INSTITUTO DE FÍSICA DE SÃO CARLOS
UNIVERSIDADE DE SÃO CARLOS

FILIPI NASCIMENTO SILVA

Redes Complexas: Novas Metodologias e Modelagem de Aquisição de Conhecimento.

São Carlos
2009

FILIPI NASCIMENTO SILVA

Redes Complexas: Novas Metodologias e Modelagem de Aquisição de Conhecimento.

Dissertação apresentada ao Programa de Pós-graduação em Física do Instituto de Física de São Carlos da Universidade de São Paulo, para a obtenção do título de Mestre em Ciência.

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

São Carlos
2009

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

Silva, Filipi Nascimento
Redes complexas: novas metodologias e modelagem de aquisição de conhecimento / Filipi Nascimento Silva; orientador Luciano da Fontoura Costa.-- São Carlos, 2009.
185 p.

Dissertação (Mestrado em Ciência - Área de concentração: Física Aplicada – Opção: Física Computacional) – Instituto de Física de São Carlos da Universidade de São Paulo.

1. Redes complexas. 2. Sistemas complexos. 3. Visualização computacional. 4. Propriedades concêntricas. I. Título.

FOLHA DE APROVAÇÃO

Filipi Nascimento Silva

Dissertação apresentada ao Instituto de Física de São Carlos da Universidade de São Paulo para obtenção do título de Mestre em Ciências. Área de Concentração: Física Aplicada – Opção: Física Computacional.

Aprovado em: 17/12/2009

Comissão Julgadora

Prof. Dr. Roberto Fernandes Silva Andrade

Instituição: UFBA Assinatura

Prof. Dr. Alneu de Andrade Lopes

Instituição:ICMC/USP Assinatura

Prof. Dr. Luciano da Fontoura Costa

Instituição: IFSC/USP Assinatura

Dedico este trabalho a minha mãe Edna, a minha tia Eliana e a minha avó Nélia pelo amor e carinho.

Agradecimentos

- Agradeço ao professor Luciano da Fontoura Costa, pela oportunidade de realizar pesquisas científicas interessantes, por sua orientação séria e competente, pela confiança e pela possibilidade de realização deste trabalho.
- Aos colegas e amigos da graduação, Bruno, Celso, Danilo, Fabio, Mariane e Rejane; pelos momentos de descontração e sessões de RPG.
- Aos colegas e ex-colegas do grupo de visão cibernética, Bruno, Debora, Luiz, Matheus e Mauro; pelo apoio e discussões acadêmicas.
- Aos amigos *online*, Eric, Hugo, Lucas e Roger pela ajuda técnica e compartilhamento de informações.
- À Marilza da biblioteca pelo apoio e fornecimento de dados cruciais para a elaboração deste trabalho.
- À professora Maria Cristina pelas aulas de visualização computacional que estimularam a criação do visualizador de redes apresentado neste trabalho.
- À minha família pelo apoio e paciência.
- Ao Conselho Nacional de Desenvolvimento Ciêntifico e Tecnológico - CNPq, pelo apoio financeiro.
- Às bibliotecárias do IFSC, em especial à Neusa pela simpatia e ajuda nos últimos estagios de desenvolvimento deste trabalho.

“Two are better than one; because they have a good reward for their labour. For if they fall, the one will lift up his fellow: but woe to him that is alone when he falleth; for he hath not another to help him up”
— KING SOLOMON (1011 BC - 932 BC)

“If in other sciences we should arrive at certainty without doubt and truth without error, it behooves us to place the foundations of knowledge in mathematics”
— ROGER BACON (1214 - 1294)

“I do not think that the wireless waves I have discovered will have any practical application”
— HEINRICH RUDOLF HERTZ (1857 - 1894)

Resumo

Silva, F. N. *Redes complexas: novas metodologias e modelagem de aquisição de conhecimento*. 2009. 185p. Dissertação (Mestrado) - Instituto de Física de São Carlos, Universidade de São Paulo, 2009.

Estudos em redes complexas têm ganhado cada vez mais atenção devido ao seu potencial de representação simples de modelos complexos em diversas áreas de conhecimento. A obtenção de modelos quantitativos que representem fenômenos observados da natureza, assim como o desenvolvimento de metodologias de caracterização de redes complexas, tornaram-se essenciais para a compreensão e desenvolvimento de pesquisas com essas estruturas. Este trabalho tem como objetivo desenvolver e estudar alguns métodos recentes, usados para a caracterização de redes complexas, explorando-os no contexto da modelagem de conhecimento. Para isso, duas redes complexas foram geradas, uma rede de colaboração de pesquisadores da USP e outra obtida a partir do banco de dados de artigos da Wikipédia, considerando apenas aqueles da categoria de teoremas matemáticos. As medidas concêntricas, que foram recentemente formalizadas, são exploradas e aplicadas às redes descritas, assim como para diversos modelos teóricos, fornecendo informações muito relevantes sobre a topologia dessas redes. Resultados ainda mais interessantes são obtidos pela caracterização dos vértices da rede de colaboração, que revelam padrões de interdisciplinaridade entre as diferentes áreas do conhecimento. Um modelo de aquisição de conhecimento também foi proposto, aplicando a utilização de simulações de múltiplos agentes interagentes que caminham por uma rede complexa segundo uma heurística auto-esquivante. Resultados dessas simulações, realizadas para a rede da Wikipédia e outros modelos teóricos, mostram que certas configurações de parâmetros e de redes apresentam melhor desempenho na aquisição do conhecimento, com a rede de teoremas apresentando o pior deles. Entretanto, diferentemente do que era esperado, a variação da memória dos agentes pouco influência a velocidade de aquisição de conhecimento dos agentes. A frequência de acesso dos vértices pelos agentes também foi determinada e explorada superficialmente. Diversos softwares foram desenvolvidos para uso neste projeto de mestrado, dentre eles destaca-se o visualizador 3D, que se tornou indispensável para a análise das contribuições das outras propriedades apresentadas.

Palavras-chave: Redes Complexas, Sistemas Complexos, Visualização Computacional, Propriedades Concêntricas.

Abstract

Silva, F. N. *Complex Networks: New methodologies and knowledge acquisition modeling*. 2009. 185p. Dissertation (Master program) - Instituto de Física de São Carlos, Universidade de São Paulo, 2009.

Studies of complex networks have gained increasing research interest in recent years, in part due to its potential for simple representation of complex systems in various fields of science. The needs of quantitative models representing observed phenomena, as well the development of methods for the characterization of complex networks, is a essential matter for the development and understanding of scientific researches exploring such structures. This work aims to develop and study some new methods for the characterization of complex networks, exploring them in the context of knowledge modeling. Initially, two complex networks were developed, a collaborative network of researchers from the Universidade de São Paulo and the other obtained from the database of Wikipédia articles, considering only those strict related to mathematical theorems. The recently formalized concentric measurements are explored and applied to the described networks, as well to other several theoretical models, providing much more information about the topology of these networks than by the use of traditional measurements. Even more interesting results are obtained by the characterization of the vertices of the collaboration network, which reveal patterns of interdisciplinarity among the many fields of science. A model of knowledge acquisition has also been proposed by the use of simulations of multiple interacting agents walking through a complex network in self-avoiding trajectories. Results of those simulations, performed for the network of Wikipedia and other theoretical models shows that certain sets of parameters and networks perform better in the acquisition of knowledge, through the network of theorems presenting the worst of them. However, unlike what should be expected on the basis of intuition, the agents memories do not play much influence to the speed of acquisition of knowledge. The agent access frequencies of vertices was also been obtained and explored superficially in order to determine where the agents walk more ofen. Several softwares had been developed in this master's thesis project, among these, there is a complex network computational visualization tool, which had become indispensable for the many analysis of the contributions obtained by the use of the other described properties.

Keywords: Complex Networks, Complex Systems, Computational Visualization, Concentric Measurements.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

Desde o início da convivência social, a organização das tarefas e funções de cada indivíduo tornou-se essencial para a produtividade e sobrevivência de tribos primitivas. Com o crescimento da humanidade, a comunicação e as relações entre humanos tornaram-se cada vez mais *complexas*. Cada ser humano se relacionava de diferentes modos: relações familiares, de amizade, de trabalho, de vizinhança, etc. Aqueles com quem interagiam, por consequência, se relacionavam com outros, que também possuíam suas próprias conexões sociais, esta situação se repetia muitas vezes até que todos aqueles que pertenciam a uma comunidade, em algum grau, estivessem conectados.

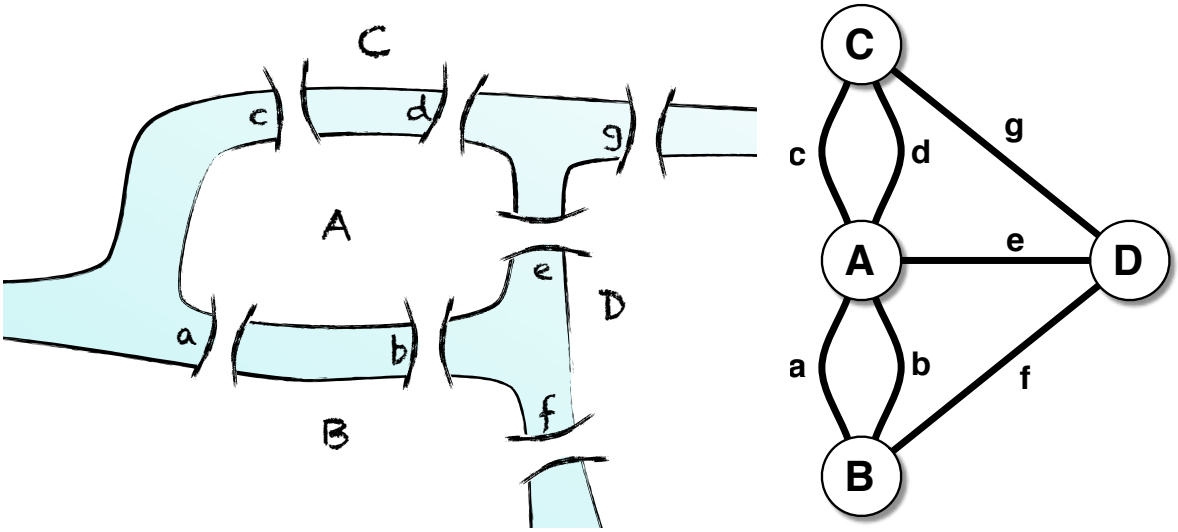
Em tempo, relações de comércio se estabeleciam entre as diferentes comunidades, conectando os aglomerados humanos entre si. Surgem, então, *redes* sociais maiores, formadas pelas conexões entre as diferentes comunidades e pelos seus indivíduos. Diversas dinâmicas acompanham o crescimento da rede, como a propagação de informação, transporte de recursos e disseminação de cultura e conhecimento. Em um curto espaço de tempo, relativo ao *tempo de vida* dessas conexões, as comunidades eram abastecidas, atualizadas e influenciadas culturalmente, por viajantes ou comerciantes, que trafegavam de vila em vila. Em contrapartida, outros seres vivos se aproveitam dessas redes, e as usam, mesmo que indiretamente, como meio de propagação de doenças, como os parasitas, vírus e bactérias.

Com o aumento da população humana e da qualidade de vida, as redes sociais também cresciam, e, com os avanços tecnológicos, cada vez mais rapidamente. Dada a importância dessas relações para o desenvolvimento e vitalidade da civilização, é interessante responder a algumas questões como: Como conter a propagação de um vírus? Como deve ocorrer a disseminação do conhecimento? Como se dá o crescimento dessas redes? Quais indivíduos são os mais importantes?

As respostas para essas perguntas não são triviais, e permitem múltiplas interpretações, entretanto, no início do século XXI, avanços na área de *redes complexas* (1, 2, 3) permitiram que algumas dessas respostas pudessem, ao menos, serem esboçadas (4, 5, 6, 7, 8). Os mecanismos descritos anteriormente ilustram algumas das características e problemas presentes no estudo de

redes complexas, que pode ser resumido como o estudo de *sistemas complexos* formados pelas relações entre seus elementos.

Costuma-se atribuir a *Leonhard Paul Euler* (9) o desenvolvimento inicial da *teoria dos grafos*, que é a base dos estudos de redes complexas. Euler, em 1735, resolveu um problema popular sobre sete pontes da cidade de Königsberg da Prússia (atualmente Kalingrad, Rússia), onde o rio Pregel divide-se em duas partes separando quatro regiões de terra (figura 1.1a). O problema baseia-se em encontrar um caminho que visite todas as pontes apenas uma vez. Euler foi o primeiro a provar matematicamente que tal caminho não existe (10, 11).



(a) Mapa da região da divisão do rio Pegel. **(b)** Representação por grafo.

Figura 1.1 – Representações do problema das sete pontes da cidade de Königsberg.

O primeiro passo de Euler para a solução do problema de Königsberg foi perceber que não se tratava de um problema geométrico, e métricas de distância em nada poderiam ajudar. Euler reduziu o problema ao eliminar as propriedades geográficas e ao introduzir o conceito de vizinhança. Cada região de terra do mapa foi considerado como um nó, ou *vértice*, e cada ponte, uma conexão, ou *aresta* conectando cada vértice, o sistema foi reduzido ao que é chamado hoje de grafo, como ilustrado na figura 1.1b.

O trunfo de Euler foi ter resolvido não somente o problema especifico das pontes, mas também, por ter generalizado a solução para qualquer caso semelhante, isto é, para qualquer grafo. A solução baseia-se no fato de que o caminho proposto somente existirá se cada vértice tiver um número par de arestas conectadas a ele, isto é, uma *conectividade* par, de modo que toda vez que o caminho passa pelo vértice haja ao menos uma entrada e uma saída. Excepcionalmente, os vértices inicial e final podem ter um número ímpar de conexões, já que estes

necessitam de apenas uma saída e uma entrada respectivamente. A solução pode ser simplificada dizendo apenas que: existirá um caminho que passe por todas as arestas de um grafo apenas uma vez se e somente se houver no máximo 2 vértices com conectividade ímpar, o inicial e o final.

A solução dos *caminhos de Euler*, que utiliza apenas a conectividade de cada vértice, despertou grande interesse de outros matemáticos da época, o que culminou na solução e surgimento de diversos outros problemas semelhantes que deram origem ao que chamamos hoje de teoria dos grafos.

As redes complexas têm sua origem remotamente relacionada aos estudos realizados por pesquisadores como Flory (12), Rapoport (13), Erdös e Rényi (14), em meados do século XX. Estes trabalhos dedicavam-se, exclusivamente, a resolver alguns problemas sobre teoria dos grafos e de redes sociais, entretanto no final do século XX e inicio do século XXI, houveram avanços significativos na área, representados principalmente pelos trabalhos de Watts e Strogatz (15), Barabási (16) e seus colaboradores, que estenderam o escopo de redes complexas a diferentes áreas do conhecimento. Muitos fatores contribuíram para essa situação, como o aumento recente da capacidade de aquisição de dados e do poder computacional, permitindo que modelos cada vez mais completos pudessem ser estudados através de simulações e investigações de medidas topológicas.

O estudo de teoria dos grafos e redes complexas distinguem-se pelo último ser amplamente aplicável a diferentes áreas do conhecimento, tornando-se uma eficaz ferramenta para a representação de uma variedade de modelos computacionais (17, 16), físicos (18, 19), biológicos (20) e sociais (21, 22, 23).

A caracterização de grafos e redes complexas é um problema antigo que existe desde os primórdios da teoria de grafos, e baseia-se, principalmente, na obtenção das propriedades de conectividade e distâncias topológicas, entretanto, as novas metodologias e aplicações desses estudos tornaram algumas dessas propriedades, consideradas *clássicas*, um tanto defasadas com relação aos avanços gerais da área. Enquanto que a teoria dos grafos trata principalmente de problemas que podem ser solucionados por metodologias analíticas, em redes complexas, métodos estatísticos são amplamente utilizados devido ao seu caráter complexo e/ou não determinístico, exigindo a aplicação de métodos de simulação e cálculo numérico com o auxílio de computadores.

Apesar de essenciais, as métricas clássicas de redes complexas se mostraram limitantes para a determinação de diversas características tanto das redes quanto de seus elementos. Somente com o avanço do poder de processamento em computadores e da capacidade de aquisição de

grande quantidade de dados, foi possível observar fenômenos de uma forma mais sistemática, revelando a necessidade de novas métricas. Nos últimos anos houve considerável interesse na busca de novas propriedades para a caracterização das redes complexas.

A busca de novas métricas em redes complexas deu origem a diversas novas propriedades, como as medidas de centralidade, desenvolvidas ainda no século XX que visam determinar a importância de um vértice de uma rede. Outras propriedades baseiam-se no desempenho ou em dinâmicas (3, 24) que ocorrem dentro de uma rede, como por exemplo dinâmicas de agentes (25, 26) e difusão (27), estas últimas não somente tem o objetivo de caracterizar a rede, como também simular alguns mecanismos do mundo real, como, por exemplo, simulações de propagação de doenças (4, 5) ou pragas virtuais (28).

Introduzidas em (29) e formalizadas em (30), as métricas *concêntricas* (ou hierárquicas) apresentaram grande potencial para completar a caracterização de redes complexas, pois estendem as métricas tradicionais ao considerar não somente propriedades locais, como também a topologia das redes por completo. Associadas às medidas de *centralidade*, ajudam a revelar características sobre o papel de cada vértice na rede.

Dada a importância da determinação de padrões das métricas concêntricas, este trabalho teve como um de seus objetivos obter e analisar as diferentes medidas concêntricas para diversas redes, tanto aquelas baseadas em modelos teóricos, quanto para redes obtidas a partir de dados reais. Um software foi especialmente desenvolvido com a finalidade de calculá-las sistematicamente para quaisquer conjuntos de redes, tornando possível a obtenção dos padrões concêntricos, mesmo para conjuntos com grande quantidade de dados.

A caracterização individual dos vértices através das medidas concêntricas também foi explorada e aplicada a uma rede de colaboração, baseada no conjunto de dados dos trabalhos acadêmicos da Universidade de São Paulo (USP), elaborando análise comparativa entre as propriedades e as diferentes áreas de conhecimento referentes aos vértices da rede.

A modelagem de conhecimento por redes complexas ganhou atenção em estudos recentes (25, 31). Apesar de redes baseadas na WWW, redes de citações e de colaboração expressarem indiretamente o conhecimento, redes semânticas (32, 26) e hierárquicas (33) apresentaram bons resultados na área de inteligência artificial (34, 35). A representação do conhecimento por redes complexas permite que dinâmicas de aquisição de conhecimento possam ser simuladas por diferentes configurações e heurísticas.

Redes reais que expressem o conhecimento através de conceitos, isto é, redes semânticas, são difíceis de serem criadas, pois exigem um mecanismo sistemático que identifique a proximidade entre conceitos, na falta desse componente a rede pode se tornar subjetiva e dependente

daquele que a criou. Neste trabalho é proposta uma nova metodologia para se obter redes reais na forma de subredes do banco de dados de artigos da Wikipédia. Uma rede de conhecimento foi gerada pelo método, baseando-se nos artigos correspondentes aos teoremas da matemática.

A aquisição de conhecimento pode ser modelada por agentes que se encontram em caminhada aleatória (36, 37) dentro de uma rede, neste trabalho são considerados apenas agentes com dinâmica de caminhada aleatória *auto-esquivante* (self-avoiding). As dinâmicas de agentes estudadas até agora tratam principalmente da coleta de informações sobre a própria rede com o objetivo de caracterizá-las (38, 39), apesar disso a dinâmica de aquisição de conhecimento pode ser simulada em termos de vértices visitados de uma rede.

O presente trabalho também considera a dinâmica de múltiplos agentes interagentes (40) que navegam em uma rede com o objetivo de simular a aquisição de conhecimento, onde cada agente possui suas próprias características como memória e qualidade (em termos da probabilidade de ocorrerem erros observacionais). Os agentes interagem por meio de uma rede complexa de interação, ou colaboração. Diferentemente dos trabalhos existentes sobre o assunto que consideram a interação limitada por uma rede regular (lattice ou grade) ou em comunicação completa (todos os agentes se comunicam com todos os outros), foram usados os modelos teóricos de redes complexas como a rede de interação.

Resultados preliminares mostram que para certas redes, como a rede de teoremas da Wikipedia, o tamanho da memória dos agentes não é relevante para a velocidade de aquisição de conhecimento, a dinâmica nessa rede também apresentou o pior desempenho de aquisição de dados com relação aos modelos teóricos. A caracterização pelas metodologias descritas neste trabalho ajudaram a identificar superficialmente quais propriedades das redes de interação e de conhecimento, influenciam a aquisição de conhecimento.

Outro problema comum em redes complexas é a apresentação das propriedades e das próprias redes, enquanto que redes com poucos elementos podem ser apresentadas graficamente na íntegra de modo relativamente simples, redes com grande quantidade de vértices representam um problema com relação a sua visualização. Nem sempre toda a informação de uma rede complexa pode ser transmitida para uma imagem, inevitavelmente, a apresentação de grafos grandes ou de suas propriedades ainda é um problema em aberto, exigindo, muitas vezes, soluções criativas e específicas para cada grafo. Entretanto metodologias sistemáticas (41, 42, 43, 44) têm apresentado bons resultados para a visualização geral de redes largas e esparsas. Neste trabalho foi desenvolvido um software para a visualização 3D interativa (45) de redes complexas usando metodologias sistemáticas baseadas em dinâmica molecular.

As visualizações geradas permitiram que diversas propriedades estudadas pudessem ser me-

lhor exploradas e devidamente apresentadas. A visualização efetiva das redes complexas foi de extrema importância por viabilizar e facilitar o desenvolvimento e interpretação dos resultados das outras contribuições desta dissertação.

Outros softwares foram desenvolvidos para o cálculo das propriedades, simulações e análise dos dados, entre eles destacam-se um software para o cálculo sistemático de propriedades concêntricas, uma biblioteca para trabalhar com grafos e redes complexas e extrair diversas propriedades e um programa para redução e classificação de dados baseado na análise de componentes principais (PCA).

Esta dissertação está dividida em 5 capítulos e 1 apêndice. O segundo capítulo, intitulado *Redes Complexas*, apresenta os principais conceitos sobre redes complexas, tais como a definição, principais modelos e algumas propriedades como centralidade e métricas concêntricas, além de alguns métodos dos métodos de análise de componente principal e variáveis canônicas. O terceiro capítulo, *Metodologia*, descreve em detalhes a metodologia usada e os principais problemas enfrentados, além de descrever os recursos computacionais usados e desenvolvidos. O quarto capítulo, *Resultados e Discussões*, apresenta os resultados obtidos mais importantes, descrevendo-os e discutindo-os. O último capítulo, *Conclusões*, resume os principais resultados e contribuições deste trabalho além de explorar as perspectivas de trabalhos futuros.

A dissertação é concluída por um apêndice que descreve em detalhes o algoritmo usado para o desenvolvimento do visualizador 3D interativo de redes complexas.

2 *Redes Complexas*

Em geral, sistemas reais podem ser modelados por propriedades intrínsecas e pela forma como interagem com outros sistemas, que por sua vez possuem suas próprias características. A composição lógica dos estudos de fenômenos característicos de cada sistema isoladamente, nem sempre reflete o comportamento real de um conjunto de sistemas. Um exemplo típico é a interação de diferentes partículas, que separadas podem se comportar de modo semelhante, no entanto, em uma molécula, apresentam comportamentos muito diferentes dependendo de como estão interagindo (ou conectadas). O termo *Sistemas Complexos* refere-se a sistemas que são descaracterizados quando estudados de uma maneira "reducionista". *Redes Complexas* (1, 2, 3) são estruturas abstratas que representam as relações entre os diversos elementos desses sistemas.

Uma rede complexa constitui-se de um conjunto numeroso de vértices(ou nós) e arestas(ligações). Cada vértice pode representar elementos como partículas, pessoas, proteínas, trabalhos acadêmicos, etc; enquanto as arestas representam as relações ou conexões entre eles, como a interação entre partículas, a amizade entre pessoas, cooperação em trabalhos e semelhança entre proteínas.

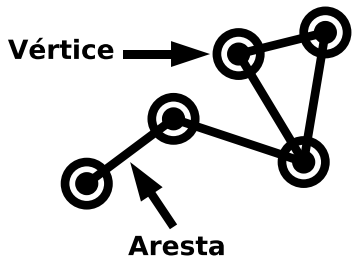


Figura 2.1 – Exemplo de Rede, sendo indicado uma aresta e um vértice.

Matematicamente, pode-se definir uma rede complexa da mesma forma que um grafo (fig. 2.1), isto é, como um conjunto de elementos $\Gamma(\mathscr{V},\mathscr{E})$, onde $\mathscr{V}$ representa um conjunto de vértices $\mathscr{V} = \{v_1, v_2, v_3, \ldots v_n\}$ e $\mathscr{E}$ um conjunto de arestas entre os vértices, $\mathscr{E} = \{(v_i, v_j) : \{v_i, v_j\} \in \mathscr{V}\}$. Uma rede complexa também pode ter valores numéricos, ou pesos, associados a cada vértice ($\mathscr{W}_v$) ou aresta ($\mathscr{W}_e$), essas, são chamadas de redes de vértices ou arestas ponderados. As redes também podem ser digrafos, quando suas arestas indicam a direção da conexão;

ou não-direcionadas, quando as arestas não indicam uma direção privilegiada.

Em termos computacionais, a rede complexa é, geralmente, representada pela sua matriz de adjacência ou por uma lista de adjacências. Uma matriz de adjacência G de uma rede $\Gamma(\mathscr{V},\mathscr{E})$ de N vértices, tem dimensão $N \times N$, onde seus elementos são definidos por G_{ij} tal que se $G_{ij} = 1$ então $\{(v_i, v_j) \in \mathscr{E}\}$; A lista de adjacência é formada pelo conjunto de vértices da rede $E = [e_1, e_2, \ldots]$, onde cada elemento e_k representa uma aresta $e_k = [v_i, v_j]$. Deve-se observar que para uma rede não direcionada, a matriz de adjacência é simétrica, portanto $G_{ij} = G_{ji}$.

2.1 Propriedades Básicas de Redes Complexas

Quantificar as informações contidas em redes complexas nem sempre é uma tarefa trivial. Diversas propriedades (46) podem ser usadas para refletir, em números, as diferentes características de seus vértices ou da rede por completo. Dentre as medidas comumente usadas destacam-se o *grau* (em inglês *vertex degree*), também chamado de *conectividade*; o *coeficiente de aglomeração* (clustering coefficient), usado para quantificar o agrupamento; e os caminhos mínimos(shortest path) ou geodésicas entre vértices.

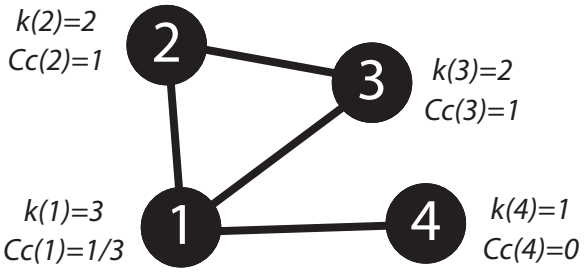


Figura 2.2 – Exemplo de rede com as propriedades de grau e coeficiente de aglomeração em destaque para cada vértice.

