apresentam a mesma centralidade. Por outro lado, a dominância de um ponto central pode ser encontrada em, por exemplo, grafos do tipo “estrela”, os quais apresentam um único vértice central que participa da maior parte dos caminhos mínimos existentes.

2.1.15 Reciprocidade

A reciprocidade ρ mede a proporção de arestas bidirecionais (de i para j e vice-versa) que existem em uma rede direcionada (35), comparando-a com a proporção análoga em uma rede aleatória de mesmo número de vértices e arestas. A definição de reciprocidade é a seguinte:

$$\rho = \frac{L^{\leftrightarrow}/L - \bar{a}}{1 - \bar{a}}, \quad (2.26)$$

onde, excluindo-se *loops*, L é o número total de arestas direcionadas, $L^{\leftrightarrow}$ é o número de arestas que possuem uma aresta recíproca (no sentido inverso) e $\bar{a}$ é a razão entre L e o número máximo possível de arestas que pode existir na rede ($N(N-1)$). A reciprocidade permite-nos distinguir redes recíprocas ($\rho > 0$, mais conexões simétricas do que em uma rede aleatória) e antirrecíprocas ($\rho < 0$, menos conexões simétricas do que em uma rede aleatória). Se todas as arestas têm um par recíproco, então $\rho = 1$. É interessante também calcular a reciprocidade mínima possível em uma rede com mesmo número de vértices e arestas:

$$\rho_{\min} = -\frac{\bar{a}}{1 - \bar{a}}, \quad (2.27)$$

ou seja, $\rho_{\min}$ fornece a reciprocidade quando nenhum par de arestas recíprocas existe na rede.

2.1.16 Componente Conexo

Um componente conexo é um subgrafo que apresenta pelo menos um caminho conectando qualquer vértice i a qualquer outro vértice j , ou seja, não existem porções isoladas de vértices (7).

Quando o maior componente conexo é a própria rede, ela é dita conexa. Em redes direcionadas, existe a distinção entre componentes fortemente e fracamente conexos (7). O componente fortemente conexo deve apresentar pelo menos um caminho conectando qualquer i a qualquer j , e vice-versa (de j para i). Já no fracamente conexo é apenas exigido que não existam porções de vértices isolados.

2.2 Dinâmica do Passeio Aleatório Simples

Neste projeto foi escolhido o passeio aleatório simples como modelo de dinâmica em redes (20, 25–27). É um processo estocástico, mais especificamente, uma cadeia de Markov de tempo discreto, definida como uma sequência de variáveis aleatórias $\{X_t\}_{t=0}^{\infty}$ em que X_t pode assumir o valor de cada um dos elementos definidos no espaço de estados $S = \{s_1, s_2, \dots, s_N\}$. No caso do passeio aleatório em grafos, $S = V = \{v_1, v_2, \dots, v_N\}$, ou seja, o espaço de estados é igual ao conjunto de vértices do grafo. O passeio aleatório simples satisfaz a seguinte propriedade Markoviana:

$$\mathbb{P}(X_{t+1} = x_{t+1} \mid X_t = x_t, X_{t-1} = x_{t-1}, \dots, X_0 = x_0) = \mathbb{P}(X_{t+1} = x_{t+1} \mid X_t = x_t), \quad (2.28)$$

para todo $t \geq 0$, onde $x_1, x_2, \dots, x_t, x_{t+1} \in S$. Ou seja, a probabilidade de ocorrência do próximo movimento depende apenas do estado atual: é um processo “sem memória” (101). A matriz de transições P de ordem $N \times N$ descreve completamente a cadeia de Markov, e é definida como:

$$P(i, j) = \mathbb{P}(X_{t+1} = s_j \mid X_t = s_i), \quad (2.29)$$

onde $P(i, j)$ é a probabilidade de se mudar para o estado s_j dado que o estado atual é s_i (102). Essa matriz pode também ser definida da seguinte maneira:

$$P(i, j) = \frac{A(i, j)}{k_i^{\text{out}}}, \quad (2.30)$$

ou seja, a probabilidade de visitar o vértice j , dado que atualmente o processo está no vértice i , é igual a $A(i, j)$ normalizado pelo grau de saída de i . Lembre-se da estrutura da matriz de

adjacências A (Seção 2.1): se a aresta que liga i a j existe, $A(i,j) = 1$, caso contrário, $A(i,j) = 0$. P é uma matriz estocástica, pois satisfaz as seguintes propriedades: (i) $P(i,j) \geq 0$ para quaisquer i e j , e (ii) $\sum_{j=1}^N P(i,j) = 1$ para todo i .

Tome o vetor $\vec{p}_t = (p_1, p_2, \dots, p_N)$ como sendo a distribuição de probabilidades de se encontrar o processo em qualquer um dos estados no tempo t (p_i é a probabilidade para o estado s_i). Para se calcular a distribuição de probabilidades de ocupação no tempo $t + 1$, basta realizar a multiplicação $\vec{p}_{t+1} = \vec{p}_t P$. Neste projeto, estamos interessados no regime estacionário do passeio aleatório, quando o vetor de ocupação atinge o equilíbrio $\vec{p}_{t+1} = \vec{p}_t$. Nesse caso, $\vec{p}$ é denotado por $\vec{\pi}$. Portanto, no regime estacionário $\vec{\pi} = \vec{\pi} P$, ou seja, $\vec{\pi}$ é o autovetor de P associado ao autovalor 1 (24). Note que $\vec{\pi}$ converge para o mesmo valor independentemente da distribuição inicial $\vec{p}_0$.

Neste projeto, o termo *atividade* de i corresponde ao valor π_i , o qual pode ser interpretado como a chance de o passeio aleatório se encontrar no vértice i no regime estacionário. Em outras palavras, a atividade representa a taxa com que cada vértice é visitado na dinâmica do passeio aleatório (no regime estacionário). Por exemplo, quando um dado vértice exibe atividade igual a 0,1, tem-se que em 10% do tempo o processo encontra-se nesse vértice (ou que existe 10% de chance de que esse vértice seja o vértice atual do passeio). Essa interpretação geral pode ser adaptada às redes reais, como, por exemplo, ao caso da WWW (a atividade é a taxa de acessos a cada página), das redes corticais (frequência de pulsos cerebrais) e das redes de comunicação (intensidade de tráfego).

Para que a solução $\vec{\pi}$ seja única, a matriz de transições P deve ser irredutível, o que significa, no caso de um passeio aleatório em grafo, que ele deve ser conexo (ou fortemente conexo, no caso específico de grafos direcionados) (101). Para fins de normalização de probabilidades, utiliza-se aqui a solução que apresenta soma unitária $\sum_{i=1}^N \pi_i = 1$. O regime estacionário também pode ser denotado pelo limite $\pi_i = \lim_{t \rightarrow \infty} v_i/t$, onde v_i representa o número de visitas ao vértice i após t passos do passeio aleatório. Para que esse limite exista, o processo deve ser aperiódico (103). A aperiodicidade é garantida quando o maior divisor comum dos comprimentos de todos os ciclos presentes na cadeia de Markov (ou no grafo) é igual a 1 (104, 105) – um

exemplo periódico é o grafo bipartido. As condições de irreducibilidade e aperiodicidade foram satisfeitas nos experimentos aqui relatados. O algoritmo para o cômputo de $\vec{\pi}$ foi implementado em Scilab utilizando-se o método das potências.

No caso específico de grafos não direcionados, a solução de $\vec{\pi}$ pode ser obtida a partir dos graus dos vértices, $\pi_i = k_i / \sum_{j=1}^N k_j$ (25), pois

$$\begin{aligned}
 \vec{\pi}P &= \frac{1}{\sum_{j=1}^N k_j} \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_N \end{bmatrix}^T \begin{bmatrix} \frac{A(1,1)}{k_1} & \frac{A(1,2)}{k_1} & \cdots & \frac{A(1,N)}{k_1} \\ \frac{A(2,1)}{k_2} & \frac{A(2,2)}{k_2} & \cdots & \frac{A(2,N)}{k_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{A(N,1)}{k_N} & \frac{A(N,2)}{k_N} & \cdots & \frac{A(N,N)}{k_N} \end{bmatrix} = \\
 &= \frac{1}{\sum_{j=1}^N k_j} \begin{bmatrix} k_1 \frac{A(1,1)}{k_1} + k_2 \frac{A(2,1)}{k_2} + \cdots + k_N \frac{A(N,1)}{k_N} \\ k_1 \frac{A(1,2)}{k_1} + k_2 \frac{A(2,2)}{k_2} + \cdots + k_N \frac{A(N,2)}{k_N} \\ \vdots \\ k_1 \frac{A(1,N)}{k_1} + k_2 \frac{A(2,N)}{k_2} + \cdots + k_N \frac{A(N,N)}{k_N} \end{bmatrix}^T = \\
 &= \frac{1}{\sum_{j=1}^N k_j} \begin{bmatrix} \sum_{i=1}^N A(i,1) \\ \sum_{i=1}^N A(i,2) \\ \vdots \\ \sum_{i=1}^N A(i,N) \end{bmatrix}^T = \frac{1}{\sum_{j=1}^N k_j} \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_N \end{bmatrix}^T = \vec{\pi},
 \end{aligned}$$

onde T denota transposição. Portanto, essa solução de $\vec{\pi}$ satisfaz a condição do regime estacionário em redes não direcionadas. Em redes direcionadas, a solução análoga pode ser obtida se, para todo vértice i , $k_i^{\text{in}} = k_i^{\text{out}}$ (34).

2.3 Dinâmica de Acessibilidade

A medida de acessibilidade de um vértice é baseada na dinâmica do passeio aleatório (36). Define-se primeiramente a matriz de probabilidade P_h , cujas posições $P_h(i,j)$ indicam a proba-

bilidade de, partindo-se do vértice i , chegar ao vértice j em h passos. Note que nessa probabilidade considera-se um passeio aleatório sem repetição de vértices, em que não é permitido retornar a vértices já visitados – é, portanto, diferente do passeio aleatório simples. Por exemplo, na Figura 2.1 são exibidos dois grafos (A e B) de sete vértices cada, onde é possível observar os diferentes valores das probabilidades $P_h(i, j)$, para $h \leq 2$ e $i = 1$ (ou seja, partindo-se do vértice 1 em até 2 passos). Em ambos os grafos, o vértice 1 possui dois vizinhos, portanto, $P_1(1, j) = 1/2$ para qualquer vértice j que seja vizinho do vértice 1 ($j \in \{2, 4\}$ no grafo A e $j \in \{2, 5\}$ no grafo B). Quando $h = 2$, as possibilidades de acesso do vértice 1 mudam, apesar de existirem, em ambos os grafos, quatro vértices distantes dois passos do vértice inicial. No grafo B, os vértices 3, 4, 6 e 7 podem ser acessados a partir do vértice 1 com a mesma probabilidade ($1/4$); já no grafo A, as conexões no passo 2 são mais heterogêneas: o vértice 3 tem probabilidade de acesso igual a $1/2$, enquanto os vértices 5, 6 e 7 apresentam probabilidade $1/6$. Esse último caso ajuda a ilustrar de maneira sucinta o método de cálculo das probabilidades: se no passo h tem-se probabilidade de acesso $P_h(i, j)$ a um determinado vértice j , então a probabilidade de acesso a um vizinho não visitado k de j será igual a $P_{h+1}(i, k) = P_h(i, j) / \eta_j$, onde η_j é o número de vizinhos não visitados de j . Neste projeto, foi implementado em linguagem C um algoritmo exato e determinístico para o cômputo da matriz P_h , o qual baseia-se no método de percurso de grafos em profundidade (7). Para redes com mais de 1.000 vértices, empregou-se um algoritmo aproximado* em que um grande número de passeios aleatórios é simulado, para a seguir as probabilidades serem estimadas a partir do número de visitas a cada vértice (36).

A acessibilidade $Ac_i^{(h)}$ do vértice i no passo h é calculada a partir da matriz P_h da seguinte maneira (36):

$$Ac_i^{(h)} = -\frac{1}{\log(N-1)} \sum_{j=1, j \neq i}^N P_h(i, j) \log(P_h(i, j)) . \quad (2.31)$$

Note que a acessibilidade é uma medida de entropia das probabilidades de acesso a partir do vértice i em um número fixo de passos h . A acessibilidade do vértice 1 nos grafos A e B da Figura 2.1 é igual a 0,39, considerando-se $h = 1$. As acessibilidades são iguais pois em ambos os grafos o vértice 1 pode acessar seus vizinhos com as mesmas probabilidades. Já quando

* Implementado por Matheus Palhares Viana.

Figura 2.1 – Exemplo de cálculo da acessibilidade em dois grafos A e B com mesmo número de vértices. As probabilidades $P_h(i, j)$ são calculadas a partir de $i = 1$. As acessibilidades no grafo A são iguais a $Ac_1^{(1)} = 0,39$ e $Ac_1^{(2)} = 0,69$. Já no grafo B são iguais a $Ac_1^{(1)} = 0,39$ e $Ac_1^{(2)} = 0,77$.

$h = 2$, existe uma maior probabilidade de acesso ao vértice 3 do que aos vértices 5, 6, e 7 (grafo A). Ou seja, existe um desbalanço de acesso. No grafo B, a probabilidade de acesso é a mesma para os quatro vértices 3, 4, 6 e 7. Portanto, a acessibilidade do vértice 1 à distância $h = 2$ é maior no grafo B (0,77) do que no grafo A (0,69), pois no grafo B existe uma maior uniformidade de acesso aos vizinhos distantes 2 passos do vértice de partida.

A acessibilidade, além de ser uma medida relacionada a um processo dinâmico, pode ser associada ao conceito de *borda* de redes, em oposição ao *núcleo* (36). Vértices no núcleo tendem a apresentar maior acessibilidade do que os vértices posicionados na periferia da rede. Quando a acessibilidade é pequena, o vértice de partida está na borda e tende a acessar uma porção da rede de maneira desbalanceada, ou seja, alguns vértices têm maior probabilidade de serem acessados do que outros. Já quando a acessibilidade de um vértice é grande, ele está no núcleo da rede, e apresenta possibilidades de acesso a um (relativamente) grande número de vértices de maneira balanceada.

2.4 Dinâmica de Transporte

Em um processo de transporte realizado em grafos, dados são movidos de um ponto a outro da rede por meio de sua estrutura de vértices e arestas. Usualmente, um processo desse tipo envolve vértices de origem (os que produzem os dados) e vértices de destino (os que recebem e utilizam os dados). Os vértices geralmente atuam simultaneamente como origens e destinos. Uma característica importante da dinâmica de tráfego é o método de roteamento dos dados, o qual define para qual vértice os dados devem ser repassados sendo que atualmente estejam alocados em um vértice específico. Neste projeto, foi implementado em linguagem C um modelo de transporte inspirado na transmissão de pacotes de dados em redes de computadores (106). Entretanto, o modelo pode ser considerado um modelo geral e, portanto, pode ser interpretado como uma aproximação para outros casos, como transporte viário ou transmissão de energia elétrica.

A cada passo de tempo discreto t existe uma probabilidade λ de cada vértice gerar um pacote. O destino do pacote é escolhido de maneira uniformemente aleatória, com a restrição de que seja diferente da origem. Existe uma fila de pacotes em cada vértice, sendo que quando um pacote é criado, ele é colocado no final da fila do respectivo vértice criador. Um vértice i tem uma capacidade de processamento de pacotes C_i proporcional ao seu grau k_i :

$$C_i = 1 + \text{int}(\beta k_i) . \quad (2.32)$$

onde β é um parâmetro de controle variando entre 0 e 1, e int é uma função que retorna a parte inteira de um número real. Portanto, cada vértice i processa os primeiros C_i pacotes de sua fila em cada passo de tempo. Os pacotes processados são removidos da fila e são repassados a algum vértice vizinho de acordo com o algoritmo de roteamento, ou são removidos do tráfego se já atingiram seu destino. Quando um vértice recebe um pacote proveniente de um de seus vizinhos, o pacote é adicionado ao fim da sua fila de processamento. O roteamento é realizado por meio do caminho mínimo existente entre a origem e o destino. Se, em um determinado passo de tempo, existir mais de uma opção de roteamento para um dado pacote (ou seja, se

existir mais de um caminho de comprimento mínimo), o pacote é repassado ao vizinho que pertença a um caminho mínimo e que apresente o menor número de pacotes em sua fila de processamento.

O tempo necessário para que um pacote atinja seu destino depende do comprimento do caminho a ser percorrido e também do número de pacotes já posicionados ao longo do caminho – em outras palavras, depende fortemente do nível de congestionamento na rede. Para valores pequenos de λ , o número de pacotes existentes na rede $n(t)$ em determinado instante atinge, após um regime transiente, um regime de flutuação em torno de um valor constante (vide exemplos na Figura 2.2). Em outras palavras, a taxa de criação de pacotes é aproximadamente igual à taxa de entrega de pacotes. A partir de um λ crítico (identificado por λ_c) o número de pacotes $n(t)$ sempre cresce ao longo do tempo, o que indica iminência de congestionamento (Figura 2.2). Quando a fila de processamento em cada vértice tem tamanho máximo limitado, o congestionamento pode provocar falhas em vértices que tenham esse limite ultrapassado, podendo até levar a falhas em cascata na rede. Esse não é o caso do modelo utilizado neste projeto, em que o tamanho das filas é (teoricamente) ilimitado. Portanto, o parâmetro aqui estudado é $n(t)$ e sua variação ao longo do tempo para diferentes λ , o que permite identificar a taxa crítica λ_c de congestionamento.

2.5 Dinâmica de Sincronização

Considera-se aqui o modelo de Kuramoto com osciladores posicionados em cada vértice do grafo (19). As arestas, consequentemente, representam a existência de acoplamentos entre os osciladores. Os estados de todos os osciladores i (ou seja, cada fase $\phi_i(t)$ no tempo t) são descritos pelo conjunto de N equações diferenciais que definem a dinâmica do sistema ($1 \leq i \leq N$):

$$\dot{\phi}_i(t) + \frac{K}{\langle k \rangle} \sum_{j \in \eta_i} \sin(\phi_i(t) - \phi_j(t)) = \omega_i, \quad (2.33)$$

Figura 2.2 – Número de pacotes de dados $n(t)$ ao longo do tempo t na dinâmica de transporte realizada em uma rede pequeno mundo (Seção 2.6.2) de $N = 1.000$ e $\langle k \rangle = 6$. Diferentes taxas de geração de pacotes λ foram utilizadas, sendo que as curvas em vermelho indicam estado de congestionamento, e as curvas em preto indicam taxa de geração de pacotes aproximadamente igual à taxa de entrega de pacotes.

onde η_i é o conjunto dos vizinhos do vértice i , $\langle k \rangle$ é o grau médio da rede e K é a intensidade de acoplamento do sistema. Já ω_i representa a frequência natural de oscilação no vértice i , definida aleatoriamente seguindo uma distribuição Gaussiana de variância unitária (107). O objetivo é avaliar o comportamento global de sincronia na rede ao longo do tempo, dado um fator de acoplamento K . Foi utilizado o método de Heun para resolver o conjunto de Equações 2.33 numericamente em passos de tempo $\Delta t = 0,05$ (107).

O estado coletivo de sincronia dos osciladores é descrito pela equação:

$$r_{\text{sync}}(t) = \left| \frac{1}{N} \sum_{j=1}^N e^{i\phi_j(t)} \right| = \left| \frac{1}{N} \sum_{j=1}^N [\cos(\phi_j(t)) + i \sin(\phi_j(t))] \right|, \quad (2.34)$$

onde o exponencial de um número complexo segue a formula de Euler ($e^{ix} = \cos(x) + i \sin(x)$), e r_{sync} indica o nível da sincronização de todas as fases ϕ_j da rede no tempo t , ou seja, é um indicador macroscópico do nível de sincronia entre os vértices da rede. O modelo de sincronização de Kuramoto foi implementado na linguagem Python*, tendo sido utilizado neste projeto

* Implementado por João Luiz Bunoro Batista.

da seguinte maneira: dado um fator de acoplamento K , o nível global de sincronia r_{sync} foi avaliado ao longo do tempo em diferentes redes.

2.6 Modelos de Redes

A implementação de cada um dos modelos utilizados foi feita em linguagem Scilab. Todos são modelos estocásticos que, partindo de um conjunto de parâmetros inicial, levam à geração de uma rede com características peculiares. Essas características são especificadas a seguir.

2.6.1 Erdős-Rényi (ER)

Nesse modelo, cada par de vértices é conectado com probabilidade uniforme p (3, 103). Redes aleatórias apresentam uma distribuição de graus do tipo Poisson, apresentando portanto um grau médio característico, ou seja, os valores dos graus não fogem demasiadamente de $\langle k \rangle$. Uma rede ER tem grau médio dado por $\langle k \rangle = p(N - 1)$ no limite $N \rightarrow \infty$.

2.6.2 Watts-Strogatz (WS)

O modelo WS foi criado para reproduzir as características das redes pequeno mundo: alto coeficiente de agrupamento e baixo comprimento dos caminhos mínimos (12). Para se obter uma rede WS, inicia-se com um grafo circular de N vértices onde cada vértice é conectado aos seus κ vizinhos imediatos em ambas as direções. A seguir, uma ponta de cada aresta é movida para outro vértice com probabilidade p . Se $p \rightarrow 0$, a rede mantém-se próxima da rede regular inicial; se $p \rightarrow 1$, ela torna-se uma rede aleatória. Em nossos experimentos, utilizamos $p = 0.2$. Por fim, o grau médio de uma rede WS é dado por $\langle k \rangle = 2\kappa$.

2.6.3 Barabási-Albert (BA)

Uma rede BA começa com um conjunto de m_0 vértices e, a cada passo, a rede cresce com a adição de um novo vértice com m arestas. Os vértices que recebem as novas arestas são escolhidos por meio de uma regra linear de ligação preferencial. A probabilidade de um novo vértice i conectar-se a um vértice preexistente j é proporcional ao grau de j : $\mathbb{P}(i \leftrightarrow j) = k_j / \sum_u k_u$ (13). Uma rede BA apresenta sua distribuição de graus na forma de uma lei de potência. Essa distribuição mostra que a rede contém *hubs* (poucos vértices com alto grau) coexistindo com um grande número de vértices pouco conectados. O grau médio de uma rede BA é $\langle k \rangle = 2m$.

2.6.4 Geográfico (GG)

Um modelo para redes geográficas considera a posição espacial dos vértices a fim de estabelecer conexões entre eles. Mais especificamente, vértices próximos uns dos outros tendem a conectar-se, enquanto vértices distantes apresentam menores chances de se conectar (93). Note que essas estruturas são também conhecidas por redes espaciais (108). Apesar de diversos modelos para redes geográficas existirem, propusemos um modelo simples que funciona da seguinte maneira: vértices são distribuídos aleatoriamente em um plano finito de tamanho $L \times L$, e arestas são criadas entre vértices separados por uma distância menor ou igual a R . O grau médio de uma rede geográfica é aproximado por $\langle k \rangle \approx \pi R^2 N / L^2$.

2.7 Redes Reais

Nesta seção são descritas diversas redes reais coletadas por terceiros e que foram utilizadas neste projeto. As Tabelas 2.1 e 2.2 contêm informações básicas a respeito de cada uma dessas redes.

2.7.1 Redes Sociais

A rede de colaboração NetScience codifica as relações de coautoria entre cientistas que publicaram artigos na área de redes complexas (109). Foi compilada utilizando-se as referências citadas em dois artigos de revisão da área (5, 6), mais algumas referências manualmente incluí-

Tabela 2.1 – Redes reais utilizadas neste projeto em ordem crescente de tamanho. Para cada rede, são exibidos: a seção que contém sua descrição, seu número de vértices N , número de arestas L e grau médio $\langle k \rangle$. Redes direcionadas são marcadas com asterisco (*) ao lado do respectivo nome. A abreviação “max. comp.” significa “maior componente conexo”, e “metab.” representa “metabolismo”.

Rede	Seção	N	L	$\langle k \rangle$
*Macaco’30	2.7.5	30	311	10,37
*Macaco’47	2.7.5	47	505	10,74
*Gato’52	2.7.5	52	818	15,73
*Gato’53	2.7.5	53	826	15,58
*Macaco’70	2.7.5	70	745	10,64
*Macaco’85	2.7.5	85	2.356	27,72
*Gato’95	2.7.5	95	2.126	22,38
*Kyeburn	2.7.5	98	629	6,42
*Canton	2.7.5	110	713	6,48
* <i>C. elegans</i>	2.7.5	239	1.912	8,00
NetScience	2.7.1	379	914	4,82
* <i>E. coli</i> MSFN	2.7.5	437	397	0,91
<i>E. coli</i> (metab.)	2.7.5	763	2.412	6,32
*Blogs	2.7.4	793	15.783	19,90
*Roget (max. comp.)	2.7.2	904	4.831	5,34
*Roget	2.7.2	1.022	5.075	4,97
E-mails	2.7.1	1.133	5.451	9,62
*UCI-Students	2.7.1	1.294	19.026	14,70
* <i>E. coli</i> TRN	2.7.5	1.521	3.529	2,32
*ODLIS	2.7.2	1.802	14.166	7,86
<i>E. coli</i> PPIN	2.7.5	1.930	9.720	10,07
*P2P’8	2.7.3	2.068	9.313	4,50
*P2P’9	2.7.3	2.624	10.776	4,11
*P2P’5	2.7.3	3.234	13.453	4,16
*USFFAN	2.7.2	4.845	61.563	12,71
Power Grid	2.7.3	4.941	6.594	2,67
*EAT	2.7.2	7.754	247.172	31,88
*Massey	2.7.4	10.810	158.102	14,63
*FOLDLOC	2.7.2	13.274	119.485	9,00
*Flinders	2.7.4	13.579	152.273	11,21
*Nottingham	2.7.4	22.523	270.751	12,02
Internet	2.7.3	22.963	48.436	4,22
*Epinions	2.7.1	32.223	443.506	13,76

das. Cada cientista é representado(a) por um vértice na rede, e uma aresta não direcionada é criada entre dois cientistas que publicaram pelo menos um artigo em conjunto. NetScience possui 1.589 vértices, dos quais 379 encontram-se no maior componente conexo (porção utilizada em nossos experimentos). Essa rede apresenta originalmente pesos em suas arestas (frequência de coautoria), os quais não foram aqui considerados.

A rede social Epinions (www.epinions.com) contém críticas escritas por consumidores a

Tabela 2.2 – Disponibilidade das redes reais utilizadas neste projeto (endereço WWW de coleta). As redes estão dispostas em ordem alfabética. A abreviação “max. comp.” significa “maior componente conexo”, e “metab.” representa “metabolismo”.

Rede	Disponível em (acesso em 09/2011)
Blogs	http://www-personal.umich.edu/~mejn/netdata
<i>C. elegans</i>	http://cdg.columbia.edu/cdg/datasets (<i>endereço desativado após coleta</i>)
Canton	http://www.nceas.ucsb.edu/interactionweb/html/thomps_towns.html
<i>E. coli</i> (metab.)	http://www.nd.edu/~networks/resources.htm
<i>E. coli</i> MSFN	Obtida por comunicação pessoal
<i>E. coli</i> PPIN	Obtida por comunicação pessoal
<i>E. coli</i> TRN	Obtida por comunicação pessoal
E-mails	http://deim.urv.cat/~aarenas/data/welcome.htm
EAT	http://vlado.fmf.uni-lj.si/pub/networks/data/dic/eat/Eat.htm
Epinions	http://snap.stanford.edu/data/soc-Epinions1.html
Flinders	http://cybermetrics.wlv.ac.uk/database
FOLDDOC	http://vlado.fmf.uni-lj.si/pub/networks/data/dic/foldoc/foldoc.htm
Gato'52	http://sites.google.com/a/brain-connectivity-toolbox.net/bct/datasets
Gato'53	Obtida por comunicação pessoal
Gato'95	http://sites.google.com/a/brain-connectivity-toolbox.net/bct/datasets
Internet	http://www-personal.umich.edu/~mejn/netdata
Kyeburn	http://www.nceas.ucsb.edu/interactionweb/html/thomps_towns.html
Macaco'30	http://sites.google.com/a/brain-connectivity-toolbox.net/bct/datasets
Macaco'47	http://sites.google.com/a/brain-connectivity-toolbox.net/bct/datasets
Macaco'70	http://sites.google.com/a/brain-connectivity-toolbox.net/bct/datasets
Macaco'85	http://www.biological-networks.org/?page_id=25
Massey	http://cybermetrics.wlv.ac.uk/database
NetScience	http://www-personal.umich.edu/~mejn/netdata
Nottingham	http://cybermetrics.wlv.ac.uk/database
ODLIS	http://vlado.fmf.uni-lj.si/pub/networks/data/dic/odlis/Odlis.htm
P2P'5	http://snap.stanford.edu/data/index.html#p2p
P2P'8	http://snap.stanford.edu/data/index.html#p2p
P2P'9	http://snap.stanford.edu/data/index.html#p2p
Power Grid	http://cdg.columbia.edu/cdg/datasets (<i>endereço desativado após coleta</i>)
Roget	http://vlado.fmf.uni-lj.si/pub/networks/data/dic/roget/Roget.htm
Roget (max. comp.)	http://vlado.fmf.uni-lj.si/pub/networks/data/dic/roget/Roget.htm
UCI-Students	http://toreopsahl.com/datasets/#online_social_network
USFFAN	http://vlado.fmf.uni-lj.si/pub/networks/data/dic/fa/FreeAssoc.htm

respeito de diversos produtos disponíveis no mercado. Os consumidores cadastrados têm a possibilidade de indicar outros usuários cujas críticas consideram confiáveis, formando portanto uma “rede de confiança” (110). A escolha de quais críticas devem ser exibidas a um determinado consumidor são então baseadas nessas relações de confiança. O maior componente fortemente conexo dessa rede (que é direcionada) foi empregado nesta tese, o qual contém $N = 32.223$ vértices.