O grau (ou conectividade), $k(i)$, de um vértice i, é a medida topológica mais simples que se pode obter em uma rede, sendo definido como o número de arestas conectadas a este vértice. Em termos da matriz de adjacência G, o grau de um vértice i é a soma de todos os elementos de sua linha ou coluna correspondente (equação 2.1):

$$k(i) = \sum_{j=1}^{n} G_{ij} \tag{2.1}$$

A conectividade de rede corresponde quantitativamente ao grau médio, $\langle k \rangle$. Na próxima seção (2.2) será mostrado que enquanto uma rede aleatória é caracterizada por um pequeno

valor de desvio padrão do grau, redes reais ou modelos livres de escala tendem a apresentar valores mais altos, portanto a distribuição do grau fornece informações relevantes para a sua caracterização em um desses diferentes modelos. Na rede da figura 2.2, por exemplo, o grau médio é $\langle k \rangle = (2+2+3+1)/4 = 2$.

O coeficiente de aglomeração, $Cc(i)$, de um vértice i, é definido como a razão do número de conexões entre os vértices da primeira vizinhança de i, $e_1(i)$, pelo maior número de conexões, $e_{max}(i)$, que seriam possíveis entre os vizinhos. Como $e_{max}(i) = n_1(i)(n_1(i)-1)/2$, onde $n_1(i)$ é o número de primeiros vizinhos de i, pode-se usar a equação 2.2:

$$Cc(i) = \frac{e_1(i)}{e_{max}(i)} = 2\frac{e_1(i)}{n_1(i)(n_1(i)-1)} \tag{2.2}$$

Quando a conectividade na vizinhança é máxima, o valor de $Cc(i)$ é 1, e 0 quando a conectividade é mínima. O coeficiente de aglomeração fornece um forte indicativo sobre a conectividade local em uma rede, mas, por limitar-se em considerar apenas os vizinhos mais próximos a um vértice, pode não caracterizar totalmente a rede. A solução para este problema é o coeficiente de aglomeração concêntrico que é apresentado no seção 2.3.

O mínimo caminho médio, l, é uma importante propriedade em redes complexas, é definido como a média das distâncias mínimas (geodésicas), d_{ij} para todos os pares de vértices, como na equação 2.4:

$$l = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij} \tag{2.3}$$

Essa propriedade é válida somente quando a rede considerada é conectada, isto é, quando há caminhos entre todos os pares de vértices da rede. Este problema pode ser resolvido tomando-se a média harmônica (equação 2.4), no entanto, para este trabalho, são considerados apenas os componentes conectados das redes.

$$\frac{1}{l^*} = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{ij}} \tag{2.4}$$

2.2 Modelos de Redes Complexas

Um dos principais problemas em redes complexas é encontrar redes que modelem fenômenos ou estruturas da natureza. Redes reais são obtidas através de medidas experimentais e observações da natureza, e tem como objetivo, limitar o estudo a um conjunto de sistemas e as relações entre

eles. Muitas vezes, é necessário obter redes que não necessariamente correspondem a sistemas físicos reais, mas, a uma aproximação representativa de uma categoria de sistemas, que possuem propriedades em comum. Diversos modelos teóricos foram desenvolvidos com essa finalidade, dentre eles destacam-se os modelos regulares, Erdős-Rényi (aleatória) (14), Watts-Strogatz (pequeno mundo) (15, 47), Barabási-Albert (livre de escala) (48) e geográficos (49).

Os modelos teóricos permitem a criação de redes com parâmetros controlados, fixando o número de nós e a conectividade, por exemplo; também permitem a obtenção de resultados analíticos de algumas propriedades que, muitas vezes, podem ser estendidas a redes reais.

Em geral, as redes não necessariamente pertencem a uma única categoria de modelos teóricos, estudos anteriores mostram que a maioria das redes reais apresentam comportamento híbrido (50), por exemplo, podem apresentar características de pequeno-mundo e livres de escala ao mesmo tempo, ou ainda apresentar regiões aleatórias e regiões livres de escala.

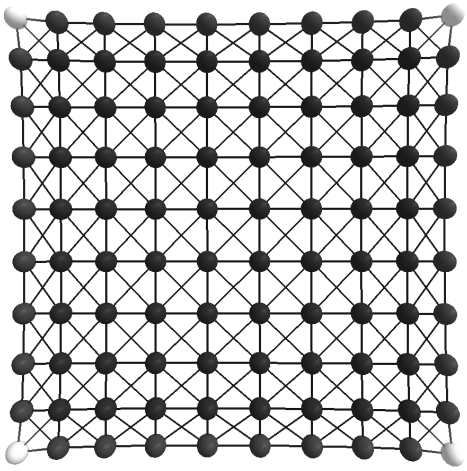
A seguir os principais modelos teóricos de redes complexas serão apresentados, destacando as propriedades caraterísticas de cada modelo

2.2.1 Redes Regulares

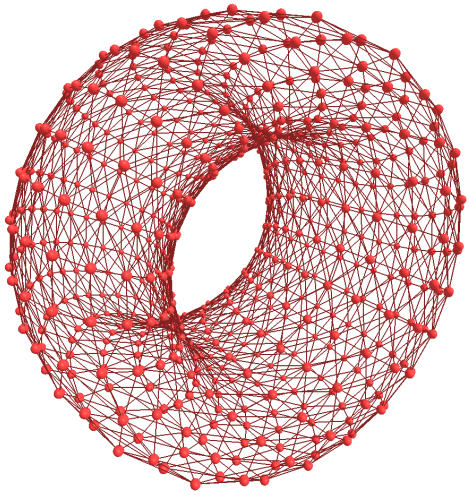
Redes com características regulares (figuras 2.3a e 2.3b) não representam muitos sistemas reais, ainda assim, são muito estudados na teoria de grafos. Uma rede regular possui todos os vértices com mesma conectividade e apresenta uma dimensão característica.

Devido ao caráter infinito das redes regulares, em geral, é necessário truncá-la limitando o estudo a um grupo finito de vértices. O truncamento pode ser realizado usando diferentes formas topológicas, mas convém limitar a rede por um quadrado, cubo ou hipercubo n-dimensional dependendo da dimensão característica.

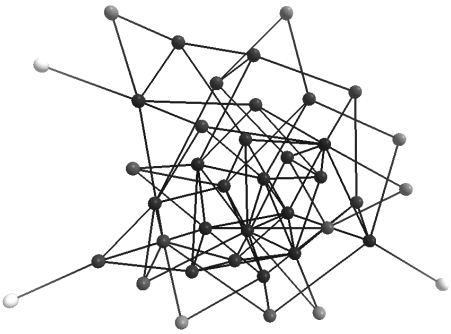
Além de definir a forma, é necessário determinar como os vértices da borda devem ser truncados: A borda pode ser preservada eliminando as conexões que não pertencem ao componente removido, como na rede regular bidimensional da figura 2.4a, apresentando vértices com conectividade diferente dos demais. Também é possível eliminar a borda por completo reconectando os elementos ciclicamente como se a rede se repetisse infinitamente pelo espaço, como a figura 2.4b, para este caso todos os vértices apresentam exatamente o mesmo grau e são degenerados.



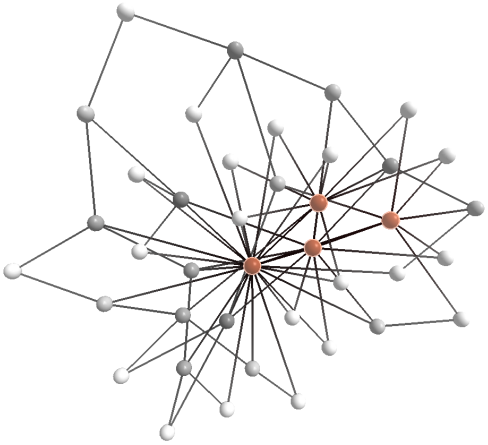
(a) Regular 2D



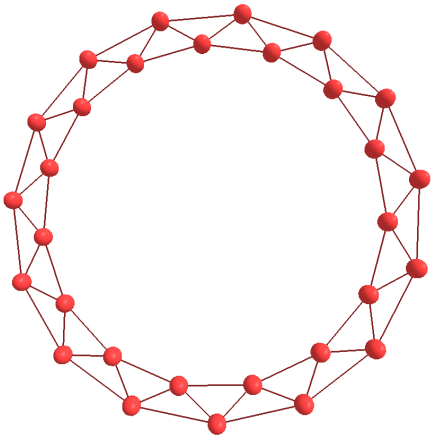
(b) Regular 2D - sem bordas



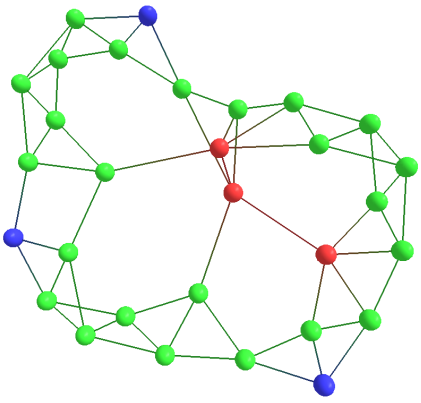
(c) Aleatória



(d) Barabási-Albert



(e) Regular 1D (Lattice)



(f) Pequeno-Mundo

Figura 2.3 – Modelos de redes complexas.

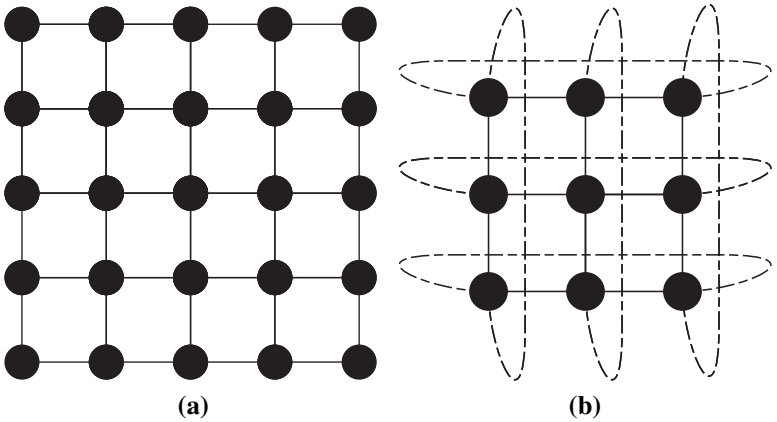


Figura 2.4 – Dois modelos de redes regulares: (a) com efeito de bordas e (b) sem efeito de bordas.

2.2.2 Redes Aleatórias

O modelo de redes aleatórias, inicialmente desenvolvido por Anatol Rapoport (13) e independentemente por Paul Erdős e Alfréd Rényi (14), foi uma das primeiras tentativas de modelagem de redes reais. Também conhecidas como redes Erdõs-Rényi (ER), essas redes apresentam tendências gerais para a maioria das propriedades, portanto, apesar de representarem poucos sistemas reais, são muito importantes para a análise do comportamento médio dessas propriedades em um conjunto de redes.

As redes aleatórias também são de grande ajuda para o estudo analítico ou estatístico de redes complexas, sendo usadas como a base para muitas outras redes sofisticadas.

Em uma rede Aleatória (fig. 2.3c), os N vértices, inicialmente desconectados, são ligados ou não dependendo de um parâmetro de probabilidade p. Para redes com grande número de vértices a distribuição do grau é binomial (distribuição de Poisson) centrada no valor de conectividade média $\langle k \rangle = pN$, como na Figura 2.5.

2.2.3 Modelo Barabási-Albert

Desenvolvido por Albert-László Barabási e Réka Albert (48), o modelo Barabási-Albert (BA) é um dos mais bem sucedidos métodos para a geração de redes *livres de escala*. Essas redes apresentam distribuição de grau que obedece a uma lei de potência (figura 2.6) $P(k) \propto k^{-3}$.

Redes baseadas nesse modelo representam muitos sistemas reais devido ao fato de que a

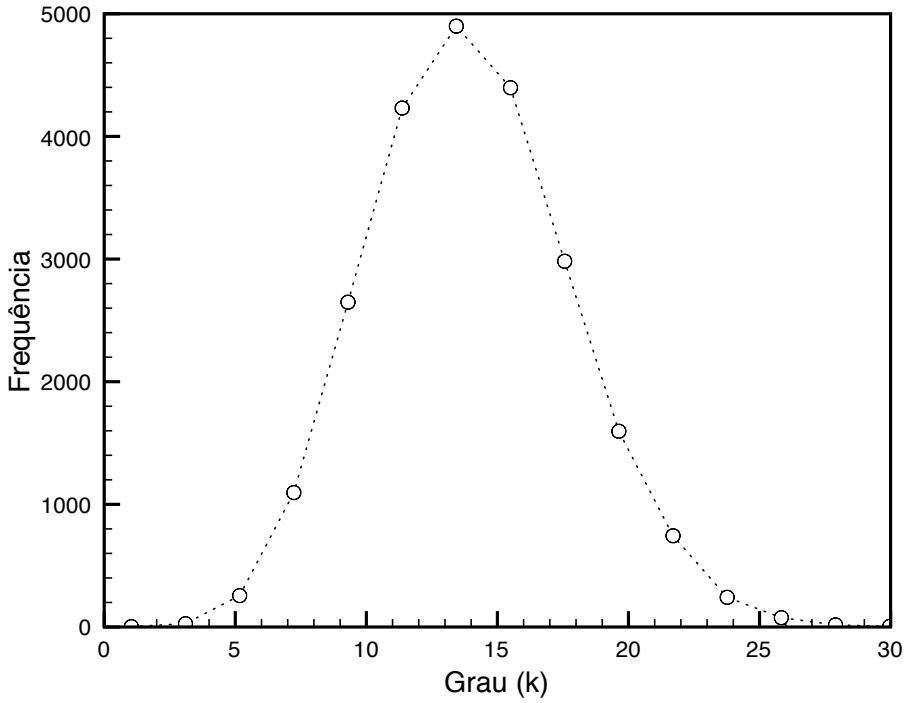


Figura 2.5 – Distribuição do grau para uma rede aleatória (ER).

grande maioria são livres de escala (16). O modelo também caracteriza o crescimento dessas redes, tornando-se uma um bom modelo para entender como essas redes podem ser geradas.

Diferentemente das redes aleatórias, onde os vértices apresentam conectividade semelhante, em uma rede livre de escala é comum o surgimento de vértices com número de conexões muito maior do que a média, a figura 2.3d ilustra esse fato.

O termo *hub* costuma ser usado para denominar os vértices com maior número de conexões em uma rede livre de escala, no entanto, não é possível defini-lo formalmente, pois nessas redes, não há um valor característico de conectividade que permita a existência de um limiar intrínseco. Neste trabalho o termo hubs é usado para definir uma pequena fração dos vértices de maior conectividade de uma rede ou vértices com número de conexões de ordens de grandeza superior à média.

O modelo Barabási-Albert baseia-se no conceito de que redes reais possuem tanto uma dinâmica de crescimento quanto um mecanismo preferencial de ligação. Esses 2 fatores combinados são essenciais para a geração de redes livre de escala (48, 2).

Uma rede BA é gerada a partir um conjunto de vértices m_0 (≥ 2) que podem estar desconectados ou aleatoriamente conectados. Progressivamente, t novos vértices são adicionados

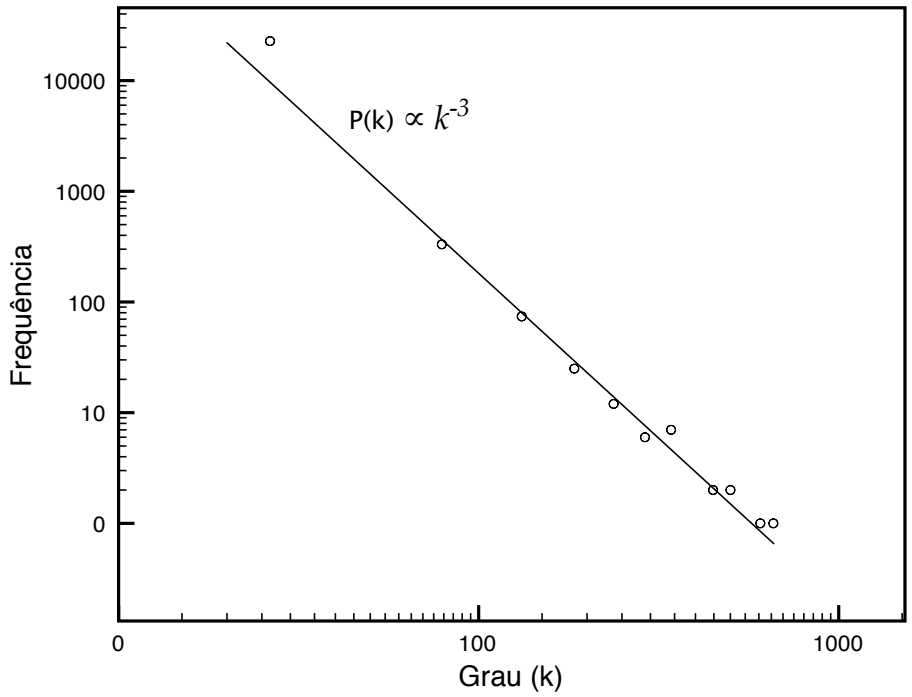


Figura 2.6 – Distribuição de grau para uma rede Barabási-Albert (BA).

ligando-se a m ($\geq m_0$) vértices dos V_t já presentes na rede. A escolha dos vértices é aleatória obedecendo a uma probabilidade dependente linearmente da conectividade do vértice destino e(eq. 2.5):

$$P(e) = \frac{k(e)}{\sum_{j \in V_t} k(j)} \tag{2.5}$$

O processo resulta em uma rede de tamanho $N = m_0 + t$ vértices com $\langle k \rangle = 2m$. Como a cada passo os novos vértices são ligados, preferencialmente, aos mais conectados, estes tendem a adquirir ainda mais conexões.

2.2.4 Redes Watts-Strogatz

Pequeno-mundo é o nome dado a redes caracterizadas por terem menor caminho médio (média das geodésicas) muito baixo, isto é, seus vértices podem, em geral, serem acessados com poucos passos partindo de outro vértice qualquer. Esse fenômeno costuma aparecer em diversas redes reais e é muito presente em redes sociais (22). Outra característica de redes pequeno-mundo é a existência de muitas conexões do tipo triângulo, isto é, alto valor do coeficiente de aglomeração. Em uma rede social, por exemplo, a existência dessas conexões é comum pois

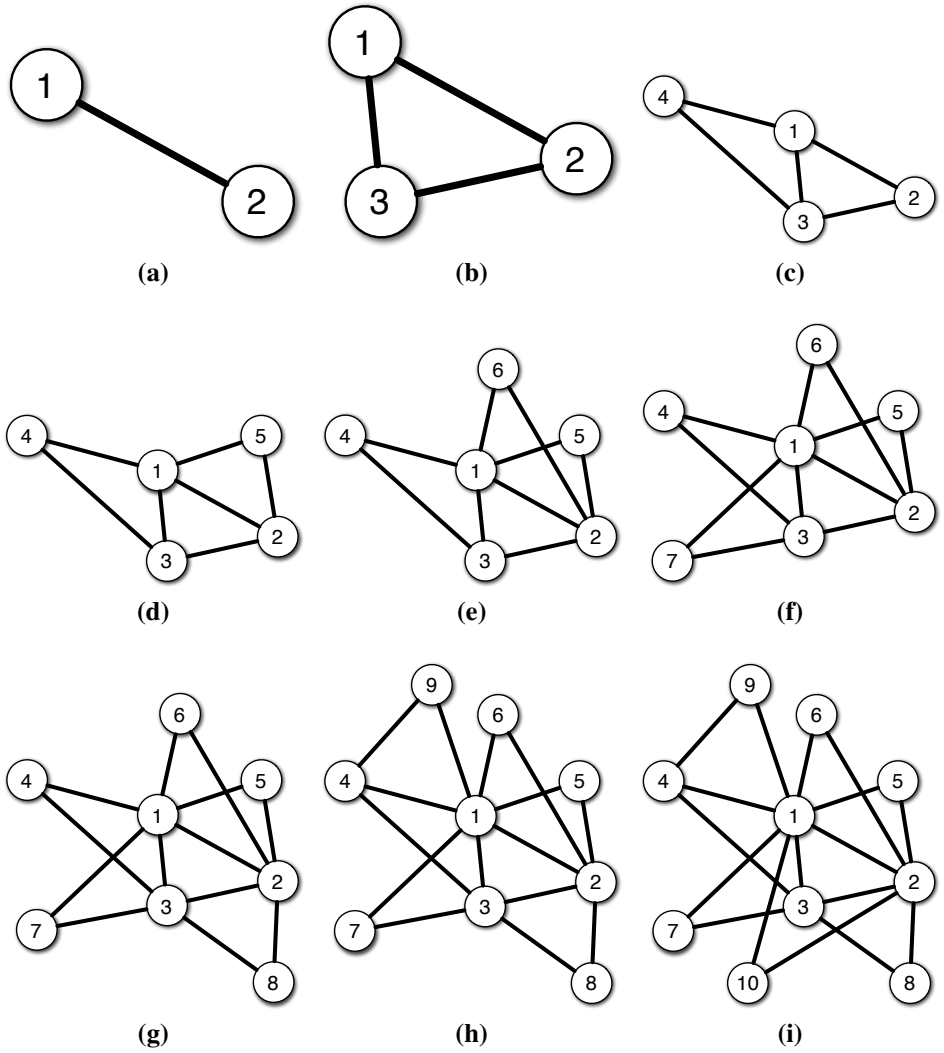


Figura 2.7 – Estágios para a geração de uma rede BA com 10 vértices e $\langle k \rangle = 4$.

amigos de uma pessoa tem grandes chances de serem amigos entre si.

Um modelo que representa bem uma rede pequeno mundo foi desenvolvido por Duncan J. Watts e Steven Strogatz (15). Nessa rede, chamada de Watts-Strogatz (WS), uma rede regular tem algumas de suas suas arestas modificadas de modo que possam conectar qualquer outro par de vértices escolhido aleatoriamente. Dependendo da quantidade de arestas trocadas obtém-se redes que apresentam características que variam de redes regulares a redes aleatórias. Uma rede WS começa a apresentar o fenômeno de pequeno-mundo logo que alguns poucos vértices são trocados.

A construção de uma rede WS começa a partir de uma regular unidimensional (figura 2.8a), progressivamente, todas as arestas são consideradas aleatoriamente, com probabilidade p, para trocarem seus vértices de origem e destino por outros também selecionados aleatoriamente. A figura 2.8 ilustra o processo de construção.

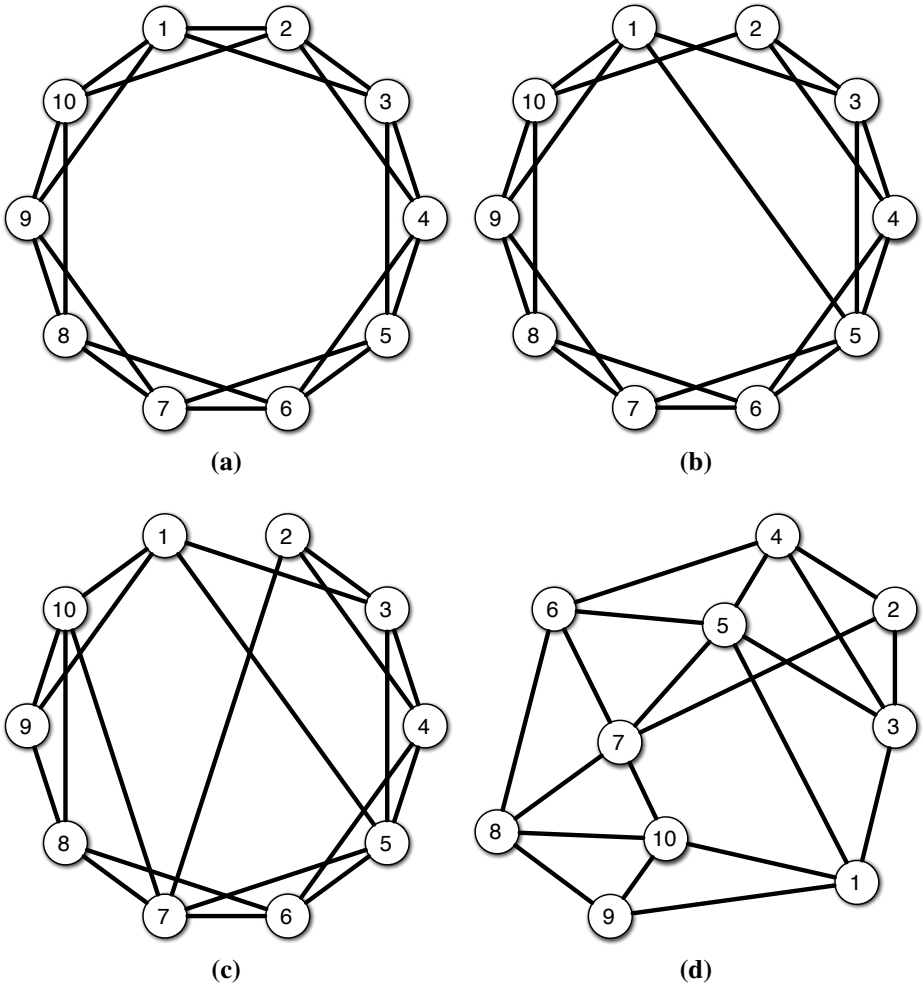


Figura 2.8 – Construção de uma rede Watts-Strogatz com 10 vértices.

2.3 Níveis e Propriedades Concêntricas

Propriedades básicas de redes complexas, como a conectividade e o coeficiente de aglomeração, fornecem informações importantes para a caracterização dos vértices e da própria rede, no entanto, consideram apenas a estrutura da vizinhança imediata de cada vértice. As propriedades topológicas locais de um vértice podem não ser suficientes para descrever seu papel na rede, necessitando de mais informações sobre a estrutura geral da rede. Uma alternativa para completar essas propriedades é estende-las ao considerar os vizinhos mais distantes través do conceito de *níveis concêntricos* (ou hierárquicos) (29) formalizado em (30).

Os níveis concêntricos podem ser definidos como um número inteiro, d_{ij}, associado a cada par de vértices, i e j, com o valor do menor caminho entre eles, para redes não direcionadas $d_{ij} = d_{ji}$. Pode-se organizar os vértices da rede de acordo com os níveis concêntricos em subredes chamadas de *anéis*.

Cada anel, $R_d(i)$, é definido pelo nível concêntrico d e por um vértice central i; e contém todos os vértices $j \in V$ tal que $d_{ij} = d$, isto é, o anel de nível 1, $R_1(i)$, contém todos os vértices que são vizinhos imediatos ao vértice central i, o anel de nível 2, $R_2(i)$, é formado por todos os vértices vizinhos aos do nível 1 que não pertençam a $R_1(i)$. O mesmo raciocínio pode ser usado para determinar os vértices de um anel $R_{n+1}(i)$ conhecendo-se $R_n(i)$. A figura 2.9 exemplifica a definição dos níveis concêntricos mostrando os 3 anéis de uma rede.

A estrutura hierárquica pode ser obtida realizando uma busca em largura na árvore espalhada mínima da rede partindo do vértice central, como mostra a figura 2.10. Por questões de simplificação, o termo nível concêntrico (ou hierárquico) é usado para caracterizar os vértices de um anel. Nota-se que a estrutura hierárquica pode ser obtida somente para componentes conexos da rede, outra observação importante é a de que o anel de nível 0 contém apenas o vértice central.

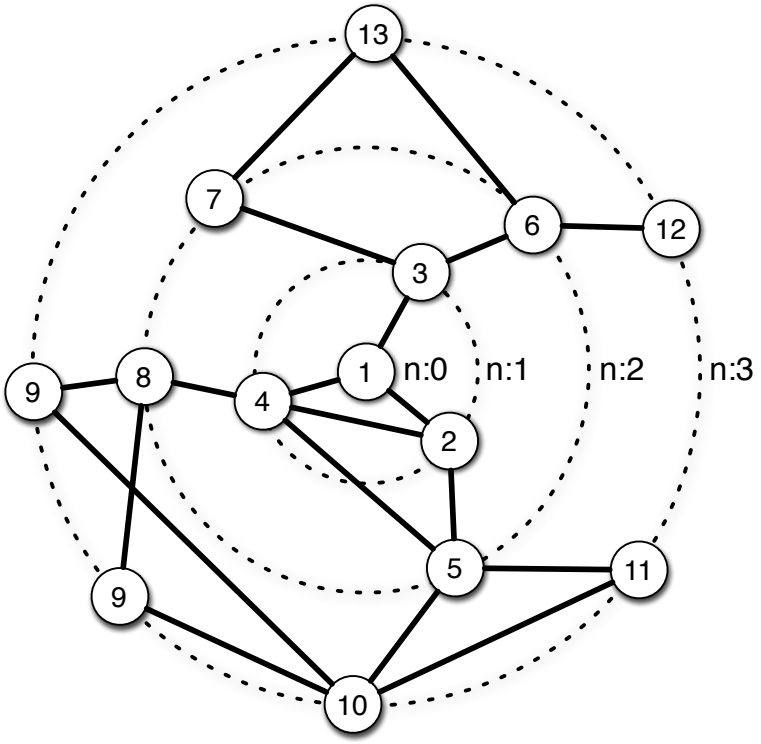


Figura 2.9 – Exemplo de anéis concêntricos centrados em 1 para uma rede de 13 vértices.

2.3.1 Medidas Concêntricas

Aplicando o conceito de níveis concêntricos é possível estender as propriedades clássicas de modo a considerarem diversos anéis centrados no vértice estudado, dando origem às *pro-*

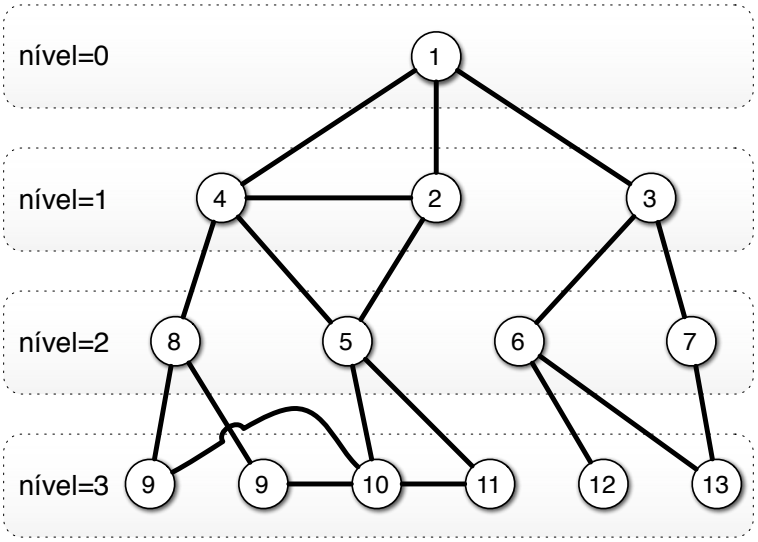


Figura 2.10 – Árvore espalhada para a rede da figura 2.9 centrada em 1 com os respectivos níveis hierárquicos.

priedades concêntricas. As medidas concêntricas (também chamadas de medidas hierárquicas) são definidas para cada anel ao longo dos níveis concêntricos.

Apesar de estenderem as medidas tradicionais, as medidas concêntricas não se derivam trivialmente delas, por exemplo, a conectividade se estende em 4 novas propriedades e 2 outras não possuem análogo tradicional. As propriedades concêntricas propostas são: número de nós (number of nodes), número de arestas em um anel (number of edges on a ring), grau concêntrico (concentric degree), grau entre-nível (inter-ring degree), grau intra-nível (intra-ring degree), grau concêntrico comum (concentric common degree), coeficiente de aglomeração concêntrico (concentric clustering coefficient) e taxa de convergência (convergence ratio). As propriedades estão definidas para cada anel centrado em um vértice da rede e dependem exclusivamente dos vértices que estão no anel e das conexões internas ou entre os anéis vizinhos. As propriedades são definidas em detalhes a seguir considerando um anel $R_d(i)$, os exemplo das medidas referem-se à figura 2.9.

Número de nós, $n_d(i)$:

É a mais simples das medidas concêntricas e denota exatamente o número de vértices que pertencem ao anel $R_d(i)$, isto é, a cardinalidade do conjunto representado pelo anel. Por exemplo, na figura 2.9 o número de nós dos 4 níveis concêntricos são respectivamente: $n_0(1)$=1, $n_1(1)$=3, $n_2(1)$=4 e $n_3(1)$=6.

Número de arestas em um anel, $e_d(i)$:

Consiste no número de arestas exclusivamente contidas no anel $R_d(i)$, isto é, todas as arestas que conectam vértices com o mesmo nível hierárquico d. Por exemplo, $e_0(1) = 0$, $e_1(1) = 1$, $e_2(1) = 0$ e $e_3(1) = 3$.

Grau concêntrico, $k_d(i)$:

O grau concêntrico é a extensão direta do grau clássico aplicada aos níveis hierárquicos, definindo o número de conexões que um anel $R_d(i)$ possui com o seguinte $R_{d+1}(i)$. Para a rede da figura 2.9, $k_0(1) = 3$, $k_1(1) = 5$, $k_2(1) = 7$ e $k_3(1) = 0$.

Grau intra-nível, $A_d(i)$:

Outro método de se estender a conectividade, o grau intra-nível é determinando através do valor médio do grau obtido dos vértices do anel, mas considerando apenas as conexões internas, isto é, ignorando conexões com os vértices dos anéis anterior, $R_{d-1}(i)$, e subsequente, $R_{d+1}(i)$. A propriedade também pode ser definida em termos das outras propriedades concêntricas pela equação 2.6:

$$A_d(i) = \frac{2e_d(i)}{n_d(i)} \tag{2.6}$$

Nota-se que para o anel de nível 0 o valor seu valor não está definido e por conveniência usa-se $A_0(i) = 0$. Na figura de exemplo tem-se: $A_1(1) = 1/3$, $A_2(1) = 0$ e $A_3(1) = 1/2$.

Grau entre-níveis, $E_d(i)$:

É obtida pelo número médio de arestas que conectam o anel $R_d(i)$ ao seguinte, $R_d(i)$ e é definido, em termos das outras propriedades pela equação 2.7:

$$E_d(i) = \frac{k_d(i)}{n_d(i)} \tag{2.7}$$

Por exemplo, para a rede da figura 2.9, tem-se: $E_0(1) = 3$, $E_1(1) = 5/3$, $E_2(1) = 7/4$ e $E_3(1) = 0$.

Coeficiente de aglomeração concêntrico, $Cc_d(i)$:

Estende a propriedade do coeficiente de aglomeração tradicional. É obtido através da equação 2.8:

$$Cc_d(i) = 2\frac{e_d(i)}{n_d(i)(n_d(i)-1)} \tag{2.8}$$

Nota-se que o coeficiente de aglomeração concêntrico preserva a mesma propriedades da medida tradicional, variando entre 0 e 1. Por exemplo, para a rede da figura, tem-se: $Cc_0(1) = 0$, $Cc_1(1) = 1/3$, $Cc_2(1) = 0$ e $Cc_3(1) = 1/5$.

Taxa de convergência, $C_d(i)$:

É obtido pela razão do número de arestas conectando vértices pertencentes ao nível d aos vértices do nível concêntrico seguinte pelo número de vértices desse. Pode ser definida através da equação 2.9:

$$C_d(i) = \frac{k_d(i)}{n_{d+1}(i)} \tag{2.9}$$

Para a rede exemplificada, $C_0(1) = 1$, $C_1(1) = 5/4$ e $C_2(1) = 1$.

Grau concêntrico comum, $H_d(i)$:

O grau concêntrico comum é a média dos graus tomada para todos os vértices pertencentes ao anel $R_d(i)$, considerando, inclusive, as conexões que compartilham níveis concêntricos distintos. Em termos das propriedades concêntricas pode ser definido pela equação 2.10:

$$H_d(i) = \frac{k_d(i) + k_{d-1}(i) + 2e_d(i)}{n_d(i)} \tag{2.10}$$

Para a rede da figura de exemplo, $H_0(1) = 1$, $H_1(1) = 10/3$, $H_2(1) = 3$ e $H_3(1) = 11/6$.

2.4 Outras Propriedades

Redes complexas também podem apresentar outras características importantes que não são evidentes através do estudo pelas propriedades tradicionais ou hierárquicas. Determinar a centralidade, ou importância, de um vértice na rede, por exemplo, é um desses problemas e algumas medidas especificas foram desenvolvidas para isso. Outro problema é determinar as melhores combinações de medidas para se agrupar ou caracterizar vértices ou redes. Este último pode ser

resolvido através de métodos de análise multi-variável como *Análise de Componentes Principais* ou simplesmente *PCA* (do inglês Principal Component Analysis).

A seguir há uma breve descrição de quatro medidas de centralidade, a *centralidade de auto-vetor*, (eigenvector centrality), *centralidade de proximidade* (closeness centrality), *centralidade de intermediação* (betweenness centrality) e *centralidade de subgrafos* (sub-graph centrality). Um método recentemente formalizado para se determinar a importância de vértices também é sumariamente descrito, assim como a utilização de análise multiváriavel para a classificação de redes.

2.4.1 Centralidade e Importância

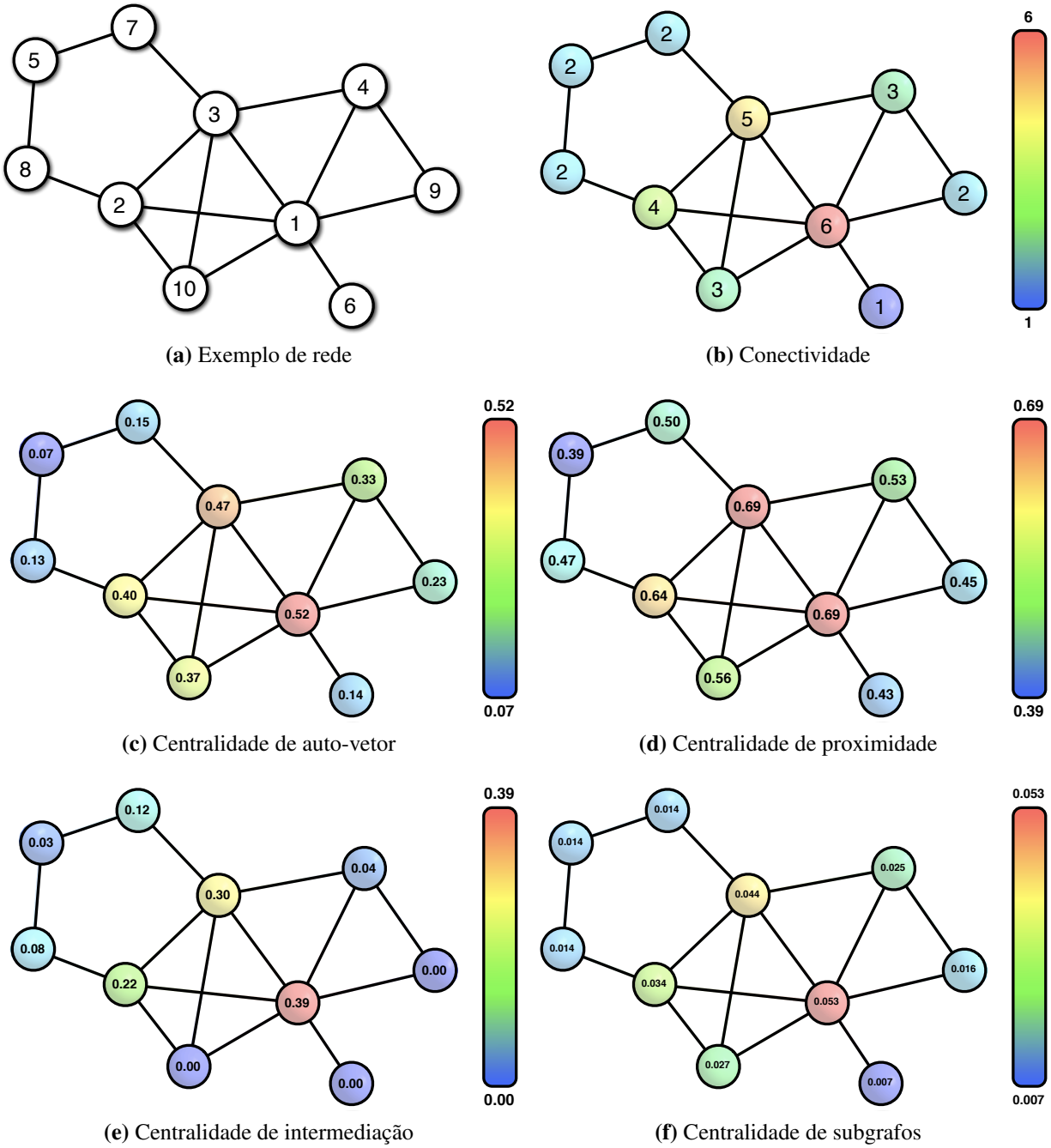
Determinar os nós mais importantes em uma rede não é uma tarefa trivial, primeiramente é necessário definir qual é o tipo de importância que deve ser comparada, por exemplo, vértices com grande número de conexões podem ser importantes distribuidores de informação em uma rede de computadores, enquanto que pontes, nessa mesma rede, também tem sua importância por serem críticas para o fluxo máximo de informação.

Muitas vezes, utiliza-se a suposição de que os vértices *centrais* de uma rede são os mais importantes, este conceito está correto para grande parte das situações, como no caso de agentes em uma caminhada aleatória que tenderão a permanecer mais tempo nesses vértices do que nos vértices das *bordas*. A conectividade tradicional também é considerada uma medida de centralidade, mesmo sendo extremamente localizada e não considerar a topologia da vizinhança do vértice. Para se determinar os vértices centrais em uma rede pode-se usar as métricas de centralidade tradicionais, descritas a seguir:

Centralidade de auto-vetor, $x(i)$:

Baseia-se na idéia de atribuir pontuações de importância a cada vértice da rede, aqueles com maior pontuação tendem a também ter vizinhos com alta pontuação (51). Para se conhecer as pontuações de um vértice é necessário saber as pontuações de seus vizinhos, que por consequência também necessitam da pontuação do vértice original. O cálculo da propriedade pode parecer paradoxal, mas ao ser analisado matematicamente, é um simples problema de álgebra linear que necessita a resolução de um sistema de N equações e N incógnitas.

O valor da centralidade de auto-vetor, $x(i)$, para um vértice i pode ser escrita como a soma



(a) Exemplo de rede

(b) Conectividade

(c) Centralidade de auto-vetor

(d) Centralidade de proximidade

(e) Centralidade de intermediação

(f) Centralidade de subgrafos

Figura 2.11 – Medidas de centralidade para uma rede pequena.

de todas as pontuações dos vizinhos, em termos da matriz de adjacência G_{ij} de uma rede de tamanho N é dada pela equação 2.11:

$$x(i) = \frac{1}{\alpha} \sum_{j=1}^{N} x(j) G_{ij} \tag{2.11}$$

A equação 2.11 pode ser rescrita em notação matricial, resultando na equação de auto-vetor 2.12:

$$\vec{x}G = \alpha\vec{x} \tag{2.12}$$

Considerando que são desejáveis soluções de auto-vetores com elementos positivos, pelo teorema de Perron-Frobenius (52), há apenas um auto-vetor nessas condições, aquele de maior auto-valor, α, também chamado de auto-vetor principal. A solução pode ser facilmente obtida pelo método iterativo das potências. A figura 2.11c ilustra a propriedade em cores e na etiqueta dos vértices para uma rede pequena.

Centralidade de proximidade, $C_C(i)$:

Utiliza o conceito de que vértices mais importantes são aqueles que estão mais próximos de todos os outros (53, 54). Essa propriedade pode ser quantificada considerando o tamanho médio dos caminhos mínimos partindo do vértice de referência i a todos os outros da rede. Sendo d_{ij} a distância entre um par de vértices da rede, a centralidade de proximidade para um vértice i é definida pela equação 2.13:

$$C_C(i) = \frac{N-1}{\sum_{j=1}^{N} d_{ij}} \tag{2.13}$$

Um exemplo da propriedade pode ser visto na figura 2.11d.

Centralidade de intermediação, $C_B(i)$:

Vértices que pertencem a muitos caminhos mínimos que ligam outros vértices também podem ser considerados importantes, a centralidade de intermediação mensura essa característica (55). É definido, para um vértice i, como a soma das razões do número total de caminhos mínimos que atravessam i, $\sigma_{st}(i)$, pelo número total de caminhos mínimos, σ_{st}, considerando todos os pares de vértices s e t pertencentes à rede, como na equação 2.14:

$$C_B(i) = \sum_{s}^{N} \sum_{t \neq s}^{N} \frac{\sigma_{st}(i)}{\sigma_{st}} \tag{2.14}$$

A figura 2.11e ilustra a propriedade.

Centralidade de subgrafos, $SC(i)$:

Recentemente introduzido em (56), baseia-se na hipótese de que vértices mais importantes tendem a pertencer a maior quantidade de subgrafos fechados, isto é, subgrafos compostos pelos vértices que representam um ciclo que inicia e finaliza no mesmo vértice. É definido como a soma do número de subgrafos fechados, $\mu_k(i)$, iniciando e finalizando em i em k passos. O valor de $\mu_k(i)$ pode ser obtido de forma simples em termos da diagonal de potências da matriz de adjacência G, como na equação 2.15:

$$\mu_k(i) = (G^k)_{ii} \tag{2.15}$$

A soma de todos os subgrafos ao longo dos valores de passos em cada ciclo, k, é infinita ($\sum_{k=0}^{N} \mu_k(i) = \infty$). Para solucionar esse problema, aplica-se a hipótese de que ciclos de menor tamanho são aqueles que mais contribuem na caracterização da importância de um vértice, portanto a soma deve ser tomada atribuindo a cada elemento um peso que decresça rapidamente com k, como $1/k!$. Pode-se, então, definir a centralidade de subgrafos pela equação 2.16:

$$SC(i) = \sum_{k=0}^{\infty} \frac{\mu_k(i)}{k!} \tag{2.16}$$

O exemplo pode ser visto na rede da figura 2.11f.

2.4.2 Análise dos componentes principais

Diferentes métricas revelam diferentes propriedades em redes complexas. No entanto, a composição dessas diferentes métricas pode apresentar redundância ou informações não relevantes para a caracterização de seus vértices ou das próprias redes. Outro problema é determinar quais propriedades apresentam maior quantidade informação, de forma a reduzir o tamanho do espaço de dados, permitindo a aplicação de métodos de visualização e segmentação de dados. A análise multiváriavel baseada nos componentes principais, ou PCA (57, 58), pode ajudar na solução deste problema.

A metodologia PCA tem como objetivo obter bases ordenadas por relevância compostas por combinações lineares do conjunto de dados e, simultaneamente, eliminar as possíveis redundâncias reduzindo a dimensionalidade. A aplicação do método é simples, e baseia-se na obtenção da base dos auto-vetores mais importantes da matriz de covariância; e na projeção dos dados nessa nova base.

Supondo um conjunto de dados, $\mathscr{X}$, composto por N observações de M diferentes variáveis, em notação vetorial, $\vec{X}_k = \{X_{1k}, X_{2k}, \ldots X_{Mk}\}$, com $k = \{1, 2, 3 \ldots N\}$; deseja-se reduzir a dimensionalidade de $\mathscr{X}$ para L, com $L < M$, mantendo a maior quantidade da informação.

Os elementos da matriz de covariância, Σ_{ij}, de tamanho $M \times M$; são definidos pelos valores das covariância dos respectivos vetores observados, X_{ki}, onde X é a matriz composta pelos N vetores $\vec{X}_k$ agrupados . A equação 2.17 define a matriz de covariância:

$$\Sigma_{ij} = cov(X_i, X_j) = \langle (X_i - \mu_i)(X_j - \mu_j) \rangle \tag{2.17}$$

onde $\langle \ldots \rangle$ significa o valor esperado, ou valor médio ao longo das N observações e

$$\mu_i = \langle X_i \rangle = \frac{1}{N} \sum_{k=1}^{N} X_{ik} \tag{2.18}$$

Uma forma mais elegante para se definir a matriz de covariância é através do *produto externo* (ou produto tensorial) dos vetores de desvio da média, $\vec{D}_k = \{D_{1k}, D_{2k}, \ldots D_{Mk}\}$ definidos pelos elementos:

$$D_{kj} = X_{ki} - \mu_i \tag{2.19}$$

A matriz de covariância pode ser calculada pela média em k do produto externo de $\vec{D}_k$ por ele mesmo:

$$\Sigma = \langle \vec{D} \otimes \vec{D} \rangle = \langle \vec{D} \cdot \vec{D}^* \rangle \tag{2.20}$$

onde $\vec{D}^*$ é a matriz adjunta do vetor $\vec{D}$. Como o vetor é composto por números reais, $\vec{D}^* = \vec{D}^T$.

A matriz de covariância é simétrica, portanto hermitiana, o que garante que todos os auto-valores são reais e os auto-vetores são ortogonais. Com o objetivo de eliminar a covariância entre as diferentes variáveis é desejável que a matriz de covariância para os dados transformados seja diagonal, isto é possível projetando os vetores dos dados originais na base ortonormal composta pelos auto-vetores de Σ, $\vec{v}_i$ com $i = \{1, 2, 3, \ldots M\}$.

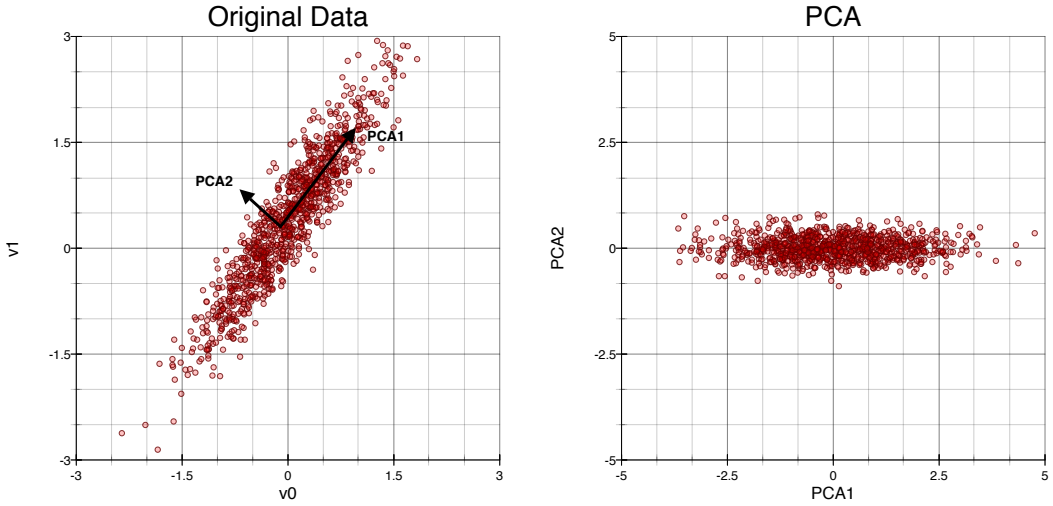
A cada auto-vetor $\vec{v}_i$ está associado um auto-valor λ_i, que corresponde ao elemento $\Sigma_{ii}^{(\text{transformada})}$ da matriz de covariância dos vetores projetados, $\vec{P}_k$. Os elementos da diagonal da matriz de covariância equivalem à variância:

$$\lambda_i = \Sigma_{ii}^{(\text{transformada})} = var(P_i) = var(\vec{v}_i \cdot \vec{X}) \tag{2.21}$$

portanto, os vetores projetados mais significativos são aqueles compostos pelos auto-vetores de maiores auto-valores. Pode-se, então, ordenar os auto-valores e os correspondentes auto-vetores em ordem decrescente tal que $\lambda_1 > \lambda_2 > \lambda_3 > \ldots \lambda_M$.

A projeção das variáveis na nova base pode ser feita membro a membro como indicado pela equação 2.22

$$P_{ik} = \vec{v}_i \cdot \vec{X}_k \tag{2.22}$$



(a) Dados originais e seus auto-vetores da matriz de covariância.

(b) Projeção usando PCA.

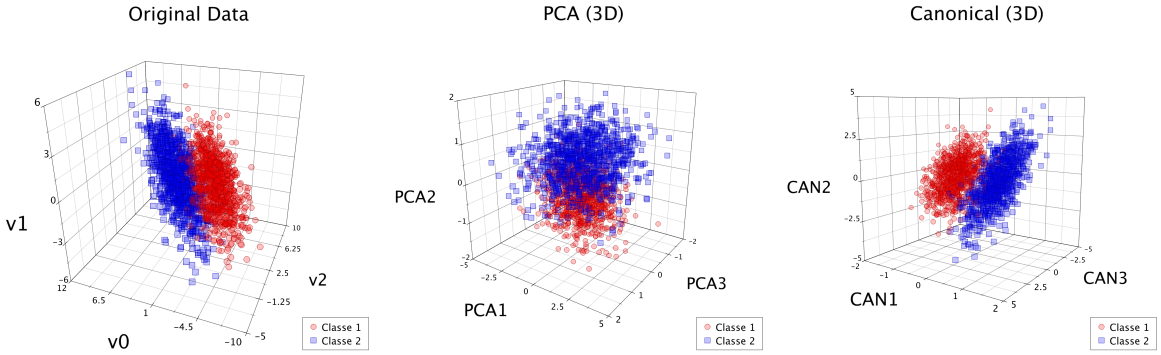
Figura 2.12 – Exemplo de PCA para um conjunto de dados de 2 variáveis distribuído de modo normal ao longo de uma reta.

Geometricamente, os valores originais foram submetidos a uma rotação multidimensional (59). A figura 2.12 ilustra um exemplo clássico onde os valores do conjunto de dados estão distribuídos ao redor de uma reta. Verifica-se que as variáveis originais, ”v0” e ”v1” são altamente correlacionadas. O auto-vetor de maior auto-valor, equivalente a nova variável chamada de ”PCA1” ajusta-se perfeitamente na direção de maior variância, enquanto que para o outro auto-vetor, ”PCA2”, é ortogonal a primeira e apresenta muito menos variância, e consequentemente menos informação. Neste caso, a redução de dimensionalidade para L poderia ser feita simplesmente descartando-se as variáveis menos importantes (menores auto-valores). Nota-se que após a transformação, como esperado, as variáveis não estarão correlacionadas.