A rede UCI-Students representa a estrutura de trocas de mensagens efetuadas pelos estudantes da Universidade da Califórnia, Irvine, EUA, de abril a outubro de 2004, utilizando uma plataforma *online* de interação social (111). As arestas representam a direção das mensagens: uma ou mais mensagens enviadas pelo estudante i ao j resulta na criação de uma aresta direcionada $i \rightarrow j$. Os pesos das arestas não foram aqui considerados, os quais representam o total de mensagens enviadas de um estudante a outro. O maior componente fortemente conexo dessa rede ($N = 1.294$) foi empregado nesta tese.

O fluxo de mensagens eletrônicas trocadas entre os membros da Universitat Rovira i Virgili (Espanha) foi codificado em uma rede (112), a qual é aqui identificada simplesmente pelo nome *E-mails*. Ela possui um único componente conexo em que cada endereço de *e-mail* é identificado por um vértice (1.133 no total), e cada mensagem enviada do endereço i ao endereço j é representada por uma aresta não direcionada.

2.7.2 Redes Semânticas

O tesouro (do inglês *thesaurus*) criado por Peter M. Roget serviu de base para a compilação dessa rede. A versão aqui utilizada é a de 1979, editada pelo seu filho John L. Roget. Cada vértice dessa rede representa uma das 1.022 categorias de conceitos contidas no tesouro, e conexões direcionadas são criadas para categorias utilizadas na definição de outras categorias (j recebe uma conexão de i se j é utilizado na definição de i). Outro critério para conexão é o relacionamento direto entre categorias definido por Roget em sua obra por meio da posição das

categorias no texto. O tesauro Roget já foi utilizado, por exemplo, em (113). Adicionalmente, o maior componente dessa rede, com $N = 904$ vértices, foi empregado nas análises do Capítulo 3.

O EAT (Edinburgh Associative Thesaurus) é formado por um conjunto de palavras associadas semanticamente. Ao contrário do tesauro Roget, as associações no EAT não foram criadas de maneira controlada, e sim de maneira livre por um grupo de 100 pessoas (114). Uma palavra foi exibida a um indivíduo (estímulo), que então forneceu outra palavra (resposta) que tenha vindo à sua mente logo após receber o estímulo. Dessa maneira, constrói-se uma rede em que arestas são criadas dos estímulos para as respostas. O maior componente fortemente conexo dessa rede apresenta $N = 7.754$ vértices, formando portanto a porção utilizada nesta tese. Outra rede de associação de palavras aqui utilizada é a USFFAN (University of South Florida Free Association Norms) (115). Seu maior componente fortemente conexo apresenta $N = 4.845$ palavras, e foi construída de maneira similar à rede EAT.

FOLDOC (Free On-line Dictionary of Computing) é um dicionário, mantido desde 1985, com definições de termos relacionados à computação (116). Cada termo é representado por um vértice, e conexões são criadas do vértice i ao j se o termo j for utilizado na definição do termo i . Foi empregado neste projeto o maior componente conexo dessa rede, com $N = 13.274$ vértices.

O dicionário ODLIS (Online Dictionary for Library and Information Science) é uma fonte de referência para bibliotecários e profissionais da ciência da informação (117). Os termos foram incluídos no dicionário de acordo com a respectiva importância ou utilidade para a área dos referidos profissionais. A rede proveniente desse dicionário codifica relações entre termos, sendo que existe uma ligação do termo i ao j se o termo j for utilizado na descrição do termo i . O maior componente fortemente conexo dessa rede foi aqui utilizado, com $N = 1.802$ vértices.

2.7.3 Redes de Infraestrutura

As conexões que associam os sistemas autônomos da Internet foram consideradas na construção dessa rede. Um sistema autônomo é um grupo de redes de computadores que compartilham uma

dada política de roteamento e que possuem administração centralizada. As conexões foram coletadas por meio do protocolo BGP (Border Gateway Protocol) em 22 de julho de 2006. Foram identificados 22.963 sistemas autônomos formando uma rede conexa e não direcionada.

As conexões entre clientes e servidores na rede Gnutella de compartilhamento de arquivos foram coletadas, sendo que, a cada procedimento de coleta, um grafo foi gerado (118). Cada *host* representa um vértice e arestas direcionadas representam a conexão dos servidores para os clientes. Foram utilizados três grafos desse tipo: P2P'5, coletado em 5 de agosto de 2002; P2P'8, coletado em 8 de agosto de 2002; e P2P'9, coletado em 9 de agosto de 2002, onde P2P significa *peer-to-peer*. O tamanho de cada uma é, respectivamente, $N = 3.234$, $N = 2.068$ e $N = 2.624$. Somente os maiores componentes fortemente conexos foram considerados em cada caso.

A rede Power Grid representa a conectividade entre elementos da rede elétrica do oeste dos EUA, considerando os caminhos que vão desde as usinas até os consumidores (12). Nesse caso, a rede é conexa com 4.941 vértices interconectados por arestas não direcionadas.

2.7.4 World Wide Web

Redes obtidas a partir de páginas Web foram utilizadas, sendo que cada uma delas refere-se a um domínio específico. São eles: www.massey.ac.nz, www.flinders.edu.au e www.nottingham.ac.uk, identificados a partir deste ponto apenas por Massey, Flinders e Nottingham, respectivamente. As páginas localizadas dentro de cada domínio representam os vértices da respectiva rede, e seus *hyperlinks* representam as arestas direcionadas que apontam da página origem para a página destino (119). Essas redes formam componentes fortemente conexos de tamanhos $N = 10.810$ (Massey), $N = 13.579$ (Flinders) e $N = 22.523$ (Nottingham).

Na rede de *blogs*, cada *blog* representa um vértice, e os *hyperlinks* criados entre eles (somente os da página principal foram considerados) definem as arestas (direcionadas) de um *blog* a outro (120). Os *blogs* são limitados aos que tratam de assuntos políticos, mais especifica-

mente, os que apresentam conteúdo sobre a eleição presidencial dos EUA em 2004. Essa rede (maior componente) codifica as interligações entre $N = 793$ blogs.

2.7.5 Redes Biológicas

Foram utilizadas duas redes que representam cadeias alimentares em ambientes distintos da Nova Zelândia (Canton e Kyeburn) (121). As relações tróficas entre as espécies (vértices) foram armazenadas na forma de arestas direcionadas nas redes Canton e Kyeburn, que apresentam 110 e 98 vértices, respectivamente.

Foram empregadas neste projeto diversas redes corticais de conectividade em larga escala, em que vértices representam áreas corticais e arestas representam tratos córtico-corticais. Quatro redes referem-se ao macaco do gênero *Macaca* da família Cercopithecidae: (i) uma considera a conectividade entre 70 regiões corticais (122), (ii) outra apresenta a conectividade em apenas um hemisfério, com 85 regiões mapeadas (123), (iii) outra inclui apenas o córtex visual, com 30 regiões consideradas (124) e (iv) outra considera 47 regiões das áreas visual e sensorio-motora (50). Essas redes são identificadas, respectivamente, pelos termos macaco'70, macaco'85, macaco'30 e macaco'47. As outras redes são provenientes do córtex do gato: (i) uma que considera 52 áreas corticais (49), (ii) outra que mapeia 53 áreas corticais (125) e (iii) uma terceira que contém todas as áreas corticais e talâmicas (95 no total) (49). Essas redes são aqui identificadas por gato'52, gato'53 e gato'95, respectivamente. A última rede refere-se ao verme *Caenorhabditis elegans*. É uma estrutura coletada a nível neuronal, ou seja, cada vértice representa um neurônio (126). Uma aresta nessa rede conecta dois neurônios se existir uma sinapse ou uma junção comunicante entre eles. Em todos os casos de redes corticais/neuronais empregadas neste projeto, foram utilizados os maiores componentes fortemente conexos.

A rede de regulação de transcrição (TRN, do inglês *Transcriptional Regulatory Network*) possui arestas direcionadas que representam fatores de transcrição (TF's, do inglês *Transcrip-*

tion Factors) que regulam genes que por sua vez codificam proteínas na bactéria *Escherichia coli*. Os TF's contém regiões do DNA (sequências de aminoácidos conhecidas) que podem funcionar como regiões ativadoras ou repressoras da transcrição. Essa rede foi obtida a partir do banco de dados RegulonDB (74). Note que, por definição, alguns genes que não codificam para TF's não regulam outros genes e são chamados de genes-alvo (TG's, do inglês *Target Genes*). Essa rede possui $N = 1.521$ vértices (1.352 TG's, 169 TF's).

Outra rede da *E. coli* considerada neste projeto é a de vias metabólicas e de sinalização (MSFN, do inglês *Metabolic and Signaling Feedback Network*) (88), uma rede direcionada com $N = 437$ vértices. Essa rede é formada por genes sinalizadores que possuem a habilidade de produzir sinais celulares (metabólitos transportados do exterior da célula ou produzidos dentro da célula) e também de fosforilar um regulador de resposta – a fosforilação é a adição de uma molécula de fósforo a uma proteína ou outra molécula. A maioria das proteínas reguladoras possui ligação com o DNA quando estão fosforiladas, funcionando então como fatores de transcrição que levam à indução ou repressão da transcrição de genes específicos. Já os genes sinalizadores são enzimas, transportadores ou histidinas quinases responsáveis por produzir sinais (histidinas quinases são proteínas “sensoras” que servem como receptores de sinais externos e que os retransmitem para as proteínas reguladoras de resposta). Dessa maneira, os genes da rede MSFN são responsáveis por modular a atividade dos TF's.

O conjunto de reações bioquímicas que constitui o metabolismo da *E. coli* foi considerado em outra rede. Os substratos são representados pelos vértices, e arestas (não direcionadas) são criadas entre todos os edutos de uma reação e seus respectivos produtos (81). Como essa rede não estava disponível publicamente, foi utilizada uma rede similar que apresenta direção nas arestas (dos edutos para os produtos) e outros vértices representando elementos adicionais tais como enzimas. Removeu-se portanto todos os vértices adicionais e a direção das arestas, para que fosse possível comparar os resultados relatados no Capítulo 5 (análise de subgrafos) com outros previamente divulgados na literatura. Obteve-se uma rede com $N = 763$ vértices, bastante similar à utilizada em (81).

Outro tipo de rede aqui utilizado é o das interações proteína-proteína (PPIN, do inglês *Protein-Protein Interaction Network*) que ocorrem na *E. coli* (89): genes que codificam proteínas, as quais interagem entre si, têm seus vértices conectados na rede. Essas interações ocorrem quando duas ou mais proteínas são acopladas a fim de executar uma determinada função biológica. Diversos processos celulares, como os de ação enzimática, são regidos por grandes estruturas moleculares formadas por um grande número de proteínas, as quais são organizadas por interações proteína-proteína. Essa rede é não direcionada e apresenta $N = 1.930$ vértices.

2.8 Análise de Componentes Principais

Uma técnica bastante conhecida de redução de dimensionalidade é a análise de componentes principais (PCA, do inglês *Principal Components Analysis*) (127). Para cada rede/gráfico n são computadas M medidas e um vetor de atributos $\vec{f}_n = (f_n^1, f_n^2, \dots, f_n^M)$ contendo essas medidas é formado para posterior análise de componentes principais. O método PCA então assume como entrada uma matriz $R \times M$ (R é o número total de vetores/redes) tal que:

$$\Omega = \begin{bmatrix} \leftarrow \vec{f}_1 \rightarrow \\ \leftarrow \vec{f}_2 \rightarrow \\ \vdots \\ \leftarrow \vec{f}_R \rightarrow \end{bmatrix} = \begin{bmatrix} f_1^1 & f_1^2 & \cdots & f_1^M \\ f_2^1 & f_2^2 & \cdots & f_2^M \\ \vdots & \vdots & \ddots & \vdots \\ f_R^1 & f_R^2 & \cdots & f_R^M \end{bmatrix}. \quad (2.35)$$

A seguir, de cada dimensão j é subtraída a respectiva média entre todos os vetores existentes $\langle f^j \rangle = \sum_{i=1}^R f_i^j / R$, resultando em uma nova matriz Ω_0 :

$$\Omega_0 = \begin{bmatrix} (f_1^1 - \langle f^1 \rangle) & (f_1^2 - \langle f^2 \rangle) & \cdots & (f_1^M - \langle f^M \rangle) \\ (f_2^1 - \langle f^1 \rangle) & (f_2^2 - \langle f^2 \rangle) & \cdots & (f_2^M - \langle f^M \rangle) \\ \vdots & \vdots & \ddots & \vdots \\ (f_R^1 - \langle f^1 \rangle) & (f_R^2 - \langle f^2 \rangle) & \cdots & (f_R^M - \langle f^M \rangle) \end{bmatrix}. \quad (2.36)$$

Note que a média de cada nova coluna é zero, ou seja, o vetor média é movido para a origem do espaço M -dimensional. Os autovalores e autovetores da matriz de covariância Σ de Ω_0 são então computados. Considere a seguir que os autovalores $\lambda_1, \lambda_2, \dots, \lambda_M$ estão arranjados em ordem decrescente, e os respectivos autovetores $\vec{\gamma}_1, \vec{\gamma}_2, \dots, \vec{\gamma}_M$ estejam arranjados em uma nova matriz da seguinte maneira:

$$\Gamma = \begin{bmatrix} \leftarrow \vec{\gamma}_1 \rightarrow \\ \leftarrow \vec{\gamma}_2 \rightarrow \\ \vdots \\ \leftarrow \vec{\gamma}_M \rightarrow \end{bmatrix}. \quad (2.37)$$

Por fim, a projeção PCA é realizada aplicando-se a seguinte transformação linear em Ω_0 :

$$\Omega_{PCA} = (\Gamma \Omega_0^T)^T, \quad (2.38)$$

onde T denota transposição. A matriz Ω_{PCA} de ordem $R \times M$ contém os novos vetores $\vec{F}_i$ que são projeções PCA dos vetores originais $\vec{f}_i$:

$$\Omega_{PCA} = \begin{bmatrix} \leftarrow \vec{F}_1 \rightarrow \\ \leftarrow \vec{F}_2 \rightarrow \\ \vdots \\ \leftarrow \vec{F}_R \rightarrow \end{bmatrix}, \quad (2.39)$$

com a propriedade de que as maiores variâncias são apresentadas pelas primeiras dimensões (o autovalor λ_i é igual à variância da i -ésima coluna de Ω_{PCA}). Em nossas análises, utilizamos os dois principais componentes, ou seja, a primeira e a segunda colunas de Ω_{PCA} , pois elas incorporam quase que totalmente a variância total dos dados originais (vide análise no Capítulo 7). Observe também que quaisquer correlações existentes entre os atributos originais são removidas após a transformação realizada por PCA. Neste projeto, a implementação do método PCA foi realizada inteiramente em ambiente Scilab, o qual utiliza o pacote LAPACK para o cômputo de autovalores e autovetores.

2.9 Análise de Variável Canônica e Classificação

A técnica estatística multivariada denominada análise de variável canônica* é uma extensão da análise de componentes principais. Ela permite uma redução de dimensionalidade do espaço original maximizando a separação entre exemplos de diferentes categorias enquanto minimiza a dispersão de exemplos dentro das categorias (93, 128, 129). Neste projeto, as categorias (ou classes) são tipos de rede (mais especificamente, modelos de rede – Seção 2.6) e cada dimensão refere-se a uma medida (Seção 2.1) calculada para uma determinada rede/exemplo. A análise de componentes principais, apesar de também realizar uma combinação linear do espaço original, não leva em consideração as categorias (130).

A análise canônica funciona do seguinte modo. O número total de categorias é denotado por C , e cada categoria é identificada por C_i ($i = 1, \dots, C$). Cada rede n é representada no espaço original pelo vetor de atributos $\vec{f}_n = (f_n^1, f_n^2, \dots, f_n^M)^T$ formado por M medidas obtidas da rede (T denota transposição). As matrizes de dispersão intraclasse e interclasse são então calculadas como funções dos vetores de atributos:

$$S_{\text{intra}} = \sum_{i=1}^C \sum_{n \in C_i} \left(\vec{f}_n - \langle \vec{f} \rangle_i \right) \left(\vec{f}_n - \langle \vec{f} \rangle_i \right)^T, \quad (2.40)$$

$$S_{\text{inter}} = \sum_{i=1}^C N_i \left(\langle \vec{f} \rangle_i - \langle \vec{f} \rangle \right) \left(\langle \vec{f} \rangle_i - \langle \vec{f} \rangle \right)^T, \quad (2.41)$$

onde N_i é o número de redes pertencentes à categoria i , $\langle \vec{f} \rangle$ é o vetor médio que considera os valores de todos os vetores de atributos, e $\langle \vec{f} \rangle_i$ é o vetor médio correspondente apenas aos exemplos da classe i . A decomposição espectral do produto $S_{\text{intra}}^{-1} S_{\text{inter}}$ é utilizada na projeção canônica, em que os seus primeiros d autovetores (correspondentes aos d maiores autovalores em módulo) são arranjados em uma matriz para serem então aplicados na transformação e

* Daqui em diante denominada somente por *análise canônica*. É também conhecida por análise discriminante de Fisher ou análise discriminante linear (128).

projeção do espaço original, como segue:

$$\Gamma = \begin{bmatrix} \uparrow & \uparrow & \dots & \uparrow \\ \vec{\gamma}_1 & \vec{\gamma}_2 & \dots & \vec{\gamma}_d \\ \downarrow & \downarrow & \dots & \downarrow \end{bmatrix}, \quad (2.42)$$

$$\vec{F}_n = \Gamma^T \vec{f}_n, \quad (2.43)$$

onde o autovetor $\vec{\gamma}_i$ está associado ao autovalor λ_i e todos os autovetores estão arranjados em ordem decrescente pelos respectivos autovalores. Essa combinação linear garante uma distribuição ótima dos exemplos com relação às suas categorias: a dispersão interclasse é maximizada enquanto a dispersão intraclasse é minimizada.

Os conjuntos de redes utilizadas neste projeto apresentam uma excelente separação entre suas categorias quando seus atributos são projetados em $d = 2$ dimensões. Essa redução da dimensionalidade do espaço de medidas apresenta diversas vantagens: (i) é possível visualizar a projeção e, conseqüentemente, os exemplos pertencentes a cada classe juntamente com as fronteiras de decisão de um classificador (vide exemplo na Figura 6.3, p. 124), (ii) a aplicação de algoritmos de classificação outrora demasiadamente custosos é facilitada pela diminuição do número de dimensões e (iii) evita-se a “maldição da dimensionalidade”, já que o tamanho de uma amostra precisa crescer exponencialmente com o número de dimensões para que a densidade média de pontos seja mantida (131). O item (iii) é exemplificado pelo classificador aqui utilizado. Para que a classificação de máximo a posteriori seja efetuada, é necessário estimar as distribuições de probabilidades de se encontrar um membro da classe C_i no espaço multidimensional utilizado. Como o classificador é aplicado após projeção canônica, a estimativa consegue ser mais representativa pois em duas dimensões a densidade de exemplos é maior do que no espaço original.

O classificador aqui utilizado baseia-se na lei de Bayes, que fornece a probabilidade condicional de um evento A ocorrer dado que o evento B já ocorreu:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}, \quad (2.44)$$

onde B tem probabilidade não nula de ocorrência. Se considerarmos A igual ao evento “ocorrência da classe C_i ” e B igual ao evento “ocorrência do vetor de atributos $\vec{F}_n$ ”, tem-se que:

$$\mathbb{P}(C_i|\vec{F}_n) = \frac{\mathbb{P}(\vec{F}_n|C_i)\mathbb{P}(C_i)}{\mathbb{P}(\vec{F}_n)}. \quad (2.45)$$

$\mathbb{P}(C_i|\vec{F}_n)$ é chamada probabilidade *a posteriori* de ocorrência da classe C_i dado que o vetor de atributos $\vec{F}_n$ tenha sido observado. Já $\mathbb{P}(C_i)$ é a probabilidade *a priori* de observação da classe C_i . As outras probabilidades são: probabilidade de ocorrência do vetor de atributos $\vec{F}_n$ dado que a classe C_i é conhecida, $\mathbb{P}(\vec{F}_n|C_i)$, e probabilidade não condicional de observação de $\vec{F}_n$ na população, $\mathbb{P}(\vec{F}_n)$. O classificador de probabilidade máxima a posteriori usa o seguinte critério para definição de uma classe (93, 127):

$$\max_{C_i} \{\mathbb{P}(C_i|\vec{F}_n)\} = \max_{C_i} \{\mathbb{P}(\vec{F}_n|C_i)\mathbb{P}(C_i)\}. \quad (2.46)$$

Note que a divisão por $\mathbb{P}(\vec{F}_n)$ no segundo termo não é necessária, pois essa probabilidade é constante no caso da classificação de uma observação $\vec{F}_n$. Portanto, dado $\vec{F}_n$, a classe com maior probabilidade a posteriori de ocorrência é a escolhida pelo método. A Figura 2.3 mostra um exemplo hipotético de regiões de decisão no caso unidimensional envolvendo três distribuições do tipo $\mathbb{P}(\vec{F}_n|C_i)\mathbb{P}(C_i)$ (uma para cada classe), onde $\vec{F}_n$ contém apenas a variável aleatória x . Um exemplo análogo para duas dimensões pode ser visto na Figura 2.4.

Neste projeto escolheu-se uma abordagem frequentista para estimar a probabilidade de máximo a posteriori. As probabilidades $\mathbb{P}(C_i)$ foram estimadas simplesmente computando a proporção de cada categoria no conjunto de dados de treinamento. Já as probabilidades condicionais $\mathbb{P}(\vec{F}_n|C_i)$ foram estimadas pela abordagem não paramétrica chamada janelas de Parzen (127). Primeiramente, calcula-se as distribuições d -dimensionais discretas (histogramas) dos vetores de atributos $\vec{F}_n$ – note que existe uma distribuição para cada categoria (como na Figura 2.4). A seguir, as distribuições são representadas como somas de deltas de Dirac que, convoluídos com uma função Gaussiana normalizada, resultam em uma aproximação de $\mathbb{P}(\vec{F}_n|C_i)$. Portanto, essas estimativas podem ser consideradas interpolações suavizadas dos respectivos histogramas.

Figura 2.3 – Classificação de máximo a posteriori unidimensional (figura adaptada de Antiqueira et al. (51)). As distribuições de probabilidades são exibidas para três categorias, C_1 , C_2 e C_3 , e as regiões de decisão indicadas abaixo do eixo- x são obtidas por probabilidade máxima a posteriori.

O processo de redução de dimensionalidade de vetores de medidas de redes e classificação de categorias de redes desconhecidas, inteiramente implementado em Scilab, pode ser assim resumido (Figura 2.5): (i) gera-se um conjunto de R realizações de cada modelo de rede considerado (ER, WS, BA e GG – Seção 2.6) com o mesmo número de vértices e arestas da(s) rede(s) de categoria(s) desconhecida(s); (ii) calcula-se um conjunto de M atributos (medidas da Seção 2.1) para cada rede considerada; (iii) projeta-se, por meio de análise canônica, os vetores

Figura 2.4 – Classificação de máximo a posteriori bidimensional. As distribuições de probabilidades são exibidas para três categorias, C_1 , C_2 e C_3 , e as regiões de decisão são indicadas por cores.

Figura 2.5 – Visão geral do método de classificação de redes (figura adaptada de Antiqueira et al. (51)). Engloba o cálculo de vetores de atributos (medidas) para cada rede, projeção canônica e classificação da(s) rede(s) de categoria(s) desconhecida(s).

de atributos das redes com categoria conhecida; (iv) obtêm-se as regiões de decisão no espaço projetado utilizando o classificador de máximo a posteriori; e (v) classifica-se as rede(s) com categoria(s) desconhecida(s) verificando-se em que região de decisão encontram-se seus vetores projetados. Adicionalmente, em alguns experimentos relatados no Capítulo 6, comparou-se o classificador de máximo a posteriori com o classificador do vizinho mais próximo (127, 132). Os resultados obtidos pelos dois métodos foram bastante parecidos entre si, pois a projeção canônica já discrimina satisfatoriamente os exemplos das classes aqui utilizadas.

3 Correlações entre Grau e Atividade

Este capítulo contém um estudo das correlações estrutura-dinâmica em termos do grau e da atividade calculados para cada vértice de uma dada rede. Em redes não direcionadas, essas correlações são bem conhecidas; entretanto, no caso direcionado, pouco sabe-se a respeito. As correlações em redes direcionadas podem apresentar comportamentos bastante intrincados, e formam o foco da pesquisa apresentada neste capítulo. Entender essas correlações tem importantes consequências, tanto teóricas quanto práticas, já que um indicador simples de conectividade (o grau) seria nesse caso suficiente para uma descrição completa da atividade na rede. A seguir, são descritos em detalhes os tipos de correlação estudados e os resultados obtidos. As medidas de reciprocidade e acessibilidade são centrais nessa análise, já que permitiram uma maior compreensão do relacionamento grau-atividade em redes direcionadas.

3.1 Tipos de Correlação

Estamos interessados em analisar o relacionamento entre grau (estrutura) e atividade (dinâmica), ou seja, estudar as correlações entre grau de entrada k^{in} e atividade π e entre grau de saída k^{out} e atividade π – a atividade é definida na Seção 2.2 (p. 49). Se a estrutura é fortemente correlacionada com a dinâmica, uma dessas propriedades pode ser obtida a partir da outra com grande precisão. Correlações perfeitas sempre ocorrem em grafos (conexos) não direcionados (25), onde a atividade de um vértice pode ser exatamente obtida a partir do seu grau: $\pi_i = k_i / \sum_{j=1}^N k_j$ (a Figura 3.1a contém um exemplo de grafo não direcionado e a Figura 3.1c apresenta a respectiva correlação entre grau e atividade). No caso de grafos direcionados, a correlação perfeita ocorre quando $k_i^{\text{in}} = k_i^{\text{out}}$ para todo vértice i (34) – o que raramente ocorre em redes reais. Os graus de entrada e de saída são em geral fracamente correlacionados com a

Figura 3.1 – Exemplos de correlações grau-atividade. (a) Grafo não direcionado com indicações dos graus dos vértices (em vermelho, dentro de cada vértice) e de suas atividades (abaixo de cada vértice). (b) Grafo direcionado resultante da exclusão de algumas direções das arestas de (a), com graus de entrada/saída indicados em vermelho e atividades abaixo de cada vértice. (c) Diagrama de dispersão (*scatter-plot*) grau contra atividade para o grafo de (a). Diagramas de dispersão para grau de entrada contra atividade (correlação tipo A) (d) e grau de saída contra atividade (correlação tipo B) (e) para o grafo de (b).

atividade. A Figura 3.1b contém um exemplo de grafo direcionado, cujos diagramas de dispersão $\vec{k}^{\text{in}}$ vs. $\vec{\pi}$ (tipo A) e $\vec{k}^{\text{out}}$ vs. $\vec{\pi}$ (tipo B) são exibidos nas Figuras 3.1d e 3.1e, respectivamente. Para mensurar a intensidade dessas correlações, utilizamos o coeficiente de correlação linear de Pearson ρ_{corr} (133), sendo que correlações fortes apresentam $|\rho_{\text{corr}}| \rightarrow 1$ e correlações fracas resultam em $|\rho_{\text{corr}}| \rightarrow 0$. O sinal de ρ_{corr} indica se a correlação é positiva ou negativa. Observa-se que os diagramas de dispersão tipo A e B são sempre inspecionados neste trabalho a fim de complementar a informação dada pelo coeficiente de Pearson, como nos exemplos da Figura 3.1.

3.2 Reciprocidade e Correlações Tipo B

As correlações tipo A e B foram obtidas para dezenas de redes reais direcionadas – a lista de redes pode ser observada na Figura 3.2, as quais foram descritas na Seção 2.7, a partir da p. 58 (com exceção da rede integrada da bactéria *Escherichia coli*, cujo processo de construção faz parte do desenvolvimento desta tese e está descrito na Seção 7.1, p. 135). Primeiramente, foi observado se a reciprocidade (Seção 2.1.15, p. 48) tem relação com a intensidade das correlações. Mais especificamente, procurou-se descobrir se a queda na reciprocidade, partindo das redes não direcionadas (reciprocidade máxima) para diferentes configurações de redes direcionadas, provoca uma queda na intensidade das correlações do tipo A ou B. De fato, as correlações do tipo B (grau de saída vs. atividade) perdem suas intensidades conforme a reciprocidade diminui (Figura 3.2). Ou seja, alta reciprocidade indica que quase todas as arestas da rede são simétricas (como em redes não direcionadas), resultando em fortes correlações. Exemplos são as redes corticais Gato'95, Macaco'47 e Macaco'70, e a rede de troca de mensagens UCI-Students. Nessas redes, portanto, o grau de saída de um vértice já indica com grande precisão sua atividade. Isso não acontece em redes com baixa reciprocidade, as quais apresentam fracas correlações do

Figura 3.2 – Intensidades das correlações do tipo B como função da reciprocidade em 23 redes reais. O relacionamento entre essas duas variáveis resulta em um alto coeficiente de Pearson: $\rho_{corr} = 0,89$.

Figura 3.3 – Intensidades das correlações do tipo A como função da reciprocidade em 23 redes reais. De maneira geral, as redes apresentam alta correlação do tipo A, com exceção das redes cujos pontos são exibidos em laranja.

tipo B – por exemplo, nas redes *peer-to-peer* P2P'5, P2P'8 e P2P'9.

Tal relacionamento não foi verificado entre a reciprocidade e as correlações do tipo A (vide Figura 3.3). De maneira geral, as redes consideradas apresentam fortes correlações grau de entrada vs. atividade, com exceção de algumas redes indicadas em laranja na Figura 3.3. Essas redes são os mapas WWW (Massey, Flinders e Nottingham), a rede neuronal *C. elegans*, a rede cortical Macaco'85 e o tesouro Roget (essas duas últimas com correlações não tão baixas quanto as exibidas pelas quatro primeiras). O comportamento verificado nessas redes pode ser melhor entendido utilizando-se a medida de acessibilidade, conforme explicado na próxima seção.

3.3 Acessibilidade e Correlações Tipo A

Antes de iniciar a discussão dos resultados restantes, é apresentada uma motivação para o tipo de análise realizada nas correlações tipo A. Foi observado que nas redes cujos pontos são exibidos em laranja na Figura 3.3 ocorre um espalhamento do diagrama grau de entrada vs. atividade, conforme ilustrado na Figura 3.4a. Esse exemplo mostra, por meio de pontos arbitrários, o tipo

Figura 3.4 – Ilustração do tipo de análise realizada nas correlações do tipo A. (a) Diagrama de dispersão original, mostrando um relacionamento hipotético entre duas variáveis X e Y (nos experimentos, usa-se grau de entrada e atividade, respectivamente). (b) Transformação para coordenadas polares dos pontos de (a), onde exibe-se raio vs. ângulo. (c) Histograma da distribuição de pontos de (b) ao longo do eixo das ordenadas (ângulo).

de espalhamento que ocorre nessas redes: conjuntos de pontos separados tendem a aparecer, os quais distribuem-se de maneira radial a partir da origem. O objetivo é, portanto, encontrar alguma propriedade da rede que tenha relação com esse comportamento. Para facilitar essa tarefa, optou-se por transformar o conjunto de pontos (x,y) original para coordenadas polares (ρ, θ) :

$$\rho = \sqrt{x^2 + y^2}$$

$$\theta = \arctan\left(\frac{y}{x}\right),$$

onde ρ é o raio – distância do ponto (x,y) até a origem $(0,0)$ – e θ é o ângulo da inclinação da reta que liga os pontos (x,y) e $(0,0)$. Um exemplo dessa transformação é exibido na Figura 3.4b, onde é possível observar que a transformação polar permite uma clara separação entre os conjuntos de pontos por meio da coordenada θ . A seguir, um histograma de θ é computado (Figura 3.4c), o qual serve de base para a análise subsequente, em que verifica-se qual característica da rede reflete a separação observada na distribuição de pontos.

Pôde-se verificar que a medida de acessibilidade (Seção 2.3, p. 51) apresenta intenso relacionamento com o espalhamento de pontos observados na correlação do tipo A. Vide, por exemplo, a Figura 3.5, a qual mostra a análise realizada para a rede Macaco'85. Na parte (a) dessa figura, é exibido o relacionamento entre grau de entrada e atividade formado por dois

conjuntos de pontos com cores diferentes, os quais representam a borda (acessibilidade menor que um limiar T , em vermelho) e o núcleo da rede (acessibilidade maior que um limiar T , em verde). Portanto, existem dois regimes de correlação nessa rede, um para os vértices da borda e outro para os vértices do núcleo – alguns *outliers* aparecem na Figura 3.5a e serão comentados em breve. Por ora, uma explicação mais detalhada do papel da acessibilidade neste experimento é necessária. O limiar de acessibilidade T é escolhido automaticamente como sendo o que causa o menor *erro* no histograma da Figura 3.5c. Para um dado T , obtém-se a divisão borda-núcleo dos vértices e, portanto, é possível gerar separadamente histogramas para a borda e para o núcleo. O erro é então dado pela área relativa comum aos dois histogramas (intersecção dos histogramas – observe as figuras ao longo desta seção). Note também que: (i) a acessibilidade foi obtida para passos h de 2 até 6, ou até $\lfloor \ell \rfloor$ quando $\lfloor \ell \rfloor > 6$ (ℓ é a media dos comprimentos dos caminhos mínimos na rede); (ii) o cálculo do erro considera todos os passos h de acessibilidade calculados para uma determinada rede; (iii) o grau de entrada e a atividade são normalizados de modo que seus respectivos valores máximos sejam iguais a 1; e (iv) é dada maior ênfase, no cômputo dos histogramas, aos pontos mais distantes da origem na Figura 3.5a

Macaco'85

Figura 3.5 – Correlações do tipo A na rede Macaco'85. Pontos vermelhos referem-se aos vértices da borda, enquanto os verdes referem-se aos vértices do núcleo da rede. (a) Grau de entrada vs. atividade, com intensidades das correlações na rede toda, somente na borda ou somente no núcleo; (b) Transformação polar dos pontos de (a), com indicação do limiar de acessibilidade utilizado; e (c) histograma do raio ρ , com indicação do erro do particionamento resultante da escolha do limiar de acessibilidade.

ou, equivalentemente, aos pontos com maior ρ (raio) na Figura 3.5b. Isso é feito pois os pontos próximos da origem, principalmente nas redes com um grande número de vértices, acarretam em um ruído muito grande no cômputo do erro – optou-se pelo termo “ruído” pois pretende-se aqui analisar as correlações em termos de altos valores de grau e de atividade. Dessa maneira, para cada ponto nas coordenadas polares, o histograma considera um peso que cresce linearmente de 0 a 1 com o tamanho do raio ρ , ao invés de considerar sempre uma unidade por ponto existente.

As correlações na rede Macaco’85 dentro e fora da borda apresentam altos coeficientes de correlação ($\rho_{\text{corr}} = 0,95$ e $0,85$, respectivamente), ambos maiores do que o coeficiente para a rede toda ($\rho_{\text{corr}} = 0,66$). Na borda, a atividade cresce mais rapidamente conforme o grau de entrada aumenta, quando comparada com o crescimento da atividade no núcleo. Alguns *outliers*, marcados como vértices do núcleo na Figura 3.5a, posicionam-se em um ângulo intermediário, um tanto quanto próximos dos pontos referentes aos vértices da borda. Não foi possível, somente com a acessibilidade, identificar as causas do aparecimento desses *outliers*. No entanto, é possível explicar, em grande parte, o aparecimento de correlações distintas na rede cortical Macaco’85.

A rede neuronal *C. elegans* também apresenta correlações mais fortes na borda ($\rho_{\text{corr}} = 0,43$) e no núcleo ($\rho_{\text{corr}} = 0,54$) do que na rede toda ($\rho_{\text{corr}} = 0,14$) – vide Figura 3.6a. Alto grau de entrada não acarreta em alta atividade nessa rede; vértices com alta atividade apresentam graus de entrada moderados e localizam-se na borda. Um comportamento parecido pode ser observado na rede Roget (Figura 3.7a). De maneira geral, o espalhamento que ocorre nos diagramas do tipo A são majoritariamente causados pelos vértices da borda. Apesar de não apresentarem altos graus de entrada, alguns desses vértices conseguem ser bastante ativos. Lembre-se que na borda, por definição, estão localizados os vértices com menores acessibilidades. Ao retomar-se a definição de acessibilidade (Seção 2.3, p. 51), para então associá-la à dinâmica do passeio aleatório, nota-se que existe uma certa “dificuldade” em se deixar os vértices da borda a fim de que sejam alcançados os demais vértices da rede, justamente pela baixa acessibilidade dos vértices da borda. Essa propriedade indica que a dinâmica do passeio

C. elegans

Figura 3.6 – Correlações do tipo A na rede *C. elegans*. Essa figura está organizada da mesma maneira que a Figura 3.5.

aleatório apresenta uma tendência a permanecer na borda uma vez estando lá.

Todas as redes analisadas até o momento têm relação com a dinâmica cerebral: duas são diretamente obtidas a partir do sistema nervoso do macaco e do verme *C. elegans*, e a outra (Roget) codifica a estrutura de relações entre termos de um tesauro, mostrando, de maneira indireta, a estrutura da associação de termos/conceitos no cérebro. Os resultados mostram que existe uma tendência em existir um comportamento peculiar nos vértices de borda dessas redes. Vértices com alto grau de entrada, os quais recebem fluxo de diversos outros vértices, aparen-

Roget

Figura 3.7 – Correlações do tipo A na rede Roget. Essa figura está organizada da mesma maneira que a Figura 3.5.

temente deveriam apresentar altas atividades. Ao invés, são os vértices da borda que mostram esse comportamento, mesmo não apresentando um grande grau de entrada – principalmente nas redes *C. elegans* e Roget. Portanto, como aqui evidenciado, regiões corticais ou neurônios mais ativos podem não ser os que recebem mais conexões, pois a dinâmica nesses vértices depende da estrutura global da rede. No caso da rede Roget, pode-se tentar interpretar a atividade como a frequência de uso dos termos/conceitos na produção escrita e falada em língua natural. Ao se verificar quais as seis palavras mais ativas nessa rede (*cry*, *stridor*, *ululation*, *junction*, *loudness* e *jurisdiction*), torna-se difícil aceitar essa interpretação, pois palavras usadas na prática de maneira mais frequente apresentam menor atividade (tais como *time* e *height*). Um modelo de uso das palavras deveria incorporar outros mecanismos, tais como saltos aleatórios, já que o uso dos conceitos em nosso cérebro não está limitado somente aos conceitos recentemente utilizados. De qualquer maneira, o conhecimento a respeito do relacionamento grau-atividade não deixa de avançar nesse caso, e pode fornecer subsídios à compreensão dessas correlações em outras redes direcionadas.

Nas redes WWW, o espalhamento na correlação tipo A é muito mais acentuado (Figuras 3.8–3.10). As atividades para vértices da borda chegam a ser muito maiores que as atividades para os vértices do núcleo. Particularmente na rede Massey, a separação borda/núcleo

Figura 3.8 – Correlações do tipo A na rede Massey. Essa figura está organizada da mesma maneira que a Figura 3.5.

Flinders

Figura 3.9 – Correlações do tipo A na rede Flinders. Essa figura está organizada da mesma maneira que a Figura 3.5.

Nottingham

Figura 3.10 – Correlações do tipo A na rede Nottingham. Essa figura está organizada da mesma maneira que a Figura 3.5.

Figura 3.11 – Correlações do tipo A (visualização parcial) nas redes WWW. Somente os pontos próximos à origem são exibidos.

claramente distingue os dois regimes de correlação observados (Figuras 3.8a e 3.11a). A correlação linear dentro do núcleo é bem forte ($\rho_{\text{corr}} = 0,92$), ainda mais quando comparada à correlação entre todos os vértices da rede ($\rho_{\text{corr}} = 0,15$). Nas redes Flinders (Figura 3.9) e Nottingham (Figura 3.10) as correlações não chegam a ser tão fortes na borda ou no núcleo (vide também as Figuras 3.11b-c). É possível também observar nesses casos que os vértices mais ativos tendem a se posicionar na borda. No caso dos mapas WWW aqui empregados, esses vértices não representam necessariamente as páginas mais acessadas na prática, pois a dinâmica do passeio aleatório simples é uma aproximação do fluxo de visitas realizado somente por meio de *hyperlinks*. Para um modelo mais realista de navegação na Web, seria necessário, no mínimo, incorporar um mecanismo de saltos aleatórios. Contudo, o objetivo neste capítulo é avançar a compreensão das correlações grau-atividade, onde a atividade é modelada pelo passeio aleatório simples, e não criar modelos de dinâmica para as diversas redes de diferentes áreas aqui consideradas.

A Figura 3.3 foi reconstruída (vide Figura 3.12) levando-se em consideração as intensidades das correlações do tipo A na borda e no núcleo das seis redes analisadas em termos da acessibilidade. De maneira geral, as correlações na borda ou no núcleo são mais fortes do que

Figura 3.12 – Intensidades das correlações do tipo A como função da reciprocidade após a análise borda/núcleo. As redes representadas por pontos em laranja na Figura 3.3 são aqui exibidas por pontos vermelhos (B: borda) e verdes (N: núcleo).

quando a rede completa é considerada. Na maior parte das redes, a correlação no núcleo é maior do que na borda. De fato, conforme já discutido, alguns vértices da borda apresentam alta atividade mesmo sem exibir alto grau de entrada, resultando em coeficientes de correlação mais fracos. Em suma, correlações do tipo A aparentemente fracas podem ser divididas em termos do conceito de acessibilidade, permitindo assim uma interpretação mais completa das correlações grau-atividade em redes direcionadas.

4 Conectividade Guiada pela Atividade

O modelo aqui introduzido, identificado pela sigla APA (do inglês *Activity-based Preferential Attachment*), apresenta estrutura e dinâmica fortemente interligados, já que a criação de conexões nesse modelo é diretamente guiada pela atividade de cada vértice (definida na Seção 2.2, p. 49). APA foi criado para redes direcionadas, ao contrário do modelo BA (Seção 2.6.3, p. 58), que apenas trata de redes não direcionadas. Portanto, arestas de entrada e de saída são criadas de maneira independente no modelo APA. Além disso, esse modelo pode ser considerado uma generalização do modelo BA, pois o segundo é exatamente reproduzido quando apenas arestas recíprocas são criadas no primeiro. Quando essa restrição não é imposta, as redes APA têm estrutura próxima das de algumas redes corticais reais. O modelo é detalhado na seção seguinte, enquanto suas propriedades básicas e sua comparação com redes reais são apresentadas nas Seções 4.2 e 4.3, respectivamente.

4.1 Modelo Proposto

O modelo APA começa no tempo $T = 0$ com uma rede aleatória direcionada e fortemente conexa de n_0 vértices e distribuição inicial de atividade $\vec{\pi}^0$. A cada instante de tempo subsequente T ($T \leq N - n_0$), um novo vértice i é adicionado à rede e arestas direcionadas são criadas independentemente partindo de m vértices “antigos” ($m \ll n_0$) para o novo vértice i (arestas de entrada) e do vértice i para m outros vértices (arestas de saída). O conceito geral por trás desse modelo é baseado na premissa de que um novo vértice tem preferência em estabelecer arestas de entrada com vértices altamente ativos, para que possa receber uma boa quantidade de atividade logo no início de sua existência. Portanto, os vértices que fornecem arestas de entrada para i são escolhidos seguindo uma probabilidade de ligação preferencial $\mathbb{P}(j \rightarrow i) = \pi_j^{T-1}$, ou

seja, proporcional à atividade de cada vértice j no tempo $T - 1$. Para as arestas de saída não foi possível definir uma regra intuitiva que pudesse contribuir para uma alta atividade do vértice i . Portanto, propôs-se duas abordagens: a probabilidade preferencial por atividade, conforme já definida para as arestas de entrada, e uma probabilidade uniforme $\mathbb{P}(i \rightarrow j) = 1/(n_0 + T - 1)$ de conexão. No último caso, o modelo é identificado por APA'. A Figura 4.1 ilustra uma rede sendo construída pelo modelo APA com $m = 2$, onde um novo vértice tende a se conectar com vértices altamente ativos independentemente do sentido da aresta.

Note que os graus médios de entrada e de saída, tanto no modelo APA quanto na variante APA', são definidos como no modelo BA: $\langle k^{\text{in}} \rangle = \langle k^{\text{out}} \rangle = 2m$. O modelo BA é exatamente reproduzido pelo modelo APA quando arestas recíprocas são consideradas – ou seja, quando as m arestas de entrada e de saída são criadas aos pares entre os mesmos vértices (de i para j e vice-versa). Isso acontece pois, em redes não direcionadas, a atividade é proporcional ao grau: $\mathbb{P}(i \leftrightarrow j) = \pi_j^{T-1} = k_j / \sum_u k_u$. Como o modelo BA é baseado na ligação preferencial pelo grau, o modelo APA pode ser considerado uma generalização do modelo BA.

Figura 4.1 – Passo T ao longo do crescimento de uma rede APA (figura adaptada de Antiqueira et al. (51)). Dez vértices já existem na rede, cujas atividades π_i^{T-1} são indicadas por tons de cinza. O novo vértice tende a formar conexões com vértices altamente ativos.

4.2 Propriedades do Modelo

Simulações foram realizadas a fim de caracterizar o modelo APA e sua variante APA', ambos implementados em linguagem Scilab. Um total de 150 realizações para cada um foram geradas, com $N = 2.500$ e $\langle k^{\text{in}} \rangle = \langle k^{\text{out}} \rangle = 6$. O modelo APA resulta em uma lei de potência para os graus de entrada e de saída (Figuras 4.2a-b, como no modelo BA) – exceto para altos valores de grau, provavelmente pelo tamanho limitado das redes simuladas. Contudo, ainda é possível observar a existência de *hubs* nesses casos. No caso das redes APA', o componente de ligação uniforme faz com que as distribuições dos graus desviem consideravelmente de uma lei de potência. Além disso, os resultados da Figura 4.2 evidenciam que as redes APA' não apresentam vértices tão bem conectados quanto os *hubs* das redes APA. Com relação às distribuições de atividade, as redes APA tendem a se aproximar de uma lei de potência, apresentando, portanto, vértices que concentram atividade (algo como *hubs* de atividade). A rede BA também apresenta uma lei de potência, inclusive com inclinação próxima da distribuição de atividade da rede APA, mas com altura mais curta. APA' apresenta um decaimento mais rápido na distribuição de atividade, novamente devido ao mecanismo de ligação uniforme incorporado ao modelo. Resultados semelhantes aos aqui mostrados foram obtidos para redes de diferentes tamanhos e graus médios (resultados não exibidos).

Figura 4.2 – Distribuições médias do (a) grau de entrada, (b) grau de saída e (c) atividade nos modelos APA, APA' e BA considerando 150 realizações de cada modelo. Em (a) e (b) a distribuição do modelo APA está quase inteiramente sobreposta pela distribuição do modelo BA. Os respectivos desvios-padrão foram computados, embora sejam pequenos demais para visualização.

Outras medidas foram obtidas das redes simuladas: (i) coeficiente de agrupamento médio $\langle cc \rangle$, (ii) comprimento médio dos caminhos mínimos ℓ , (iii) assortatividades entre os graus de entrada r^{ii} e de saída r^{oo} e (iv) reciprocidade ρ com sua respectiva reciprocidade mínima $\rho_{\min}$. Todas essas medidas são explicadas na Seção 2.1 (iniciada na p. 39). Como foram geradas 150 realizações de cada modelo, obteve-se médias e desvios-padrão dessas medidas (Tabela 4.1, com valores também para o modelo BA para fins de comparação). Ambos os modelos APA e APA' apresentam baixos coeficientes de agrupamento (principalmente APA') e caminhos mínimos curtos (comprimento aproximadamente 4), lembrando portanto características das redes BA. Quanto às assortatividades, APA praticamente não apresenta correlação grau-grau (como BA), enquanto APA' mostra correlações ligeiramente positivas. Como APA e APA' são redes direcionadas, é importante também investigar o nível de simetria das arestas. Os valores de reciprocidade próximos de zero mostram que APA e APA' não são recíprocas nem antirrecíprocas; seus níveis de reciprocidade são praticamente equivalentes ao de uma rede aleatória. Além disso, esses níveis são próximos das respectivas reciprocidades mínimas. Novamente, propriedades similares foram observadas em redes de diferentes tamanhos e densidades (resultados não exibidos).

Tabela 4.1 – Médias e desvios-padrão do coeficiente de agrupamento $\langle cc \rangle$, comprimento dos caminhos mínimos ℓ , assortatividades entre os graus de entrada r^{ii} e de saída r^{oo} e reciprocidades ρ e $\rho_{\min}$, considerando 150 realizações dos modelos APA, APA' e BA.

Medida	Modelo APA	Modelo APA'	Modelo BA
$\langle cc \rangle$	$0,0306 \pm 0,0023$	$0,0090 \pm 0,0004$	$0,0177 \pm 0,0023$
ℓ	$3,7453 \pm 0,0240$	$4,2433 \pm 0,0075$	$3,7780 \pm 0,0322$
r^{ii}	$-0,0584 \pm 0,0063$	$0,1512 \pm 0,0095$	$-0,0521 \pm 0,0073$
r^{oo}	$-0,0500 \pm 0,0071$	$0,2094 \pm 0,0107$	$-0,0521 \pm 0,0073$
ρ	$0,0123 \pm 0,0015$	$0,0063 \pm 0,0009$	$1,0000 \pm 0,0000$
$\rho_{\min}$	$-0,0024 \pm 10^{-6}$	$-0,0024 \pm 10^{-6}$	$-0,0024 \pm 9 \times 10^{-7}$

Como a definição de arestas nas redes APA e APA' é guiada pela medida de atividade, é também interessante analisar a correlação entre o número de conexões e a atividade de cada vértice ao final do processo de geração das redes. Foi utilizado o coeficiente de correlação linear de Pearson ρ_{corr} (133) para mensurar a intensidade das correlações entre grau de entrada e atividade (tipo A) e entre grau de saída e atividade (tipo B). O modelo APA é o que mais

Figura 4.3 – Histogramas dos coeficientes de correlação grau-atividade nos modelos (a) APA e (b) APA'. Médias e respectivos desvios-padrão são também exibidos.

se aproxima do BA*, com correlações muito próximas de 1, principalmente as do tipo A (Figura 4.3). Os coeficientes de correlação do tipo B apresentam uma dispersão maior que os do tipo A – correlações para três redes APA são exibidas na Figura 4.4. No caso das redes APA', as correlações do tipo B ficam abaixo de 0,8 e as do tipo A mantêm-se por volta de 0,95. Isso significa que as redes APA' não apresentam características tão próximas às das redes BA, ao contrário das redes APA. Analisando-se alguns exemplos específicos de correlações em redes APA' (Figura 4.5), percebe-se que de fato existe uma maior dispersão das nuvens de pontos com relação às respectivas retas aproximadas obtidas por mínimos quadrados.

Por fim, considerando-se todas as análises feitas nesta seção, observa-se que o modelo APA é o que apresenta características mais semelhantes às do modelo BA. Semelhanças foram observadas nas distribuições de graus, nas médias do coeficiente de agrupamento e do comprimento dos caminhos mínimos, nos valores de assortatividade e nas correlações grau-atividade. Entretanto, note que essas observações não significam que esses modelos são equivalentes, pois o modelo APA gera redes direcionadas com baixos valores de reciprocidade, enquanto o modelo BA gera redes totalmente recíprocas (ou seja, com arestas não direcionadas). A comparação com o modelo BA faz-se interessante pois a nossa proposta de construção de um novo modelo

* Note que o modelo BA apresenta correlações exatamente iguais a 1, já que a atividade é uma função linear do grau em redes não direcionadas (vide Seção 4.1).

Figura 4.4 – Correlações entre grau de entrada e atividade (Tipo A, primeira coluna) e entre grau de saída e atividade (Tipo B, segunda coluna) em três redes (a–c) do tipo APA. As retas que aproximam os conjuntos de pontos foram obtidas pelo método dos mínimos quadrados.

baseou-se nos mecanismos de crescimento e ligação preferencial presentes no modelo BA.

4.3 Comparação com Redes Reais

A caracterização dos modelos APA e APA' foi estendida por meio da comparação com outros modelos de rede e também com redes reais. Para tanto, foi utilizado um conjunto de medidas maior do que na seção anterior: (i) grau médio $\langle k \rangle$, (ii) coeficiente de agrupamento médio

Figura 4.5 – Correlações entre grau de entrada e atividade (Tipo A, primeira coluna) e entre grau de saída e atividade (Tipo B, segunda coluna) em três redes (a–c) do tipo APA'. As retas que aproximam os conjuntos de pontos foram obtidas pelo método dos mínimos quadrados.

$\langle cc \rangle$, (iii) comprimento médio dos caminhos mínimos l , (iv) assortatividades entre os graus de entrada r^{ii} e (v) de saída r^{oo} , (vi) dominância do ponto central CPD , (vii) centralidade média $\langle B^{(n)} \rangle$, (viii) coeficiente de agrupamento hierárquico médio no nível 2 $\langle cc^{(2)} \rangle$ e (ix) taxa de convergência média no nível 2 $\langle cv^{(2)} \rangle$ (definições na Seção 2.1, a partir da p. 39). Os outros modelos considerados na avaliação são: (i) Erdős-Rényi (ER), (ii) Watts-Strogatz (WS), (iii) Barabási-Albert (BA) e (iv) Geográfico (GG) – vide Seção 2.6, p. 57. Como a tarefa de classificação geralmente necessita de um conjunto significativo de exemplos a fim de minimizar flutuações estatísticas, aplicamos análise de variável canônica para redução de dimensionalidade e maxi-

mização da separação entre categorias antes da classificação por máximo a posteriori (veja Seção 2.9, p. 68).

As redes reais classificadas de acordo com um dos modelos considerados são: (i) córtex do gato com $N = 52$ áreas corticais (gato'52) (49), (ii) córtex do gato incluindo todas as áreas corticais e talâmicas ($N = 95$, gato'95) (49), (iii) córtex do macaco incluindo as áreas visual e sensório-motora ($N = 47$, macaco'47) (50), (iv) cadeia alimentar da região Canton na Nova Zelândia ($N = 110$) (121), (v) cadeia alimentar de Kyeburn, também na Nova Zelândia ($N = 98$) (121) e (vi) tesouro Roget ($N = 1.022$) (113). Todas as redes são descritas na Seção 2.7 (p. 58). Para classificação de cada rede foram geradas 100 realizações de cada modelo com mesmo número de vértices e arestas da respectiva rede real.