2.4.3 Análise por Variáveis Canônicas

É comum que conjuntos de dados apresentem, além das propriedades observadas, um atributo de classe, que determina a que tipo de grupo cada elemento pertence. Em uma rede complexa de colaboração acadêmica, por exemplo, a classe pode ser um atributo que identifica a

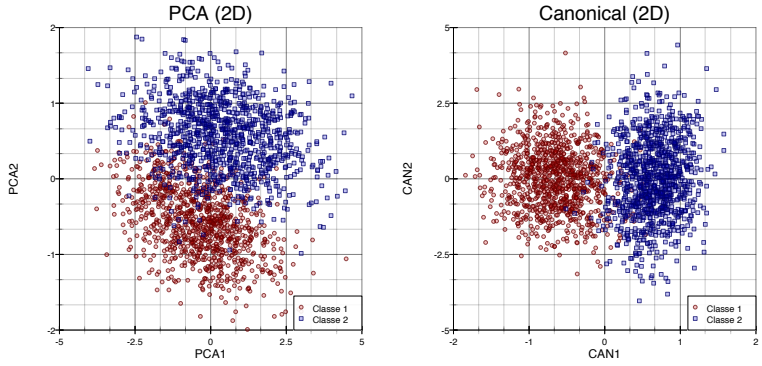
qual instituição um pesquisador (vértice) faz parte. Muitas vezes é desejável que as classes sejam consideradas ao realizar uma análise de redução de dimensionalidade ou visualização dos dados, no entanto, a metodologia PCA não é adequada a esse tipo de problema. A figura 2.13b ilustra esta deficiência, a figura mostra a projeção de um conjunto de dados composto por 2 classes usando PCA, no entanto a variável que deveria ser mais importante, ”PCA1”, não segrega os dados atribuídos a classe, enquanto que a segunda variável PCA, ”PCA2”, apresenta melhor separação entre as duas classes, portanto, se fosse realizada uma operação de redução de dimensionalidade a informação sobre a segregação das classes poderia ser perdida.



(a) Dados originais em 3D.

(b) Projeção 3D no espaço determinado por PCA.

(c) Projeção 3D no espaço canônico.



(d) Projeção 2D usando PCA.

(e) Projeção 2D por variáveis canônicas.

Figura 2.13 – Projeções de um conjunto de dados tridimensional composto por 2 classes em variáveis determinadas por PCA e por análise de variáveis canônicas.

A metodologia de análise por variáveis canônicas (59, 58) estende a PCA considerando os atributos de classe de cada elemento do conjunto de dados, para isso busca-se uma transformação que maximize a dispersão entre os elementos de classes diferentes e, ao mesmo tempo, minimize a dispersão dentro de uma mesma classe. O método foi proposto inicialmente por Fisher(60) e utiliza quase a mesma técnica do PCA exceto por considerar outra matriz a ser diagonalizada, e não a matriz de convergência.

Considerando um conjunto de dados, $\mathscr{X}$, composto por N elementos classificados em K_C classes identificadas por C_i, $i = 1,2,3,\ldots K_C$, onde o número de elementos em cada classe é K_i, e cada a cada elemento é associado um vetor de propriedades observadas, $\vec{X}_e$, com $e = 1,2,3,\ldots N$ define-se uma matriz de dispersão interna para cada classe, S_i, como:

$$S_i = \sum_{e \in C_i}^{N} (\vec{X}_e - \langle \vec{X} \rangle_i) \otimes (\vec{X}_e - \langle \vec{X} \rangle_i) \tag{2.23}$$

onde $\langle \vec{X} \rangle_i$ é a média considerando apenas os elementos pertencentes a classe i. Pode-se definir uma nova matriz de dispersão, composta por todas as dispersões internas das classes, S_{intra}, como a soma:

$$S_{\text{intra}} = \sum_{i=1}^{K_C} S_i \tag{2.24}$$

Também define-se a matriz de dispersão entre as classes, S_{inter}, como:

$$S_{\text{inter}} = \sum_{i=1}^{K_C} N_i (\langle \vec{X} \rangle_i - \langle \vec{X} \rangle) \otimes (\langle \vec{X} \rangle_i - \langle \vec{X} \rangle) \tag{2.25}$$

Nota-se que a dispersão total pode ser escrita em termos da soma de ambas as dispersões:

$$S_{\text{total}} = S_{\text{intra}} + S_{\text{inter}} \tag{2.26}$$

Como o objetivo da análise canônica é minimizar a dispersão interna de cada classe e, simultaneamente, maximizar a dispersão entre elas, pode-se buscar por uma métrica de magnitude obtida a partir das matrizes de dispersão, como o determinante ou o traço. Neste caso, o uso do traço é a melhor opção, pois além da simplicidade, pode ser escrito em termos dos auto-valores, λ_i de uma matriz quadrada A: $tr(A) = \sum_i^M \lambda_i$. Outra vantagem é o fato do traço ser uma operação linear e, portanto, quando aplicado à dispersão total, seu valor independe da transformação realizada aos dados originais e é conservado:

$$tr(S_{\text{total}}) = tr(S_{\text{intra}}) + tr(S_{\text{inter}}) \tag{2.27}$$

o que torna a razão entre os traços uma boa medida a ser maximizada. Pode-se mostrar (59) que as variáveis observadas mais adequadas ao problema são aquelas que mais contribuem para o valor do traço da matriz $S_{\text{intra}}^{-1} S_{\text{inter}}$. Portanto, a transformação mais adequada aos dados originais é aquela gerada pela base de auto-vetores de $S_{\text{intra}}^{-1} S_{\text{inter}}$, $\vec{\gamma}_i$, associados aos respectivos auto-valores λ_i, tal que $\lambda_1 > \lambda_2 > \lambda_3 > \ldots \lambda_M$.

Ao realizar a redução de dimensionalidade para L, os vetores da base devem compor o novo espaço de transformação como $\Gamma_L = \{\vec{\gamma}_1, \vec{\gamma}_2, \vec{\gamma}_3, \ldots \vec{\gamma}_L\}$, e $tr(S_{\text{intra}}^{-1} S_{\text{inter}})$ fornecerá uma métrica

da qualidade dos resultados, quanto maior seu valor, mais bem sucedida foi a separação entre as classes e minimização da dispersão interna.

A figura 2.13e mostra a projeção usando as variáveis canônicas e verifica-se que essa abordagem é muito mais adequada do que a PCA (fig. 2.13d), delimitando as regiões entre as classes usando apenas a primeira variável.

3 *Metodologia*

3.1 Recursos e Procedimentos Computacionais

Devido a complexidade ou indeterminismo de algumas propriedades em redes complexas a metodologia analítica nem sempre é possível de ser aplicada, entretanto métodos estatísticos, numéricos ou simulações computacionais podem contribuir muito para a compreensão dessas propriedades. Metodologias estatísticas são mais comuns e tendem a ser usadas para a caracterização de grande número de redes ou vértices, um exemplo dessas metodologias, é a utilização da distribuição da conectividade dos vértices, que revela o comportamento livre de escala da rede.

Com o avanço do poder computacional, ocorrido nos últimos anos, tornou-se possível a utilização de metodologias de simulação computacional aplicáveis a redes complexa. A simulação computacional baseia-se na representação dos modelos em termos das estruturas de dados presentes em um ambiente programação e da execução de rotinas e operações matemáticas, o programa, em um computador.

A abordagem computacional é responsável pelo crescente aumento do interesse de estudos em redes complexas, pois permite que grande quantidade de dados sejam analisados, resultando em propriedades que não eram visíveis ao considerar os problemas de modo ”reducionista”. Algoritmos computacionais também permitem que representações visuais automatizadas dessas redes possam ser geradas, facilitando a visualização de suas propriedades.

Esta seção apresenta algumas das ferramentas computacionais usadas no contexto deste trabalho, assim como o desenvolvimento detalhado de novos softwares, discutindo brevemente alguns problemas enfrentados e as soluções.

3.1.1 Recursos de Software

A análise de redes complexas exige diversas etapas que se iniciam com a aquisição dos dados e terminam com a visualização e publicação dos resultados, a grande maioria delas baseiam-se em cálculos ou na execução de algoritmos sobre os dados, necessitando a utilização de softwares específicos ou da implementação de programas pelo próprio pesquisador. Transformar os resultados em informação também é uma tarefa que pode ser auxiliada por utilitários computacionais, tanto na forma de imagens, quanto na criação de representações visuais interativas. A seguir, alguns dos softwares mais relevantes ao trabalho são apresentados, a tabela 3.1 mostra uma lista com aqueles mais relevantes e seus respectivos ponteiros.

Inicialmente foram necessários utilitários com rotinas para gerar os modelos de redes clássicas (ER, BA, WS e geográficas), o software *Pajek* possui parte dessas rotinas e permite selecionar parâmetros como tamanho e grau da rede, ele também permite a extração de medidas simples, como a conectividade, menores caminhos e centralidades clássicas. O software *jComplexNetworks* e o conjunto de bibliotecas *ComNetKit*, desenvolvidos especialmente para o trabalho, também incluem algumas das rotinas geradoras e rotinas para o cálculo de propriedades clássicas. Complementarmente, o jComplexNetworks permite obter as propriedades concêntricas.

A visualização de redes pode ser feita tanto pelo o Pajek quanto pelo *Cytoscape*, o primeiro possui diversas opções gráficas e de diagramação, mas, apesar disso, ambos não são apresentam visualizações interativas de redes, além de não serem adequadas a redes com grande número de vértices. Por esse motivo, foi desenvolvido um software especifico para a criação de visualizações interativas de redes complexas em duas ou três dimensões, o *Network3D*, criado sobre as bibliotecas ComNetKit, é optimizado para redes grandes, e suas visualizações podem ser vistas em qualquer computador com máquina virtual Java instalada.

Os softwares *Cluster* e *Java TreeView* compõem um conjunto de ferramentas para analise estatística e segmentação não supervisionada de dados através de algoritmos de aglomeração hierárquica (58, 57). Apesar do Cluster realizar PCA sob os dados, o software *Classification-ToolKit* foi especialmente desenvolvido para este trabalho, implementando PCA, análise por variáveis canônicas e visualização dos resultados em 2D ou 3D.

O desenvolvimento dos programas para este trabalho foram realizados em diversos ambientes. Para aqueles escritos em C ou Objective-C compiladores *GNU* (*GCC*) foram utilizados, e o ambiente de desenvolvimento foi o *Xcode*. Programas escritos em Java foram desenvolvi-

Tabela 3.1 – Softwares usados ou desenvolvidos durante a execução deste trabalho.

Ferramentas de Desenvolvimento		
Compiladores GNU	v4.2	Conjunto de compiladores C e Objective-C. http://gcc.gnu.org/
Java	v1.6	Execução e compilação de códigos Java. http://www.java.com/
Netbeans	v6.5	Ambiente para desenvolvimento Java. http://www.netbeans.org/
Python	v2.5	Interpretador da linguagem ”script” de mesmo nome. http://www.python.org/
Scilab	v5.1	Ambiente de programação rápida e ciêntifica. http://www.scilab.org/
Xcode	v3.2	Ambiente de desenvolvimento em C e Objective-C http://http://developer.apple.com/technology/xcode.html

Ferramentas de Publicação		
Latex	-	Software gerador de documentos para publicação. http://www.latex-project.org/
Papers	v1.9	Gerenciador de referências bibliográficas. http://mekentosj.com/papers/
TexShop	-	Editor de documentos Latex. http://www.uoregon.edu/˜koch/texshop/

Ambientes Operacionais		
Mac OS X	v10.5	Sistema Operacional UNIX baseado em Darwin. http://www.apple.com/macosx/
Ubuntu	v8.0	Distribuição do Sistema Operacional Linux. http://www.ubuntu.org/
Windows	XP	Sistema Operacional Comercial. http://www.microsoft.com/windows/

Softwares de Análise de Redes Complexas		
Cytoscape	-	Ferramenta de obtenção de propriedades e visualização de redes complexas. http://www.cytoscape.org/
Pajek (61)	v1.24	Ferramenta de obtenção de propriedades de redes complexas. http://pajek.imfm.si/

Softwares de Análise Estatística		
Cluster (62)	v3.0	Software para redução de dimensionalidade e aglomeração hierárquica (57). http://bonsai.ims.u-tokyo.ac.jp/˜mdehoon/software/cluster/
Java TreeView	v13	Visualizador e gerenciador de dendrogramas gerados pelo software Cluster. http://jtreeview.sourceforge.net/

Softwares desenvolvidos		
ClassificationToolKit	beta	Software para redução de dimensionalidade, classificação e visualização de dados. não disponivel no momento.
ComNetKit	v1.0	Biblioteca de códigos para redes complexas em Objective-C/Cocoa. não disponivel no momento.
jComplexNetworks	v0.8b	Software para extrair propriedades concêntricas de redes complexas. http://cyvision.if.sc.usp.br/˜bant/hierarchical/
Network3D	v1.0	Software gerador de visualização interativas de redes complexas. não disponivel no momento.

Tabela 3.2 – Linguagens de programação usadas para o desenvolvimento dos softwares.

Linguagem	Tipo	Benefícios
Objective-C	Nativa	Orientada a objetos, rápida, baseada em C, ambiente completo.
Python	Script	Facilidade, ambiente bom para processar textos e integrar aos softwares.
Java	Bytecoded	Orientada a objetos, independe da arquitetura.

dos integralmente no software *NetBeans*. Pequenos utilitários e prototipação de código foram feitos em linguagens roterizadas (*scripts*) como *Scilab* e *Python*, este último foi incorporado ao software Network3D, permitindo que este fosse roterizável.

A criação de publicações em revistas e desta dissertação foram realizadas usando softwares baseados na tecnologia *Latex*, como o editor de documentos *TexShop* e o organizador de referências bibliográficas *Papers*.

Todo o trabalho foi realizado em três diferentes sistemas operacionais, *Mac OS X*, *Windows* e *Linux*, representado pela distribuição *Ubuntu*.

As seções a seguir apresentam alguns dos problemas enfrentados e soluções para o desenvolvimento dos softwares deste trabalho, a tabela 3.2 mostra as linguagens de programação usadas, destacando algumas das vantagens.

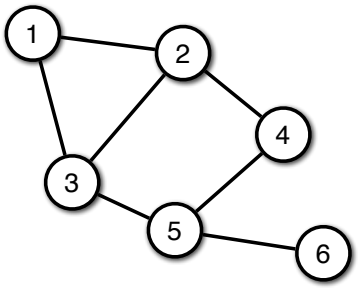
3.1.2 Representação de Redes Complexas

Representar um modelo de rede complexa em estruturas computacionais nem sempre é uma tarefa trivial. A representação mais usada, a matriz de adjacências, pode não ser adequada a certas situações, como, por exemplo, redes com grande número de nós e poucas arestas, pois há grande desperdício de espaço computacional já que a maioria de seus valores são nulos.

Para o desenvolvimento dos softwares criados durante o curso deste trabalho, foram usadas duas representações computacionais para as redes: matriz de adjacência, listas de adjacência e uma combinação de listas de adjacência com lista de arestas, chamada de lista híbrida.

Matriz de adjacência

Já descrita no capítulo anterior (cap. 2), além da simplicidade da estrutura de dados, apresenta a vantagem de permitir saber se um vértice arbitrário está conectado a outro em tempo $O(1)$.



$$G = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Figura 3.1 – Exemplo de rede representada pela matriz de adjacência.

A figura 3.1 ilustra a matriz de adjacência para uma rede de 6 vértices não direcionada. Em redes ponderadas nas arestas, costuma-se atribuir o valor da matriz como o peso da conexão, ou seja $G_{ij} = W_{ij}$, quando não há conexões, o valor do elemento correspondente é tomado como inválido.

Como mostra a tabela 3.3 matriz de adjacência apresenta como grande vantagem um tempo $O(1)$ para se descobrir se dois vértices estão ligados por uma aresta, no entanto requer um custo de memória $O(N^2)$ mesmo para redes esparsas ($E \ll N$), neste caso há grande desperdício de espaço. Determinar todas as arestas também é problemático já que toda a matriz deve ser percorrida.

Outra desvantagem dessa representação é para o caso de redes dinâmicas, pois a adição ou remoção de novos vértices necessita que a matriz de adjacências seja totalmente reestruturada, no entanto não apresenta problemas para a adição ou remoção de arestas.

Listas de adjacência

Considerando uma rede complexa $\Gamma(\mathscr{V}, \mathscr{E})$, define-se uma lista de adjacência, ξ_v, de um vértice v, como uma lista contendo todos os vértices vizinhos a v, $\{e_{va}, e_{vb}, \ldots\} \subseteq \mathscr{E}$, pode-se, então, representa-la como o conjunto $\Xi = \{\xi_1, \xi_2, \ldots \xi_N\}$ contendo todas as listas de adjacência para cada vértice.

Em termos computacionais, cada lista de adjacência pode ser implementada por uma lista encadeada de vértices, e o conjunto das listas por uma tabela hash. A vantagem dessa implementação está na possibilidade de dinamicamente expandir ou reduzir o número de vértices da rede.

A figura 3.2 ilustra a representação de uma rede complexa por listas de adjacência implementada por listas encadeadas.

As tabelas 3.3 e 3.4 comparam as representações em termos dos tempos para realizar

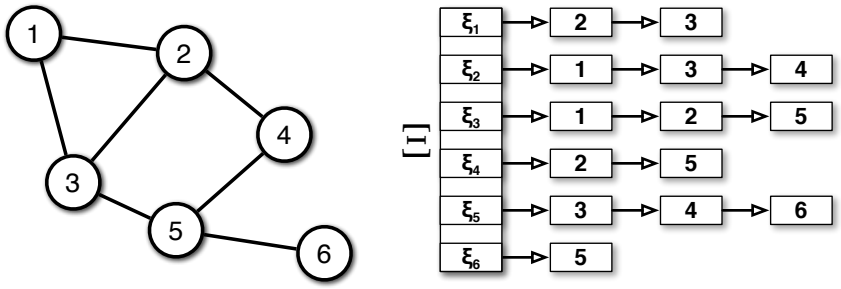


Figura 3.2 – Exemplo de rede representada por listas de adjacência.

operações e do espaço de memória para mante-las. Para redes esparsas, a utilização de listas de adjacência apresenta grande vantagem sobre as matrizes de adjacência, pois além de menor espaço de memória necessário pode-se obter os vizinhos de um vértice em tempo menor do que $O(N)$. Em geral, algoritmos que percorram a rede tenderão a ter uma performance superior usando essa estrutura de dados.

Lista híbrida

Muitas vezes, pode ser importante obter o conjunto de todas as arestas de uma rede complexa, um exemplo dessa necessidade ocorre quando há dinâmicas aplicadas diretamente sobre elas. Tanto a matriz de adjacência quanto as listas de adjacência não apresentam uma forma de enumerar as arestas. Uma solução para este problema é anexar às listas de adjacência um conjunto E contendo todas as arestas $\{e_1, e_2, \ldots\} \subseteq \mathscr{E}$, diretamente acessíveis e enumeráveis.

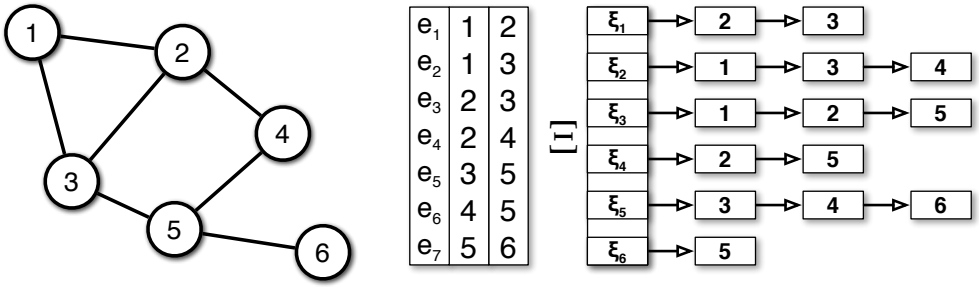


Figura 3.3 – Exemplo de rede representada por lista híbrida.

Um exemplo dessa representação pode ser encontrado na figura 3.2, e, como pode-se observar nas tabelas 3.3 e 3.4, tem como vantagem o acesso a todas as arestas em $O(E)$, além do fato de estarem enumeradas.

Tabela 3.3 – Comparação dos tempos de diferentes operações usando as três representações de redes complexas consideradas.

Operação	Matriz de adjacências	Listas de adjacências	Lista híbrida
É vizinho?	$O(1)$	$\leq O(N)$	$\leq O(N)$
Arestas saindo do vértice	$O(N)$	$\leq O(N)$	$\leq O(N)$
Arestas entrando no vértice	$O(N)$	$O(N+E)$	$O(E)$
Todas as arestas	$O(N^2)$	$O(N+E)$	$O(E)$

Tabela 3.4 – Comparação dos requisitos de espaço em memória das três representações consideradas.

	Matriz de adjacências	Listas de adjacências	Lista híbrida
Espaço	$O(N^2)$	$O(N+E)$	$O(N+2E)$

3.2 Obtenção de propriedades das redes

A partir de uma representação computacional da rede, é possível extrair as propriedades da rede, no entanto, apesar do cálculo ser trivial para propriedades locais como o grau e o coeficiente de aglutinação, pode tornar-se cada vez mais complexa ou custosa, em termos de tempo e espaço computacional, dependendo de sua abrangência. As medidas de centralidade e concêntricas consideram toda a rede e, portanto, são as mais custosas, necessitando de algoritmos mais sofisticados.

O software Pajek já possui rotinas embutidas para o cálculo de algumas propriedades, como as centralidades de intermediação e proximidade, no entanto para as outras propriedades duas centralidade, centralidade de subgrafos e de auto-vetor, foram criadas rotinas em Scilab e Python para a realização do cálculo. A centralidade de auto-vetor pode ser obtida facilmente pelo método das potências (63), iniciando de um vetor arbitrário, y_0, realiza-se sucessivas operações de multiplicação pela matriz de adjacência, G, da rede, seguida de normalização,

$$\vec{y}_{k+1} = \frac{G\vec{y}_k}{\|G\vec{y}_k\|} \tag{3.1}$$

o vetor resultante que corresponde ao k, tal que $\|\vec{y}_{k+1} - \vec{y}_k\| < \varepsilon$, será o valor aproximado do auto-vetor principal de G, $\vec{x}$, com erro máximo da ordem de ε.

A centralidade de subgrafos pode ser calculada diretamente em termos de potências da matriz de adjacência, como descrito na seção 2.4.

Dois softwares com a finalidade de calcular outras propriedades de redes complexas foram desenvolvidos, o jComplexNetworks e o ComNetKit, o primeiro é um programa destinado ao usuário final que permite calcular e visualizar as propriedades hierárquicas de redes complexas, o segundo é um conjunto de bibliotecas escritas em linguagem Objective-C orientada a objetos.

A biblioteca ComNetKit é melhor descrita na seção 3.5, mas implementa as mesmas técnicas descritas para o jComplexNetworks.

3.2.1 Cálculo das propriedades concêntricas.

Determinar as propriedades concêntricas de um vértice necessita, inicialmente, que a rede seja organizada em sub-redes, ou anéis, tal que cada uma deve conter apenas vértices que distanciam pela mesma quantidade do vértice de referência. A determinação dessas sub-redes pode se feita usando uma busca em largura na rede, partindo do vértice de referência, e para cada nível de profundidade gerar um anel, como mostra o algoritmo em pseudo-código 3.2.1.

Algoritmo 3.2.1: ANÉIS($\mathscr{V}, \mathscr{E}, ref$)

// onde ref é o vértice de referencia.

$aneis \leftarrow \{\emptyset, \emptyset, \emptyset, \ldots\}$
$visitados \leftarrow \emptyset$
$fila \leftarrow \{ref\}$
$filanivel \leftarrow \{0\}$
enquanto $fila$ não está vazia
faça
- $v \leftarrow fila$.próximo()
- $nivel \leftarrow filanivel$.próximo()
- $aneis[nivel]$.adicionar(v)
- **para cada** $\{u \in \mathscr{V} : \{v,u\} \in \mathscr{E}\}$
  - **faça**
    - **se** $u \notin visitados$
      - **então**
        - $visitados$.adicionar(u)
        - $fila$.adicionar(u)
        - $filanivel$.adicionar($nivel+1$)

return ($aneis$)

Após a obtenção das sub-redes, o cálculo das propriedades concêntricas reduz-se em determinar localmente as propriedades de cada um desses anéis. O algoritmo para cada vértice pode ser executado com ordem de tempo $O(N+E)$, e portanto, para se obter o perfil concêntrico completo da rede é necessário um tempo de ordem $O(N^2+NE)$.

3.2.2 jComplexNetworks

Desenvolvido para obter as propriedades concêntricas de redes complexas, é um software destinado ao usuário final. Foi escrito em linguagem Java e, portanto, pode ser executado na maioria das plataformas atuais. Ele baseia-se em um fluxo de dados linear, isto é, um conjunto de dados, a rede, é importado ou gerado e este passa por diversas operações até se obter os resultados finais com certa interação com o usuário.

A figura 3.4 apresenta um esquema simplificado do fluxo de dados do software. Inicialmente o usuário é apresentado a um painel(figura 3.5a) contendo diversas opções de parâmetros de entrada. O usuário pode optar por gerar uma rede complexa de acordo com os modelos teóricos discutidos anteriormente ou importar a rede de um arquivo no formato *.net* (usado principalmente pelo software Pajek). O próximo passo é o cálculo das propriedades, para isso há alguns parâmetros a serem fornecidos pelo usuário como o máximo número de anéis concêntricos que devem ser considerados, se as propriedades devem ser normalizadas pelo padrão de uma rede aleatória e se as propriedades devem ser consideradas também para bolas concêntricas.

O resultado final é exportado automaticamente na forma de diversos documentos associando, a cada vértice, uma distribuição das propriedades concêntricas ao longo dos níveis.

O software também permite a visualização das propriedades concêntricas e a geração de gráficos personalizáveis para anexar a publicações, a exemplo do painel apresentado na figura 3.5b.

3.2.3 Software para estudo PCA

Apesar de softwares como o Cluster realizarem estudos usando análise por componentes principais, apresentam pouca ou nenhuma capacidade de interação com o usuário para gerar imagens ou redução de dimensionalidade. Em geral, também não apresentam estudos por variáveis canônicas, nem permitem a classificação supervisionada de dados. Por estes motivos um software, ClassificationToolKit, dedicado exclusivamente ao estudo de dados por PCA foi desenvolvido para usuários finais, em linguagem Java, implementando os métodos descritos na seção 2.4.2 usando um fluxo de dados linear, como exemplificado na figura 3.6.