As cadeias alimentares e o tesouro Roget foram classificados como WS (Figura 4.6d, resul-

Figura 4.6 – Projeção canônica e classificação das redes corticais do gato (a,b), do macaco (c) e também do tesouro Roget (d) – figura adaptada de Antiqueira et al. (51).

tados não exibidos para as cadeias alimentares), como seria de se esperar considerando trabalhos anteriores (113, 134). Conforme já visto na seção anterior, APA e APA' apresentam algumas características semelhantes às do modelo BA. De fato, as projeções canônicas mostram sistematicamente que esses três modelos estão próximos uns dos outros. Entretanto, as particularidades do modelo APA' fazem com que ele seja o mais provável na classificação das três redes corticais (Figura 4.6a–c). Embora seja argumentado que redes neuronais podem ser bem reproduzidas

Figura 4.7 – Correlações grau de entrada-atividade (primeira coluna) e grau de saída-atividade (segunda coluna) nas três redes corticais classificadas como APA': (a) macaco, (b) gato'52 e (c) gato'95.

pelo modelo WS (135–137), os resultados aqui obtidos mostram que o modelo APA' é mais preciso na realização dessa tarefa. Esse experimento pode ser, portanto, interpretado como uma validação de APA' como um modelo de rede cortical. Note que essa é uma validação realizada em termos estruturais, ou seja, por meio da comparação da estrutura de conexões das redes corticais com a das redes APA'.

Sugere-se, portanto, que o cérebro seja organizado de modo a favorecer conexões com regiões altamente ativas. Ou seja, ao invés de serem guiadas por características estruturais (como o grau no modelo BA), as conexões no cérebro tendem a ser atraídas pela dinâmica de ativação dos vértices. Esse processo tem relação com a teoria Hebbiana (47), em que os vértices mais visitados (ou mais ativos) são os que recebem estímulos repetida e persistentemente e, assim, tendem a estabelecer um grande número de conexões. De fato, nas três redes corticais analisadas, os vértices mais ativos são os que recebem o maior número de arestas (Figura 4.7, primeira coluna). Em outras palavras, a atividade está fortemente correlacionada ao grau de entrada nessas redes. No caso da correlação entre grau de saída e atividade (Figura 4.7, segunda coluna) esse comportamento não é tão claro para todas as redes. Portanto, sugere-se que as conexões de saída estejam relacionadas à função de integração cortical, propriedade que pode ter sido captada pelo mecanismo de conexão com probabilidade uniforme para as arestas de saída do modelo variante APA'.

5 Caracterização da Distribuição de Subgrafos

O método apresentado neste capítulo foi criado para que, dado um subconjunto de vértices qualquer, sua disposição na rede seja estudada. O foco da análise pode ser qualquer grupo de vértices que apresente uma característica de interesse, como a participação em motivos (*motifs*). Como alguns desses vértices podem já estar interconectados, considera-se aqui a análise mais geral de *subgrafos*, e não somente de vértices isolados, de modo que os subgrafos são formados exatamente pelos componentes conexos existentes entre os vértices selecionados. A Figura 5.1 ajuda a ilustrar a formação de subgrafos. Nela está contida uma rede com um determinado número de vértices de interesse (em preto), sendo que diversos deles já estão interconectados,

Figura 5.1 – Exemplo de um grafo e oito subgrafos (delimitados em azul), formados pelos vértices de interesse em preto (figura adaptada de Antiqueira e Costa (58)). Com a metodologia proposta, procura-se caracterizar a distribuição dos subgrafos na rede e suas inter-relações (indicadas por setas bidirecionais).

formando componentes conexos. Note que, como vértices isolados são também subgrafos (chamados subgrafos triviais (98)), a metodologia aqui apresentada pode ser de fato referenciada como uma metodologia para análise de subgrafos.

O objetivo é avaliar como os subgrafos estão dispostos na rede e suas inter-relações. Primeiramente, considera-se as distâncias entre todos os pares de subgrafos, quantificadas pelo comprimento dos menores caminhos mínimos. Tais distâncias são organizadas em um histograma, fornecendo portanto informações úteis quanto à distribuição topológica dos subgrafos. Além disso, os tamanhos dos subgrafos (em número de vértices) são também organizados em um histograma, indicando assim se os vértices de interesse tendem a formar subgrafos ou a permanecer isolados. Embora distâncias e tamanhos sejam informações valiosas, a maior contribuição do método proposto é a possibilidade de progressivamente unir subgrafos, ou seja, de gradativamente adicionar vértices das vizinhanças dos subgrafos até que todos os subgrafos formem um único componente conexo. O critério aqui utilizado para a expansão dos subgrafos permite a análise de suas vizinhanças de acordo com as respectivas *relevâncias* de seus vértices. A relevância é definida como a importância de um vértice para a conectividade entre subgrafos (detalhes na Seção 5.1.2), sendo que o algoritmo de expansão de subgrafos é apenas uma aplicação dos valores de relevância.

Note que o método criado fornece uma caracterização *estrutural* de subgrafos. Adicionalmente, analisa-se na Seção 5.2 a relação entre as características dos subgrafos e o comportamento de dois tipos de processos *dinâmicos*: o de transporte e o de espalhamento. Para tanto, foram definidos subgrafos específicos, formados por vértices com alto coeficiente de agrupamento, em redes artificiais e reais.

5.1 Método Proposto

Assume-se que o grafo $G(V, E)$ em estudo é conexo, não direcionado e sem pesos nas arestas. Seja um conjunto C de subgrafos $G_i = (V_i, E_i)$, $1 \leq i \leq C$ definido da seguinte maneira:

- (i) $V_i \subset V$,
- (ii) $V_i \neq \emptyset$,
- (iii) $V_i \cap V_j = \emptyset$ para todo par $i \neq j$,
- (iv) Se $v_a \in V_i$ e $v_b \in V_j$, $i \neq j$, então $(v_a, v_b) \notin E$,
- (v) $(v_i, v_j) \in E_i$ se e somente se $v_i \in V_i$, $v_j \in V_i$ e $(v_i, v_j) \in E$,
- (vi) G_i é um componente conexo.

Para que um subgrafo G_i seja criado, é suficiente definir um V_i válido, já que E_i deve conter todas as arestas de E que conectam pares de vértices em V_i (regra v). Considera-se um V_i “válido” como o subconjunto não vazio de V (regras i e ii) que resulta em um componente conexo (regra vi) que não compartilha qualquer vértice ou aresta com outro subgrafo, ou seja, cada componente conexo precisa ser maximal (regras iii e iv)*. Neste trabalho, os subconjuntos de vértices V_i são formados pelos vértices de interesse selecionados. Note também que a existência de subgrafos triviais é naturalmente permitida. Na Figura 5.2 é exibido um grafo arbitrário G com $N = 64$ vértices e quatro subgrafos $G_1, \dots, G_4$. Esse grafo é usado ao longo desta seção como exemplo de aplicação do método aqui proposto.

A escolha dos vértices de interesse que formam os subgrafos depende do estudo em questão. Por exemplo, vértices que apresentam valores extremos de uma determinada medida podem ser selecionados, como os que apresentam alto grau ou alto coeficiente de agrupamento. Nesse caso, os subgrafos correspondem aos componentes conexos formados por todos os vértices selecionados. Outra opção seria utilizar padrões bem conhecidos de subgrafos, os chamados motivos. De maneira geral, qualquer critério pode ser utilizado na seleção de vértices, contanto que cada subgrafo seja um componente conexo maximal. Na Seção 5.2 são fornecidos exemplos de definição e análise de subgrafos.

* Na teoria dos grafos, o termo subgrafo tem uma definição menos restrita. Como as seis regras aqui consideradas ainda resultam em subgrafos de acordo com a definição original, preferiu-se manter o termo tradicional, embora deva-se ter em mente que os subgrafos aqui utilizados apresentam uma definição mais particular.

5.1.1 Tamanhos de (e Distâncias entre) Subgrafos

Primeiramente, é importante conhecer a estrutura básica dos subgrafos antes que suas inter-relações sejam analisadas. Todos os tamanhos de subgrafos (ou seja, número de vértices $N_i = |V_i|$) são organizados em um histograma, possibilitando assim avaliar se os vértices de interesse tendem a ficar isolados ou se tendem a formar componentes conexos. Dependendo do caso, especialmente se o histograma for normalizado, é interessante fornecer o número de subgrafos C junto com o histograma. Exemplos desse histograma são fornecidos na Seção 5.2.

Talvez a maneira mais direta de se analisar a distribuição de subgrafos em uma rede seja utilizando a distância, a qual é definida como sendo o comprimento do caminho mínimo existente entre um par de subgrafos. Portanto, a Equação 2.19 (p. 45) pode ser utilizada para esse fim. Note que existem ao menos $|V_i||V_j|$ caminhos diferentes entre dois subgrafos G_i e G_j , já que um caminho pode ser iniciado em qualquer vértice do subgrafo de origem e terminar em qualquer vértice do subgrafo de destino. O comprimento do caminho mínimo entre dois subgrafos é,

Figura 5.2 – Grafo $G = (V, E)$ com $N = 64$ vértices, cada qual identificado por um número v (figura adaptada de Antiqueira e Costa (58)). Este grafo apresenta quatro subgrafos G_1 , G_2 , G_3 e G_4 compostos pelos seguintes vértices (áreas em azul): $V_1 = \{6, 7, 8, 15, 16, 23\}$, $V_2 = \{18, 19, 20, 27, 28, 29, 37\}$, $V_3 = \{26, 33, 34\}$ e $V_4 = \{48, 55, 56, 64\}$. Os subgrafos correspondem aos quatro componentes conexos com vértices em preto na figura.

portanto, definido como:

$$d_{G_i, G_j} = \min_{v \in V_i, w \in V_j} \{d_{vw}\} - 1, \quad (5.1)$$

onde impõe-se $d_{G_i, G_j} = 0$ sempre que $i = j$. Do comprimento do caminho mínimo entre todos os pares de vértices v e w é subtraído 1, modificando assim uma distância baseada no número de arestas para uma distância que considera o número de vértices presentes no caminho*. Embora utilize-se aqui uma distância baseada no menor caminho existente entre dois subgrafos, outro tipo de distância poderia ser definido, tal como a média dos comprimentos dos diferentes caminhos que ligam dois subgrafos. A matriz de distâncias D_s , de ordem $C \times C$, é então definida com elementos $D_s(i, j) = d_{G_i, G_j}$. Por exemplo, a matriz de distâncias entre os subgrafos da Figura 5.2 é igual a:

$$D_s = \begin{bmatrix} 0 & 1 & 5 & 4 \\ 1 & 0 & 2 & 2 \\ 5 & 2 & 0 & 5 \\ 4 & 2 & 5 & 0 \end{bmatrix},$$

sendo que D_s é simétrica, pois G é um grafo não direcionado. Se essas distâncias (excluindo-se a diagonal) forem dispostas em um histograma, a proximidade entre subgrafos pode ser avaliada mais facilmente do que somente observando-se D_s – vide experimentos relatados na Seção 5.2.

5.1.2 Expansão de Subgrafos

Por meio do procedimento detalhado nesta seção, procura-se gradualmente expandir os subgrafos G_i dentro do grafo G , dando ao mesmo tempo especial atenção aos relacionamentos entre subgrafos. Essa expansão progressiva foi implementada por meio do crescimento gradual dos subgrafos, ou seja, por meio da adição de vértices e arestas que ainda não pertençam a subgrafo algum. O critério de adição é baseado na importância que um determinado vértice tem no estabelecimento da conectividade entre subgrafos; uma aresta, por sua vez, é adicionada quando

* Suponha que exista apenas uma aresta entre dois subgrafos. A distância entre eles será zero e, na verdade, os dois subgrafos fariam parte do mesmo componente conexo (por definição, parte do mesmo subgrafo).

os dois vértices por ela conectados já estão inseridos em algum subgrafo. Grande importância é associada aos vértices que participam de algum caminho curto que conecte pares de subgrafos. Mais especificamente, para um vértice v que não pertença a qualquer subgrafo, calcula-se o comprimento do caminho mínimo, entre todos os pares de diferentes subgrafos, que necessariamente passe por v . Esse valor é considerado a *relevância* de um vértice na expansão de subgrafos.

Os caminhos mínimos podem ser indiretamente obtidos ao aplicar-se dilatações consecutivas em cada subgrafo. A dilatação é uma operação morfológica $\delta(g)$ definida sobre um subgrafo g de G , que resulta em outro subgrafo igual à união de g e seus vizinhos em G , mais as respectivas arestas (138, 139). A Figura 5.3 ilustra a dilatação do subgrafo G_1 da Figura 5.2, formado pelos vértices $V_1 = \{6, 7, 8, 15, 16, 23\}$. A dilatação $\delta(G_1)$ resulta em um subgrafo cujos vértices são $V_1 \cup \{14, 22, 24\}$ (os vértices $\{14, 22, 24\}$ são vizinhos de G_1), juntamente com as respectivas arestas que conectam esses vértices em G .

As dilatações funcionam como um passo intermediário na expansão de subgrafos. Quando $\delta(G_i)$ é aplicada recursiva e sequencialmente, até que dilatações adicionais não sejam mais possíveis, um mapa de distância é traçado entre o subgrafo G_i e os outros vértices de G . A

Figura 5.3 – Dilatação $\delta(G_1)$ do subgrafo G_1 da Figura 5.2 (figura adaptada de Antiqueira e Costa (58)). G_1 contém os vértices $V_1 = \{6, 7, 8, 15, 16, 23\}$ (área em azul claro), e sua dilatação é formada pelos vértices $V_1 \cup \{14, 22, 24\}$ (área em tom de azul mais escuro).

dilatação recursiva é denotada por:

$$\delta_d(G_i) = \underbrace{\delta(\dots(G_i)\dots)}_{d \text{ vezes}}. \quad (5.2)$$

Os vértices incluídos em $\delta_d(G_i)$, mas não em $\delta_{d-1}(G_i)$, estão distantes d arestas de G_i . Para $d = 0$, a dilatação recursiva é definida como $\delta_0(G_i) = G_i$ e, como a dilatação $\delta_{-1}(G_i)$ não é morfologicamente possível, os vértices dentro de G_i naturalmente recebem o valor de distância 0. Note que essas distâncias são diferentes daquelas definidas na Seção 5.1.1, as quais são calculadas apenas entre subgrafos, e não entre um subgrafo e todos os demais vértices do grafo.

Como todos os C subgrafos serão dilatados, é necessário aplicar a dilatação $\delta_d(G_i)$ desconsiderando os vértices dos outros subgrafos G_j , $i \neq j$. Esse comportamento é desejável pois a expansão não é feita em direção ao interior dos subgrafos. Portanto, ao se dilatar um subgrafo G_i , algum outro subgrafo G_j pode bloquear o acesso a alguns vértices. Por exemplo, na Figura 5.2, o subgrafo G_2 não pode se comunicar com os vértices 25 e 41, pois o subgrafo G_3 encontra-se bloqueando o caminho. Ainda nesse exemplo, apenas o subgrafo G_3 pode se comunicar com os vértices 25 e 41, ou, de maneira inversa, esses vértices podem acessar apenas o subgrafo G_3 . A seguir, é definido o conjunto de subgrafos acessíveis a partir de um vértice v :

$$Q(v) = \{G_i \mid \text{Exista uma dilatação } \delta_d(G_i) \text{ que contenha } v\}, \quad (5.3)$$

de modo que o número total de tais subgrafos é:

$$q(v) = |Q(v)|, \quad (5.4)$$

onde $|Q(v)|$ é a cardinalidade do conjunto $Q(v)$. Note que $q(v) \geq 1$ para qualquer v , pois G é um grafo conexo.

O conjunto completo de dilatações, o qual considera todo subgrafo e toda dilatação possível* iniciando-se em $d = 1$, permite a definição de uma matriz de distâncias D_δ , de ordem $N \times C$. Um elemento $D_\delta(v, i)$ contém a distância entre o vértice v e o subgrafo G_i , fornecida pela dilatação desse subgrafo. Contudo, uma exceção requer tratamento especial: para um vértice v não

* Até que nenhum outro vértice possa ser adicionado à dilatação.

acessível a partir de um subgrafo G_i (ou seja, $G_i \notin Q(v)$), define-se $D_\delta(v, i) = d_{\max} + 1$, onde $d_{\max} = N - 1$ é a maior distância possível em um grafo com N vértices. A definição da matriz D_δ é o último passo antes da especificação dos valores de relevância para cada vértice. O comprimento do caminho mínimo entre dois subgrafos G_i e G_j , $i \neq j$, que necessariamente passa pelo vértice v , é então considerado como a relevância $r(v)$ de v . Consequentemente, quanto menor $r(v)$, maior a relevância de v no estabelecimento da conectividade entre subgrafos. Mais formalmente, $r(v)$ é dado por:

$$r(v) = \begin{cases} \min_{1 \leq i, j \leq C, i \neq j} \{D_\delta(v, i) + D_\delta(v, j)\} - 1, & \text{se } q(v) > 1 \\ \min_{1 \leq i \leq C} \{D_\delta(v, i)\}, & \text{se } q(v) = 1 \end{cases} \quad (5.5)$$

de modo que o segundo caso é uma exceção que ocorre quando o vértice v pode acessar apenas um subgrafo (ou ele já encontra-se em um subgrafo) e, portanto, $r(v)$ é igual a $D_\delta(v, i)$, onde i é o índice do único subgrafo acessível a partir de v . Nesse caso especial, $r(v)$ é definida apenas pelas dilatações de um único subgrafo. Observe também que, no primeiro caso, o resultado da função *mínimo* é subtraído em uma unidade, já que, caso contrário, $r(v)$ seria sempre maior que 1 para vértices com $q(v) > 1$ (lembre-se que $q(v)$ começa em zero). Em outras palavras, após a subtração de uma unidade de distância, a relevância contém o número de vértices no caminho que passa por v , e não mais o número de arestas. Dessa maneira, a Equação 5.5 fornece valores de relevância maiores ou iguais a 1 para vértices que encontram-se inicialmente fora de qualquer subgrafo, enquanto relevância 0 é reservada para os vértices pertencentes a algum subgrafo. Na Figura 5.4 são exibidos os valores de relevância $r(v)$ para todos os vértices do grafo da Figura 5.2. Nesse exemplo, as relevâncias são exibidas de forma numérica e gráfica. No segundo caso, os vértices são preenchidos por tons de cinza com intensidades proporcionais a $r(v)$, variando do preto (quando $r(v) = 0$) ao branco (quando $r(v) = 8$). Portanto, os vértices em tons mais escuros estão posicionados nos caminhos mais curtos que ligam os subgrafos $G_1, \dots, G_4$, com exceção dos vértices em preto ($r(v) = 0$) que já encontram-se inseridos em algum subgrafo.

Por fim, a *expansão* dos subgrafos é realizada por meio da limiarização dos valores de

Figura 5.4 – Valores de relevância $r(v)$ para cada vértice do grafo G introduzido na Figura 5.2 (figura adaptada de Antigueira e Costa (58)). O número abaixo de cada vértice denota $r(v)$, e as respectivas cores são derivadas de uma escala de tons de cinza que varia linearmente do preto ($r(v) = 0$) ao branco ($r(v) = 8$). As áreas em azul incluem os vértices dos subgrafos $G_1, \dots, G_4$.

relevância:

$$V^+ = \{v \mid r(v) \leq T\}, \quad (5.6)$$

onde $T \geq 0$ é um inteiro. Assim, um novo grupo de vértices V^+ é selecionado. Além disso, C^+ subgrafos $G_i^+ = (V_i^+, E_i^+)$, $1 \leq i \leq C^+$, são criados tal que:

- (i) $V_1^+ \cup V_2^+ \cup \dots \cup V_{C^+}^+ = V^+$,
- (ii) $V_i^+ \neq \emptyset$,
- (iii) $V_i^+ \cap V_j^+ = \emptyset$ para todo $i \neq j$,
- (iv) Se $v_a \in V_i^+$ e $v_b \in V_j^+$, $i \neq j$, então $(v_a, v_b) \notin E$,
- (v) $(v_i, v_j) \in E_i^+$ se e somente se $v_i \in V_i^+$, $v_j \in V_i^+$ e $(v_i, v_j) \in E$,
- (vi) G_i^+ é um componente conexo.

Essas regras são similares às definidas para os subgrafos originais G_i , com a diferença de que os subgrafos agora são restritos aos vértices pertencentes a V^+ . Em suma, é suficiente considerar que os subgrafos expandidos são os componentes conexos que aparecem quando os vértices $v \notin V^+$ (e suas arestas) são excluídos de G . A limiarização com $T = 0$ resulta

nos subgrafos originais, enquanto limiares maiores possibilitam a inclusão de novos vértices e, conseqüentemente, a união entre dois ou mais subgrafos. Um exemplo de expansão que resulta na união de subgrafos é exibido na Figura 5.5, com $T = 2$ aplicado nos valores de relevância da Figura 5.4. Os vértices em preto pertencem a V^+ , e formam uma expansão em que todos os subgrafos são unidos em um único componente conexo.

Note que as dilatações recursivas poderiam ser utilizadas como método de expansão de subgrafos, e não somente como um passo intermediário, sendo que cada subgrafo seria dilatado até que todo o grafo fosse coberto. Entretanto, essa abordagem não discriminaria a importância dos vértices na conexão entre subgrafos – ou seja, não teríamos valores de relevância, e sim apenas indicadores de proximidade a algum subgrafo. A Figura 5.3 indica que a primeira dilatação do subgrafo G_1 incorpora os vértices 14 e 24, enquanto o procedimento de expansão de subgrafos não inclui nenhum dos dois vértices até um limiar de relevância igual a $T = 3$ (Figura 5.5). Embora os vértices 14 e 24 sejam vizinhos de G_1 , eles estão relativamente mais distantes de outros subgrafos, e, portanto, não participam de caminhos de comprimento curto conectando pares de subgrafos. Essa comparação mostra que as dilatações por si só não discriminam a

Figura 5.5 – Expansão dos subgrafos G^+ dentro do grafo G da Figura 5.2, considerando um limiar de relevância $T = 2$ (figura adaptada de Antikeira e Costa (58)). Os valores de relevância $r(v)$ são exibidos abaixo de cada vértice, os vértices dos subgrafos $G_1, \dots, G_4$ estão delimitados em azul claro, e os vértices pertencentes à expansão estão delimitados por um tom de azul mais escuro.

posição relativa dos vértices entre subgrafos, e essa é a razão pela qual foram definidas a matriz de distâncias D_δ e os valores de relevância $r(v)$.

O processo chamado de *expansão gradual* é aqui definido como a sequência de expansões para limiares crescentes, começando em $T = 0$, até que V^+ seja igual a V (ou seja, até que os subgrafos se expandam totalmente e formem um único componente conexo que englobe todo o grafo). O número C^+ de subgrafos na expansão pode então ser monitorado até o final da limiarização gradual, quando C^+ é igual a 1 (C^+ é função decrescente de T). Em uma expansão gradual, é possível quantificar o nível de proximidade entre subgrafos ao se observar o quão rápido C^+ cai para 1. Essa abordagem complementa os histogramas de distâncias (Seção 5.1.1) ao ressaltar uma fração dos caminhos mínimos existentes entre todos os subgrafos. Em outras palavras, são levados em consideração somente os caminhos necessários para que todos os subgrafos sejam unidos. Além disso, o histograma de distâncias não fornece informação alguma a respeito dos vértices posicionados entre subgrafos, ao passo que a expansão gradual permite identificar o núcleo de vértices que une os subgrafos em um único componente conexo. De maneira geral, quaisquer variações no conjunto V^+ para limiares crescentes T podem ser úteis na análise de relevância. Na seção seguinte, são relatados experimentos que ilustram a aplicação da metodologia completa de análise de subgrafos, tanto para redes artificiais quanto para redes do mundo real.

5.2 Subgrafos com Alto Agrupamento e Processos Dinâmicos

Foram geradas 100 realizações dos modelos (i) Erdős-Rényi (ER), (ii) Watts-Strogatz (WS), (iii) Barabási-Albert (BA) e (iv) Geográfico (GG) (detalhes na Seção 2.6, p. 57), cada um com $N = 1.000$ vértices e grau médio $\langle k \rangle = 6$. As redes reais utilizadas, introduzidas na Seção 2.7 (p. 58), são as seguintes: (i) NetScience, uma rede de colaboração científica, (ii) *E. coli*, uma rede metabólica, (iii) E-mails, uma rede de troca de mensagens eletrônicas, (iv) Power Grid,

uma rede de infraestrutura elétrica e (v) Internet, uma rede de sistemas autônomos. Seus tamanhos variam de 379 a 22.963 vértices, com graus médios entre 2,67 e 9,62 – a Tabela 2.1 (p. 59) contém mais detalhes a respeito.

Um passo importante no planejamento dos experimentos é a definição dos subgrafos a serem analisados. Escolheu-se a medida de coeficiente de agrupamento cc_i (Seção 2.1.1, p. 40) como critério de seleção de vértices, sendo que os 2,5% de vértices com os maiores cc_i são selecionados para a especificação dos subgrafos. Os subgrafos correspondem, portanto, aos componentes conexos formados pelos vértices com maiores coeficientes de agrupamento. Esses vértices são interessantes pois tendem a apresentar um padrão de conexões mais coeso do que nos outros vértices. Em outras palavras, em um subgrafo formado por vértices com alto coeficiente de agrupamento, existem diversos caminhos alternativos, e curtos, que ligam esses vértices entre si. A caracterização da conectividade entre subgrafos desse tipo podem ajudar no aperfeiçoamento de estratégias de transporte e de distribuição. A fim de clarificar esse último ponto, foram efetuadas simulações do modelo de transporte definido na Seção 2.4 (p. 54), com $\beta = 0,3$, a fim de se descobrir a taxa crítica de geração de pacotes λ_c que induz ao estado de congestionamento. O passeio aleatório simples (Seção 2.2, p. 49) também foi executado a fim de se avaliar o tempo de cobertura dos vértices, ou seja, o tempo necessário para que todos os vértices da rede sejam visitados pelo menos uma vez. Todos esses procedimentos foram implementados em Scilab, com exceção da dinâmica de transporte, implementada em linguagem C. Nas próximas subseções são relatados e discutidos os resultados da análise dos subgrafos com alto cc_i em redes artificiais e reais, primeiramente utilizando os histogramas de tamanhos e distâncias. No procedimento de expansão gradual dos subgrafos, monitora-se o número de subgrafos C^+ e o número de vértices $|V^+|$ enquanto o limiar T cresce. Os resultados para os modelos de redes são exibidos em termos de médias e respectivos desvios-padrão, pois são analisadas 100 realizações de cada modelo.

5.2.1 Resultados para Modelos de Redes

Todos os modelos, especialmente o BA, apresentam subgrafos em sua maioria de tamanho 1 (Figura 5.6). Portanto, o número de subgrafos no modelo BA é praticamente o maior possível nesse experimento ($C \approx 25$, em média). O modelo ER apresenta o menor número de subgrafos (em média, aproximadamente 18), já que nele não é raro encontrar subgrafos de tamanho 2 ou 3. Os subgrafos no modelo ER tendem a ficar separados por caminhos de comprimento 3, e, no BA, por caminhos de comprimento 2 ou 3 (Figura 5.6), ou seja, seus subgrafos estão muito próximos uns dos outros. Já o modelo WS, apesar de apresentar o efeito pequeno mundo, resulta em distâncias maiores (próximas de 5), mas ainda relativamente pequenas quando comparadas ao tamanho das redes.

Os valores do número de subgrafos C^+ e de vértices $|V^+|$ (expansão gradual, Figura 5.7) para as redes ER, WS e BA reforçam a observação de que os subgrafos com alto coeficiente de agrupamento tendem a ficar próximos uns dos outros. Redes ER e BA têm um único subgrafo ($C^+ = 1$) quando $T = 2$, englobando menos de 100 vértices na expansão – o aqui chamado *núcleo*. Portanto, nesses dois modelos, os subgrafos são rapidamente unidos. Contudo, um limiar $T = 1$ une a maior parte dos subgrafos no modelo BA, indicando que os subgrafos nesse modelo são ainda mais coesos do que no modelo ER. Os subgrafos nas redes WS são unidos quando $T = 4$, com aproximadamente 230 vértices incluídos na expansão, demandando portanto mais vértices para unir todos os subgrafos do que nas redes ER e BA. Em suma, os modelos ER, WS e BA tendem a conter subgrafos, em sua maioria triviais, próximos uns dos outros, com menos de 23% dos vértices dentro do núcleo que une todos os subgrafos – essa tendência é ainda mais marcante nos modelos ER e BA. Além disso, o limiar T necessário para que a expansão cubra todos os N vértices não é maior que 9 nesses três modelos, indicando que não somente o núcleo está próximo dos subgrafos, mas também todo o restante dos vértices.

Note que a expansão gradual permitiu-nos identificar quais as distâncias da Figura 5.6 que contribuem para uma efetiva comunicação entre os subgrafos (em termos da união em um único componente conexo). Por exemplo, no modelo WS, alguns subgrafos chegam a estar separados

Figura 5.6 – (a) Distribuições médias dos tamanhos dos subgrafos e (b) das distâncias entre subgrafos com seus respectivos desvios-padrão, considerando 100 realizações de cada modelo de rede (ER, WS, BA e GG) – figura adaptada de Antiqueira e Costa (58). Médias e desvios-padrão do número de subgrafos C são exibidas juntamente com os gráficos da primeira coluna.

por 6 vértices, enquanto um limiar de relevância $T = 4$ é suficiente para que os subgrafos sejam unidos – esse limiar indica que caminhos de comprimento no máximo 4 integram todos os subgrafos. Adicionalmente, o procedimento de expansão gradual possibilita a identificação de quais vértices realizam a integração dos subgrafos em um componente conexo.

Figura 5.7 – (a) Número médio de subgrafos C^+ e (b) número médio de vértices $|V^+|$ na expansão gradual efetuada em 100 realizações dos modelos ER, WS, BA e GG – figura adaptada de Antieira e Costa (58).

Subgrafos no modelo GG também tendem a incluir apenas um vértice, e suas distâncias chegam a valores muito maiores que nos demais modelos (≤ 60 , Figura 5.6). Um pico no intervalo 20–30 foi encontrado no respectivo histograma de distâncias, com um decaimento

lento para distâncias maiores. De fato, as redes GG apresentam características muito distintas dos outros modelos, inclusive quanto a expansão gradual é considerada (Figura 5.7): um limiar $T \approx 25$ é necessário para unir todos os subgrafos, quando quase todos os vértices (≈ 950) são incluídos na expansão. Além disso, a rede completa é incluída na expansão somente quando $T \approx 40$. Todos esses resultados mostram que vértices com alto coeficiente de agrupamento estão esparsamente distribuídos nas redes GG.

A proporção de triângulos (ciclos de comprimento 3) é alta em torno dos vértices com alto coeficiente de agrupamento. Os triângulos estão relacionados à propriedade da transitividade: se o vértice i está conectado ao vértice j , e j está conectado ao vértice k , então i está conectado a k . Dessa maneira, triângulos contribuem para a redundância em uma rede, já que existem diversos caminhos alternativos dentro de uma região formada por vértices com alto coeficiente de agrupamento. Portanto, as redes ER, WS e BA apresentam redundância em localidades próximas umas das outras, interconectadas por um núcleo de no máximo 23% dos vértices. Além disso, como todos os vértices são rapidamente cobertos pelo algoritmo de expansão, vértices fora do núcleo posicionam-se perto da fronteira do núcleo.

Esses resultados sugerem que processos dinâmicos podem tirar proveito do fácil acesso às regiões com alta redundância de conectividade. Por exemplo, congestionamentos podem ser evitados se caminhos alternativos estiverem prontamente disponíveis sempre que necessário – note que o coeficiente de agrupamento já foi previamente associado ao incremento da eficiência de algoritmos de transporte (140). De fato as redes ER, WS e BA só chegam ao estado de congestionamento com taxas de geração de pacotes λ_c muito maiores do que nas redes GG (Figura 5.8). Em outras palavras, as redes ER, WS e BA são capazes de suportar, sem apresentar congestionamento, uma carga maior de fluxo do que as redes GG. Esse resultado não é de se surpreender, já que as redes GG apresentam subgrafos esparsamente distribuídos, mostrando que sua redundância de conectividade não é concentrada nem facilmente acessível por outros vértices. Portanto, a rede é mais restrita quanto à disponibilidade de caminhos alternativos, o que indica uma estrutura propensa a congestionamentos. A redundância de conexões também pode ajudar a prevenir falhas em cascata, quando vértices ou arestas tornam-se indisponíveis

devido a, por exemplo, um congestionamento. Adicionalmente, verificou-se que a dinâmica de espalhamento é mais lenta nas redes GG do que nas demais redes: é necessário aproximadamente um tempo 10 vezes maior em redes GG para que todos os vértices sejam visitados pelo menos uma vez (Figura 5.9). Dessa maneira, essa topologia pode ser útil se for a estrutura base em um processo de propagação de doenças; contudo, pode ser pouco eficiente ao desejar-se uma propagação rápida de, por exemplo, notícias. O oposto ocorre, portanto, nas redes ER, WS e BA. É importante salientar que os comportamentos observados nas dinâmicas de transporte e de espalhamento não são unicamente causados pela distribuição dos subgrafos analisados. O comprimento médio dos caminhos mínimos entre todos os pares de vértices pode também ter grande influência no comportamento das dinâmicas. Entretanto, conforme já mencionado, o coeficiente de agrupamento tem papel importante na eficiência de transporte (140), sendo as-

Figura 5.8 – Histogramas das taxas críticas de geração de pacotes λ_c para os modelos ER, WS, BA e GG. Foram realizadas 10 simulações do processo de transporte (com λ variando) para cada rede, totalizando, portanto, 1.000 realizações do processo para cada modelo de rede (pois foram geradas 100 redes para cada modelo). Médias e respectivos desvios-padrão são exibidos em cada histograma.

sim um parâmetro que pode ser associado às propriedades de processos dinâmicos por meio da distribuição de seus subgrafos.

5.2.2 Resultados para Redes Reais

Os subgrafos com alto coeficiente de agrupamento nas redes reais também apresentam, em sua maioria, apenas um vértice, principalmente nas redes *E. coli*, E-mails e Internet (Figura 5.10). As redes NetScience e Power Grid contêm alguns subgrafos maiores, embora a maioria seja de tamanho 1. Na Internet, que contém quase 23 mil vértices, subgrafos estão distantes por no máximo 2 vértices. As redes *E. coli* e E-mails apresentam subgrafos um pouco mais distantes uns dos outros (no máximo 4). Já as redes Power Grid e NetScience têm seus grafos distribuídos mais esparsamente, principalmente a primeira, com distâncias no máximo 35, e picos em 9

Figura 5.9 – Curvas médias para a proporção de cobertura (vértices já visitados) em função do número de passos no passeio aleatório simples para os modelos ER, WS, BA e GG. Foram efetuadas 10 simulações do passeio aleatório para cada rede, totalizando 1.000 simulações para cada modelo de rede. Desvios-padrão também foram incluídos e são representados por barras verticais.

Figura 5.10 – (a) Distribuição dos tamanhos de subgrafos e (b) das distâncias entre subgrafos nas redes reais (figura adaptada de Antiqueira e Costa (58)). Note que o primeiro gráfico também contém o número de subgrafos C em cada rede. Além disso, as curvas das redes *E. coli*, E-mails e Internet estão sobrepostas no primeiro gráfico.

e 18. NetScience apresenta distâncias variando entre 1 e 9, valores que podem ser considerados moderados dado o pequeno tamanho dessa rede ($N = 379$).

Subgrafos na Internet são unidos quando o limiar de relevância é $T = 2$ (Figura 5.11). Apenas 653 vértices (2,84% of N) são incluídos nesse ponto da expansão, formando assim o núcleo da Internet. Em outras palavras, apenas 0,34% de N , ou 81 vértices, são capazes de unir todos os 572 subgrafos da Internet – note que 2,5% dos vértices foram inicialmente selecionados para compor os subgrafos. Outras redes também não requerem altos limiares de relevância para chegar a $C^+ = 1$, embora incluam relativamente mais vértices no núcleo: *E. coli* tem seus subgrafos unidos quando $T = 3$, utilizando 345 vértices (45% de N); subgrafos na rede E-mails são unidos quando $T = 3$, incorporando 368 vértices na expansão (32% de N); e NetScience mantém seus subgrafos desconectados até $T = 5$, quando 90 vértices são incluídos na expansão (24% de N). Note que, embora os subgrafos da rede NetScience estejam separados por no máximo caminhos de comprimento 9 (Figura 5.10), um limiar de relevância de

Figura 5.11 – (a) Número de subgrafos C^+ e (b) de vértices $|V^+|$ na expansão gradual efetuada nas redes reais (figura adaptada de Antiqueira e Costa (58)).

praticamente metade desse valor é suficiente para que se alcance $C^+ = 1$. Isso mostra que esses subgrafos estão na verdade mais próximos uns dos outros do que o histograma de distâncias mostra, já que apenas os caminhos que passam pelo núcleo são considerados na união dos subgrafos. Contudo, fora do núcleo, é possível encontrar valores de relevância maiores (até 13, Figura 5.11), um valor relativamente muito maior do que nas outras redes, dado que essa é a menor rede do experimento.

De maneira geral, as redes Internet, *E. coli* e E-mails mostram comportamentos similares aos observados para os modelos ER, WS e BA, já que em todos esses casos $C^+ = 1$ é alcançado a limiares baixos, e nenhuma dessas redes requer altos limiares para que toda a rede seja incluída na expansão (principalmente o modelo BA e a Internet). A interpretação dos resultados obtidos é similar à discussão da seção anterior, no caso dos modelos de rede. Na Internet, a transmissão de pacotes de dados pode tirar proveito da redundância de conexões a fim de que sejam evitados congestionamentos e falhas em cascata, já que caminhos alternativos (triângulos) são frequentemente encontrados ao se percorrer a rede. Embora a Internet seja geograficamente disposta,

ela apresenta resultados bem diferentes dos apresentados pelo modelo GG. Na verdade, ela é intimamente associada ao modelo BA, já que ambas as suas distribuições de graus seguem uma lei de potência (13, 14). De fato, os resultados aqui obtidos corroboram essa associação.

Assumindo que a rede E-mails codifique relacionamentos próximos entre indivíduos, possivelmente a nível pessoal, pode-se concluir que triângulos de amizade (a conhece b que conhece c , então a conhece c) são bem distribuídos pela rede – ou seja, não existem grupos distantes de pessoas bem agrupadas. Portanto, a dinâmica de espalhamento (difusão) de informações é beneficiada por essa estrutura. No caso da rede *E. coli*, os vértices (substratos) com alto coeficiente de agrupamento são conhecidos por relacionarem-se a funções metabólicas específicas (81). Já que a conectividade entre módulos é uma característica importante nesse caso, a proximidade entre subgrafos mostra que a rede *E. coli* é especialmente arranjada para maximizar a comunicação entre módulos. Esse não é o caso da rede NetScience, em que subgrafos apresentam-se mais distantes dos vértices que se encontram fora do núcleo, separando a rede em uma região de laços sociais mais desenvolvidos, e outra onde é mais difícil encontrar um colaborador comum a um par de cientistas (ou seja, um ciclo de comprimento 3). Nesse caso, a distribuição de conhecimento (dinâmica de espalhamento) fica prejudicada, demorando mais tempo para chegar a todos os cientistas.

A distribuição de graus da rede *E. coli* segue uma lei de potência (69), assim como o modelo BA, mas ela é melhor associada a um modelo hierárquico determinístico que reproduz satisfatoriamente muitas de suas propriedades, incluindo a distribuição de graus (81). No modelo hierárquico, vértices altamente agrupados são distribuídos nos níveis hierárquicos mais baixos, enquanto vértices altamente conectados (os *hubs*) integram recursivamente os níveis mais baixos até o ponto em que todos os vértices estejam dentro da hierarquia. No experimento aqui reportado, os vértices com alto coeficiente de agrupamento estão todos separados uns dos outros, desviando assim do modelo hierárquico em que alguns desses vértices são diretamente conectados. Contudo, já que há evidência de uma organização hierárquica em redes metabólicas (81), os resultados desta seção mostram que os níveis mais baixos da hierarquia não somente incluem vértices com alto coeficiente de agrupamento, mas também outros com

coeficiente mais baixo.

Um comportamento completamente diferente foi encontrado na rede Power Grid, onde os subgrafos só são unidos quando $T = 9$ (Figura 5.11). Adicionalmente, a expansão gradual necessita de um alto limiar de relevância ($T = 38$) para que a rede seja completamente incluída na expansão. Observa-se que essa rede apresenta resultados similares aos do modelo GG. De fato, a rede Power Grid é disposta geograficamente, mas ela tem sido associada na literatura ao modelo WS por também apresentar o efeito pequeno mundo (12). Dada a associação aqui observada entre a rede Power Grid e o modelo GG, a dinâmica de transporte em GG (Figura 5.8) pode ser considerada uma aproximação da dinâmica de distribuição de energia elétrica. Portanto, os resultados obtidos sugerem que a rede Power Grid seja sensível a congestionamento (sobrecargas) em pontos com baixa redundância de conexões e, conseqüentemente, seja suscetível a falhas em cascatas. Entretanto, nota-se que no presente estudo não foi levada em consideração a capacidade de transmissão de cada elemento da infraestrutura elétrica, um passo necessário em uma análise mais aprofundada dessa rede.

6 Efeitos da Amostragem de Redes Corticais

Ao longo dos anos, a teoria dos grafos tornou-se uma abordagem popular no estudo de redes funcionais e anatômicas do sistema nervoso central. As redes neuronais do sistema nervoso do nemátodo *C. elegans* e as redes corticais do gato e do macaco exibem estrutura consistente com as chamadas redes complexas (vide revisão acerca desse assunto em (141)). A organização dessas estruturas mostrou-se compatível com o paradigma do efeito pequeno mundo, uma mistura de rede aleatória e regular*. Vários estudos baseados em (f)MRI (do inglês *Functional Magnetic Resonance Imaging*) revelaram padrões compatíveis com o efeito pequeno mundo em humanos (142–145), que também apareceram em redes funcionais construídas a partir de experimentos baseados em MEG/EEG (146–148). Em alguns distúrbios neurológicos o efeito pequeno mundo desaparece, dando lugar a uma estrutura mais parecida com a aleatória (149–153). Devido à importância desses estudos para a área de neurociência, é necessário avaliar as limitações metodológicas de coleta de dados que, se não devidamente entendidas, podem levar a uma interpretação equivocada dos dados.

Pretende-se aqui avaliar se a estrutura de uma rede amostrada é equivalente à estrutura da rede subjacente. O exemplo da Figura 6.1 ilustra como a técnica de amostragem pode levar a conclusões errôneas. A rede original (Figura 6.1a) é uma rede circular uniforme com conexões ligando os vizinhos mais próximos. Algumas arestas foram criadas aleatoriamente, criando portanto o efeito pequeno mundo (seguiu-se o modelo de Newman-Watts (154)). Essa rede é amostrada duas vezes por meio de dez sensores (identificados por triângulos em cinza, Figura 6.1b–c), sendo que cada sensor engloba cinco vértices da rede original. Arestas são criadas na amostra sempre que um subgrupo de cinco vértices conecta-se a outro na rede original. A

* Note que na Seção 4.1 (p. 85) propôs-se um modelo mais compatível com algumas redes corticais do que o modelo pequeno mundo de Watts-Strogatz (12).

Figura 6.1 – Ilustração do processo de amostragem de redes (figura de Antiqueira et al. (66)). A rede original (a) é amostrada duas vezes (b,c), sendo que os sensores (triângulos) são rotacionados de uma amostra para outra. A primeira rede amostrada (d) resulta em uma estrutura similar à da original. Já a segunda amostra (e) leva a uma rede diferente, com um vértice bem mais conectado que os outros.

primeira amostragem leva a uma rede que reproduz satisfatoriamente a estrutura da rede original (Figura 6.1d), em que vértices vizinhos são conectados e algumas arestas de longa distância existem. Na segunda amostragem, os sensores foram levemente rotacionados (Figura 6.1c), combinando portanto diferentes grupos de vértices com relação à primeira amostragem (Figura 6.1b). Embora os números de vértices das duas amostras sejam idênticos, a segunda faz com que apareça um *hub*, isto é, um vértice muito mais conectado que os demais (Figura 6.1e). Portanto, a técnica de coleta de dados utilizada pode nos levar a conclusões equivocadas. Geralmente, as implicações do processo de amostragem são analisadas por meio de métodos teóricos ou simulações, devido à própria natureza do problema – ou seja, não tem-se acesso à rede completa. Contudo, muito pode ser feito dessa maneira, como no presente trabalho, onde analisou-se a amostragem de redes corticais por meio de modelos e simulações.

6.1 Modelo de Rede Neuronal

A fim de simular a estrutura de uma rede neuronal, utilizamos um modelo simples que incorpora conexões locais e globais: os vértices são densamente conectados com sua vizinhança (conexões locais ou de curto alcance) e esparsamente conectados com vértices distantes (conexões globais ou de longo alcance). Os níveis de conectividade local e global podem ser ajustados para que diferentes topologias sejam obtidas. Essa propriedade reflete a natureza geográfica do córtex, em que vértices espacialmente mais próximos têm maior probabilidade de se conectar do que vértices distantes uns dos outros*. Uma semiesfera de raio unitário forma o espaço tridimensional onde os vértices são aleatoriamente posicionados. A seguir, definem-se esferas de raio $R \ll 1$ centradas em cada vértice i , sendo que todos os outros vértices que encontram-se dentro dessa esfera passam a se conectar ao vértice i – note que arestas não direcionadas são criadas. Dessa maneira, conexões locais são criadas levando a um grau médio de curto alcance aproximadamente igual a $\langle k_{sr} \rangle \approx 2R^3 N$ – fórmula não exata devido a efeitos de borda†. Analogamente ao modelo pequeno mundo de Newman e Watts (154), E_{lr} arestas são aleatoriamente adicionadas à rede, criando, portanto, arestas de longo alcance. Essas arestas são criadas até que o grau médio total desejado seja obtido. Portanto, o grau médio total é dado por $\langle k \rangle = \langle k_{sr} \rangle + \langle k_{lr} \rangle$, onde $\langle k_{lr} \rangle = 2E_{lr}/N$ é o grau médio de longo alcance. Um exemplo desse modelo neuronal está ilustrado na Figura 6.2a.

6.2 Amostragem da Rede Neuronal

O método de amostragem foi desenvolvido como uma primeira aproximação da encefalografia convencional. Neste trabalho sabe-se, por meio de simulações, o arranjo da rede original e as posições dos pontos de amostragem, os quais representam sensores MEG ou eletrodos EEG. A

* Note que, para fins de comparação, o modelo geográfico apresentado na Seção 2.6.4 (p. 58) não apresenta ajustes de conectividade local e global.

† A aproximação é dada pelo volume de N esferas de raio R ($\frac{4}{3}\pi R^3 N$) dividido pelo volume da semiesfera de raio unitário ($\frac{1}{2}\frac{4}{3}\pi 1^3$).

Figura 6.2 – Ilustrações do modelo neuronal e do método de amostragem (figura de Antigueira et al. (66)). (a) Rede neuronal com $N = 300$ vértices e graus médios de curto e longo alcance iguais a $\langle k_{sr} \rangle = 7$ e $\langle k_{lr} \rangle = 1$, respectivamente. (b) Distribuição de 50 pontos de amostragem na superfície do cérebro (representado por uma semiesfera) e respectivas linhas de amostragem, sendo que três regiões de amostragem são exibidas (volumes cônicos). (c) Rede com $N^{(s)} = 50$ vértices e $T = 1$ amostrada da rede (a) utilizando os pontos de amostragem de (b).

disposição do conjunto de pontos de amostragem segue a metodologia real de monitoramento por sensores, ou seja, os pontos são posicionados na superfície da semiesfera – optou-se aqui pela distribuição uniforme (equidistante) de sensores. Uma ilustração com 50 pontos de amostragem encontra-se na Figura 6.2b. As linhas radiais nessa figura representam *linhas de amostragem* utilizadas na discriminação de regiões cônicas definidas dentro da semiesfera. Mais especificamente, cada linha forma o eixo de uma *região de amostragem* tridimensional, partindo de um ponto de amostragem até o centro da semiesfera. Cada vértice da rede original faz parte de uma única região, aquela cujo eixo encontra-se mais próximo. Portanto, o espaço da rede original (e, conseqüentemente, seu conjunto de vértices) é particionado de acordo com a

disposição dos pontos de amostragem.

O número total de vértices na rede amostrada $N^{(s)}$ é igual ao número de pontos de amostragem. As respectivas arestas são criadas de acordo com as interações que ocorrem entre diferentes regiões de amostragem, e o principal indicador da interação entre as regiões é o número de arestas (da rede original) que conectam regiões distintas. Mais formalmente, utiliza-se o peso $w_i^{(n)}$ de cada vértice i da rede original, definido como uma função inversamente proporcional à distância $d_i^{(n)}$ de i à superfície da semiesfera. Ou seja, quanto mais próximo da superfície, maior a contribuição do vértice para a intensidade do sinal captado pelos sensores. Utilizou-se aqui dois diferentes tipos de pesos para simular o decaimento do sinal: um linear $w_i^{(n)} = 1 - d_i^{(n)}$ e outro, mais realista, pelo inverso do quadrado da distância $w_i^{(n)} = 1/d_i^{(n)2}$. O peso $w_{ij}^{(e)}$ de uma aresta da rede original que conecta os vértices i e j pertencentes a regiões distintas é definido como o menor peso $w_{ij}^{(e)} = \min \{w_i^{(n)}, w_j^{(n)}\}$ entre os dois vértices. A seguir, uma matriz $W^{(s)}$ de ordem $N^{(s)} \times N^{(s)}$ é criada com elementos $W^{(s)}(a,b)$ contendo a soma dos pesos das arestas que conectam os vértices das regiões a e b . Por fim, a matriz de adjacência $A^{(s)}$ da rede amostrada é obtida ao se limitar os valores de $W^{(s)}$ da seguinte maneira: $A^{(s)}(a,b) = 1$ se $W^{(s)}(a,b) \geq T$, ou $A^{(s)}(a,b) = 0$ caso contrário. O limiar $T > 0$ é escolhido de modo que se obtenha uma rede amostrada com um determinado grau médio pré-fixado.

A rede amostrada possui, portanto, apenas arestas com peso mínimo T . Dessa maneira, o modelo de amostragem simula uma captação mais intensa de sinal das arestas e vértices mais próximos da superfície. A Figura 6.2c mostra um exemplo com $N^{(s)} = 50$ vértices amostrados da rede da Figura 6.2a. Na próxima seção é relatada uma avaliação da similaridade entre esses dois tipos de redes, isto é, se a rede amostrada representa fielmente a estrutura da rede original.

6.3 Efeitos Estruturais e Dinâmicos da Amostragem

Foram utilizados três diferentes valores de graus médios nos experimentos reportados nesta seção: $\langle k \rangle = 14, 16, 18$. Para cada caso, foram geradas 100 realizações do modelo neuro-

nal (implementado em Scilab) com $N = 4.000$ vértices. A razão entre conexões de longo e curto alcance foi fixada em 0,25, baseando-se em estudo anterior (84) que mostra um ápice de sincronia* neural quando essa proporção é próxima de 0,2. A seguir, o método de amostragem (também implementado em Scilab) foi aplicado a cada rede neuronal considerando-se sete tamanhos diferentes de amostras: $N^{(s)} = 32, 100, 200, \dots, 600$, ou seja, gradualmente incrementados até um máximo de 15% do tamanho das redes originais. Dois tipos de simulação foram realizados para cada tamanho de amostra, um com decaimento linear de sinal e outro com decaimento proporcional ao quadrado da distância (vide Seção 6.2). Os limiares de amostragem T foram escolhidos de modo que amostras com mesmo grau médio das respectivas redes originais fossem geradas. Os efeitos da amostragem foram avaliados por meio da comparação das redes originais e suas amostras com redes geradas pelos modelos (i) Erdős-Rényi (ER), (ii) Watts-Strogatz (WS), (iii) Barabási-Albert (BA) e (iv) Geográfico (GG) – detalhes na Seção 2.6 (p. 57). Para cada um desses modelos foram geradas 100 realizações para cada diferente configuração (número de vértices e grau médio) das redes originais e suas respectivas amostras.

Vetores de atributos $\vec{x}_n$ foram computados para cada uma das redes. Cada atributo é uma medida de rede definida na Seção 2.1 (a partir da p. 39): (i) grau médio $\langle k \rangle$, (ii) grau máximo $k_{\max}$, (iii) coeficiente de agrupamento médio $\langle cc \rangle$, (iv) grau da vizinhança médio $\langle k^{(nn)} \rangle$, (v) assortatividade r , (vi) grau hierárquico médio no nível 2 $\langle k^{(2)} \rangle$, (vii) coeficiente de agrupamento hierárquico médio no nível 3 $\langle cc^{(3)} \rangle$, (viii) taxa de divergência média no nível 3 $\langle dv^{(3)} \rangle$, (ix) grau inter-anel médio no nível 2 $\langle er^{(2)} \rangle$, (x) grau intra-anel médio no nível 2 $\langle ar^{(2)} \rangle$, (xi) eficiência global E_{global} , (xii) eficiência local E_{local} , (xiii) centralidade de vértices média $\langle B^n \rangle$, (xiv) centralidade de arestas média $\langle B^e \rangle$ e (xv) dominância do ponto central CPD . Para uma comparação entre os diversos modelos de redes complexas aqui utilizados, foram incluídos na Tabela 6.1 os vetores de atributos médios para o conjunto de redes neuronais e também para as redes dos modelos ER, WS, BA e GG, com $\langle k \rangle = 14$.

* Nesse caso, o nível de sincronia está relacionado ao ritmo oscilatório em um modelo de neurônios excitatórios e inibitórios.

Tabela 6.1 – Médias e respectivos desvios-padrão dos vetores de medidas (detalhes no texto) considerando o modelo neuronal mais os modelos ER, WS, BA e GG com $N = 4.000$ vértices e grau médio $\langle k \rangle = 14$.

	ER	WS	BA	GG	Modelo neuronal
$\langle k \rangle$	$14,01 \pm 0,07$	$14,00 \pm 0,00$	$14,00 \pm 0,00$	$14,00 \pm 0,07$	$14,00 \pm 0,00$
$k_{\max}$	$29,34 \pm 1,65$	$20,55 \pm 0,70$	$294,95 \pm 26,54$	$28,11 \pm 1,38$	$30,18 \pm 1,86$
$\langle cc \rangle$	$0,00 \pm 0,00$	$0,36 \pm 0,00$	$0,02 \pm 0,00$	$0,60 \pm 0,00$	$0,28 \pm 0,00$
$\langle k^{(m)} \rangle$	$15,00 \pm 0,07$	$14,19 \pm 0,01$	$35,72 \pm 0,87$	$14,45 \pm 0,08$	$14,85 \pm 0,02$
r	$0,00 \pm 0,01$	$0,00 \pm 0,01$	$0,32 \pm 0,01$	$0,59 \pm 0,02$	$0,30 \pm 0,02$
$\langle k^{(2)} \rangle$	$195,40 \pm 1,91$	$120,04 \pm 0,62$	$457,62 \pm 10,55$	$80,97 \pm 0,93$	$142,41 \pm 0,84$
$\langle cc^{(3)} \rangle$	$0,00 \pm 0,00$	$0,04 \pm 0,00$	$0,01 \pm 0,00$	$0,25 \pm 0,00$	$0,03 \pm 0,00$
$\langle dv^{(3)} \rangle$	$0,98 \pm 0,00$	$0,68 \pm 0,00$	$0,93 \pm 0,00$	$0,36 \pm 0,00$	$0,79 \pm 0,00$
$\langle er^{(2)} \rangle$	$13,95 \pm 0,07$	$8,58 \pm 0,04$	$34,48 \pm 0,87$	$5,75 \pm 0,04$	$10,16 \pm 0,06$
$\langle ar^{(2)} \rangle$	$0,05 \pm 0,00$	$4,61 \pm 0,04$	$0,24 \pm 0,01$	$7,69 \pm 0,06$	$3,69 \pm 0,06$
E_{global}	$0,30 \pm 0,00$	$0,26 \pm 0,00$	$0,34 \pm 0,00$	$0,08 \pm 0,00$	$0,27 \pm 0,00$
E_{local}	$0,50 \pm 0,00$	$0,68 \pm 0,00$	$0,51 \pm 0,00$	$0,80 \pm 0,00$	$0,64 \pm 0,00$
$\langle B^u \rangle$	$4884,12 \pm 10,99$	$6084,37 \pm 17,05$	$4152,49 \pm 16,98$	$1,45 \cdot 10^8 \pm 1,22 \cdot 10^9$	$5678,01 \pm 21,66$
$\langle B^e \rangle$	$982,94 \pm 6,28$	$1154,84 \pm 2,44$	$879,08 \pm 2,43$	$2,16 \cdot 10^7 \pm 1,82 \cdot 10^8$	$1096,79 \pm 3,09$
CPD	$1,49 \cdot 10^4 \pm 2,02 \cdot 10^3$	$1,84 \cdot 10^4 \pm 2,58 \cdot 10^3$	$5,09 \cdot 10^5 \pm 6,44 \cdot 10^4$	$1,69 \cdot 10^{10} \pm 1,27 \cdot 10^{11}$	$2,01 \cdot 10^4 \pm 3,35 \cdot 10^3$

Figura 6.3 – Classificação das redes neuronais (pontos vermelhos) e respectivas regiões de decisão considerando os modelos ER, WS, BA e GG (figura adaptada de Antiqueira et al. (66)). Todas as redes possuem o mesmo grau médio: $\langle k \rangle = 14$.