Inicialmente, um arquivo formatado é lido e dele são extraídas a tabela de dados e, se hou-

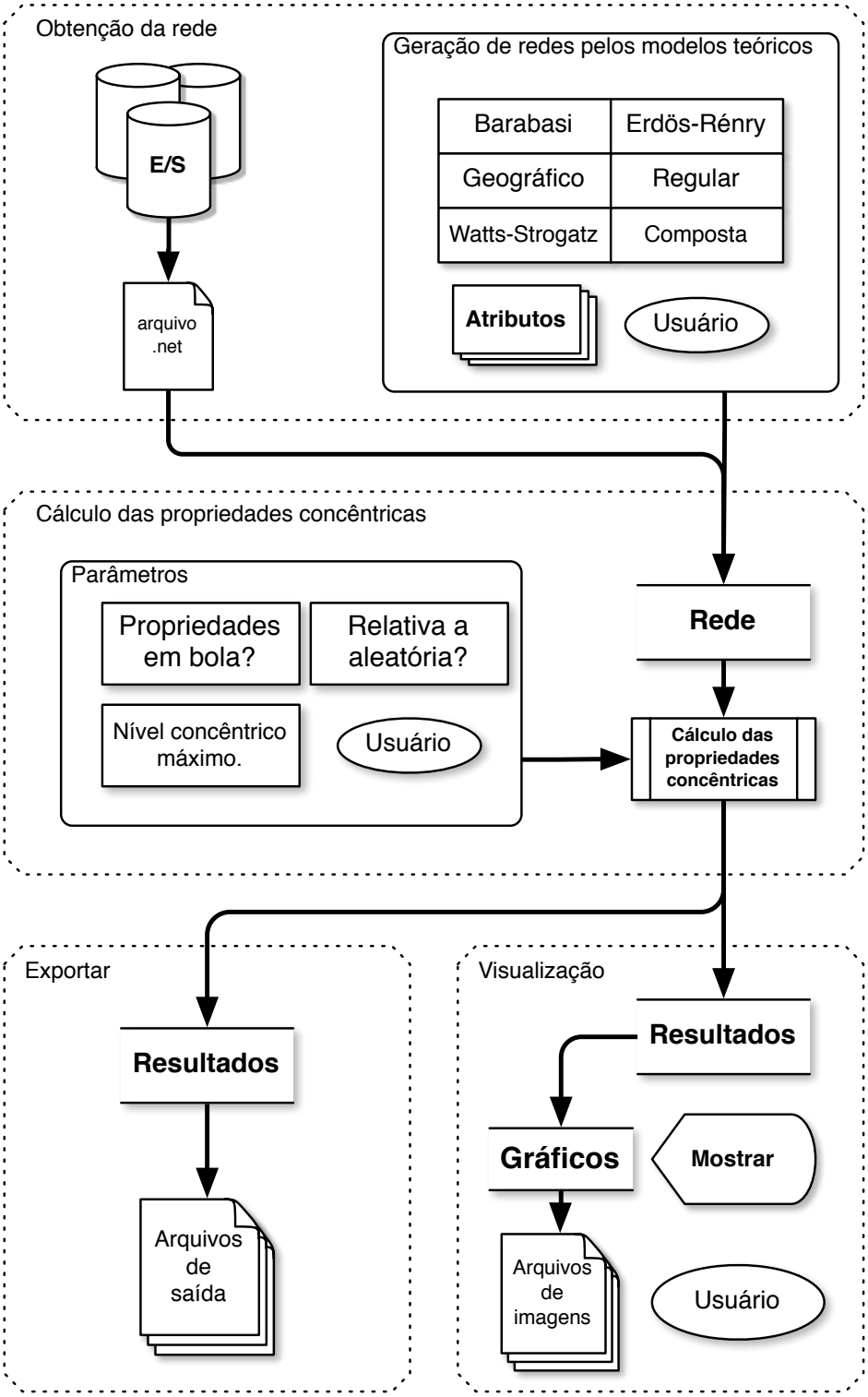
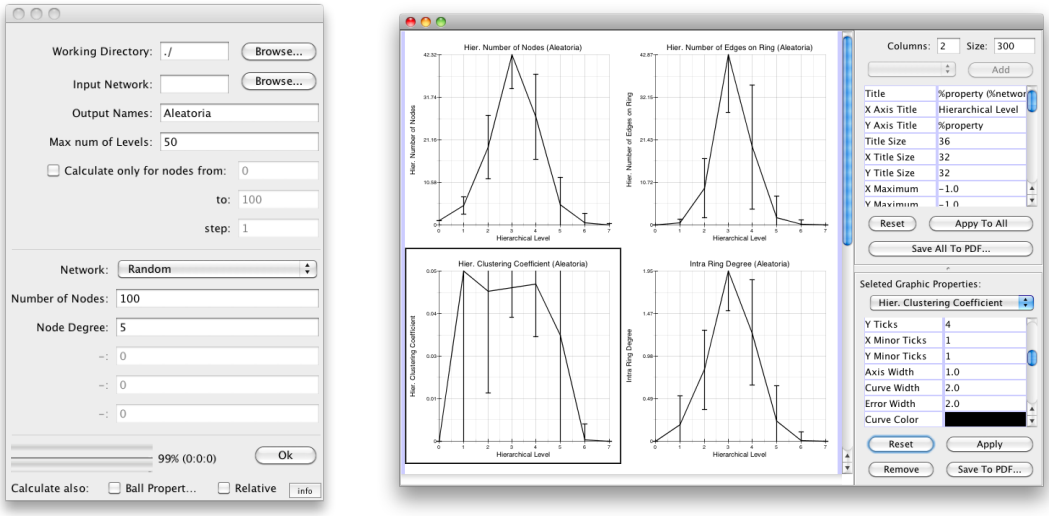


Figura 3.4 – Esquema do fluxo de dados do software jComplexNetworks.



(a) Painel inicial. **(b)** Seleção de Gráficos.

Figura 3.5 – Capturas de tela do software jComplexNetworks.

ver, a lista de categorias informando a quais cada amostra pertence. O painel inicial (fig. 3.7a) do software pergunta ao usuário por parâmetros e quais operações devem ser realizadas e permite que os dados sejam editados. Os parâmetros são o número de dimensões a serem reduzidas e variáveis a serem consideradas para a visualização. O usuário pode optar por normalizar os dados em termos dos desvios padrão (estandardização), calculando-se o *Z-Score* (64) dos dados (equação 3.3).

$$\vec{Z} = \frac{\vec{X} - \langle \vec{X} \rangle}{\sqrt{\langle (\vec{X} - \langle \vec{X} \rangle)^2 \rangle}} \tag{3.2}$$

Caso sejam fornecidos os dados correspondentes as categorias, o usuário pode escolher entre realizar o estudo PCA tradicional ou por variáveis canônicas. Dois métodos de classificação supervisionada foram implementados, a metodologia *k-vizinhos* e a por *máxima verossimilhança* (maximum likehood), ambos estão disponíveis ao usuário quando há dados não categorizáveis.

Enquanto o método k-vizinhos baseia-se, simplesmente, em classificar as amostras na categoria que possui maior quantidade de k vizinhos mais próximos (em termos de distância euclidiana), o método de máxima verossimilhança (57) utiliza a idéia de que cada categoria pode ser representada por uma distribuição gaussiana, n-dimensional, de tal modo que, por análise bayesiana, obtém-se um mapa de regiões especificas a cada de categoria, determinados por aquela com maior valor de sua distribuição no ponto. A aproximação da distribuição dos dados por uma função gaussiana n-dimensional, $p_C(\vec{X})$, é definida como:

$$p_C(\vec{X}) = \frac{1}{(2\pi)^n \sqrt{|\Sigma|}} \exp\left\{ -\frac{1}{2} (\vec{X} - \langle \vec{X} \rangle_C)^T \Sigma^{-1} (\vec{X} - \langle \vec{X} \rangle_C) \right\} \tag{3.3}$$

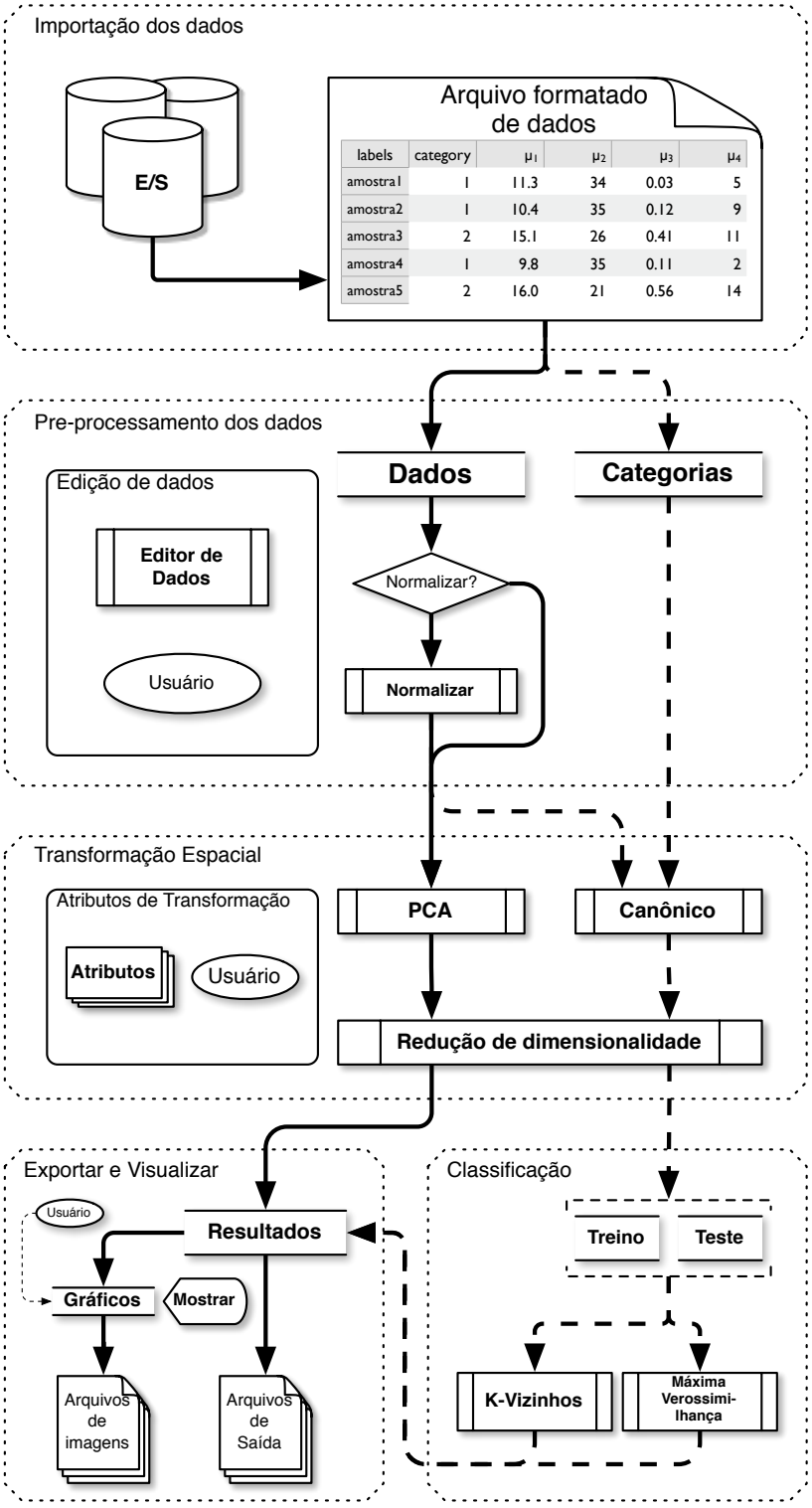
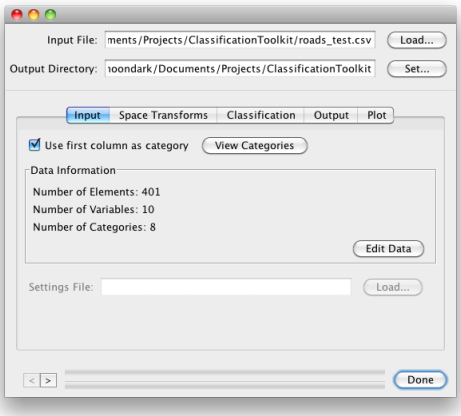
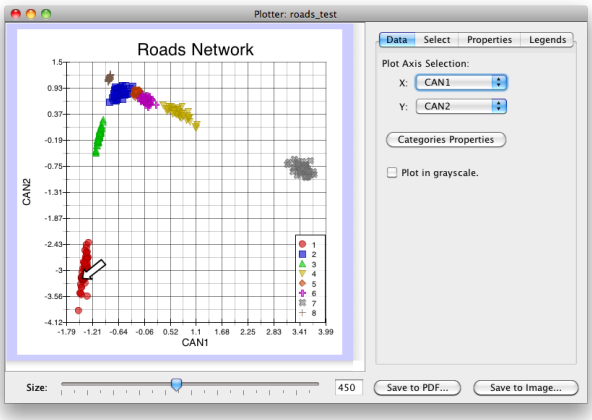


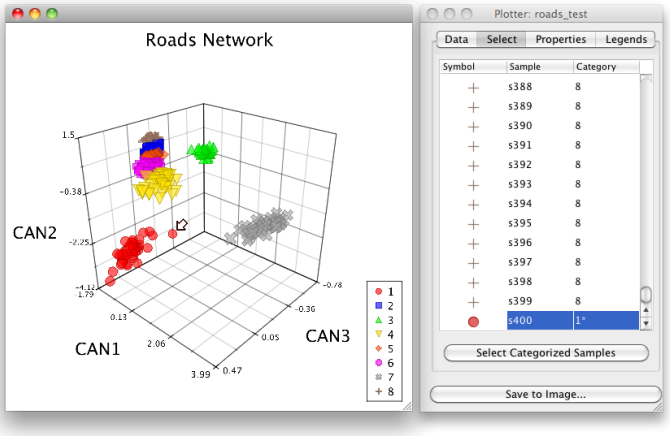
Figura 3.6 – Esquema do fluxo de dados do software ClassificationToolKit. Os caminhos pontilhados correspondem ao fluxo de dados para a análise por variáveis canônicas.



(a) Painel inicial.



(b) Visualização 2D dos resultados.



(c) Visualização 3D dos resultados.

Figura 3.7 – Capturas de tela do software ClassificationToolKit.

onde $\langle \vec{X} \rangle_C$ é a média do vetor de dados considerando apenas as amostras da classe C e Σ é a matriz de covariância.

Ao final do processo, os dados são exportados em arquivos formatados e o usuário é apresentado a um painel contendo a visualização interativa 2D (fig. 3.7b) ou 3D (fig. 3.7c) dos resultados, permitindo que este altere os parâmetros de visualização e exporte as imagens em diversos formatos prontos para divulgação.

3.2.4 Visualização de Redes Complexas

O crescente aumento do interesse em estudos sobre redes complexas resultou na necessidade de aprimoramentos e de novos métodos aplicados a visualização desse tipo de informação, deixando de ser apenas uma forma de divulgação para servir como uma ferramenta importante

para a obtenção e validação de certas propriedades.

Apesar de surgirem cada vez mais metodologias poderosas para a análise de redes complexas, a visualização gráfica ainda é uma forma simples e muitas vezes eficaz para apresentar propriedades da rede.

No início do desenvolvimento e implementação dessas aplicações, devido ao baixo poder de processamento da época e da falta de algoritmos aprimorados, a apresentação visual de redes possuía pouca qualidade e fornecia pouca informação sobre a topologia geral, pois eram limitadas a poucos vértices e arestas. Outro problema relacionado a apresentação de redes com grande número de vértices está relacionado com a quantidade de informação que deve ser apresentada. Enquanto redes pequenas podem ser apresentadas na íntegra, redes grandes devem ter sua informação reduzida.

Soluções para a visualização de redes grandes somente apareceram recentemente, devido ao avanço do poder computacional e da criação de novas metodologias, assim como visualizações interativas com o objetivo de reduzir a quantidade de informação apresentada de uma vez.

Uma forma amplamente usada para visualizar grafos é a distribuição, sob um espaço 2D ou 3D, de símbolos (como círculos), representando os vértices, e linhas(ou curvas) correspondentes às arestas. Propriedades dos vértices ou arestas podem ser observadas atribuindo cores ou etiquetas textuais, outra alternativa é a utilização de diferentes símbolos e linhas para identificalas.

Network3D

O software Network3D foi criado para gerar visualizações interativas de redes complexas através do uso de um algoritmo de posicionamento baseado em interações de particulas por forças eletromagnéticas descrito no apêndice A. O software permite que redes de grande tamanho sejam visualizadas tanto em 2D como em 3D, assim como suas propriedades na forma de cores. Ele baseia-se na biblioteca ComNetKit e foi escrito em Objective-C, podendo ser executado em ambiente operacional Mac Os X, outras plataformas devem ser suportadas em breve.

Diferentemente dos outros softwares criados para o projeto, o Network3D (fig. 3.8) não trabalha com um fluxo de dados linear, pois, praticamente, todos os estágios são interativos. Após a importação de uma rede complexa, em formato *".net"* ou *".xnet"*, o maior componente conectado é extraído e a simulação do posicionamento dos vértices é iniciada em modo interativo, permitindo que o usuário dinamicamente altere os valores dos parametros do algoritimo.

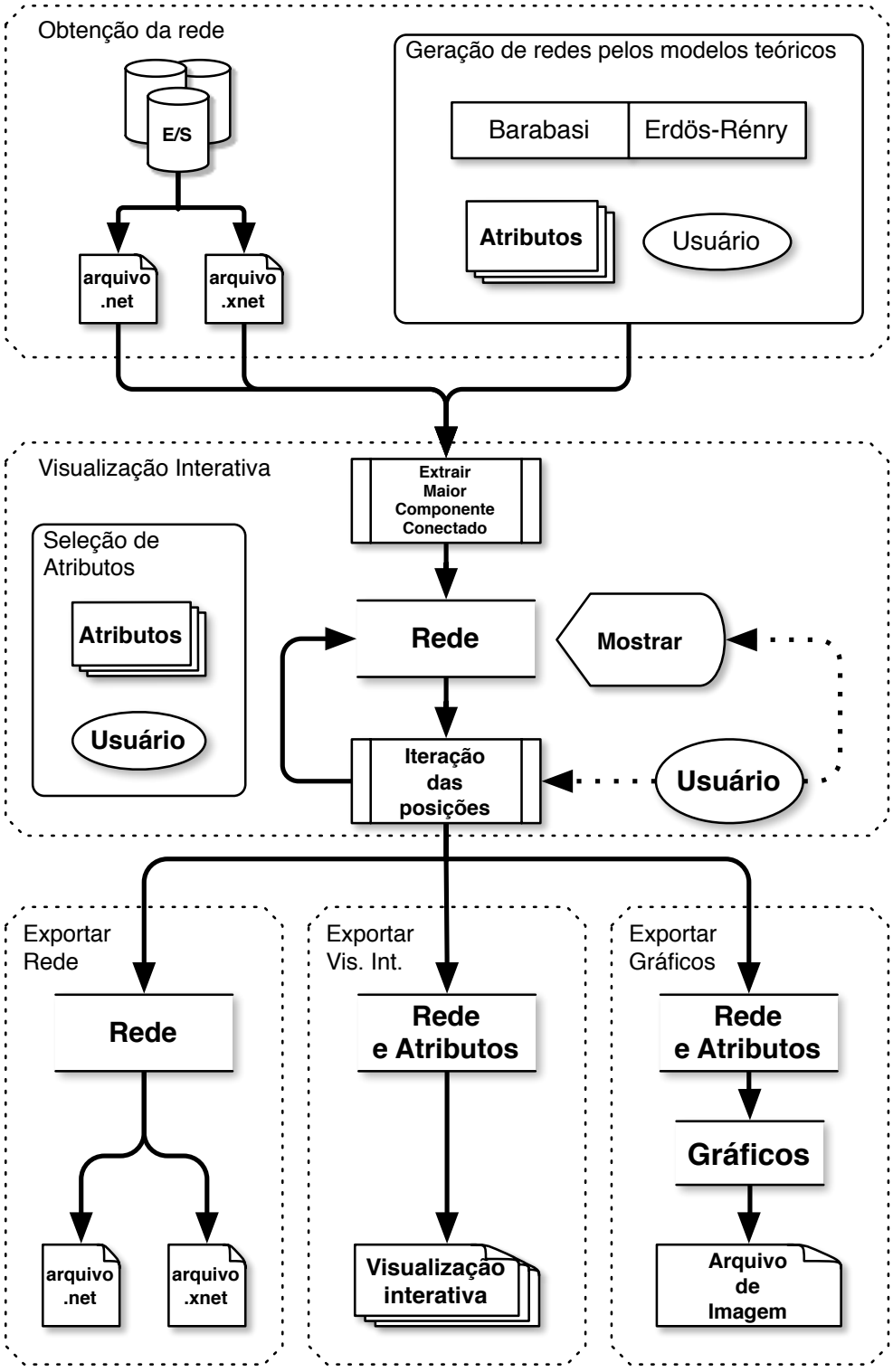


Figura 3.8 – Esquema do fluxo de dados do software Network3D.

A visualização interativa permite que o usuário realize operações de movimentação, rotação e aproximação do mapa de projeção 2D ou 3D, além de fornecer detalhes das propriedades de cada vértice ao seleciona-los. As propriedades são apresentadas por cores e o usuário tem opção de usar dois mapas de cores, o espectral (ou *JetColorMap*) ou tonalidade de cinza. As funções de transferência para as propriedades também podem ser escolhidas, com opções de linear, exponencial ou logarítmica.

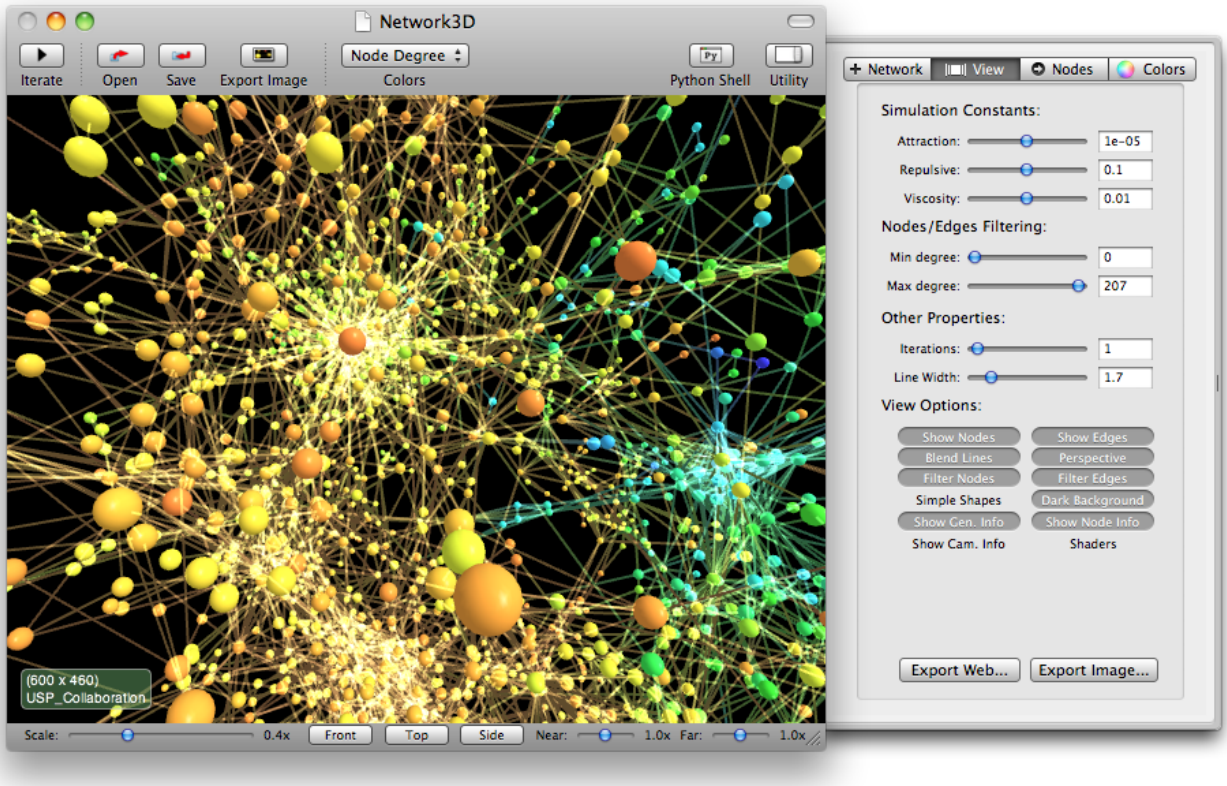


Figura 3.9 – Captura de tela do painel principal do software Network3D.

A interface é composta por um único painel contendo a visualização ao lado de seus parâmetros (fig. 3.9). Por padrão as redes são apresentadas com os vértices coloridos por seus respectivos valores de conectividade. Esta configuração resultou ser agradável visualmente para a maioria das redes.

Ao software foi adicionada a capacidade de usar roteiros de execução escritos em Python, permitindo que a maioria dos aspectos do programa pudessem ser controlados por códigos extras. Este recurso foi implementado de modo que todas as rotinas da biblioteca de manipulação de redes, ComNetKit e a maioria dos elementos de interface possam ser executados por códigos em linguagem Python. Um editor de roteiros (fig. 3.10) também está embutido ao software, assim como um *shell* de comandos.

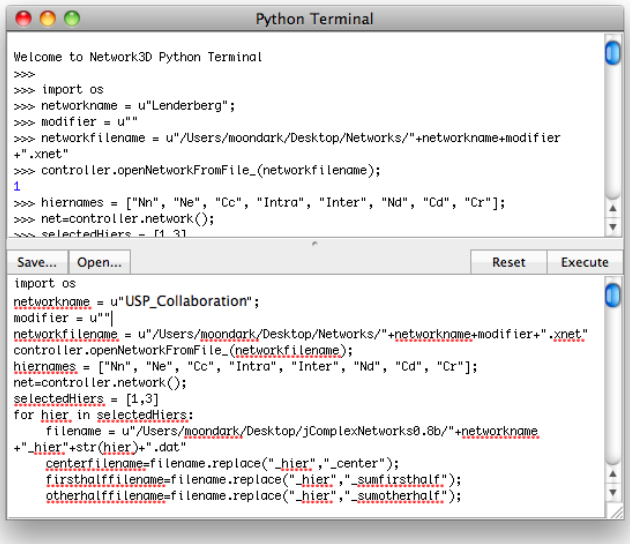


Figura 3.10 – Captura de tela do painel de editor de roteiros Python embutido no software Network3D.

Para divulgação dos resultados há duas formas de distribuição, a primeira é através de figuras, ideal para artigos científicos, permitindo que as imagens das redes complexas sejam exportadas em alta qualidade e em diferentes formatos, incluindo formatos vetoriais como *PDF* e *SVG*. O segundo método de distribuição é para a WEB, o software pode gerar uma visualização interativa que roda dentro dos navegadores da web, e pode ser publicado em um website. A visualização interativa para a web utiliza a tecnologia Java, permitindo que o mesmo código rode em qualquer arquitetura ou plataforma.

A versão para a web também traz algumas opções, como buscas interativas e um recurso para mostrar, ao lado da rede, informações mais detalhadas, fornecidas por algum website da WWW, sobre o vértice selecionado, como mostra o exemplo da figura 3.11.