Os vetores de atributos foram então utilizados na análise de variável canônica dos quatro modelos de referência (ER, WS, BA e GG) e posterior classificação das redes neuronais e respectivas amostras utilizando-se os classificadores de máximo a posteriori e do vizinho mais próximo* (vide Seção 2.9, p. 68). O modelo neuronal mostrou-se mais próximo do modelo pequeno mundo de Watts-Strogatz, como seria de se esperar pela coexistência de arestas de curto e longo alcance em sua estrutura (Figura 6.3, $\langle k \rangle = 14$). Para os outros valores de grau médio considerados (16 e 18) a mesma classificação foi obtida – resultados não exibidos. A classificação do modelo neuronal como uma rede WS forma a base para a posterior análise de suas respectivas amostras. Note que isso não implica que os modelos são equivalentes, mas sim que apresentam uma semelhança considerável no contexto aqui definido. Por exemplo, o grau máximo $k_{\max}$ e o índice de assortatividade r apresentam diferenças significativas entre as redes neuronais e as redes WS (Tabela 6.1).

Na coleta de amostras é de se esperar que, quanto menor o número de sensores, maiores as dificuldades em se reproduzir as características da rede original. De fato, as menores amostras aqui analisadas ($N^{(s)} = 32$) não puderam ser classificadas como um único modelo (Figuras 6.4a

* Como os resultados são muito semelhantes, apenas os obtidos por classificação do vizinho mais próximo são exibidos.

Figura 6.4 – Classificação das redes amostradas (pontos em azul) com decaimento linear de sinal, $\langle k \rangle = 14$ e diferentes tamanhos de amostras: (a) $N^{(s)} = 32$, (b) $N^{(s)} = 100$, (c) $N^{(s)} = 200$, (d) $N^{(s)} = 300$, (e) $N^{(s)} = 400$, (f) $N^{(s)} = 500$ e (g) $N^{(s)} = 600$. As regiões de decisão foram obtidas pelo classificador do vizinho mais próximo (o classificador de máximo a posteriori apresenta resultados similares) após análise de variável canônica – figura adaptada de Antiqueira et al. (66).

Figura 6.5 – Classificação das redes amostradas (pontos em azul) com decaimento quadrático de sinal (figura adaptada de Antikeira et al. (66)). Outras informações sobre as redes e sobre a organização da figura podem ser consultadas na legenda da Figura 6.4.

e 6.5a) – resultados obtidos considerando-se os dois tipos de decaimento de sinal explicados na Seção 6.2. As amostras de tamanho $N^{(s)} = 100$ (Figuras 6.4b e 6.5b) foram classificadas como ER, o que mostra que suas propriedades estruturais são bem diferentes das exibidas pelas redes neuronais originais. Considerando apenas o decaimento linear de sinal, amostras com 100–300 vértices foram todas classificadas como ER. Adicionalmente, as Figuras 6.4b–g complementam essa informação pois mostram que as amostras tendem a se aproximar ou até mesmo a ser incluídas na região de classificação WS conforme o número de pontos de amostragem aumenta. Entretanto, mesmo quando algumas instâncias são classificadas como WS (ou seja, a mesma da rede original), elas ainda se encontram próximas da região de classificação ER. Quando o decaimento quadrático é utilizado, toda amostra com tamanho maior que $N^{(s)} = 32$ é classificada como ER (Figura 6.5), embora a tendência de uma aproximação para a região WS ainda aconteça. Em suma, os resultados demonstram que, independentemente do tipo de decaimento do sinal, as redes amostradas apresentam características similares às das redes aleatórias ER, com uma tendência a se aproximarem das propriedades das redes originais (classificação WS) conforme o tamanho das amostras aumentam. Observa-se que os resultados para diferentes graus médios coincidem em grande parte com os aqui exibidos.

A metodologia para análise de subgrafos descrita no Capítulo 5 foi também utilizada na avaliação do efeito da amostragem de redes corticais. Foram selecionados os vértices com os maiores coeficientes de agrupamento, totalizando 5% do total de vértices de cada rede, na formação de subgrafos. A análise descrita no Capítulo 5 mostra que a distribuição desse tipo de subgrafo tem influência na dinâmica de transporte e de espalhamento, possibilitando, portanto, que as redes neuronais e suas amostras sejam analisadas de maneira análoga. Os resultados mostram que o crescimento do número de vértices na expansão gradual dos subgrafos nas redes originais é bem reproduzido nas amostras (Figura 6.6b, $\langle k \rangle = 14$) – resultados semelhantes foram obtidos para outros graus médios. A única exceção é a curva para as amostras de tamanho 32 obtidas com decaimento quadrático de sinal. Apesar de o número de subgrafos apresentar uma dinâmica mais peculiar nas redes originais (Figura 6.6a), pode-se concluir que em todos os casos os subgrafos são unidos rapidamente, com um limiar na maior parte das vezes igual

Figura 6.6 – Distribuição dos subgrafos com alto coeficiente de agrupamento nas redes originais ($N = 4.000$) e nas respectivas amostras ($N^{(s)} = 32, 100, \dots, 600$) para $\langle k \rangle = 14$, considerando os dois tipos de decaimento de sinal. (a) Número de subgrafos (com relação a N , em porcentagem) em função do limiar de relevância e (b) vértices contidos na expansão gradual (%) para diferentes valores de relevância. O Capítulo 5 contém os detalhes dessa metodologia. Note que em (a) as curvas descrescem para zero a fim de uniformizar a visualização para redes de diferentes tamanhos.

a 3. Portanto, a distribuição dos subgrafos nas redes neuronais é bem reproduzida nas amostras, apesar de diferenças marcantes já terem sido observadas na estrutura dessas redes (Figuras 6.4 e 6.5). Conforme análise reportada na Seção 5.2 (p. 105), subgrafos com alto coeficiente de agrupamento, quando dispostos de maneira coesa na rede, formam uma estrutura propícia às dinâmicas de transporte e de espalhamento. Isso mostra que as redes neuronais aqui analisadas, bem como suas amostras, estão organizadas de modo a favorecer esses tipos de dinâmica.

A fim de que fossem quantificados outros padrões dinâmicos das redes neuronais e de suas

amostras, utilizou-se o modelo de sincronização de Kuramoto (Seção 2.5, p. 55), pois o processo de sincronização é particularmente relevante no caso das redes corticais (85). O nível global de sincronia r_{sync} foi medido para todas as redes originais e suas respectivas amostras, considerando todos os graus médios e os diferentes tipos de decaimento de sinal utilizados na coleta das amostras. Na Figura 6.7 são exibidos os resultados para $\langle k \rangle = 18$ – as observações feitas valem também para os demais graus médios. Nota-se que o nível de sincronia atingido após o regime transiente é sempre menor nas redes originais (traço pontilhado) do que nas amostras. Além disso, conforme o tamanho das amostras aumenta, os respectivos níveis de sincronia aproximam-se do nível das redes originais (tanto no decaimento linear quanto no decaimento quadrático de sinal). A duração do regime transiente nas amostras também aproxima-se do observado nas redes originais conforme o tamanho das amostras aumenta. Isso mostra que as amostras de maior tamanho exibem um comportamento de sincronia, principalmente após o regime transiente, muito próximo do exibido pelas redes originais. Portanto, apesar de as amostras não apresentarem estrutura tão próxima das exibidas pelas redes originais, é possível nelas reproduzir uma propriedade importante do processo de sincronização dos sinais neuronais.

Outra propriedade dinâmica aqui avaliada foi a de acessibilidade (Seção 2.3, p. 51) de cada vértice em passos h iguais a 2 e 3. No caso das redes corticais, a acessibilidade mede o quão importante é um dado neurônio (ou região cortical) no processo de ativação dos demais neurônios (ou regiões corticais) que não sejam seus vizinhos imediatos. Os resultados para $\langle k \rangle = 14$ e decaimento quadrático de sinal são exibidos nas Figuras 6.8 e 6.9, respectivamente, para $h = 2$ e $h = 3$. Quando $h = 2$, as distribuições de acessibilidade sofrem aumento de dispersão (desvios-padrão de 0,009 para 0,066) e diminuição na média (de 0,979 para 0,752) conforme o tamanho das amostras aumenta. Ou seja, essas distribuições movem-se em direção à distribuição das redes originais (com média 0,540), mas não se aproximam o suficiente. Além disso, o aumento na dispersão não corresponde à dispersão da acessibilidade no caso das redes originais (desvio-padrão 0,036). Conclui-se, entretanto, que existe uma tendência de aumento da precisão dos valores de acessibilidade conforme o tamanho das amostras aumenta, sem, contudo, ser suficiente para reproduzir a distribuição da acessibilidade nas redes originais. No caso

Figura 6.7 – Níveis médios de sincronia $\langle r_{\text{sync}} \rangle$ ao longo do tempo nas redes originais ($N = 4.000$) e nas respectivas amostras ($N^{(s)} = 32, 100, \dots, 600$) para $\langle k \rangle = 18$. Os resultados são exibidos para as amostras coletadas com decaimento quadrático de sinal (a) e para as coletadas com decaimento linear de sinal (b). Foi feita uma realização do processo de sincronização para cada rede, totalizando, portanto, 100 realizações para cada tipo de rede (neuronal ou amostra). No total, considerando-se todos os graus médios e tipos de amostragem diferentes, foram efetuadas 4.800 realizações do processo de sincronização.

Figura 6.8 – Histogramas de acessibilidade para $h = 2$ obtidos das (a–g) amostras de diferentes tamanhos ($N^{(s)} = 32, 100, \dots, 600$) e das (h) redes originais ($N = 4.000$). Para cada histograma, as respectivas médias e desvios-padrão são também exibidos.

Figura 6.9 – Histogramas de acessibilidade para $h = 3$ obtidos das (a–g) amostras de diferentes tamanhos e das (h) redes originais. A organização desta figura é a mesma da Figura 6.8.

$h = 3$, uma tendência similar pôde ser observada, apresentando-se, porém, de maneira muito mais sutil. Os valores de acessibilidade ainda diminuem conforme o tamanho das amostras aumenta; entretanto, isso ocorre de maneira mais suave (médias variam de 0,989 para 0,942). Os resultados para decaimento linear de sinal mostram que as amostras tendem a apresentar acessibilidades um pouco menores do que quando obtidas pelo decaimento quadrático (resultados não exibidos). Para os demais graus médios, observou-se que o incremento do grau médio provoca um aumento geral da acessibilidade, tanto para as amostras quanto para as redes originais (resultados não exibidos). Contudo, nesses casos, mantém-se a observação de que quanto maior o tamanho das amostras, maior a aproximação (mesmo que insuficiente) em direção às acessibilidades exibidas pelas redes originais.

Não foram geradas amostras com mais de 600 vértices pois a razão entre o número de pontos de amostragem e o tamanho da rede original (15%) já é mais do que suficiente para representar as técnicas atuais de amostragem cerebral, tais como (f)MRI e, em particular, MEG/EEG. Enquanto amostras maiores provavelmente levariam a estruturas mais precisas, uma proporção de 15% já pode ser considerada inverossímil dado o gigantesco número de neurônios existentes no cérebro.

7 Rede *Escherichia coli*: Uma Abordagem Integrada

Conforme já mencionado na introdução, os dados de regulação da transcrição gênica na bactéria *Escherichia coli* foram arranjados em uma rede e complementados por informações a respeito de vias metabólicas/sinalizadoras e de interações entre proteínas. Primeiramente, foi realizada uma análise global dessa rede integrada, bem como das três redes que a compõem, por meio do cálculo de diversas medidas de rede. As correlações entre grau e atividade na rede integrada foram analisadas, o que possibilitou a detecção de diversos genes *outliers* (ou seja, com propriedades discrepantes ou não usuais) que apresentam alto grau e baixa atividade. Esses *outliers* têm diversas outras características: propriedades estruturais também distintas do restante da rede, importante participação em motivos (do inglês, *motifs*), grande acessibilidade e intensa expressão gênica. Além disso, os *outliers* são formados por um grupo de genes reguladores globais da *E. coli*. Mais detalhes são fornecidos a seguir, iniciando-se com o processo de construção da rede integrada.

7.1 Integração dos Dados

A regulação da expressão gênica é predominantemente controlada no nível da transcrição em procariontes. A transcrição é, por sua vez, mediada por um conjunto de TF's. O estado ativo ou inativo dos TF's em células em crescimento é normalmente dirigido por metabólitos, os quais são produzidos na célula ou obtidos do ambiente pela atividade do produto de genes efetores. A fim de entender a dinâmica regulatória dos TF's e seu relacionamento com outros componentes celulares, foram aqui integradas interações transcritivas, proteicas e metabólicas que mediam a atividade fisiológica dos TF's na bactéria *Escherichia coli*. Essa integração permitiu, pela

primeira vez em uma bactéria, o estudo dos diferentes tipos de interações e seus respectivos papéis na mediação da atividade dos TF's.

Primeiramente, tomou-se a rede de transcrição gênica TRN (Seção 2.7.5, p. 64) como base do processo de integração. A partir dessa rede, foram inseridas conexões provenientes da rede de vias metabólicas e de sinalização (rede MSFN, Seção 2.7.5). Apenas conexões entre genes já existentes na rede TRN foram consideradas, resultando em 240 arestas da rede MSFN incluídas no processo de integração. A rede não direcionada de interações entre proteínas PPIN (Seção 2.7.5) também foi considerada, sendo que suas arestas de natureza recíproca $i \leftrightarrow j$ foram separadas em arestas direcionadas independentes $i \rightarrow j$ e $i \leftarrow j$, a fim de serem incluídas na rede integrada que já possui arestas direcionadas das redes TRN e MSFN. Consequentemente, 2.620 arestas direcionadas provenientes da rede PPIN foram adicionadas à rede integrada. A rede construída possui $N = 1.521$ vértices e grau médio $\langle k^{\text{in}} \rangle = \langle k^{\text{out}} \rangle = 4,20$. Ela possui um componente fortemente conexo com 42% dos vértices, enquanto os outros componentes têm no máximo 3 vértices. Além disso, 97% dos componentes possuem apenas 1 vértice, o que evidencia uma alta fragmentação da rede fora do maior componente. Portanto, utilizamos apenas o maior componente em nossas análises, o qual apresenta $N = 635$ vértices (525 TG's e 110 TF's) e grau médio $\langle k^{\text{in}} \rangle = \langle k^{\text{out}} \rangle = 6,67$.

A Tabela 7.1 apresenta as médias e desvios-padrão de diversas medidas estruturais (definidas na Seção 2.1, p. 39) calculadas para as redes TRN, MSFN e PPIN. A rede MSFN é a mais esparsa, altamente desconectada, com graus bastante baixos e medidas hierárquicas nulas, provavelmente devido à coleta manual de dados (88). Por outro lado, a rede mais densa é a PPIN com hierarquias de segundo e terceiro níveis muito bem conectadas. Entretanto, os caminhos tendem a ser longos nessa rede, com baixo valor médio de centralidade. A rede TRN apresenta caminhos ainda maiores, com centralidade baixíssima. Ela é moderadamente interligada ao longo de suas hierarquias, sendo que as hierarquias de saída são as relativamente mais esparsas, ou seja, os respectivos coeficientes de agrupamento são baixos.

As mesmas medidas são exibidas na Tabela 7.2 para a rede integrada e seu maior compo-

Tabela 7.1 – Medidas computadas para as redes TRN, MSFN e PPIN antes do processo de integração.

São exibidos: o tamanho N de cada rede junto das médias e desvios-padrão dos graus de entrada e de saída (k^{in} e k^{out}), dos graus hierárquicos de entrada e de saída nos níveis 2 e 3 ($k^{(2|\text{in})}$, $k^{(2|\text{out})}$, $k^{(3|\text{in})}$ e $k^{(3|\text{out})}$), dos coeficientes de agrupamento de entrada e de saída (cc^{in} e cc^{out}), dos coeficientes hierárquicos de agrupamento de entrada e de saída no nível 3 ($cc^{(3|\text{in})}$ e $cc^{(3|\text{out})}$), do comprimento do caminho mínimo (ℓ) e da centralidade ($B_i^{(n)}$).

	TRN	MSFN	PPIN
N	1.521	437	1.930
$\langle k^{\text{in}} \rangle$	$2,32 \pm 1,79$	$0,91 \pm 2,29$	$10,07 \pm 22,40$
$\langle k^{\text{out}} \rangle$	$2,32 \pm 18,17$	$0,91 \pm 0,68$	$10,07 \pm 22,40$
$\langle k^{(2 \text{in})} \rangle$	$2,96 \pm 5,00$	0 ± 0	$541,07 \pm 879,57$
$\langle k^{(2 \text{out})} \rangle$	$2,96 \pm 27,86$	0 ± 0	$541,07 \pm 879,57$
$\langle k^{(3 \text{in})} \rangle$	$2,17 \pm 2,97$	0 ± 0	$2.626,81 \pm 1.579,02$
$\langle k^{(3 \text{out})} \rangle$	$2,43 \pm 25,02$	0 ± 0	$2.626,81 \pm 1.579,02$
$\langle cc^{\text{in}} \rangle$	$0,13 \pm 0,21$	0 ± 0	$0,11 \pm 0,19$
$\langle cc^{\text{out}} \rangle$	$2,00 \times 10^{-3} \pm 2,32 \times 10^{-2}$	0 ± 0	$0,11 \pm 0,19$
$\langle cc^{(3 \text{in})} \rangle$	$0,10 \pm 0,21$	0 ± 0	$0,08 \pm 0,10$
$\langle cc^{(3 \text{out})} \rangle$	$2,22 \times 10^{-4} \pm 3,23 \times 10^{-3}$	0 ± 0	$0,08 \pm 0,10$
$\langle \ell \rangle$	$1.513,16 \pm 64,26$	$436,09 \pm 0,68$	$96,12 \pm 289,30$
$\langle B_i^{(n)} \rangle$	$4,41 \times 10^{-6} \pm 4,97 \times 10^{-5}$	0 ± 0	$1,23 \times 10^{-3} \pm 5,01 \times 10^{-3}$

nente fortemente conexo. Conforme já observado, 42% de todos os vértices estão no maior componente e os vértices restantes estão praticamente desconexos. Essa observação é reforçada pelas seguintes medidas: os graus em todos os níveis hierárquicos e a centralidade são maiores no maior componente do que na rede integrada completa, enquanto o comprimento dos caminhos mínimos é menor no maior componente. Adicionalmente, quase todas as arestas provenientes das redes MSFN e PPIN são preservadas no maior componente da rede integrada, enquanto mais da metade das arestas de TRN são desconsideradas. Isso mostra (i) a importância das redes MSFN e PPIN na formação do maior componente (ou seja, interconectando uma grande parte dos vértices) e (ii) fora do maior componente, na porção fragmentada, praticamente apenas arestas da rede TRN permanecem. A última observação suporta a crença de que existem diversos módulos periféricos de regulação desconectados do núcleo da rede *E. coli*, devido à coleta incompleta de dados ou aos papéis distintos desempenhados na periferia, em contraste com o realizado no metabolismo central, como notado previamente (155).

7.2 Relacionamento Estrutura-Dinâmica: *Outliers*

Ao investigar-se as propriedades da rede integrada (daqui em diante, considera-se apenas seu maior componente fortemente conexo), especial atenção foi dada aos vértices cujas propriedades desviam das características mais comumente observadas. Inicialmente, foram analisadas as correlações entre estrutura e dinâmica, sendo que a estrutura é representada pelo grau e a dinâmica pela atividade (Seção 2.2, p. 49). Em termos biológicos, a atividade pode ser entendida como a taxa com que cada gene é regulado, sendo que as interações reguladoras ocorrem aleatoriamente em uma rede conhecida que codifica as regulações possíveis. Observa-se que o relacionamento entre grau de entrada e atividade (tipo A) apresenta forte correlação ($\rho_{\text{corr}} = 0,97$, Figura 7.1a), ou seja, quanto maior o grau de entrada, maior a atividade de um vértice. Quanto ao relacionamento entre grau de saída e atividade (tipo B), alguns vértices fogem da tendência observada no restante da rede de aumentar a atividade conforme o grau de saída aumenta,

Tabela 7.2 – Medidas estruturais obtidas para a rede integrada e seu maior componente fortemente conexo. As medidas são as mesmas descritas na legenda da Tabela 7.1, com a adição do número de arestas provenientes de cada uma das redes utilizadas no processo de integração.

	Rede integrada	Rede integrada (maior componente)
N	1.521	635
Arestas TRN	3.529	1.428
Arestas MSFN	240	205
Arestas PPIN	2.620	2.600
$\langle k^{\text{in}} \rangle$	$4,20 \pm 6,05$	$6,67 \pm 8,45$
$\langle k^{\text{out}} \rangle$	$4,20 \pm 19,36$	$6,67 \pm 15,31$
$\langle k^{(2 \text{in})} \rangle$	$51,52 \pm 92,23$	$98,98 \pm 125,28$
$\langle k^{(2 \text{out})} \rangle$	$51,52 \pm 147,95$	$98,98 \pm 152,01$
$\langle k^{(3 \text{in})} \rangle$	$298,18 \pm 250,22$	$424,51 \pm 237,06$
$\langle k^{(3 \text{out})} \rangle$	$257,49 \pm 467,90$	$415,73 \pm 321,56$
$\langle cc^{\text{in}} \rangle$	$0,14 \pm 0,18$	$0,16 \pm 0,18$
$\langle cc^{\text{out}} \rangle$	$0,04 \pm 0,16$	$0,10 \pm 0,22$
$\langle cc^{(3 \text{in})} \rangle$	$0,10 \pm 0,13$	$0,09 \pm 0,08$
$\langle cc^{(3 \text{out})} \rangle$	$0,02 \pm 0,06$	$0,06 \pm 0,09$
$\langle \ell \rangle$	$910,54 \pm 680,72$	$3,89 \pm 0,84$
$\langle B_i^{(n)} \rangle$	$8,68 \times 10^{-4} \pm 4,05 \times 10^{-3}$	$4,57 \times 10^{-3} \pm 1,36 \times 10^{-2}$

fazendo com que a correlação total seja moderada ($\rho_{\text{corr}} = 0,52$, Figura 7.1b)*. Esses genes, chamados *outliers*, são os seguintes: *crp*, *fnr*, *ihfA*, *ihfB*, *fis*, *arcA* e *hns*. Quando os *outliers* são desconsiderados, o coeficiente de correlação entre grau de saída e atividade sobe para $\rho_{\text{corr}} = 0,89$. Esses genes são, portanto, identificados pelo termo *outliers de dinâmica*.

Esses *outliers*, além de apresentarem um comportamento peculiar quanto ao relacionamento estrutura-dinâmica, exibem outras características particulares que os diferem do restante da rede. Diversas medidas foram calculadas para cada um dos vértices da rede integrada, formando assim vetores de atributos que foram analisados por meio de PCA (Seção 2.8, p. 66). Lembre-se que a dispersão dos dados é maior nas primeiras dimensões PCA, e as correlações entre os atributos originais são completamente removidas. Na análise aqui reportada, as duas primeiras dimensões foram utilizadas, permitindo assim uma inspeção visual para detecção de *outliers*. Cada vetor de atributos é formado pela medidas: k_i^{in} , k_i^{out} , $k_i^{(2|\text{in})}$, $k_i^{(2|\text{out})}$, $k_i^{(3|\text{in})}$, $k_i^{(3|\text{out})}$,

Figura 7.1 – Identificação dos *outliers* de dinâmica. (a) Correlações tipo A ($\vec{k}^{\text{in}} \times \vec{\pi}$) e (b) correlações tipo B ($\vec{k}^{\text{out}} \times \vec{\pi}$) na rede integrada. Os pontos em destaque são *outliers*. As retas em azul foram obtidas pelo método dos mínimos quadrados considerando-se inclusive os *outliers*.

* Note que a intensidade da correlação tipo B foi associada ao nível de reciprocidade da rede no Capítulo 3 (Figura 3.2, p. 75). O estudo incluído no capítulo atual complementa o do Capítulo 3 por meio da identificação e análise de *outliers* na rede *E. coli*.

cc_i^{in} , cc_i^{out} , $cc_i^{(3|in)}$, $cc_i^{(3|out)}$, ℓ_i e $B_i^{(n)}$, as mesmas utilizadas nas Tabela 7.1, com a ressalva de que agora são obtidas para cada vértice i . As duas primeiras dimensões da projeção PCA contêm 89% da variância total dos atributos originais (Figura 7.2). Nessa projeção é possível identificar *outliers* na porção superior da figura – identificados aqui por *outliers de estrutura*. Entre eles, encontram-se os *outliers* de dinâmica mencionados previamente. Portanto, os *outliers* de dinâmica também apresentam características estruturais distintas do restante da rede. Uma inspeção mais detalhada dos atributos estruturais dos *outliers* revelou que: (i) eles apresentam grau de saída muito maior que o grau de entrada, (ii) suas hierarquias de segundo nível apresentam um grande número de arestas de saída (alto $k^{(2|out)}$) e (iii) os comprimentos dos caminhos mínimos que saem dos *outliers* é menor que nos demais vértices, ou seja, é possível alcançar outros vértices a partir dos *outliers* em poucos passos.

A acessibilidade dos vértices da rede integrada foi também calculada (Seção 2.3, p. 51), considerando-se passos até $h = 5$. A acessibilidade de um vértice é uma medida baseada na sua dinâmica de acesso aos demais vértices da rede. Observou-se que os *outliers* apresentam,

Figura 7.2 – PCA dos vértices da rede integrada. A porcentagem de variância é de 77% no primeiro eixo e de 12% no segundo. Os pontos em destaque são *outliers* de estrutura e também de dinâmica (note que existem mais *outliers* de estrutura além dos selecionados na porção superior dessa figura).

sistematicamente, maiores acessibilidades que a grande maioria dos demais vértices da rede integrada (Figura 7.3). O teste da soma dos postos de Wilcoxon resultou em valores-p aproximadamente iguais a 8×10^{-6} (quando $h = 2$); 5×10^{-6} (quando $h = 3$); 7×10^{-3} (quando $h = 4$); e 0,02 (quando $h = 5$). O menor valor-p ocorre para $h = 3$, quando existe uma separação mais acentuada entre as distribuições de acessibilidade dos *outliers* e dos não *outliers* (Figura 7.3b). Ao se retomar o conceito de borda/núcleo definido em função da acessibilidade, nota-se que os *outliers* formam portanto o núcleo da rede. Ou seja, em termos da dinâmica de acessos, os *outliers* apresentam um papel central na regulação dos demais genes, pois regulam muitos outros genes de maneira balanceada (reveja a definição da acessibilidade na Seção 2.3).

Figura 7.3 – Diagramas em caixa (*boxplots*) para a acessibilidade na rede integrada. A acessibilidade foi separada por *outliers* e não *outliers* considerando-se (a) $h = 2$, (b) $h = 3$, (c) $h = 4$ e (d) $h = 5$.

Foi também realizado um procedimento de detecção de motivos na rede integrada. Enquanto *outliers* são vértices que apresentam características peculiares, motivos são pequenos grupos de vértices com conectividade bem definida e que ocorrem na rede mais do que o esperado. Foram selecionados motivos de tamanho no máximo 4 por meio da ferramenta *mfinder* (156) – o tamanho de um motivo é igual ao seu número de vértices. O método de enumeração completa foi escolhido (vide material suplementar de (55)), com 100 realizações de redes aleatórias para fins de comparação. Para um dado motivo, seu número de ocorrências M_{int} na rede integrada foi obtido, assim como o número médio de ocorrências μ_{rand} nas redes aleatórias (além do respectivo desvio-padrão σ_{rand}). O Z-score, dado por $(M_{\text{int}} - \mu_{\text{rand}}) / \sigma_{\text{rand}}$, foi utilizado como o principal indicador de relevância de um motivo, sendo que apenas motivos com Z-score > 2 foram considerados. Outros parâmetros foram utilizados: M-factor $> 1,1$ e Uniqueness $\geq$

Figura 7.4 – Motivos identificados na rede integrada. (a) Motivos de tamanho 3 e seus Z-scores. (b) Motivos de tamanho 3 separados de acordo com o tipo de aresta (azul: aresta TRN, laranja: aresta MSFN, verde: aresta PPIN) acompanhados do número de ocorrências. (c) Motivos de tamanho 4 e seus Z-scores.

Tabela 7.3 – Dez genes mais frequentes em cada tipo de motivo de tamanho 3. Para cada gene, são fornecidas suas frequências absoluta (FAbs) e relativa (FRel, em porcentagem) de ocorrência. Os genes em negrito são os *outliers* da Figura 7.1.

Motivo Tipo I			Motivo Tipo II			Motivo Tipo III		
Gene	FAbs	FRel	Gene	FAbs	FRel	Gene	FAbs	FRel
ihfB	92	42,4	crp	170	26,4	crp	115	32,8
ihfA	91	41,9	fnr	160	24,9	fnr	82	23,4
crp	47	21,7	arcA	95	14,8	fis	52	14,8
arcA	45	20,7	aceE	74	11,5	narL	49	14,0
fis	45	20,7	aceF	53	8,2	fhlA	42	12,0
fnr	45	20,7	lpd	45	7,0	ihfA	42	12,0
gadX	13	6,0	ihfA	39	6,1	ihfB	42	12,0
hupB	11	5,1	ihfB	39	6,1	pdhR	33	9,4
hns	10	4,6	fis	37	5,8	fur	19	5,4
gadE	9	4,1	sucC	29	4,5	hyfR	16	4,6

4, os quais fornecem, respectivamente, a fração $M_{\text{int}}/\mu_{\text{rand}}$ e o número mínimo de ocorrências do mesmo motivo com conjuntos distintos de vértices.

Os resultados para motivos de tamanho 2 indicam que arestas simétricas frequentemente ocorrem entre pares de vértices (Z-score = 147,72), principalmente devido à inclusão de diversas arestas bidirecionais da rede PPIN (Tabela 7.2). Motivos de tamanho 3 e seus respectivos Z-scores são exibidos na Figura 7.4a. O motivo conhecido por *feed-forward loop* (FFL) (55) aparece com Z-score = 2,61 (motivo tipo III), e outros dois motivos de tamanho 3 com Z-scores ainda maiores também foram encontrados. Note que esses dois motivos são formados pelo FFL com uma aresta adicional. Sete motivos de tamanho 4 também foram detectados (Figura 7.4c), sendo que três deles apresentam Z-scores muito maiores que os dos motivos de tamanho 3. Arestas simétricas ocorrem em quase todos os motivos de tamanho 4, especialmente nos tipos I e V. Além disso, a maior parte dos motivos de tamanho 4 (exceto os tipos I e VI) inclui o padrão FFL. Uma análise complementar revelou que os *outliers* da rede integrada atuam como importantes elementos na formação de motivos. A Tabela 7.3 contém os 10 vértices que mais participam dos motivos de tamanho 3, sendo que a maior parte deles é formada pelos sete *outliers* *crp*, *fnr*, *ihfA*, *ihfB*, *fis*, *arcA* e *hns*. Esses *outliers* também têm grande participação nos motivos de tamanho 4 (Tabela 7.4), principalmente nos tipos II, III e VII.

Os motivos de tamanho 3 são a seguir analisados mais profundamente, pois são os que apre-

Tabela 7.4 – Dez genes mais frequentes em cada tipo de motivo de tamanho 4. A estrutura de apresentação é a mesma da Tabela 7.3.

Motivo Tipo I			Motivo Tipo II			Motivo Tipo III		
Gene	FAbs	FRel	Gene	FAbs	FRel	Gene	FAbs	FRel
rplC	3161	30,1	ihfB	3749	66,3	hupB	197	49,9
rplV	2880	27,4	ihfA	3748	66,3	ihfB	176	44,6
rplM	1696	16,1	fis	996	17,6	ihfA	171	43,3
rplS	1280	12,2	arcA	934	16,5	rplC	129	32,7
rplB	1235	11,7	crp	934	16,5	hupA	112	28,4
hupA	1229	11,7	fnr	934	16,5	fis	90	22,8
rpsC	1170	11,1	pflB	165	2,9	crp	87	22,0
exbB	1110	10,6	caiC	131	2,3	rpsJ	44	11,1
aceF	975	9,3	caiT	131	2,3	hns	42	10,6
aceE	938	8,9	prfA	131	2,3	galT	40	10,1

Motivo Tipo IV			Motivo Tipo V			Motivo Tipo VI			Motivo Tipo VII		
Gene	FAbs	FRel	Gene	FAbs	FRel	Gene	FAbs	FRel	Gene	FAbs	FRel
fnr	275	32,2	aceE	1194	28,4	fnr	1503	32,0	fnr	711	44,2
arcA	237	27,8	aceF	1163	27,6	arcA	805	17,1	narL	630	39,2
aceF	222	26,0	rplD	986	23,4	aceE	711	15,1	fis	450	28,0
aceE	214	25,1	lpd	799	19,0	crp	677	14,4	crp	299	18,6
lpd	206	24,2	rplW	620	14,7	rplV	542	11,5	ihfA	263	16,4
rplD	173	20,3	hupA	587	13,9	rplC	504	10,7	ihfB	263	16,4
crp	135	15,8	rplC	587	13,9	aceF	488	10,4	fhlA	120	7,5
rplW	110	12,9	exbB	577	13,7	exbB	482	10,3	fur	119	7,4
pdhR	82	9,6	rplV	535	12,7	lpd	454	9,7	pdhR	93	5,8
rplV	72	8,4	rplM	487	11,6	ihfB	341	7,3	nirB	80	5,0

sentam *outliers* em todos os seus tipos. Considerando-se somente os vértices que participam de algum motivo de tamanho 3 (402 no total) e suas respectivas interligações, tem-se a seguinte distribuição de arestas: 1.041 arestas da rede TRN, 713 da PPIN e 69 da MSFN, ou seja, a maior parte das arestas das redes PPIN e MSFN não participa desses motivos (a Tabela 7.2 contém as estatísticas da rede completa). O motivo tipo II apresentou uma grande quantidade de instâncias (579) com uma mescla de interações de regulação de transcrição e interações proteína-proteína. Nesse caso, os TG's controlados pelo processo de transcrição interagem fisicamente entre si (interação proteína-proteína). Embora com menor frequência (56), existem também motivos do tipo II onde os TG's interagem entre si por meio de uma via metabólica/sinalizadora e de outra interação de regulação de transcrição. O motivo tipo III (FFL) é composto somente por interações de regulação de transcrição (315 ocorrências) ou por uma via metabólica/sinalizadora produzida por um dos TG's que controla a atividade do outro TG (36 ocorrências). Também fo-

ram encontradas 20 instâncias do tipo I onde os TF's interagem fisicamente a fim de controlar o TG existente no motivo. Essas observações sugerem que existe um denso inter-relacionamento entre os três tipos de interações considerados, possibilitando assim que os TF's sejam influenciados por diversos estímulos celulares no processo de regulação da transcrição gênica.

7.3 Funções Biológicas dos *Outliers*

Cada um dos sete TF's identificados como *outliers* foi descrito como regulador global na rede de transcrição da *E. coli* em pelo menos um trabalho reportado na literatura (77, 155, 157). Adicionalmente, os *outliers* podem ser divididos em dois grupos: (a) *crp*, *ihfA*, *ihfB*, *fis* e *hns*, que apresentam alta centralidade e, portanto, participam de uma grande parte dos caminhos mínimos que existem na rede, e (b) *fnr* e *arcA*, com $k_i^{(2|in)}$ baixos, o que indica que o número de arestas de entrada não cresce normalmente (quando comparado aos demais vértices) ao se considerar um nível hierárquico mais distante. Nota-se que o primeiro grupo forma um conjunto de reguladores globais identificado em todos os trabalhos de revisão anteriores (77, 155, 157), enquanto o segundo grupo contém TF's específicos à limitação de oxigênio e/ou à condição anaeróbica, sendo portanto mais relacionados a certas condições (158, 159).

Essas observações sugerem que o método aqui utilizado pode prever com alta precisão os reguladores globais de uma dada rede de transcrição. Em particular, ele possibilitou o isolamento de reguladores globais específicos a certas condições, um resultado não obtido pelos métodos previamente desenvolvidos para predição de fatores globais. Especula-se que a principal razão para os *outliers* de estrutura também serem *outliers* de dinâmica é a baixa quantidade de arestas direcionadas a eles (até no segundo nível hierárquico, em alguns casos), embora sejam reguladores importantes apresentando altos graus de saída e estejam bem próximos dos outros vértices exibindo caminhos de comprimento curto. Contudo, mesmo sendo elementos centrais na rede, eles falham em apresentar alta atividade.

Outra análise dos *outliers* foi conduzida empregando-se dados disponíveis na base M3D

Figura 7.5 – Diagramas em caixa (*boxplots*) para a expressão gênica considerando-se separadamente *outliers* e não *outliers*. O primeiro diagrama (a) refere-se à soma das intensidades de expressão para cada gene, e o segundo (b) mostra os resultados para o número de expressões acima da média por gene.

(160) relacionados às intensidades de expressão gênica na *E. coli* em 302 diferentes condições. Procurou-se observar se os *outliers* da rede integrada são também *outliers* de expressão gênica. Para tanto, as intensidades de expressão de cada gene foram somadas ao longo de todas as diferentes condições, e a seguir os resultados foram divididos entre *outliers* e não *outliers*. Observou-se que os *outliers* são mais intensamente expressos que os demais genes (Figura 7.5a, valor- $p < 10^{-7}$, teste da soma dos postos de Wilcoxon). Esse resultado mostra uma forte ligação entre propriedades estruturais/dinâmicas incomuns e intensidades de expressão gênica. Os dados de intensidade de expressão foram também filtrados por meio de limiarização: a expressão de um gene em determinada condição somente é considerada se sua intensidade de expressão é maior que a média das intensidades na mesma condição. Nesse caso, os *outliers* ainda são mais expressos que os demais genes, embora com significância estatística menor (Figura 7.5b, valor- $p \approx 0.05$).

Portanto, os métodos de integração de dados e de detecção de *outliers* por meio da seleção de vértices com comportamento peculiar no relacionamento estrutura-dinâmica permitiram a

identificação de genes reguladores globais com intensa expressão gênica. Além disso, a metodologia utilizada para análise de *outliers* permitiu o estudo das propriedades desses vértices e do papel fundamental desempenhado por eles na formação de motivos.

8 Conclusões

Diferentes correlações entre grau de entrada e atividade foram descobertas, as quais ocorrem dentro e fora da borda em diversas redes direcionadas (Capítulo 3). Consequentemente, é possível melhor relacionar a estrutura e a dinâmica nessas redes: na borda (vértices com menor acessibilidade) a atividade cresce mais rápido, conforme o grau de entrada aumenta, do que no núcleo (vértices com maior acessibilidade). Esse efeito é exibido de maneira mais intensificada nas redes WWW. Com esses resultados, é possível prever a atividade de um vértice observando-se apenas seu grau de entrada e seu nível de acessibilidade. Trabalhos adicionais devem melhor esclarecer as relações das propriedades da borda e do núcleo com os diferentes tipos de correlação observados entre grau de entrada e atividade. Já a medida de reciprocidade permitiu explicar a intensidade das correlações entre grau de saída e atividade: quanto maior a reciprocidade, mais forte a correlação. Com esses resultados, foi possível apresentar uma explicação bastante completa para as intrincadas correlações que existem em redes direcionadas, as quais têm sido negligenciadas na literatura de redes complexas. Futuramente, outros modelos de dinâmica podem ser utilizados, possivelmente mais intimamente relacionados a cada tipo de rede utilizada (por exemplo, modelo de navegação na WWW e de sincronização em redes neuronais). O papel da estrutura borda/núcleo pode ser melhor compreendida em experimentos com modelos clássicos de redes, ou com novos modelos criados especialmente para exibir níveis controlados dessa estrutura. Adicionalmente, é interessante no futuro observar as transições de correlação por meio de métodos de perturbação (criação ou remoção de vértices ou arestas).

O modelo APA, proposto no Capítulo 4, é baseado no crescimento do número de vértices e na ligação preferencial, dois mecanismos fundamentais do modelo clássico de Barabási e Albert. Entretanto, duas propriedades importantes do modelo APA o distinguem do modelo BA:

suas arestas são direcionadas e a criação de novas arestas é regida pela atividade dos vértices, ao invés do grau. Essas propriedades fazem com que o modelo APA seja uma generalização do modelo BA para redes direcionadas pois, quando a direção das arestas não é considerada no modelo APA, o modelo BA é exatamente reproduzido. O modelo proposto foi inspirado na existência de redes reais com forte relacionamento estrutura-dinâmica, como, por exemplo, a WWW, em que páginas bastante visitadas costumam ser referenciadas por um grande número de *hyperlinks*. Uma variante do modelo também foi definida, denominada APA', incorporando uma mudança no mecanismo de ligação preferencial. Ao se comparar as propriedades dos modelos ER, WS, BA, GG, APA e APA' com as de diversas redes reais, descobriu-se que a variante APA' é o modelo mais compatível com três redes corticais: duas do gato e uma do macaco. Ou seja, entre os modelos comparados, o modelo APA' é o que melhor representa (no contexto do experimento realizado) as redes corticais analisadas. Dessa maneira, esses resultados indicam que regiões corticais tendem a estabelecer interconexões com regiões altamente ativas, reforçando assim a teoria Hebbiana. Futuros avanços incluem a análise de outros tipos de redes corticais, como as que apresentam conexões em nível neuronal. Adicionalmente, outros tipos de dinâmica devem ser empregados na definição de variantes do modelo APA, tais como dinâmicas de transporte e de sincronização*.

A metodologia criada para caracterização da distribuição de subgrafos (Capítulo 5) engloba histogramas de medidas tradicionais e o conceito de relevância dos vértices, que permite unir subgrafos progressivamente na chamada expansão gradual, além de fornecer um quantificador da importância dos vértices no estabelecimento da conectividade entre subgrafos. Pretendeu-se analisar a dispersão de grupos de vértices com propriedades similares (de acordo com qualquer critério desejado) independentemente da quantidade de vértices em cada grupo. A metodologia proposta foi ilustrada por meio da análise de subgrafos cujos vértices apresentam alto coeficiente de agrupamento em redes artificiais e reais. Além de ilustrar a aplicação do método, esses experimentos possibilitaram uma maior compreensão das redes consideradas, inclusive com in-

* Após a publicação desse trabalho (51), um modelo para a WWW com ligação preferencial baseada nos valores de PageRank dos vértices foi divulgado (161). PageRank é um modelo de dinâmica utilizado pelo Google na ordenação de seus resultados de busca.

interpretações a respeito da relação desses resultados com algumas propriedades dos processos dinâmicos de transporte e de espalhamento. Para trabalhos futuros, nota-se que o método de expansão gradual pode também ser aplicado na análise multinível de redes, onde um conjunto de medidas seria aplicado aos subgrafos e, a cada limiar de relevância do procedimento de expansão, uma análise dos novos subgrafos seria realizada. Consequentemente, poderia ser traçada a variação das propriedades dos subgrafos. Outra possibilidade para a dinâmica de união de subgrafos é a obtenção de dendrogramas que representem toda a hierarquia de subgrafos ao longo dos diferentes limiares de relevância. Essa estrutura hierárquica possibilitaria uma visão complementar da organização dos subgrafos, em que seriam identificados quais subgrafos formam subgrafos maiores ao longo da expansão. Medidas da hierarquia resultante, como a altura da árvore de subgrafos, poderiam ser utilizadas como métricas adicionais para caracterização de redes. Pretende-se também adaptar a metodologia para a análise de redes direcionadas, possibilitando assim o estudo de grupos de vértices que apresentam diferentes tipos de correlação entre grau e atividade (vide Capítulo 3).

A coleta precisa e completa de redes é praticamente impossível em diversos casos, o que pode levar a conclusões errôneas em estudos baseados nessas redes. Tratou-se aqui (Capítulo 6) da amostragem de redes corticais por meio de um pequeno número de sensores posicionados na superfície do crânio, e do quanto essas amostras podem ter diferentes propriedades da rede neuronal original. Utilizando um modelo de rede cerebral em que neurônios são posicionados na superfície de uma semiesfera, e um modelo de amostragem inspirado nos métodos de encefalografia EEG/MEG, pôde-se verificar que a estrutura das amostras tende a se distanciar da estrutura da rede original, aproximando-se das propriedades das redes aleatórias. Em outras palavras, as características da rede original não puderam ser determinadas com precisão verificando-se somente a estrutura de suas amostras. Estudos anteriores têm associado às redes corticais o modelo WS, embora os resultados aqui reportados sugiram que, se o cérebro apresentasse uma estrutura pequeno mundo, as amostras seriam mais próximas do modelo ER. Amostras maiores tendem a se aproximar do modelo WS, mas o número de sensores é fisicamente limitado a uma pequena porcentagem do total de neurônios existentes no cérebro. A

reprodução das propriedades de processos dinâmicos foi também avaliada. O nível de sincronização nas amostras aproxima-se do observado nas redes originais conforme o número de sensores aumenta. Entretanto, a dinâmica de acessibilidade não é reproduzida nas amostras. Até que alguma metodologia para estimativa dos erros de amostragem seja formulada, inferências feitas a partir de amostras devem ser realizadas com cuidado. As simulações efetuadas podem ser futuramente adaptadas para que amostras com arestas direcionadas sejam obtidas, englobando assim as redes corticais como as utilizadas no Capítulo 4, as quais tiveram suas propriedades comparadas com as do modelo APA e sua variante APA'. A incorporação de redes direcionadas permitirá também que as correlações grau-atividade (Capítulo 3) sejam avaliadas em função do processo de amostragem. Adicionalmente, existe a possibilidade de que outras alterações sejam realizadas na metodologia apresentada, como a inclusão de outras medidas de redes e a comparação com outros modelos existentes, embora as dificuldades enfrentadas na obtenção de amostras precisas já tenham sido aqui evidenciadas.

Além de evidenciar a importância da integração de múltiplos tipos de dados provenientes de diferentes processos biológicos, o Capítulo 7 ilustra como o relacionamento estrutura-dinâmica pode ser utilizado em uma aplicação prática, mais especificamente, no estudo da rede integrada da bactéria *Escherichia coli*. Diversos tipos de interações foram consideradas em uma única rede: as proteína-proteína, as de regulação de transcrição gênica e as de sinalização/metabólicas. Genes identificados como *outliers* na correlação entre grau de saída e atividade foram identificados como reguladores globais na rede de transcrição da *E. coli*, os quais também apresentam intensa expressão gênica em centenas de diferentes condições. Outras propriedades desses *outliers* foram identificadas por meio da análise de diversas medidas de redes, da dinâmica de acessibilidade e de motivos. Particularmente, foi observado um importante papel dos *outliers* na formação de motivos. Outra contribuição do Capítulo 7 é a ilustração de como técnicas variadas de análise de redes complexas podem ser aplicadas a um mesmo objeto de estudo. Essa exemplificação é reforçada pelos resultados relevantes obtidos e pelo ineditismo da abordagem biológica integrada utilizada. Futuramente, o método aqui utilizado para identificação e análise de *outliers* deve ser aplicado a redes integradas de outros organismos.

Uma importante contribuição desta tese é a integração de diferentes abordagens no estudo do relacionamento estrutura-dinâmica em redes complexas. Resultados relevantes foram obtidos em diversos casos, incluindo investigações gerais (Capítulos 3 a 5) e aplicações (Capítulos 6 e 7). Em particular, todos esses casos apresentam avanços em termos de novos métodos e/ou novos resultados. Adicionalmente, trabalhou-se nesta tese com vários aspectos/conceitos a respeito de redes direcionadas, as quais são muito comuns e importantes para a área de redes complexas, mas que têm sido pouco estudadas na literatura. Em suma, as principais contribuições desta tese são:

1. Maior compreensão das correlações entre grau de saída e atividade em termos da medida de reciprocidade;
2. Idem para grau de entrada e atividade em termos da dinâmica de acessibilidade;
3. Criação de uma metodologia, utilizando a acessibilidade, para separação automática da borda e do núcleo em redes direcionadas;
4. Desenvolvimento do modelo APA, que generaliza o modelo BA para redes direcionadas;
5. Aproximação da estrutura de redes corticais pelo modelo variante APA’;
6. Desenvolvimento de uma metodologia para caracterização da distribuição de subgrafos;
7. Aplicação do método de análise da distribuição de subgrafos no estudo de grupos de vértices com alto agrupamento local, e relacionamento dos respectivos resultados com as dinâmicas de transporte e de espalhamento;
8. Elaboração de uma metodologia para simulação do processo de amostragem de redes corticais;
9. Quantificação dos efeitos da amostragem em termos de diversas medidas estruturais, da distribuição de subgrafos, do processo de sincronização e da dinâmica de acessibilidade;
10. Criação de uma rede integrada para a bactéria *E. coli* combinando-se três tipos de interações celulares;

11. Identificação de reguladores globais na *E. coli* por meio da detecção de *outliers* no relacionamento grau-atividade;
12. Caracterização desses *outliers* por meio da análise de diversas medidas estruturais, da dinâmica de acessibilidade e de motivos.

Por fim, ressalta-se que este projeto de doutoramento resultou em artigos divulgados nos seguintes periódicos internacionais: *Journal of Statistical Mechanics* (51), *New Journal of Physics* (58), *NeuroImage* (66) e *Advances in Physics* (91). Este último refere-se a um artigo de revisão da área de redes complexas, e foi elaborado em conjunto com diversos membros do grupo de pesquisa onde esta tese foi desenvolvida. Parte deste doutorado foi também apresentada no *2nd Workshop on Complex Networks* (CompleNet'2010) (37). Adicionalmente, está disponível um manuscrito (92) resultante da participação na Complex Systems Summer School, organizada anualmente pelo Santa Fe Institute na cidade de Santa Fe, EUA. Esse trabalho, no qual relaciona-se propriedades do passeio aleatório com a organização hierárquica de redes, tem íntima relação com os desenvolvimentos apresentados nesta tese. Entretanto, esse manuscrito está em um estágio inicial de desenvolvimento. Outros artigos estão em elaboração ou foram submetidos a periódicos visando uma completa divulgação dos resultados obtidos neste projeto.

Referências

- 1 BARABÁSI, A. L. *Linked: the new science of networks*. Cambridge, MA: Perseus, 2002.
- 2 HARARY, F. *Graph theory*. Reading, MA: Addison-Wesley, 1969.
- 3 ERDÖS, P.; RÉNYI, A. On random graphs I. *Publicationes Mathematicae Debrecen*, v. 6, p. 290–297, 1959.
- 4 ALBERT, R.; BARABÁSI, A. L. Statistical mechanics of complex networks. *Reviews of Modern Physics*, v. 74, n. 1, p. 47–97, 2002.
- 5 BOCCALETTI, S.; LATORA, V.; MORENO, Y.; CHAVEZ, M.; HWANG, D.-U. Complex networks: structure and dynamics. *Physics Reports*, v. 424, n. 4-5, p. 175–308, 2006.
- 6 NEWMAN, M. E. J. The structure and function of complex networks. *SIAM Review*, v. 45, n. 2, p. 167–256, 2003.
- 7 CORMEN, T. H.; LEISERSON, C. E.; RIVEST, R. L.; STEIN, C. *Introduction to algorithms*. New York: The MIT Press, 2001.
- 8 MILGRAM, S. The small world problem. *Psychology Today*, v. 1, n. 1, p. 61–67, 1967.
- 9 WALHOUT, A. J. M.; VIDAL, M. Protein interaction maps for model organisms. *Nature Reviews Molecular Cell Biology*, v. 2, n. 1, p. 55–63, 2001.
- 10 SCHLITT, T.; BRAZMA, A. Current approaches to gene regulatory network modelling. *BMC Bioinformatics*, v. 8, n. 6, p. S9, 2007.
- 11 MAZURIE, A.; BONCHEV, D.; SCHWIKOWSKI, B.; BUCK, G. Evolution of metabolic network organization. *BMC Systems Biology*, v. 4, n. 1, p. 59, 2010.

- 12 WATTS, D. J.; STROGATZ, S. H. Collective dynamics of ‘small-world’ networks. *Nature*, v. 393, n. 6684, p. 440–442, 1998.
- 13 BARABÁSI, A. L.; ALBERT, R. Emergence of scaling in random networks. *Science*, v. 286, n. 5439, p. 509–512, 1999.
- 14 FALOUTSOS, M.; FALOUTSOS, P.; FALOUTSOS, C. On power-law relationships of the Internet topology. In: CONFERENCE ON APPLICATIONS, TECHNOLOGIES, ARCHITECTURES, AND PROTOCOLS FOR COMPUTER COMMUNICATION, SIGCOMM-99, 1999, Cambridge, MA. *Proceedings...* Cambridge, MA: ACM Press, 1999. p. 251–262.
- 15 ALBERT, R.; JEONG, H.; BARABÁSI, A.-L. Error and attack tolerance of complex networks. *Nature*, v. 406, n. 6794, p. 378–382, 2000.
- 16 PASTOR-SATORRAS, R.; VESPIGNANI, A. Epidemic spreading in scale-free networks. *Physical Review Letters*, v. 86, n. 14, p. 3200–3203, 2001.
- 17 BOGUÑÁ, M.; KRIOUKOV, D.; CLAFFY, K. C. Navigability of complex networks. *Nature Physics*, v. 5, n. 1, p. 74–80, 2009.
- 18 TADIĆ, B.; RODGERS, G. J.; THURNER, S. Transport on complex networks: flow, jamming and optimization. *International Journal of Bifurcation and Chaos*, v. 17, n. 7, p. 2363–2385, 2007.
- 19 ARENAS, A.; DÍAZ-GUILERA, A.; KURTHS, J.; MORENO, Y.; ZHOU, C. Synchronization in complex networks. *Physics Reports*, v. 469, n. 3, p. 93–153, 2008.
- 20 BARRAT, A.; BARTHÉLEMY, M.; VESPIGNANI, A. *Dynamical processes on complex networks*. Cambridge: Cambridge University Press, 2008.
- 21 GUIMERÀ, R.; SALES-PARDO, M. Form follows function: the architecture of complex networks. *Molecular Systems Biology*, v. 2, p. 42, 2006. DOI: 10.1038/msb4100082.
- 22 ROSVALL, M.; BERGSTROM, C. T. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, v. 105, n. 4, p. 1118–1123, 2008.

- 23 NYKTER, M.; PRICE, N. D.; LARJO, A.; AHO, T.; KAUFFMAN, S. A.; YLI-HARJA, O.; SHMULEVICH, I. Critical networks exhibit maximal information diversity in structure-dynamics relationships. *Physical Review Letters*, v. 100, n. 5, p. 058702, 2008.
- 24 LOVÁSZ, L. Random walks on graphs: A survey. In: MIKLÓS, D.; SÓS, V. T.; SZŐNYI, T. (Ed.). *Combinatorics, Paul Erdős is eighty*. Budapest: János Bolyai Mathematical Society, 1996. v. 2, p. 353–398.
- 25 NOH, J. D.; RIEGER, H. Random walks on complex networks. *Physical Review Letters*, v. 92, n. 11, p. 118701, 2004.
- 26 BURIONI, R.; CASSI, D. Random walks on graphs: ideas, techniques and results. *Journal of Physics A*, v. 38, n. 8, p. R45–R78, 2005.
- 27 SAMUKHIN, A. N.; DOROGOVTSSEV, S. N.; MENDES, J. F. F. Laplacian spectra of, and random walks on, complex networks: are scale-free architectures really important? *Physical Review E*, v. 77, n. 3, p. 036115, 2008.
- 28 ALMAAS, E.; KULKARNI, R. V.; STROUD, D. Scaling properties of random walks on small-world networks. *Physical Review E*, v. 68, n. 5, p. 056105, 2003.
- 29 BOLLT, E. M.; BEN-AVRAHAM, D. What is special about diffusion on scale-free nets? *New Journal of Physics*, v. 7, n. 1, p. 26, 2005.
- 30 GALLOS, L. K. Random walk and trapping processes on scale-free networks. *Physical Review E*, v. 70, n. 4, p. 046116, 2004.
- 31 JASCH, F.; BLUMEN, A. Trapping of random walks on small-world networks. *Physical Review E*, v. 64, n. 6, p. 066104, 2001.
- 32 PARRIS, P. E.; KENKRE, V. M. Traversal times for random walks on small-world networks. *Physical Review E*, v. 72, n. 5, p. 056119, 2005.
- 33 SOOD, V.; REDNER, S.; BEN-AVRAHAM, D. First-passage properties of the Erdos-Renyi random graph. *Journal of Physics A*, v. 38, n. 1, p. 109–123, 2005.