A figura 3.12 mostra algumas imagens geradas pelo software Network3D.

3.3 Obtenção de Redes Complexas

Para este trabalho, foram usadas algumas redes de modelos teóricos e outras obtidas de dados reais. Enquanto as primeiras podem ser geradas com algoritmos pré determinados, fornecendo alguns parâmetros, as outras apresentam um novo problema a cada rede, já que diferentes tipos de relações podem ser extraídas de diferentes tipos de dados.

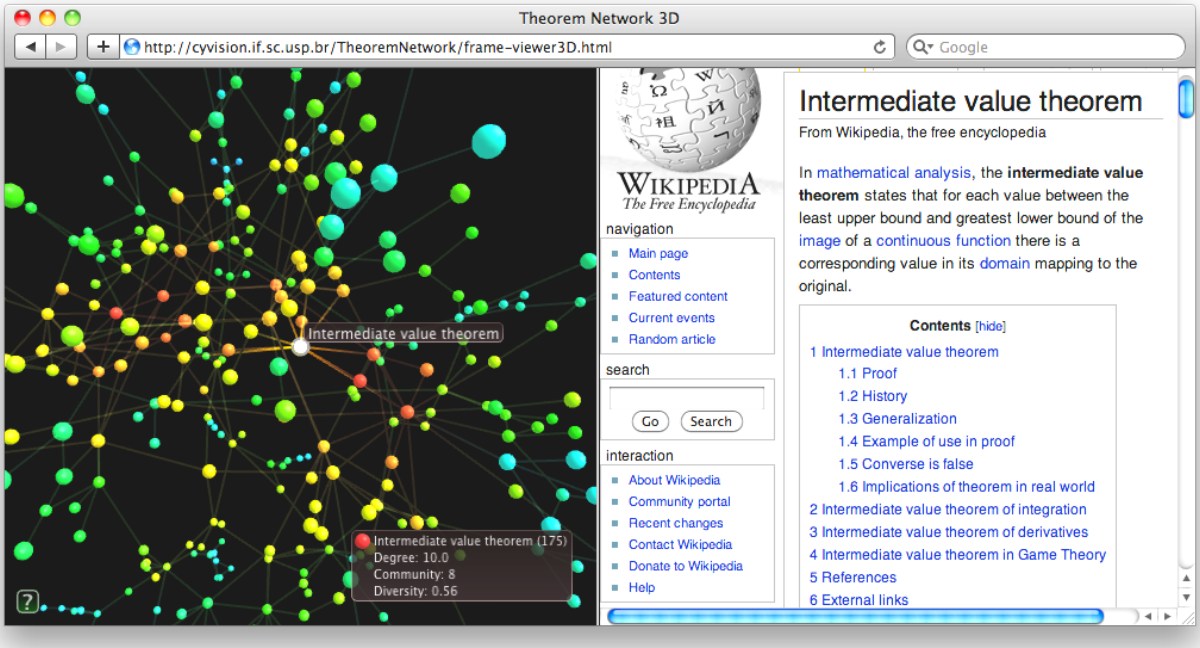
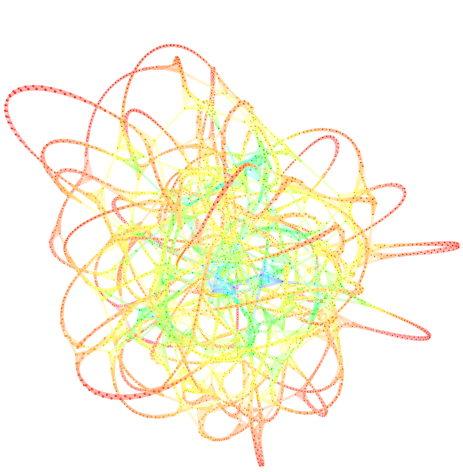


Figura 3.11 – Visualizador 3D de redes complexas gerado para a web com o software Network3D.

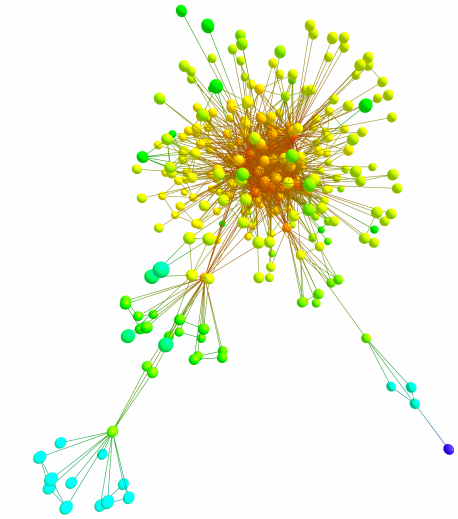
Há sistemas que intrinsecamente apresentam estruturas de relações entre elementos, facilitando a extração de redes complexas, como no caso de páginas de internet, onde cada página pode ser reduzida a um vértice e cada ponteiro a uma aresta. Em contrapartida, grande parte dos sistemas não apresentam essas relações claramente, ou ainda, apresentam relações que exigem parâmetros seletores, como no caso de uma rede social, fica claro que cada pessoa é um vértice, mas que tipo de arestas devem ser levadas em conta? Aquelas que ligam pessoas que se conhecem ou aquelas que são amigas entre si? A resposta para esta pergunta é que ambas as possibilidades são aceitáveis e podem gerar redes complexas independentes com propriedades diferentes, dependendo apenas da interpretação de cada uma.

Em especial, redes que representam o conhecimento ou a pesquisa em determinadas áreas do conhecimento têm sido alvo de estudos, principalmente com modelos de redes de colaboração de trabalhos acadêmicos (65, 66, 67, 68, 69, 70). Essas redes podem ser criadas considerando uma lista de publicações, por exemplo, e seus respectivos autores, de modo que cada vértice corresponde a um autor e cada aresta indica uma publicação escrita por 2 autores (fig. 3.13).

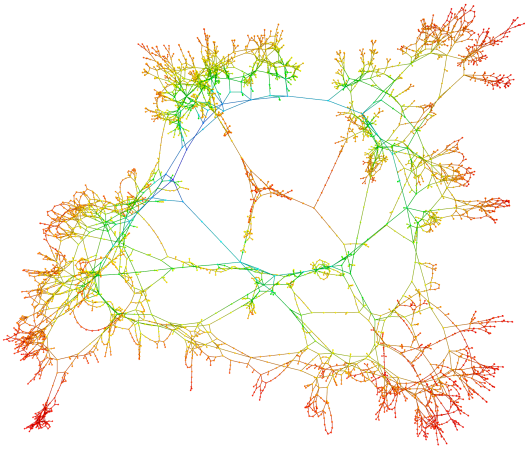
Outra forma de representar conhecimento por redes complexas é através de redes de citações (71, 72, 73, 74), nelas, cada vértice corresponde a um documento e uma aresta existe entre dois documentos quando há uma citação entre eles. Redes desse tipo, são, em geral, direcionadas, no entanto, em algumas situações pode-se reduzi-las a redes não direcionadas. Um exemplo típico



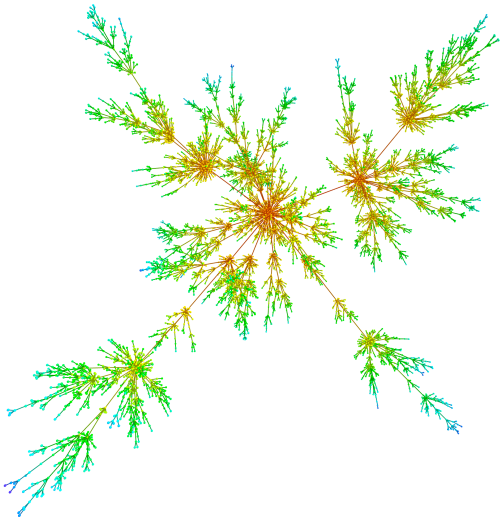
(a) Pequeno mundo



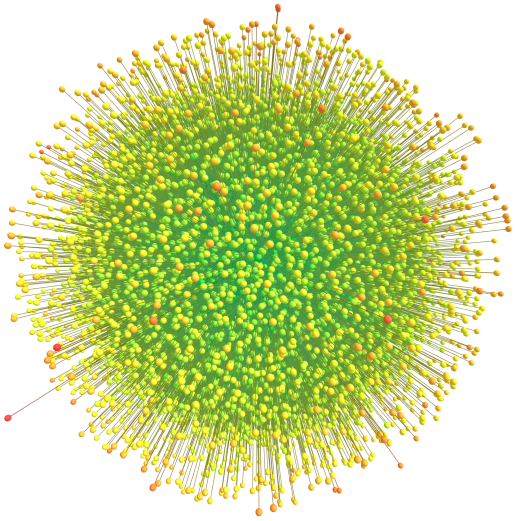
(b) Aeroportos EUA



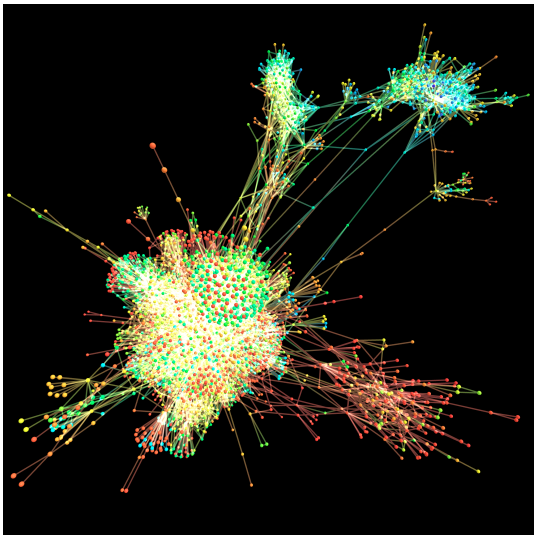
(c) Rede de alta tensão dos EUA



(d) Árvore livre de escala



(e) Associação de palavras



(f) Rede de citações.

Figura 3.12 – Exemplos de imagens obtidas de diversas redes complexas pelo software Network3D.

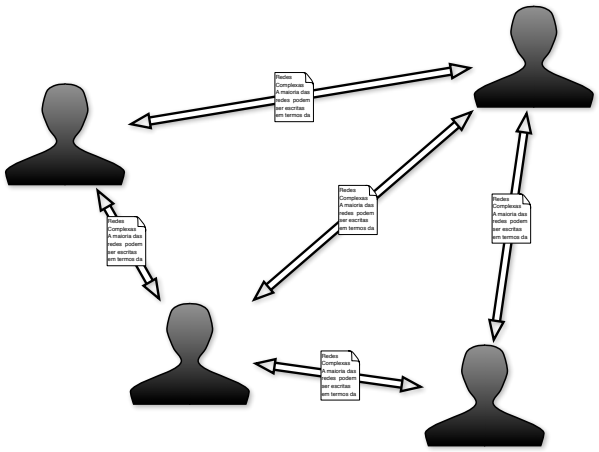


Figura 3.13 – Rede de colaboração, onde cada autor está conectado a outro somente se apresentam algum trabalho em comum.

dessa categoria de redes são redes de citação acadêmicas, onde cada página corresponde a um vértice e as citações a outros artigos às arestas entre eles, sub-redes da WWW (48) também podem ser representadas dessa forma, sendo compostas por páginas e pelos ponteiros entre elas. Diversas redes reais foram usadas, duas delas foram desenvolvidas durante progresso deste trabalho, a rede de colaboração de trabalhos científicos da Universidade de São Paulo e redes baseadas em artigos da Wikipédia.

3.3.1 Rede de colaboração da USP

Em colaboração com o Sistema Integrado de Bibliotecas da Universidade de São Paulo (SIBi-USP), uma rede de colaboração foi obtida a partir do conjunto de dados* correspondente às publicações científicas dos pesquisadores da Universidade de São Paulo (*USP*), cobrindo os anos de 2003 e 2004.

Os dados foram fornecidos em formato de texto estruturado, onde cada linha continha a informação de dois pesquisadores da USP que tivessem colaborado em algum trabalho acadêmico, apresentando os respectivos números funcionais internos e unidades as quais pertenciam. Os dados foram processados por um roteiro escrito em Python que extraiu a rede e as unidades correspondentes. A rede foi filtrada, eliminando as conexões a instituições externas e dados redundantes, assim como pequenos erros existentes nos dados originais.

Cada vértice da rede obtida corresponde a um pesquisador e a ele está associado a unidade

* Agradecimentos a Adriana Cybele Ferrari, Edna Knorich, Marilza A. Rodrigues Tognetti e ao SIBi-USP pelo fornecimento dos dados.

correspondente, cada aresta representa um trabalho que dois pesquisadores colaboraram. Para trabalhos com mais de dois colaboradores, todos os pesquisadores envolvidos foram conectados entre si. Usando a informações das unidades correspondentes a cada pesquisador, foi possível categorizá-los por cidades ou pelas grandes áreas de conhecimento (humanas, biológicas ou exatas), referentes àquela que melhor caracteriza cada unidade.

As representações percentuais correspondentes ao número de vértices de cada unidade, área e cidade presentes na rede, foram obtidas para a rede completa e também para as sub-redes obtidas pela extração do maior componente conectado e para aqueles não conectados a ele.

A visualização computacional da rede foi usada para mapear bidimencionalmente os vértices da rede em diferentes grupos representados por cores. As propriedades clássicas de redes complexas também foram obtidas para a rede de colaboração, assim como a distribuição de grau. A caracterização da interdisciplinaridade para cada grupo é discutida superficialmente.

3.3.2 Redes de Teoremas da Wikipédia

A Wikipêdia* é um projeto sem fins lucrativos que disponibiliza um banco de dados público de artigos informativos, criados por voluntários, em geral, os mesmos que a acessam. Há uma grande quantidade de artigos classificados em categorias, permitindo que uma rede complexa possa ser gerada a partir dos artigos e de seus ponteiros a outros artigos, entretanto, considerando apenas aqueles pertencentes um mesmo conjunto de categorias.

Um roteiro escrito em Python foi criado extrair redes baseadas em categorias da Wikipédia. Em caráter semi-automático foi gerada uma rede complexa (fig. 3.14) a partir dos artigos de lingua inglesa pertencentes a categoria de teoremas matemáticos†.

Devido à natureza da rede, não é possível confiar na direcionalidade das conexões em termos de alguma ordem natural ou cronológica, isto porque a rede é incompleta e pode apresentar erros. Por exemplo, considerando um teorema A, que origina B, com artigos da Wikipédia a e b, respectivamente, se b cita a, mas não há citações de a para b (pelo fato da rede estar incompleta), a implicação de que B originou A não está correta. Para solucionar este problema de modo simples, todas as arestas são consideradas como não direcionadas.

Outras redes baseadas na Wikipédia podem ser obtidas pelo conjunto de ferramentas de-

* http://wikipedia.org
† http://en.wikipedia.org/wiki/Category:Mathematical_theorems

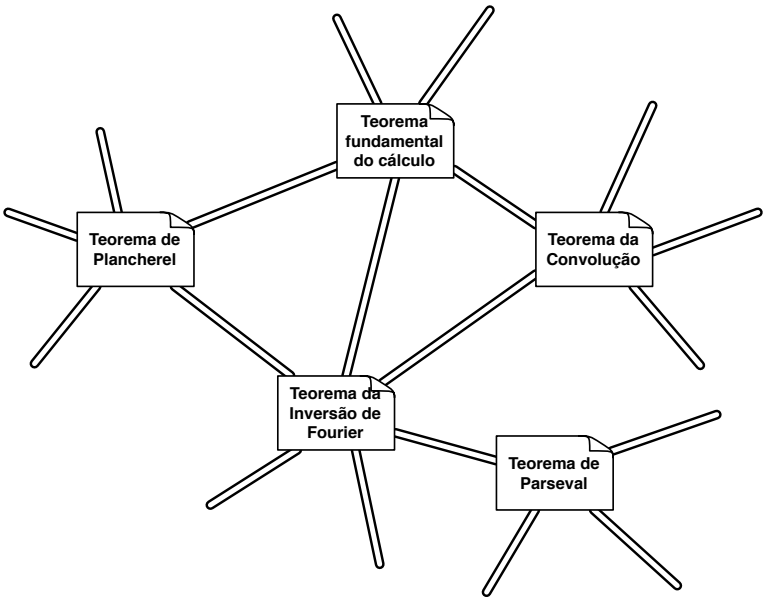


Figura 3.14 – Exemplo de rede baseada em teoremas, onde cada teorema é representado por um vértice e cada aresta uma citação de seus respectivos artigos na Wikipédia.

senvolvidas, no entanto, apenas a rede de teoremas matemáticos é considerada neste trabalho.

A rede de teoremas foi brevemente caracterizada em termos das propriedades clássicas e através da visualização computacional.

3.4 Caracterização Concêntrica de Redes Complexas

Com o objetivo de ilustrar o uso das medidas concêntricas em redes complexas, estas, foram aplicadas a diferentes redes complexas, tanto àquelas obtidas de modelos teóricos quanto às redes reais.

As redes de modelos clássicos, *Barabási-Albert (BA)* (fig. 3.15), *Erdös-Rényry (ER)*(fig. 3.16), *Watts-Strogatz (WS)*(fig. 3.17), geradas de acordo como foram descritas na seção 2.2); e o modelo geográfico(GEO)(fig. 3.18), descrito em (49); delas foram extraídas todas as propriedades concêntricas usando o software *jComplexNetworks*.

O mesmo procedimento aplicado aos modelos teóricos, foi realizado para as seguintes redes reais: de colaboração da USP (*USPCollaboration*) descrita na subseção 3.3.1; rede de teoremas da Wikipédia (*WikiTheorems*), descrita na subseção 3.3.2; rede de aeroportos dos EUA em

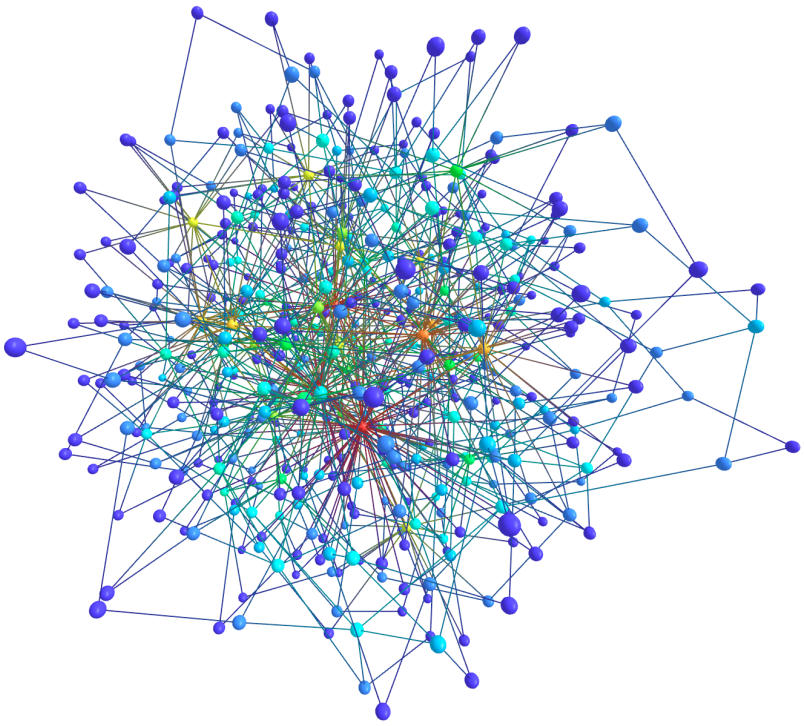


Figura 3.15 – Exemplo de rede Barabási-Albert usada para a comparação de resultados das medidas concêntricas. É possível observar a existência de alguns poucos hubs na região interna da rede.

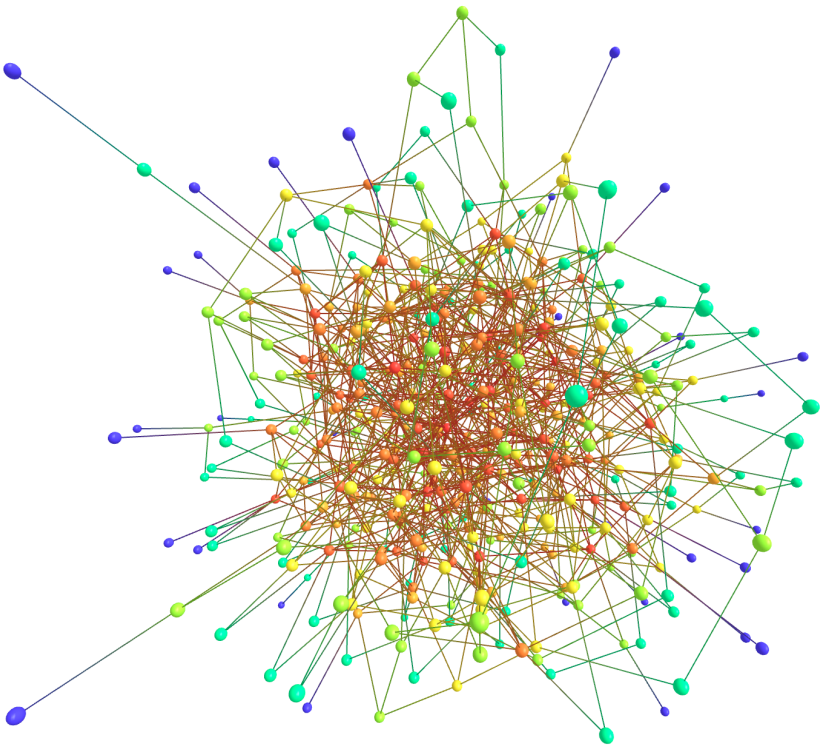


Figura 3.16 – Exemplo de rede Erdös-Rényry usada para a comparação de resultados das medidas concêntricas. Diferentemente da rede BA, os vértices mais internos apresentam a mesma conectividade.

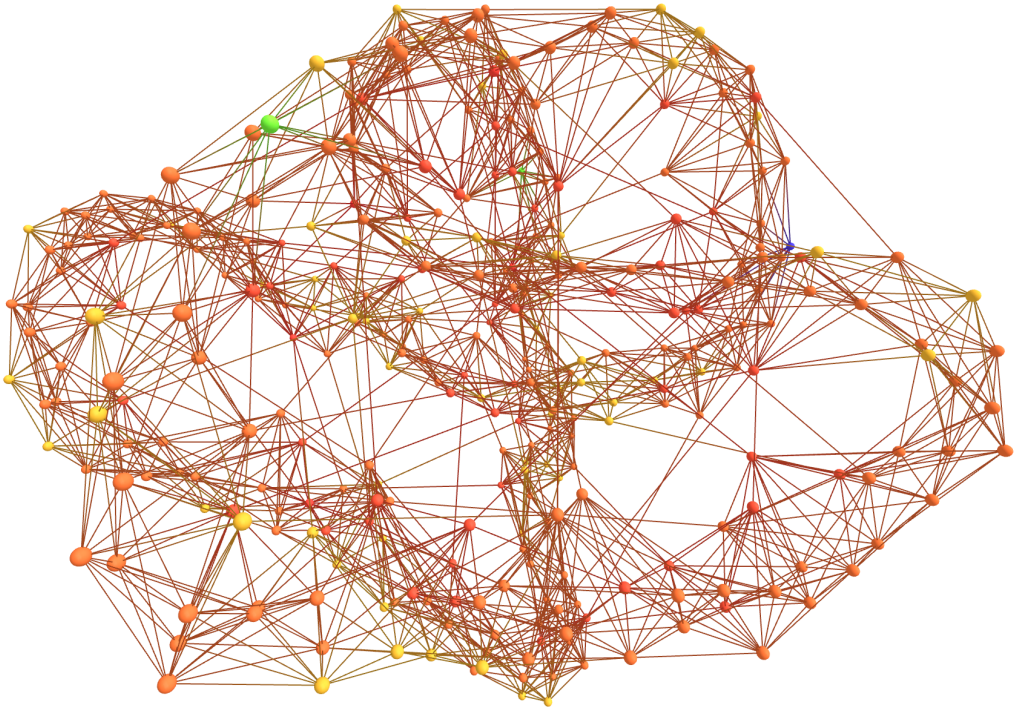


Figura 3.17 – Exemplo de rede Watts-Strogatz usada para a comparação de resultados das medidas concêntricas.

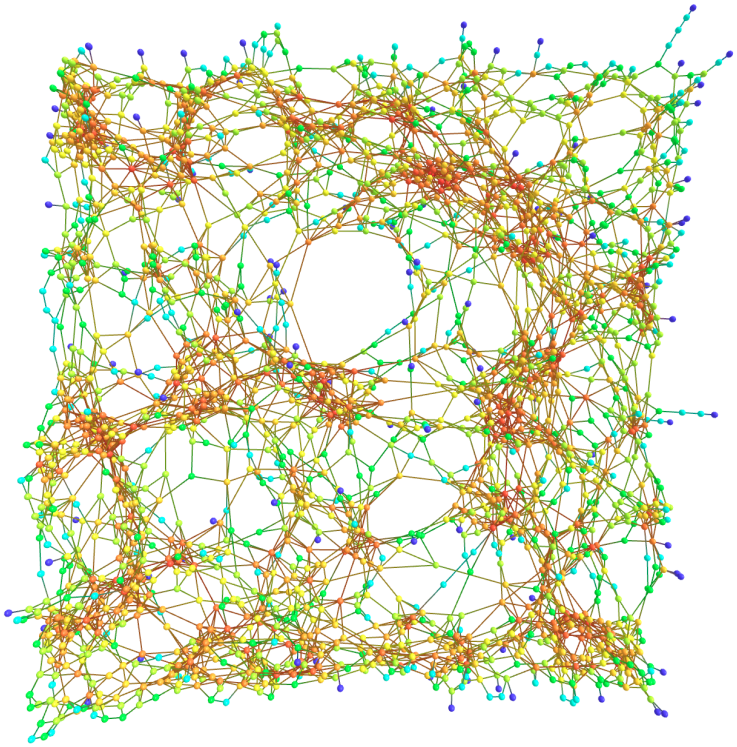


Figura 3.18 – Exemplo de rede geográfica usada para a comparação de resultados das medidas concêntricas.

1997 (*USAir97*, fig. 3.19) (75); rede de associação de palavras desenvolvida em Edinburgh (*EdinburghThesaurus*, fig. 3.20) (76); rede interação proteina-proteina (*Yeast*, fig. 3.21) (77); rede de fiação de alta tensão dos EUA (*USPowerGrid*, fig. 3.22) (15); e sub-rede da www ao pesquisar pela palavra "California" (*California*, fig. 3.23) (78).

A rede de aeroportos USAir97 é uma compilação dos vôos de aeroportos dos EUA em 1997, onde cada nó representa um aeroporto e cada aresta a existência de um vôo entre dois aeroportos, a rede é ponderada pelo número de vôos e tem 332 vértices, com $\langle k_{\text{topológico}} \rangle \simeq 6$ e $\langle k \rangle \simeq 26$, e já apresenta apenas um componente conectado.