- 34 COSTA, L. da F.; SPORNS, O.; ANTIQUEIRA, L.; NUNES, M. G. V.; OLIVEIRA JUNIOR, O. N. Correlations between structure and random walk dynamics in directed complex networks. *Applied Physics Letters*, v. 91, n. 5, p. 054107, 2007.
- 35 GARLASCHELLI, D.; LOFFREDO, M. I. Patterns of link reciprocity in directed networks. *Physical Review Letters*, v. 93, n. 26, p. 268701, 2004.
- 36 TRAVENÇOLO, B. A. N.; VIANA, M. P.; COSTA, L. da F. Border detection in complex networks. *New Journal of Physics*, v. 11, n. 6, p. 063019, 2009.
- 37 ANTIQUEIRA, L.; COSTA, L. da F. Structure-dynamics interplay in directed complex networks with border effects. In: WORKSHOP ON COMPLEX NETWORKS, COMPLENET'2010, 2010, Rio de Janeiro. *Proceedings...* Heidelberg: Springer-Verlag, 2011. (Communications in Computer and Information Science, v. 116), p. 46–56.
- 38 PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. *The PageRank citation ranking: bringing order to the Web*. Stanford, 1998. 17 p. Disponível em: <<http://ilpubs.stanford.edu:8090/422/>>. Acesso em: 09/2011.
- 39 KRAPIVSKY, P. L.; REDNER, S.; LEYVRAZ, F. Connectivity of growing random networks. *Physical Review Letters*, v. 85, n. 21, p. 4629–4632, 2000.
- 40 PONCELA, J.; GÓMEZ-GARDEÑES, J.; FLORÍA, L. M.; SÁNCHEZ, A.; MORENO, Y. Complex cooperative networks from evolutionary preferential attachment. *PLoS ONE*, v. 3, n. 6, p. e2449, 2008.
- 41 DONETTI, L.; HURTADO, P. I.; MUÑOZ, M. A. Network synchronization: optimal and pessimal scale-free topologies. *Journal of Physics A*, v. 41, n. 22, p. 224008, 2008.
- 42 CHEN, Q.; CHEN, S. A highly clustered scale-free network evolved by random walking. *Physica A*, v. 383, n. 2, p. 773–781, 2007.
- 43 EVANS, T. S.; SARAMÄKI, J. P. Scale-free networks from self-organization. *Physical Review E*, v. 72, n. 2, p. 026138, 2005.
- 44 SARAMÄKI, J.; KASKI, K. Scale-free networks generated by random walkers. *Physica A*, v. 341, p. 80–86, 2004. DOI: 10.1016/j.physa.2004.04.110.

- 45 SHAO, Z.; ZHOU, H. Dynamics-driven evolution to structural heterogeneity in complex networks. *Physica A*, v. 388, n. 4, p. 523–528, 2009.
- 46 HOLME, P.; KARLIN, J.; FORREST, S. An integrated model of traffic, geography and economy in the Internet. *ACM SIGCOMM Computer Communication Review*, v. 38, n. 3, p. 5–16, 2008.
- 47 HEBB, D. O. *The organization of behavior: a neuropsychological theory*. New York: John Wiley & Sons, 1949.
- 48 BAR-YAM, Y.; EPSTEIN, I. R. Response of complex networks to stimuli. *Proceedings of the National Academy of Sciences*, v. 101, n. 13, p. 4341–4345, 2004.
- 49 SCANNELL, J. W.; BURNS, G.; HILGETAG, C. C.; O'NEIL, M. A.; YOUNG, M. P. The connectional organization of the cortico-thalamic system of the cat. *Cerebral Cortex*, v. 9, n. 3, p. 277–299, 1999.
- 50 HONEY, C. J.; KÖTTER, R.; BREAKSPEAR, M.; SPORNS, O. Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proceedings of the National Academy of Sciences*, v. 104, n. 24, p. 10240, 2007.
- 51 ANTIQUEIRA, L.; RODRIGUES, F. A.; COSTA, L. da F. Modeling connectivity in terms of network activity. *Journal of Statistical Mechanics*, v. 2009, n. 9, p. L09005, 2009.
- 52 FORTUNATO, S. Community detection in graphs. *Physics Reports*, v. 486, n. 3-5, p. 75–174, 2010.
- 53 LOPEZ, E.; PARSHANI, R.; COHEN, R.; CARMI, S.; HAVLIN, S. Limited path percolation in complex networks. *Physical Review Letters*, v. 99, n. 18, p. 188701, 2007.
- 54 COSTA, L. da F.; RODRIGUES, F. A. What is there between any two nodes in a complex network? arXiv:0801.4068. 2008. Disponível em: <<http://arxiv.org/abs/0801.4068>>. Acesso em: 09/2011.
- 55 MILO, R.; SHEN-ORR, S.; ITZKOVITZ, S.; KASHTAN, N.; CHKLOVSKII, D.; ALON, U. Network motifs: simple building blocks of complex networks. *Science*, v. 298, n. 5594, p. 824–827, 2002.

- 56 COSTA, L. da F.; RODRIGUES, F. A. Seeking for simplicity in complex networks. *Europhysics Letters*, v. 85, n. 4, p. 48001, 2009.
- 57 MAKAROV, L. I. Estimations of subgraph positions in molecular graphs and their common subgraph peculiarities. *Journal of Structural Chemistry*, v. 46, n. 4, p. 738–743, 2005.
- 58 ANTIQUEIRA, L.; COSTA, L. da F. Characterization of subgraphs relationships and distribution in complex networks. *New Journal of Physics*, v. 11, n. 1, p. 013058, 2009.
- 59 STUMPF, M. P. H.; WIUF, C.; MAY, R. M. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proceedings of the National Academy of Sciences*, v. 102, n. 12, p. 4221–4224, 2005.
- 60 CLAUSET, A.; MOORE, C. Accuracy and scaling phenomena in Internet mapping. *Physical Review Letters*, v. 94, n. 1, p. 018701, 2005.
- 61 DALL’ASTA, L.; ALVAREZ-HAMELIN, I.; BARRAT, A.; VÁZQUEZ, A.; VESPIGNANI, A. Exploring networks with traceroute-like probes: theory and simulations. *Theoretical Computer Science*, v. 355, n. 1, p. 6–24, 2006.
- 62 LEE, S. H.; KIM, P.-J.; JEONG, H. Statistical properties of sampled networks. *Physical Review E*, v. 73, n. 1, p. 016102, 2006.
- 63 KIM, P.-J.; JEONG, H. Reliability of rank order in sampled networks. *The European Physical Journal B*, v. 55, n. 1, p. 109–114, 2007.
- 64 LATAPY, M.; MAGNIEN, C. Complex network measurements: estimating the relevance of observed properties. In: IEEE CONFERENCE ON COMPUTER COMMUNICATIONS, INFOCOM-2008, 2008, Phoenix, AZ. *Proceedings...* Phoenix, AZ: IEEE, 2008. p. 1660–1668.
- 65 HAN, J.-D. J.; DUPUY, D.; BERTIN, N.; CUSICK, M. E.; VIDAL, M. Effect of sampling on topology predictions of protein-protein interaction networks. *Nature Biotechnology*, v. 23, n. 7, p. 839–844, 2005.
- 66 ANTIQUEIRA, L.; RODRIGUES, F. A.; VAN WIJK, B. C. M.; COSTA, L. D. F.; DAFERTSHOFER, A. Estimating complex cortical networks via surface recordings – a critical note. *NeuroImage*, v. 53, n. 2, p. 439–449, 2010.

- 67 BARABÁSI, A.-L.; OLTVAI, Z. N. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, v. 5, n. 2, p. 101–113, 2004.
- 68 AMARAL, L. A. N.; OTTINO, J. M. Complex networks: augmenting the framework for the study of complex systems. *European Physical Journal B*, v. 38, n. 2, p. 147–162, 2004.
- 69 JEONG, H.; TOMBOR, B.; ALBERT, R.; OLTVAI, Z. N.; BARABÁSI, A. L. The large-scale organization of metabolic networks. *Nature*, v. 407, n. 6804, p. 651–654, 2000.
- 70 JEONG, H.; MASON, S.; BARABÁSI, A. L.; OLTVAI, Z. N. Lethality and centrality in protein networks. *Nature*, v. 411, n. 6833, p. 41–42, 2001.
- 71 SHEN-ORR, S. S.; MILO, R.; MANGAN, S.; ALON, U. Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics*, v. 31, n. 1, p. 64–68, 2002.
- 72 NG, A.; BURSTEINAS, B.; GAO, Q.; MOLLISON, E.; ZVELEBIL, M. Resources for integrative systems biology: from data through databases to networks and dynamic system models. *Briefings in Bioinformatics*, v. 7, n. 4, p. 318–330, 2006.
- 73 LU, H.; SHI, B.; WU, G.; ZHANG, Y.; ZHU, X.; ZHANG, Z.; LIU, C.; ZHAO, Y.; WU, T.; WANG, J.; CHEN, R. Integrated analysis of multiple data sources reveals modular structure of biological networks. *Biochemical and Biophysical Research Communications*, v. 345, n. 1, p. 302–309, 2006.
- 74 GAMA-CASTRO, S. et al. RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Research*, v. 36, n. Database issue, p. D120–D124, 2008.
- 75 SIERRO, N.; MAKITA, Y.; DE HOON, M.; NAKAI, K. DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Research*, v. 36, n. Database issue, p. D93–D96, 2008.
- 76 JANGA, S. C.; COLLADO-VIDES, J. Structure and evolution of gene regulatory networks in microbial genomes. *Research in Microbiology*, v. 158, n. 10, p. 787–794, 2007.
- 77 FREYRE-GONZALEZ, J. A.; ALONSO-PAVON, J. A.; TREVINO-QUINTANILLA, L. G.; COLLADO-VIDES, J. Functional architecture of Escherichia coli: new insights provided by a natural decomposition approach. *Genome Biology*, v. 9, n. 10, p. R154, 2008.

- 78 JANGA, S. C.; SALGADO, H.; MARTINEZ-ANTONIO, A.; COLLADO-VIDES, J. Co-ordination logic of the sensing machinery in the transcriptional regulatory network of *Escherichia coli*. *Nucleic Acids Research*, v. 35, n. 20, p. 6963–6972, 2007.
- 79 JANGA, S. C.; SALGADO, H.; MARTINEZ-ANTONIO, A. Transcriptional regulation shapes the organization of genes on bacterial chromosomes. *Nucleic Acids Research*, v. 37, n. 11, p. 3680–3688, 2009.
- 80 BABU, M. M.; TEICHMANN, S. A.; ARAVIND, L. Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *Journal of Molecular Biology*, v. 358, n. 2, p. 614–633, 2006.
- 81 RAVASZ, E.; SOMERA, A. L.; MONGRU, D. A.; OLTVAI, Z. N.; BARABÁSI, A. L. Hierarchical organization of modularity in metabolic networks. *Science*, v. 297, n. 5586, p. 1551–1555, 2002.
- 82 SCHNELL, S.; FORTUNATO, S.; ROY, S. Is the intrinsic disorder of proteins the cause of the scale-free architecture of protein-protein interaction networks? *Proteomics*, v. 7, n. 6, p. 961–964, 2007.
- 83 BUTLAND, G.; PEREGRIN-ALVAREZ, J. M.; LI, J.; YANG, W.; YANG, X.; CANADIEN, V.; STAROSTINE, A.; RICHARDS, D.; BEATTIE, B.; KROGAN, N.; DAVEY, M.; PARKINSON, J.; GREENBLATT, J.; EMILI, A. Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature*, v. 433, n. 7025, p. 531–537, 2005.
- 84 BUZSÁKI, G.; GEISLER, C.; HENZE, D. A.; WANG, X.-J. Interneuron diversity series: circuit complexity and axon wiring economy of cortical interneurons. *Trends in Neurosciences*, v. 27, n. 4, p. 186–193, 2004.
- 85 EYTAN, D.; MAROM, S. Dynamics and effective topology underlying synchronization in networks of cortical neurons. *The Journal of Neuroscience*, v. 26, n. 33, p. 8465–8476, 2006.
- 86 HARTWELL, L. H.; HOPFIELD, J. J.; LEIBLER, S.; MURRAY, A. W. From molecular to modular cell biology. *Nature*, v. 402, n. 6761, p. C47–C52, 1999.
- 87 WUCHTY, S.; OLTVAI, Z. N.; BARABÁSI, A.-L. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Genetics*, v. 35, n. 2, p. 176–179, 2003.

- 88 MARTINEZ-ANTONIO, A.; JANGA, S. C.; SALGADO, H.; COLLADO-VIDES, J. Internal-sensing machinery directs the activity of the regulatory network in *Escherichia coli*. *Trends in Microbiology*, v. 14, n. 1, p. 22–27, 2006.
- 89 HU, P. et al. Global functional atlas of *Escherichia coli* encompassing previously uncharacterized proteins. *PLoS Biology*, v. 7, n. 4, p. e96, 2009.
- 90 COSTA, L. da F.; RODRIGUES, F. A.; HILGETAG, C. C.; KAISER, M. Beyond the average: detecting global singular nodes from local features in complex networks. *Europhysics Letters*, v. 87, n. 1, p. 18008, 2009.
- 91 COSTA, L. da F.; OLIVEIRA JUNIOR, O. N.; TRAVIESO, G.; RODRIGUES, F. A.; VILLAS BOAS, P. R.; ANTIQUEIRA, L.; VIANA, M. P.; ROCHA, L. E. C. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Advances in Physics*, v. 60, n. 3, p. 329–412, 2011.
- 92 ANTIQUEIRA, L.; BLASI, D.; WRÓBEL, B. *Diffusion at the boundary: measures to characterize the diffusion process in complex networks*. 2010. 5 p. Disponível em: <<http://tuvalu.santafe.edu/events/workshops/images/d/d8/AntiqueiraBlasiWrobel.pdf>>. Acesso em: 09/2011.
- 93 COSTA, L. da F.; RODRIGUES, F. A.; TRAVIESO, G.; VILLAS BOAS, P. R. Characterization of complex networks: a survey of measurements. *Advances in Physics*, v. 56, n. 1, p. 167–242, 2007.
- 94 NEWMAN, M. E. J. Assortative mixing in networks. *Physical Review Letters*, v. 89, n. 20, p. 208701, 2002.
- 95 COSTA, L. da F. L-percolations of complex networks. *Physical Review E*, v. 70, n. 5, p. 056106, 2004.
- 96 COSTA, L. da F.; SILVA, F. N. Hierarchical characterization of complex networks. *Journal of Statistical Physics*, v. 125, n. 4, p. 841–872, 2006.
- 97 COSTA, L. da F. Learning about knowledge: a complex network approach. *Physical Review E*, v. 74, n. 2, p. 026103, 2006.
- 98 BONDY, J. A.; MURTY, U. S. R. *Graph theory with applications*. New York: Elsevier, 1976.

- 99 LATORA, V.; MARCHIORI, M. Efficient behavior of small-world networks. *Physical Review Letters*, v. 87, n. 19, p. 198701, 2001.
- 100 FREEMAN, L. C. A set of measures of centrality based on betweenness. *Sociometry*, v. 40, n. 1, p. 35–41, 1977.
- 101 MEYER, C. D. *Matrix analysis and applied linear algebra*. Philadelphia: SIAM, 2000.
- 102 GRIMMETT, G. R.; STIRZAKER, D. R. *Probability and random processes*. New York: Oxford University Press, 2001.
- 103 BOLLOBÁS, B. *Modern graph theory*. New York: Springer-Verlag, 1998.
- 104 JARVIS, J. P.; SHIER, D. R. Graph-theoretic analysis of finite markov chains. In: SHIER, D. R.; WALLENIS, K. T. (Ed.). *Applied mathematical modeling: a multidisciplinary approach*. Boca Raton, FL: CRC Press, 1999. cap. 13, p. 271–290.
- 105 NORRIS, J. R. *Markov chains*. Cambridge: Cambridge University Press, 1997.
- 106 ZHAO, L.; LAI, Y.-C.; PARK, K.; YE, N. Onset of traffic congestion in complex networks. *Physical Review E*, v. 71, n. 2, p. 026125, 2005.
- 107 HONG, H.; CHOI, M. Y.; KIM, B. J. Synchronization on small-world networks. *Physical Review E*, v. 65, n. 2, p. 026139, 2002.
- 108 BARTHÉLEMY, M. Spatial networks. *Physics Reports*, v. 499, n. 1-3, p. 1–101, 2011.
- 109 NEWMAN, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, v. 74, n. 3, p. 036104, 2006.
- 110 RICHARDSON, M.; AGRAWAL, R.; DOMINGOS, P. Trust management for the semantic web. In: INTERNATIONAL CONFERENCE ON COMPUTER AND ELECTRICAL ENGINEERING, ISWC'2003, 2003, Los Alamitos, CA. *Proceedings...* Heidelberg: Springer Berlin, 2003. (Lecture Notes in Computer Science, v. 2870), p. 351–368.
- 111 OPSAHL, T.; PANZARASA, P. Clustering in weighted networks. *Social Networks*, v. 31, n. 2, p. 155–163, 2009.

- 112 GUIMERÀ, R.; DANON, L.; DÍAZ-GUILERA, A.; GIRALT, F.; ARENAS, A. Self-similar community structure in a network of human interactions. *Physical Review E*, v. 68, n. 6, p. 065103, 2003.
- 113 HOLANDA, A. J.; PISA, I. T.; KINOCHI, O.; MARTINEZ, A. S.; RUIZ, E. E. S. Thesaurus as a complex network. *Physica A*, v. 344, n. 3-4, p. 530–536, 2004.
- 114 KISS, G. R.; ARMSTRONG, C.; MILROY, R.; PIPER, J. An associative thesaurus of english and its computer analysis. In: AITKIN, A. J.; BAILEY, R. W.; HAMILTON-SMITH, N. (Ed.). *The computer and literary studies*. Edinburgh, UK: Edinburgh University Press, 1973. p. 153–165.
- 115 NELSON, D. L.; MCEVOY, L. C.; SCHREIBER, T. A. *The University of South Florida word association, rhyme, and word fragment norms*. 1998. Disponível em: <<http://w3.usf.edu/FreeAssociation/>>. Acesso em: 09/2011.
- 116 HOWE, D. *FOLDOC*: free on-line dictionary of computing. 2002. Disponível em: <<http://www.foldoc.org>>. Acesso em: 09/2011.
- 117 REITZ, J. M. *ODLIS*: online dictionary for library and information science. 2002. Disponível em: <<http://lu.com/odlis/>>. Acesso em: 09/2011.
- 118 RIPEANU, M.; FOSTER, I.; IAMNITCHI, A. Mapping the Gnutella network: properties of large-scale peer-to-peer systems and implications for system design. *IEEE Internet Computing Journal*, v. 6, n. 1, p. 1–12, 2002.
- 119 THELWALL, M. A free database of university web links: data collection issues. *Cybermetrics*, v. 6/7, n. 1, p. 2, 2002/3.
- 120 ADAMIC, L. A.; GLANCE, N. The political blogosphere and the 2004 U.S. election: divided they blog. In: WORKSHOP ON LINK DISCOVERY, LinkKDD'05, 2005, Chicago, IL. *Proceedings...* New York, NY: ACM, 2005. p. 36–43.
- 121 JAARSMA, N. G.; DE BOER, S. M.; TOWNSEND, C. R.; THOMPSON, R. M.; EDWARDS, E. D. Characterising food-webs in two New Zealand streams. *New Zealand Journal of Marine and Freshwater Research*, v. 32, n. 2, p. 271–286, 1998.
- 122 YOUNG, M. P. The organization of neural systems in the primate cerebral cortex. *Proceedings: Biological Sciences*, v. 252, n. 1333, p. 13–18, 1993.

- 123 KAISER, M.; HILGETAG, C. C. Nonoptimal component placement, but short processing paths, due to long-distance projections in neural systems. *PLoS Computational Biology*, v. 2, n. 7, p. e95, 2006.
- 124 FELLEMAN, D. J.; ESSEN, D. C. V. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, v. 1, n. 1, p. 1–47, 1991.
- 125 COSTA, L. da F.; SPORNS, O. Correlating thalamocortical connectivity and activity. *Applied Physics Letters*, v. 89, n. 1, p. 013903, 2006.
- 126 ACHACOSO, T. B.; YAMAMOTO, W. S. *AY's neuroanatomy of C. elegans for computation*. Boca Raton, FL: CRC Press, 1992.
- 127 DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern classification*. New York: John Wiley & Sons, 2000.
- 128 FUKUNAGA, K. *Introduction to statistical pattern recognition*. New York: Academic Press, 1990.
- 129 FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, v. 7, n. 2, p. 179–188, 1936.
- 130 JOLLIFFE, I. T. *Principal component analysis*. New York: Springer, 2002.
- 131 EVERITT, B. S. *The Cambridge dictionary of statistics*. Cambridge: Cambridge University Press, 2006.
- 132 COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, v. 13, n. 1, p. 21–27, 1967.
- 133 MYERS, J. L.; WELL, A. D. *Research design and statistical analysis*. New Jersey: Lawrence Erlbaum Associates, 2003.
- 134 MONTROYA, J. M.; SOLÉ, R. V. Small world patterns in food webs. *Journal of Theoretical Biology*, v. 214, n. 3, p. 405–412, 2002.

- 135 SPORNS, O.; ZWI, J. D. The small world of the cerebral cortex. *Neuroinformatics*, v. 2, n. 2, p. 145–162, 2004.
- 136 HILGETAG, C. C.; KAISER, M. Clustered organization of cortical connectivity. *Neuroinformatics*, v. 2, n. 3, p. 353–360, 2004.
- 137 BASSETT, D. S.; BULLMORE, E. Small-world brain networks. *The Neuroscientist*, v. 12, n. 6, p. 512, 2006.
- 138 HEIJMANS, H. J. A. M.; NACKEN, P.; TOET, A.; VINCENT, L. Graph morphology. *Journal of Visual Communication and Image Representation*, v. 3, n. 1, p. 24–38, 1992.
- 139 COSTA, L. da F.; ROCHA, L. E. C. A generalized approach to complex networks. *European Physical Journal B*, v. 50, n. 1-2, p. 237–242, 2006.
- 140 ECHENIQUE, P.; GÓMEZ-GARDEÑES, J.; MORENO, Y. Improved routing strategies for Internet traffic delivery. *Physical Review E*, v. 70, n. 5, p. 056105, 2004.
- 141 BULLMORE, E.; SPORNS, O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, v. 10, n. 3, p. 186–198, 2009.
- 142 ACHARD, S.; SALVADOR, R.; WHITCHER, B.; SUCKLING, J.; BULLMORE, E. A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *The Journal of Neuroscience*, v. 26, n. 1, p. 63–72, 2006.
- 143 SALVADOR, R.; SUCKLING, J.; COLEMAN, M. R.; PICKARD, J. D.; MENON, D.; BULLMORE, E. Neurophysiological architecture of functional magnetic resonance images of human brain. *Cerebral Cortex*, v. 15, n. 9, p. 1332–1342, 2005.
- 144 HE, Y.; CHEN, Z. J.; EVANS, A. C. Small-world anatomical networks in the human brain revealed by cortical thickness from MRI. *Cerebral Cortex*, v. 17, n. 10, p. 2407–2419, 2007.
- 145 VAN DEN HEUVEL, M. P.; STAM, C. J.; BOERSMA, M.; HULSHOFF POL, H. E. Small-world and scale-free organization of voxel-based resting-state functional connectivity in the human brain. *NeuroImage*, v. 43, n. 3, p. 528–539, 2008.

- 146 BASSETT, D. S.; MEYER-LINDENBERG, A.; ACHARD, S.; DUKE, T.; BULLMORE, E. Adaptive reconfiguration of fractal small-world human brain functional networks. *Proceedings of the National Academy of Sciences*, v. 103, n. 51, p. 19518–19523, 2006.
- 147 STAM, C. J. Functional connectivity patterns of human magnetoencephalographic recordings: a ‘small-world’ network? *Neuroscience Letters*, v. 355, n. 1-2, p. 25–28, 2004.
- 148 REIJNEVELD, J. C.; PONTEN, S. C.; BERENDSE, H. W.; STAM, C. J. The application of graph theoretical analysis to complex networks in the brain. *Clinical Neurophysiology*, v. 118, n. 11, p. 2317–2331, 2007.
- 149 MICHELOYANNIS, S.; PACHOU, E.; STAM, C. J.; VOURKAS, M.; ERIMAKI, S.; TSIRKA, V. Using graph theoretical analysis of multi channel EEG to evaluate the neural efficiency hypothesis. *Neuroscience Letters*, v. 402, n. 3, p. 273–277, 2006.
- 150 PONTEN, S. C.; BARTOLOMEI, F.; STAM, C. J. Small-world networks and epilepsy: graph theoretical analysis of intracerebrally recorded mesial temporal lobe seizures. *Clinical Neurophysiology*, v. 118, n. 4, p. 918–927, 2007.
- 151 STAM, C. J.; DE HAAN, W.; DAFFERTSHOFER, A.; JONES, B. F.; MANSHANDEN, I.; VAN CAPPELLEN VAN WALSUM, A. M.; MONTEZ, T.; VERBUNT, J. P. A.; DE MUNCK, J. C.; VAN DIJK, B. W.; BERENDSE, H. W.; SCHELTENS, P. Graph theoretical analysis of magnetoencephalographic functional connectivity in Alzheimer’s disease. *Brain*, v. 132, n. 1, p. 213–224, 2009.
- 152 LIU, Y.; LIANG, M.; ZHOU, Y.; HE, Y.; HAO, Y.; SONG, M.; YU, C.; LIU, H.; LIU, Z.; JIANG, T. Disrupted small-world networks in schizophrenia. *Brain*, v. 131, n. 4, p. 945–961, 2008.
- 153 SUPEKAR, K.; MENON, V.; RUBIN, D.; MUSEN, M.; GREICIUS, M. D. Network analysis of intrinsic functional brain connectivity in Alzheimer’s disease. *PLoS Computational Biology*, v. 4, n. 6, p. e1000100, 2008.
- 154 NEWMAN, M. E. J.; WATTS, D. J. Renormalization group analysis of the small-world network model. *Physics Letters A*, v. 263, n. 4-6, p. 341–346, 1999.
- 155 MARTINEZ-ANTONIO, A.; JANGA, S. C.; THIEFFRY, D. Functional organisation of Escherichia coli transcriptional regulatory network. *Journal of Molecular Biology*, v. 381, n. 1, p. 238–247, 2008.

- 156 KASHTAN, N.; ITZKOVITZ, S.; MILO, R.; ALON, U. *mfinder*: network motif detection tool. 2005. Disponível em: <<http://www.weizmann.ac.il/mcb/UriAlon/>>. Acesso em: 09/2011.
- 157 MARTINEZ-ANTONIO, A.; COLLADO-VIDES, J. Identifying global regulators in transcriptional regulatory networks in bacteria. *Current Opinion in Microbiology*, v. 6, n. 5, p. 482–489, 2003.
- 158 COMPAN, I.; TOUATI, D. Anaerobic activation of *arcA* transcription in *Escherichia coli*: roles of Fnr and ArcA. *Molecular Microbiology*, v. 11, n. 5, p. 955–964, 1994.
- 159 SPIRO, S.; GUEST, J. R. Inactivation of the FNR protein of *Escherichia coli* by targeted mutagenesis in the N-terminal region. *Molecular Microbiology*, v. 2, n. 6, p. 701–707, 1988.
- 160 FAITH, J. J.; DRISCOLL, M. E.; FUSARO, V. A.; COSGROVE, E. J.; HAYETE, B.; JUHN, F. S.; SCHNEIDER, S. J.; GARDNER, T. S. Many microbe microarrays database: uniformly normalized affymetrix compendia with structured experimental metadata. *Nucleic Acids Research*, v. 36, n. Database issue, p. D866–D870, 2008.
- 161 GIAMMATTEO, P.; DONATO, D.; ZLATIĆ, V.; CALDARELLI, G. A PageRank-based preferential attachment model for the evolution of the World Wide Web. *Europhysics Letters*, v. 91, n. 1, p. 18004, 2010.