A rede de associação de palavras, EdinburghThesaurus, foi criada a partir de um conjunto de dados recolhidos de pessoas, que deviam informar quais palavras vinham às suas mentes após serem apresentadas a uma palavra de estimulo. Cada palavra corresponde a um vértice da rede e suas arestas a uma relação de associação entre as palavras de estimulo e a respectiva resposta. O procedimento detalhado da criação da rede pode ser visto em (76). Esta rede apresenta 23219 vértices com $\langle k \rangle \simeq 14$, a rede possui apenas um componente conectado.

A rede de interação de proteínas, Yeast, descrita em (77), é composta por vértices representando uma proteína e arestas, interações entre duas delas. A rede apresenta 2361 vértices, mas seu maior componente conectado é composto por 2224 vértices e $\langle k \rangle \simeq 6$, os outros componentes conectados são muito pequenos e, portanto, são irrelavantes.

A rede de fios de alta tensão da região oeste dos EUA, USPowerGrid, usada em (15), é composta por um único componente conectado com 4941, vértices representando as estações ou pontos de bifurcação da rede, e $\langle k \rangle \simeq 2.7$.

A sub-rede da internet, California, obtida em (78), foi criada a partir de uma busca pelo termo "California" em paginas da WWW, cada vértice representa um website e cada aresta um ponteiro entre eles. A rede original possui 9664 vértices, no entanto, apenas 5925 compõem o maior componente conectado, com $\langle k \rangle \simeq 5.4$.

A caracterização das redes complexas pelas propriedades concêntricas baseia-se em determinar suas respectivas distribuições médias ao longo dos níveis concêntricos, tomadas com relação a todos os vértices de cada rede como referência. Os resultados são apresentados na forma dos gráficos dessas distribuições, as curvas para cada propriedade são comparadas entre os redes baseadas em modelos teóricos e as redes reais.

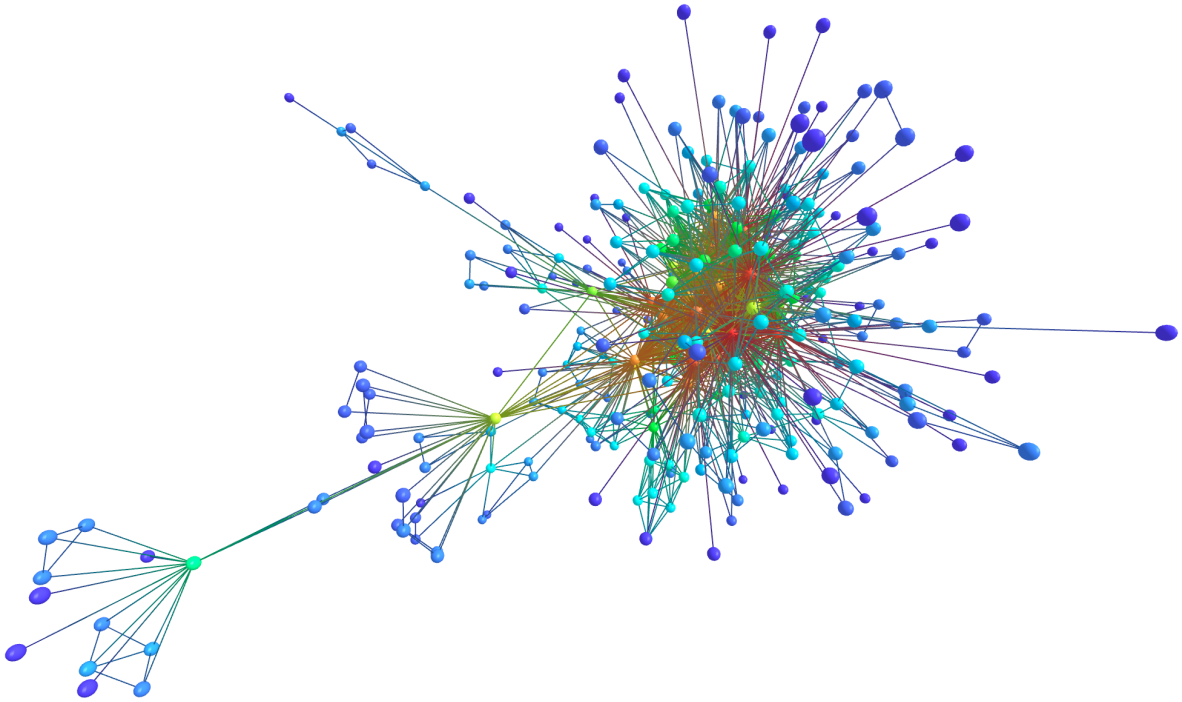


Figura 3.19 – Rede de aeroportos dos EUA, os vértices do ramo situado à esquerda inferior representam os aeroportos do Alasca.

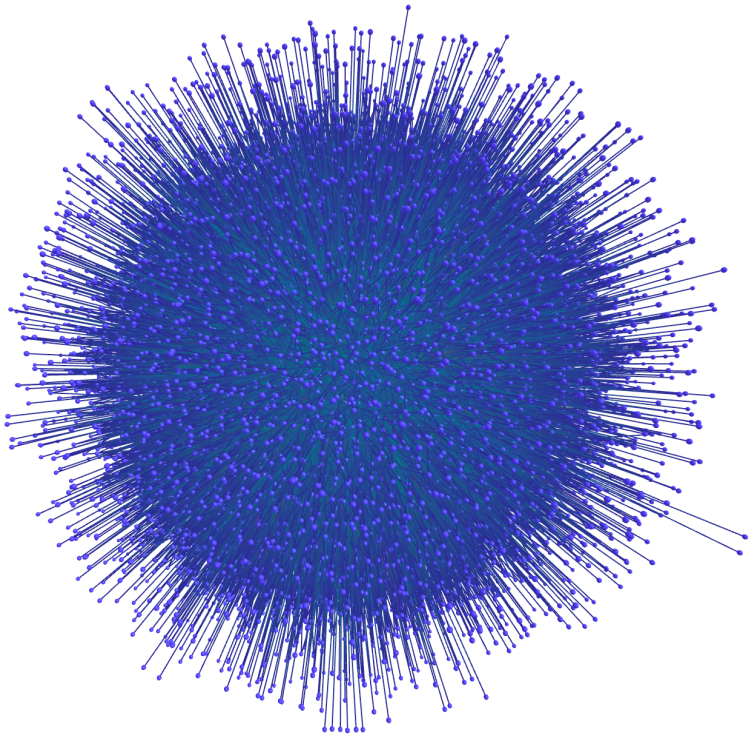


Figura 3.20 – Rede de associação de palavras, Edinburgh.

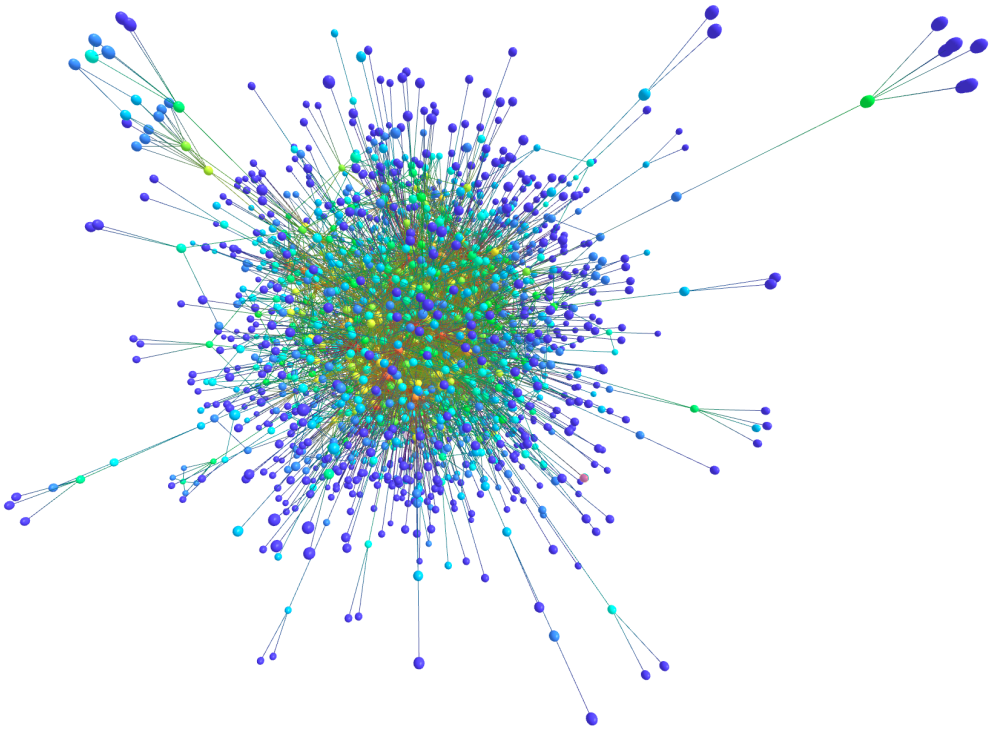


Figura 3.21 – Rede de interação de proteínas, Yeast.

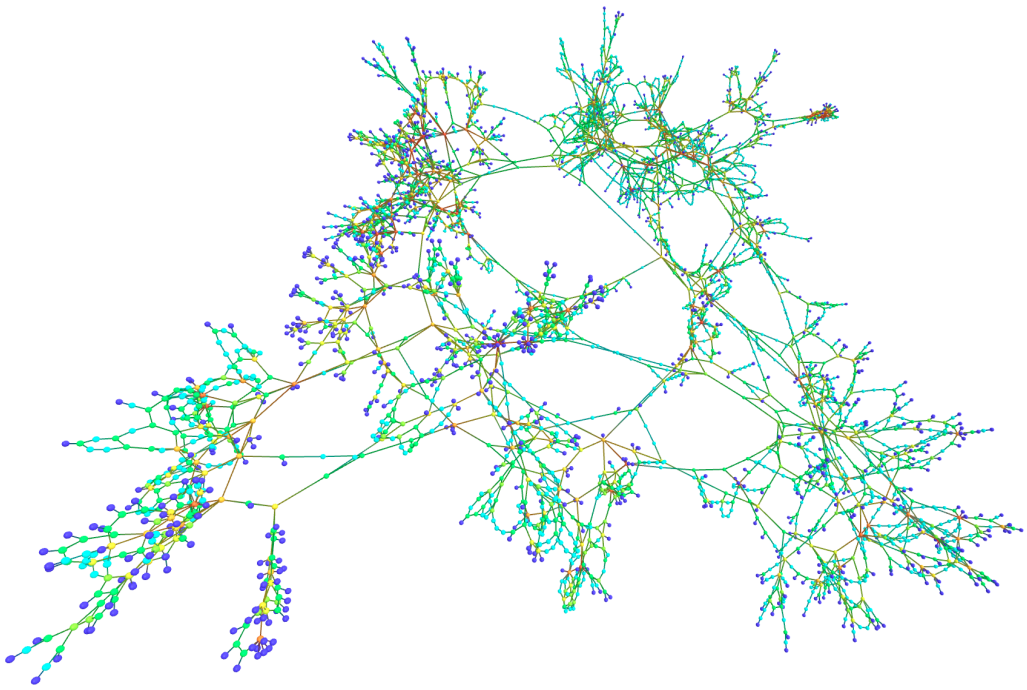


Figura 3.22 – Rede da fiação de energia elétrica de alta tensão dos EUA, nota-se a característica geográfica da rede.

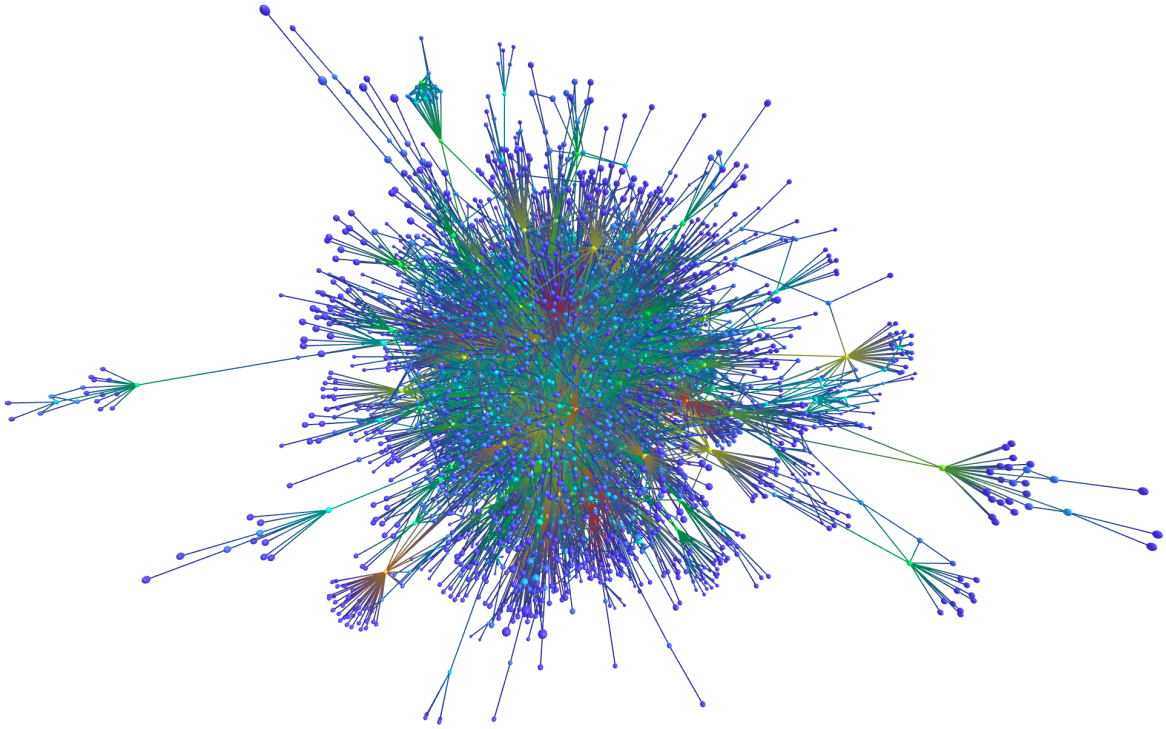


Figura 3.23 – Sub-Rede da WWW com resultados da busca por ”California”.

O poder de classificação das redes pelas medidas concêntricas redes também foi explorado por PCA para a rede de aeroportos e para a rede de proteínas, através do uso do software ClassificationToolKit.

Outro ponto estudado foi a possibilidade de classificação dos vértices das redes pelas medidas hierárquicas através da extração de características (*features*), como o centro da distribuição das propriedades em termos nos níveis concêntricos; assim como por metodologias de aglomeração hierárquica (*hierarchical clustering*) (57). A aglomeração hierárquica baseia-se em extrair uma distância entre os vetores das variáveis de um vértice (distribuição das medidas concêntricas, por exemplo), como a distância euclidiana ou coeficiente de correlação, e posteriormente os dados são aglomerados em grupos com medidas semelhantes formando um dendrograma (ou árvore) da qual é possível extrair quantas classes forem necessárias. Essa metodologia foi aplicada à rede de colaboração da USP, e os resultados foram comparados às categorias inerentes da rede: unidades e áreas do conhecimento.

3.5 Modelagem de Aquisição de conhecimento

A dinâmica de aquisição de conhecimento pode ser modelada em termos de redes complexas por agentes que caminham em uma *rede semântica*, ou de conhecimentos. Redes semânticas são compostas por vértices que representam algum conceito ou conhecimento, enquanto as arestas, alguma relação entre eles. A rede de teoremas da Wikipédia, descrita em 3.3.2, também pode ser considerada como uma rede semântica, onde cada teorema é considerado como um conhecimento da área de matemática.

Um agente é um elemento autômato que caminha pelos vértices de uma rede complexa segundo alguma heurística. Cada agente pode estar somente sobre um único vértice e pode movimentar-se apenas entre pares de vértices ligados por uma aresta. Agentes também podem apresentar diferentes propriedades ou adquirir dados sobre a rede conforme caminham.

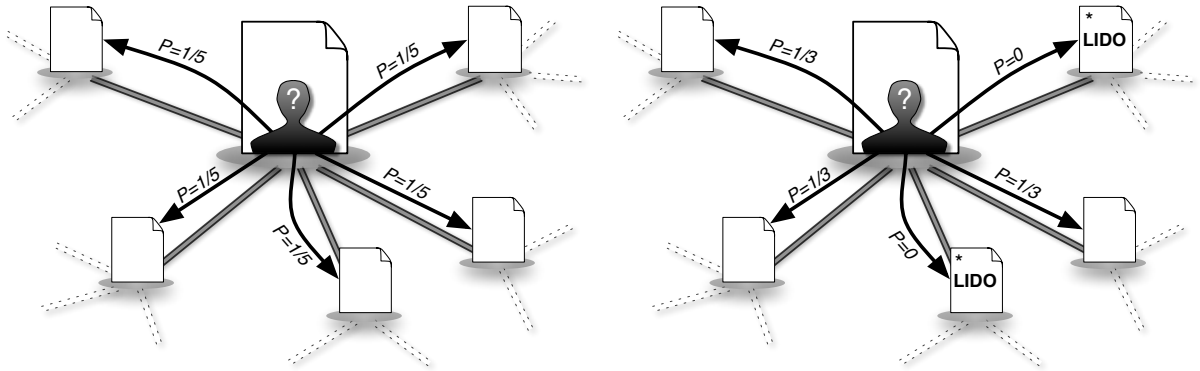
Dada a situação atual de um agente, o próximo vértice que ele visitará dependerá da heurística de escolha, que tradicionalmente pode ser *aleatória* ou *aleatória auto-esquivante* (self-avoiding). Considerando que um agente está sobre o vértice v, que possuí N vizinhos enumerados v_n, com $n \leq N$, a escolha aleatória depende apenas dos vértices vizinhos, que são escolhidos arbitrariamente com iguais probabilidades $P_v(v_n) = 1/N$; já a dinâmica de agentes auto-esquivantes baseia-se em privilegiar àqueles que ainda não foram visitados pelo agente, com probabilidade de escolha definida por:

$$P_v(v_n) = \begin{cases} 1/N^* & \text{se } v_n \text{ não foi visitado} \\ 0 & \text{se } v_n \text{ foi visitado} \end{cases} \tag{3.4}$$

onde N^* é o número de vértices ainda não visitados adjacentes a v, se todos eles já foram visitados, isto é, $N^* = 0$, então a escolha é considerada como a da caminhada aleatória. A figura 3.25 ilustra as diferentes heurísticas de caminhada.

Para este trabalho as heurísticas tradicionais de caminhada foram modificadas em termos de um número finito de vértices que um agente pode guardar. Cada agente possuí uma memória que permite guardar um certo número M de vértices que o agente já visitou, a dinâmica de escolha do próximo vértice é semelhante à auto-esquivante, entretanto depende daqueles já visitados que estão gravados na memória do agente.

A memória de um agente funciona como uma fila, quando novos vértices são visitados, eles são automaticamente adicionados à memória, enquanto aqueles já visitados e mais antigos na memória são ”esquecidos” quando a contagem de vértices guardados ultrapassa M. Nota-se que



(a) Caminhada aleatória: todos os vértices vizinhos ao vértice atual têm mesma probabilidade de serem os próximos a serem visitados.

(b) Caminhada aleatória auto-esquivante: vértices visitados (LIDOS) são evitados enquanto os não visitados possuem as mesmas chances de serem visitados.

Figura 3.24 – Ilustração das heurísticas tradicionais de caminhada dos agentes em uma rede complexa indicando as probabilidades para a escolha do próximo vértice a ser visitado.

para valores muito grandes da memória, o agente deve se comportar como um auto-esquivante, enquanto que para $M = 0$, a dinâmica se reduz à aleatória simples. A figura 3.25 ilustra a heurística.

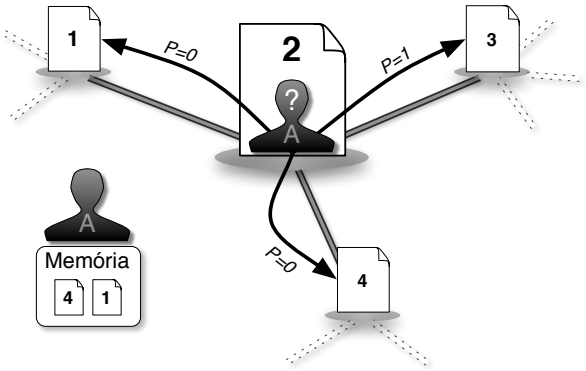


Figura 3.25 – Dinâmica de agente com memória, como os conceitos 1 e 4 já foram explorados pelo agente, apenas o conceito 3 será visitado no próximo passo. Se o tamanho da memória for $M = 2$, ao caminhar para 3, o agente esquecerá que já visitou o vértice 4.

Agentes também podem se comunicar entre si, compartilhando a informação que possuem na memória sobre os vértices já visitados, evitando que outros agentes estudem o que já foi estudado. Em trabalhos anteriores, o compartilhamento de informação era direto ou através de uma rede trivial como uma rede regular, entretanto é mais realista simular essa dinâmica através de uma rede complexa, como os modelos teóricos. O compartilhamento de informação tem como objetivo simular uma conversa ou colaboração entre dois pesquisadores, e, portanto, é normal que eles se comuniquem com aqueles mais próximos, seja devido a amizade, área acadêmica, coordenação, etc.

Dois agentes podem trocar informação entre si somente se houver uma aresta conectando-os na rede de interação. A cada estágio de caminhada, os agentes escolhem aleatoriamente, com mesma probabilidade, um conhecimento contido em sua memória e o transfere para seus vizinhos.

A simulação de múltiplos agentes pode ser caracterizada por uma rede de interação dos próprios agentes e por uma rede de conhecimento a ser analisada, enquanto que a aquisição de conhecimento é medida em termos do número de vértices visitados (simulando os conhecimentos adquiridos pelo grupo completo de agentes) a cada passo (ou iteração) de navegação dos agentes, isto é, vértices por iterações. A figura 3.26 ilustra a dinâmica de múltiplos agentes por uma rede de interação entre eles e uma rede de semântica.

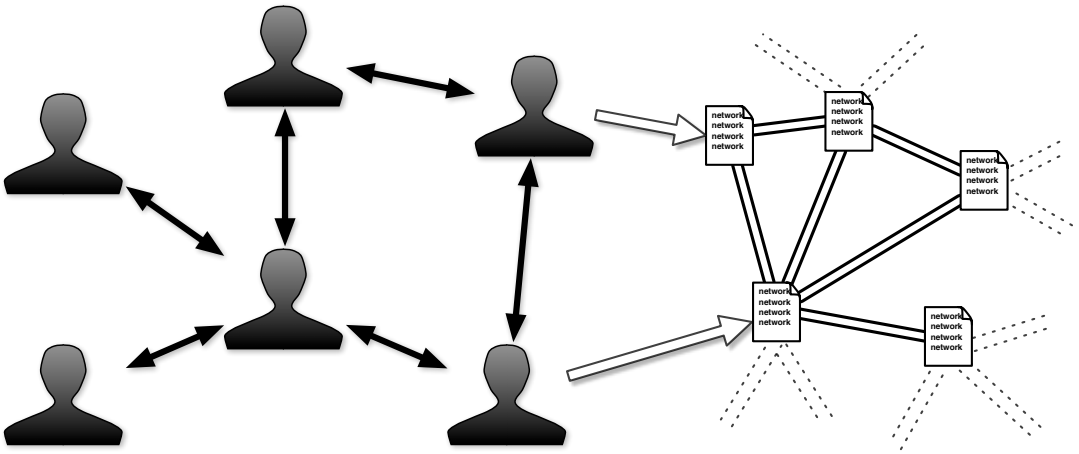


Figura 3.26 – Simulação de múltiplos agentes interagentes, à esquerda encontra-se a rede de colaboração dos agentes e à direita a rede de conhecimento por onde eles caminham.

Há ainda a possibilidade de erros na aquisição de conhecimento, ao caminhar por um vértice um agente pode erroneamente achar que adquiriu outro conhecimento, diferente daquele que realmente foi adquirido. A qualidade de um agente, com relação a susceptibilidade a erros, pode ser quantificado em termos de uma probabilidade $P_E \geq 0$ de que o vértice irá, a cada passo, adicionar à sua memória um vértice que não foi visitado por ele.

Foram estudados diferentes modelos de redes complexas usadas como redes de conhecimento e comparados à rede da teoremas da Wikipédia. Diferentes configurações de agentes foram usadas, com diferentes valores de memória e susceptibilidade a erros, duas redes de interação foram usadas: Barabási-Albert e Erdős-Rényi; com diferentes valores de grau.

Os resultados são analisados através da performance de aquisição do conhecimento e da escalabilidade com o crescimento do número de agentes. Para isso gráficos da performance contra o número de agentes foram usados. A informação sobre a frequência de acesso dos

vértices também foi analisada, embora superficialmente, e seu potencial para a caracterização dos vértices foi explorado através do uso de visualização computacional.

4 Resultados e Discussões

4.1 Rede de colaboração da USP

A aplicação da metodologia descrita em 3.3.1, para a geração de uma rede de colaboração da USP, resultou em uma rede com 5630 vértices e $\langle k \rangle \simeq 15$. Após a eliminação de vértices redundantes, erros e ligações externas, a rede reduziu-se a 3804 vértices e $\langle k \rangle \simeq 5$. O maior componente conectado foi extraído, obtendo-se uma rede complexa com 2864 vértices e $\langle k \rangle \simeq 5.2$. O segundo maior componente também foi extraído, com 62 vértices e $\langle k \rangle \simeq 7.2$. Os outros componentes menores apresentavam quantidade de vértices irrelevantes (< 10 vértices).

A tabela 4.1 apresenta as quantidades percentuais de vértices de cada unidade da rede e das sub-redes do maior componente conectado e dos componentes não conectados a ele, assim como as respectivas cidades de origem e grandes áreas do conhecimento, enumeradas Biológicas, Exatas e Humanas. A tabela 4.2 mostra as representações percentuais de cada área e cidades correspondentes aos vértices para as mesmas sub-redes consideradas anteriormente. Os resultados estão resumidos nos gráficos de pizza da figura 4.1.

Verifica-se que as unidades de maior representação percentual no maior componente conectado correspondem àquelas da área de biológicas, com 67%. Considerando as cidades, São Paulo apresenta a maior quantidade de vértices no maior componente conectado, com 56.7% dos vértices.

Os componentes desconectados apresentaram grande percentual de unidades da área de exatas, 44%, significantemente maior com relação a do maior componente conectado, 23%. A área de humanas também apresentou considerável variação da porcentagem dos vértices dos componentes desconectados, com 27%, e do maior componente conectado, 19%.

A quantidade elevada de vértices em componentes desconectados presentes na área de exa-

Tabela 4.1 – Unidades que compõem a rede de colaboração da USP e suas respectivas representações percentuais considerando a rede completa (Total), o maior componente conectado (Maior Componente) e aquelas que não estão conectadas a ele (Desconectados).

Instituto	Área	Cidade	Total %	Maior Componente %	Desconectados %
FM	Biológicas	São Paulo	9.1%	9.7%	7.0%
FMRP	Biológicas	Ribeirão Preto	8.3%	10.5%	1.8%
EP	Exatas	São Paulo	6.9%	4.2%	15.1%
ESALQ	Biológicas	Piracicaba	5.0%	5.7%	2.9%
ICB	Biológicas	São Paulo	4.1%	5.1%	0.9%
FO	Biológicas	São Paulo	3.7%	4.3%	1.8%
IF	Exatas	São Paulo	3.4%	3.3%	4.0%
FMVZ	Biológicas	São Paulo	3.3%	4.3%	0.3%
EESC	Exatas	São Carlos	3.1%	2.3%	5.2%
IQ	Exatas	São Paulo	3.1%	3.6%	1.5%
FEA	Humanas	São Paulo	2.8%	2.7%	3.2%
FOB	Biológicas	Bauru	2.7%	3.2%	1.2%
FFCLRP	Humanas	Ribeirão Preto	2.7%	3.0%	1.8%
FCF	Biológicas	São Paulo	2.6%	3.3%	0.4%
FCFRP	Biológicas	Ribeirão Preto	2.5%	3.3%	0.2%
FFLCH	Humanas	São Paulo	2.5%	0.9%	7.3%
FORP	Biológicas	Ribeirão Preto	2.5%	3.1%	0.4%
EERP	Biológicas	Ribeirão Preto	2.3%	3.0%	0.1%
IFSC	Exatas	São Carlos	2.2%	2.1%	2.3%
EE	Biológicas	São Paulo	1.9%	2.0%	1.7%
HU	Biológicas	São Paulo	1.9%	1.9%	1.8%
IB	Biológicas	São Paulo	1.9%	1.6%	2.9%
FSP	Biológicas	São Paulo	1.9%	2.0%	1.5%
IAG	Exatas	São Paulo	1.9%	0.8%	5.1%
IME	Exatas	São Paulo	1.7%	0.9%	3.9%
FAU	Humanas	São Paulo	1.6%	1.2%	2.9%
IQSC	Exatas	São Carlos	1.6%	1.9%	0.7%
ICMC	Exatas	São Carlos	1.5%	1.2%	2.7%
FE	Humanas	São Paulo	1.5%	0.8%	3.4%
FZEA	Biológicas	Pirassununga	1.4%	1.8%	0.2%
IGC	Exatas	São Paulo	1.3%	1.2%	1.7%
IO	Biológicas	São Paulo	1.1%	1.2%	0.7%
EEFE	Biológicas	São Paulo	0.9%	0.4%	2.5%
IP	Humanas	São Paulo	0.8%	0.4%	2.1%
CENA	Exatas	Piracicaba	0.8%	1.0%	0.2%
FEARP	Humanas	Ribeirão Preto	0.7%	0.6%	1.1%
HRAC	Biológicas	Bauru	0.6%	0.8%	0.2%
IEE	Exatas	São Paulo	0.6%	0.5%	1.1%
ECA	Humanas	São Paulo	0.6%	0.0%	2.3%
MAE	Humanas	São Paulo	0.3%	0.1%	1.0%
MZ	Biológicas	São Paulo	0.2%	0.1%	0.5%
CBM	Biológicas	São Sebastião	0.2%	0.0%	0.6%
FD	Humanas	São Paulo	0.2%	0.1%	0.4%
MP	Humanas	São Paulo	0.2%	0.1%	0.2%
IEB	Humanas	São Paulo	0.1%	0.0%	0.4%
COSEAS	Humanas	São Paulo	0.1%	0.1%	0.0%
SIBI	Humanas	São Paulo	0.1%	0.0%	0.3%
MAC	Humanas	São Paulo	0.1%	0.0%	0.2%

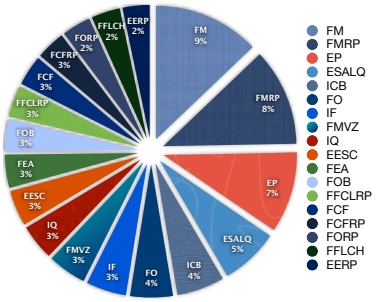
Tabela 4.2 – Contribuições percentuais das diferentes áreas do conhecimento e cidades correspondentes às unidades da rede de colaboração da USP.

Áreas do conhecimento

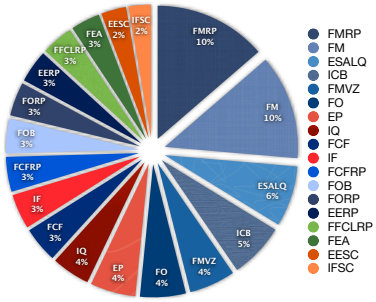
Área	Total %	Maior Componente %	Desconectados %
Biológicas	58%	67%	30%
Exatas	28%	23%	44%
Humanas	14%	19%	27%

Cidades

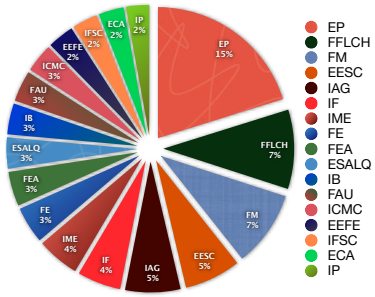
Área	Total %	Maior Componente %	Desconectados %
São Paulo	62.0%	56.7%	78.3%
Ribeirão Preto	19.0%	23.5%	5.4%
São Carlos	8.3%	7.4%	11.0%
Piracicaba	5.8%	6.6%	3.1%
Bauru	3.3%	4.0%	1.4%
Pirassununga	1.4%	1.7%	0.2%
São Sebastião	0.2%	0.0%	0.6%



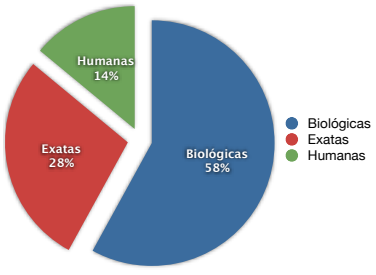
(a) Unidades na rede completa.



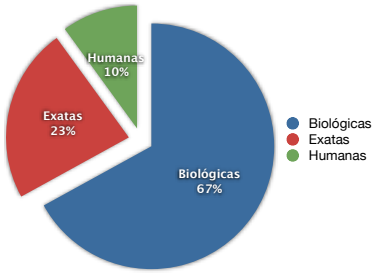
(b) Unidades no maior componente.



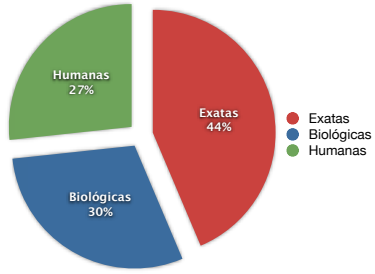
(c) Unidades não conectadas.



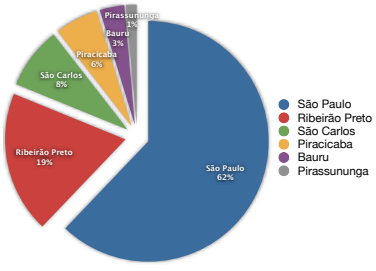
(d) Áreas na rede completa.



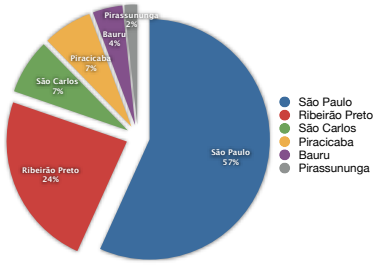
(e) Áreas no maior componente.



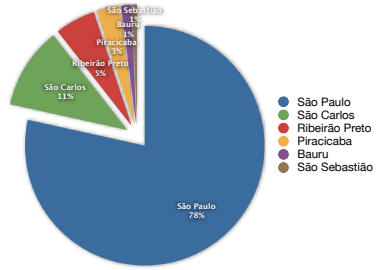
(f) Áreas não conectadas.



(g) Cidades na rede completa.



(h) Cidades no maior componente.



(i) Cidades não conectadas.

Figura 4.1 – Gráficos com as representações percentuais de cada classe dos vértices: unidades, áreas do conhecimento e cidades.

tas e humanas, pode estar refletindo uma situação de colaborações fechadas ou de pesquisas não relacionadas a outras áreas do conhecimento, ao menos para o período considerado da rede. Enquanto que a área de biológicas apresenta alto caráter interdisciplinar, fato que pode ser observado, por exemplo, pelas diferentes relações das pesquisas na área de medicina, que englobam diversos fatores sociais e éticos, característicos da área de humanas, assim como, necessitam da aplicação de metodologias sistemáticas ou utilização de ferramentas e recursos comuns à área de computação ou física.

As cidades também apresentaram significante variação da porcentagem dos vértices quando considerados os componentes desconectados e o maior componente conectado. São Paulo apresentou a maior variação, com uma participação de 78.3% dos componentes desconectados, e 56.7% do maior componente, enquanto as unidades de Ribeirão Preto representam apenas 5.4% e 23.5% respectivamente. Este resultado tem como uma de suas causas a grande quantidade de unidades da área de exatas e humanas no campus da USP de São Paulo, entretanto este comportamento é observável mesmo dentro das unidades "equivalentes"*, como para as unidades FM e FMRP, revelando que a colaboração interdisciplinar tende a ser mais frequentes para a unidade de Ribeirão Preto.

4.1.1 Caracterização pelas propriedades tradicionais

As propriedades tradicionais de redes complexas foram usadas para caracterizar superficialmente a rede de colaboração da USP. Primeiramente foi obtida a distribuição de grau da rede apresentada na figura 4.2, revelando o caráter livre de escala da rede, com coeficiente $\gamma \simeq -2.2$.

O coeficiente de aglomeração médio obtido para o maior componente conectado da rede foi $\langle Cc(i) \rangle = 0.45$ e mínimo caminho médio $l = 8.24$. Devido ao alto valor do coeficiente de aglomeração e baixo valor do mínimo caminho médio, a rede pode ser considerada como pequeno mundo. O segundo maior componente conectado apresentou coeficiente de aglomeração médio $\langle Cc_{2\text{nd}}(i) \rangle = 0.77$ e mínimo caminho médio $l_{2\text{nd}} = 3.3$.

* Equivalentes no sentido de estudarem a mesma sub-área, neste caso, medicina.

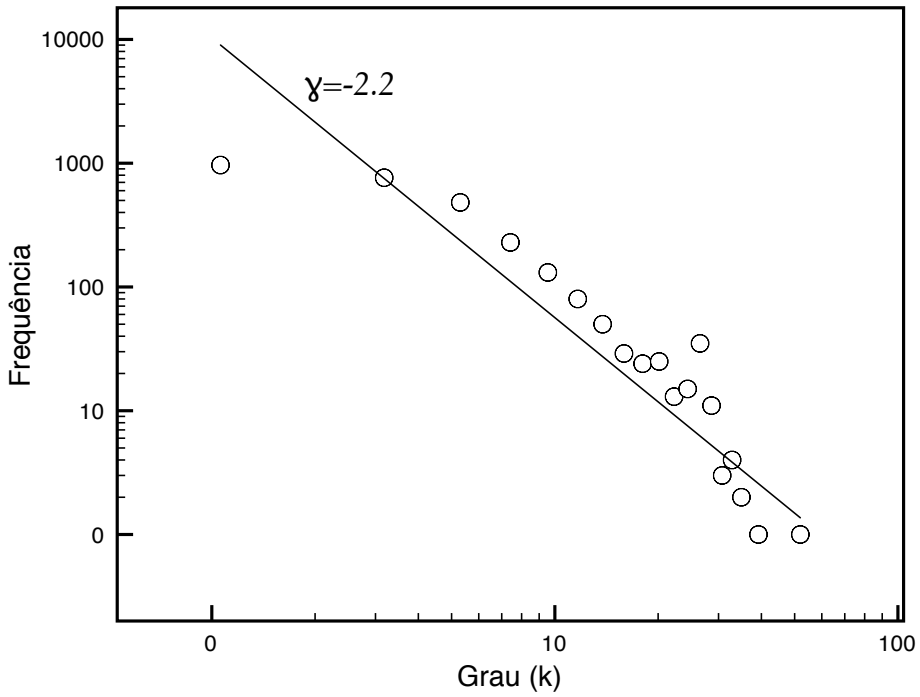


Figura 4.2 – Distribuição de grau da rede de colaboração da USP e a aproximação por lei de potência respectiva, com expoente $\gamma \simeq -2.2$.

4.1.2 Visualização da rede de colaboração da USP

A visualização do maior componente conectado da rede complexa de colaboração da USP, projetada no plano, pode ser vista nas figuras 4.3, 4.4 e 4.5, onde as cores de cada vértice representam respectivamente os valores do grau, do coeficiente de aglomeração e centralidade de proximidade, como mostrados pelas legendas. A rede é composta de um grupo bem conectado central que se expande em diversos ramos, com dois deles extensos, na região superior e inferior à esquerda; e outros dois que formam grupos que se conectam apenas por um pesquisador, localizados na região inferior direita.

A distribuição do grau para a rede de colaboração com relação à sua projeção bidimensional, parece estar sutilmente relacionada com a distância dos vértices até o centro do grande grupo conectado, com exceção para alguns grupos com muitas colaborações entre eles mesmos. Entretanto, a rede não é completamente planar e alguns grupos podem estar sobrepostos devido a projeção bidimensional, contudo, a visualização 3D da rede revelou propriedades semelhantes, e pode ser vista em versão interativa no ponteiro:
`http://cyvision.if.sc.usp.br/~filipi/networks/USP_Collaboration3D`.

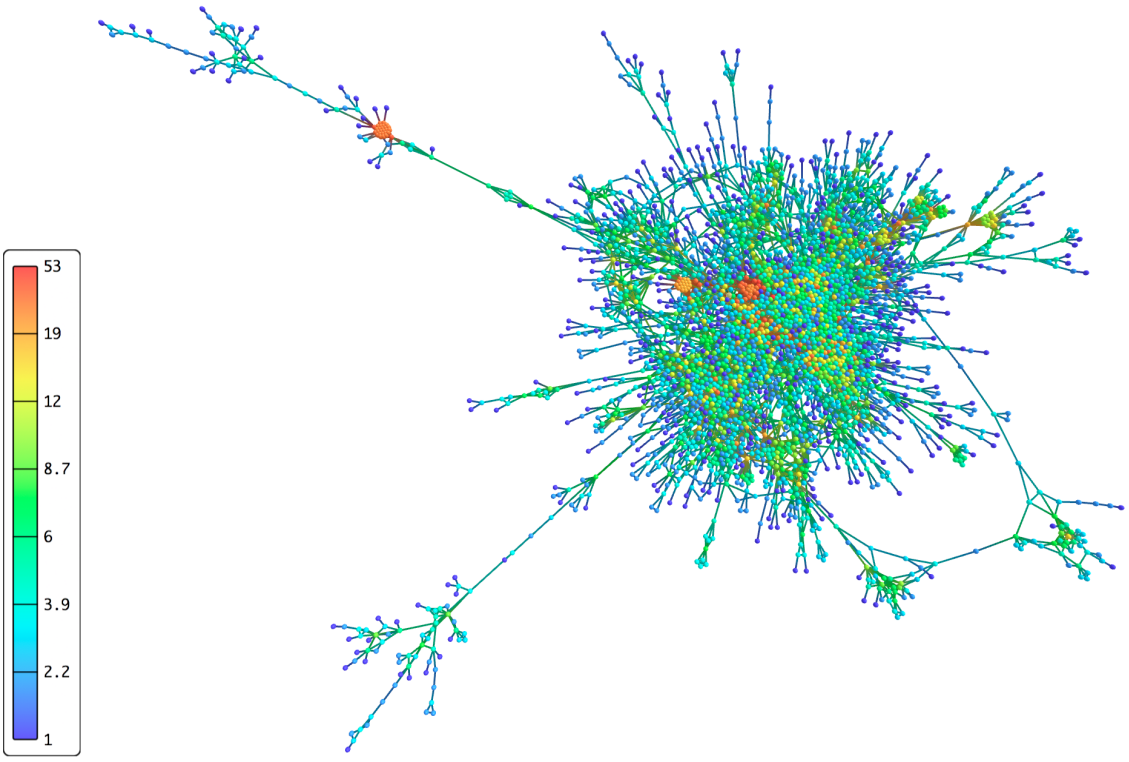


Figura 4.3 – Rede de colaboração da USP com as cores representando o valor do grau de cada vértice.

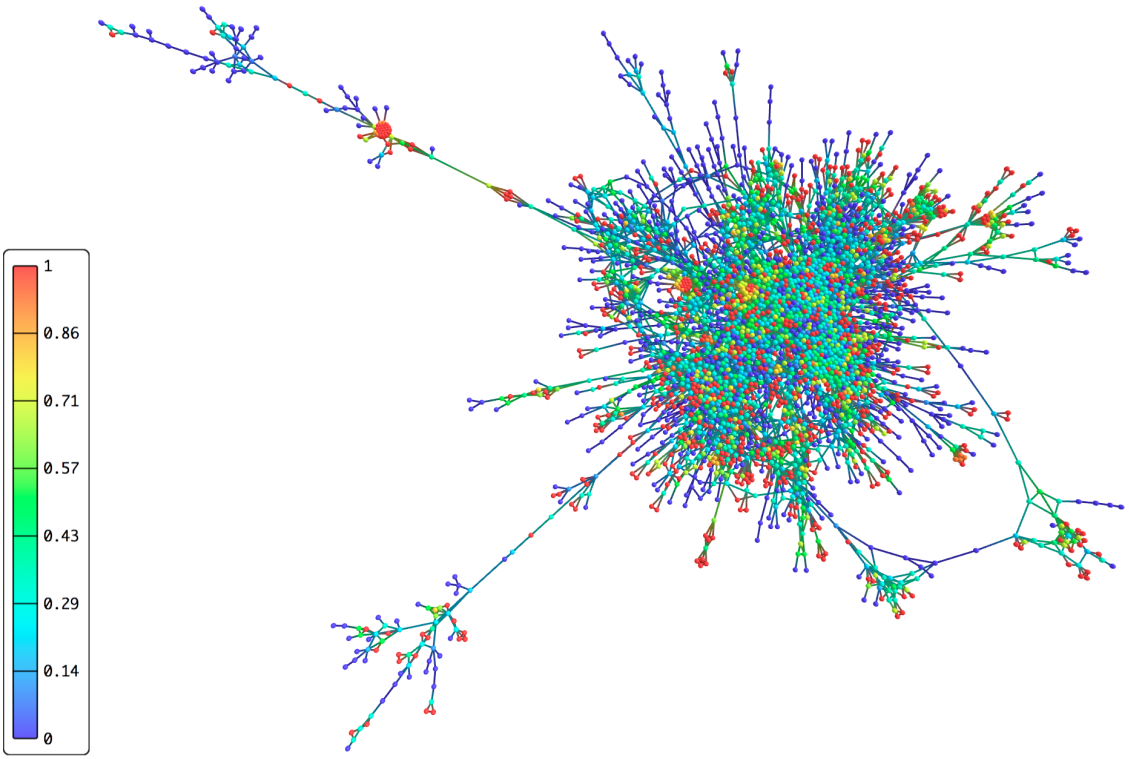


Figura 4.4 – Visualização do coeficiente de aglomeração dos vértices da rede de colaboração da USP.

O coeficiente de aglomeração revelou que existem muitos grupos espalhados pela rede que formam pequenos grupos de colaboração de modo que os pesquisadores tendem a colaborar substancialmente com aqueles com quem seus colaboradores interagem. Este comportamento é

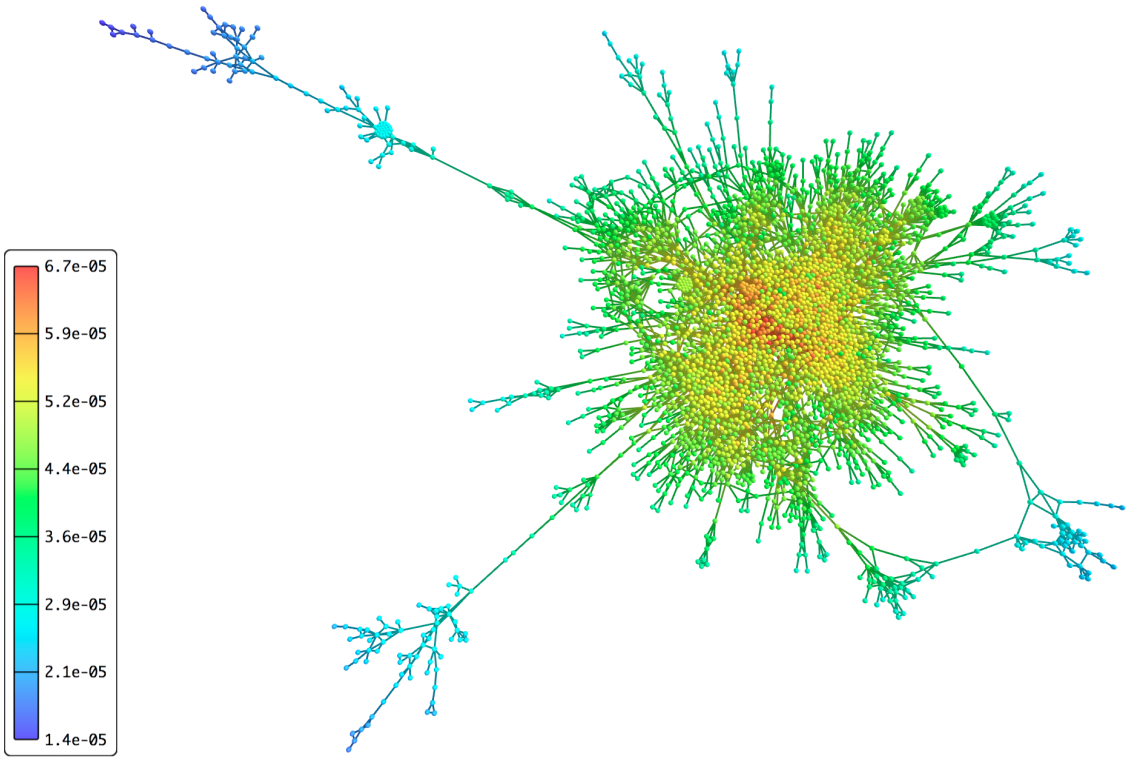


Figura 4.5 – Visualização da centralidade de proximidade obtida para a rede de colaboração da USP.

perfeitamente compreensível em uma rede de colaboração institucional, pois em cada departamento, existem diferentes grupos que pesquisam alguma área específica em comum e colaboram muito entre si. Tal comportamento parece existir mesmo para grupos distantes do centro, e pode ser caracterizado pela distribuição espacial dos pesquisadores com altos valores do coeficiente de aglomeração.

É importante notar que pesquisadores com baixos valores de coeficiente de aglomeração tendem a ser aqueles que conectam diferentes áreas ou grupos de pesquisas, podendo ser caracterizados por alta interdisciplinariedade, entretanto, devido a limitação do coeficiente de aglomeração ser considerado apenas para os vizinhos mais próximos, grupos com características interdisciplinares não podem ser caracterizados pela mesma propriedade.

O centro da rede pode ainda ser caracterizado pelas propriedades de centralidade. Aquela que apresentou melhor resultado visual foi a centralidade de proximidade, determinando os vértices centrais da rede. Entretanto, é importante notar que pode-se observar dois conjuntos de vértices com altos valores de centralidade, isto é, dois centros na rede, entretanto a visualização tridimensional elimina essa redundância, que é causada pelo fato da rede não ser completamente planar.

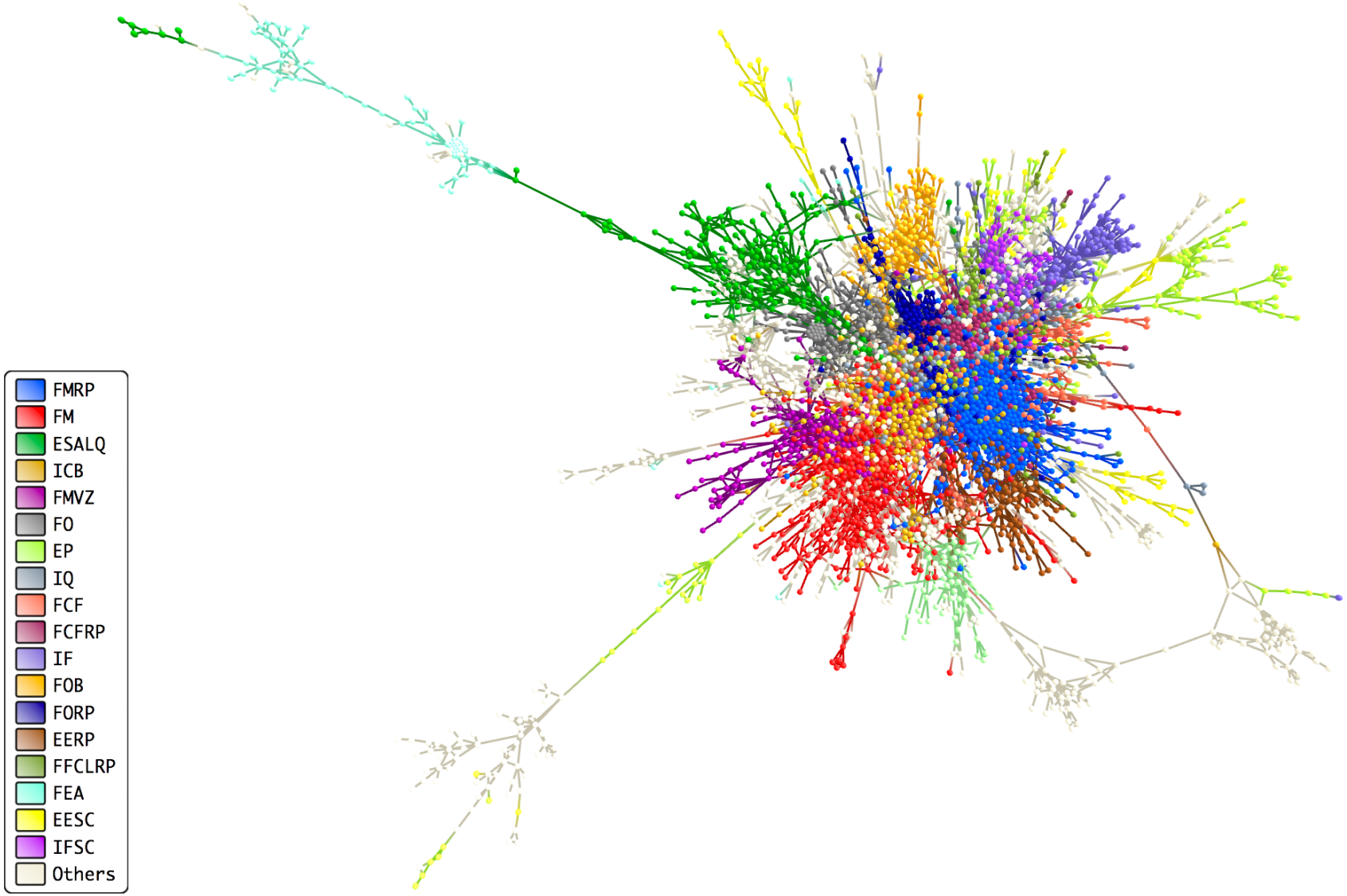


Figura 4.6 – Visualização da rede de colaboração da USP destacando as principais unidades presentes na rede. Para facilitar a visualização apenas as 18 unidades com maior percentual de vértices são apresentadas.

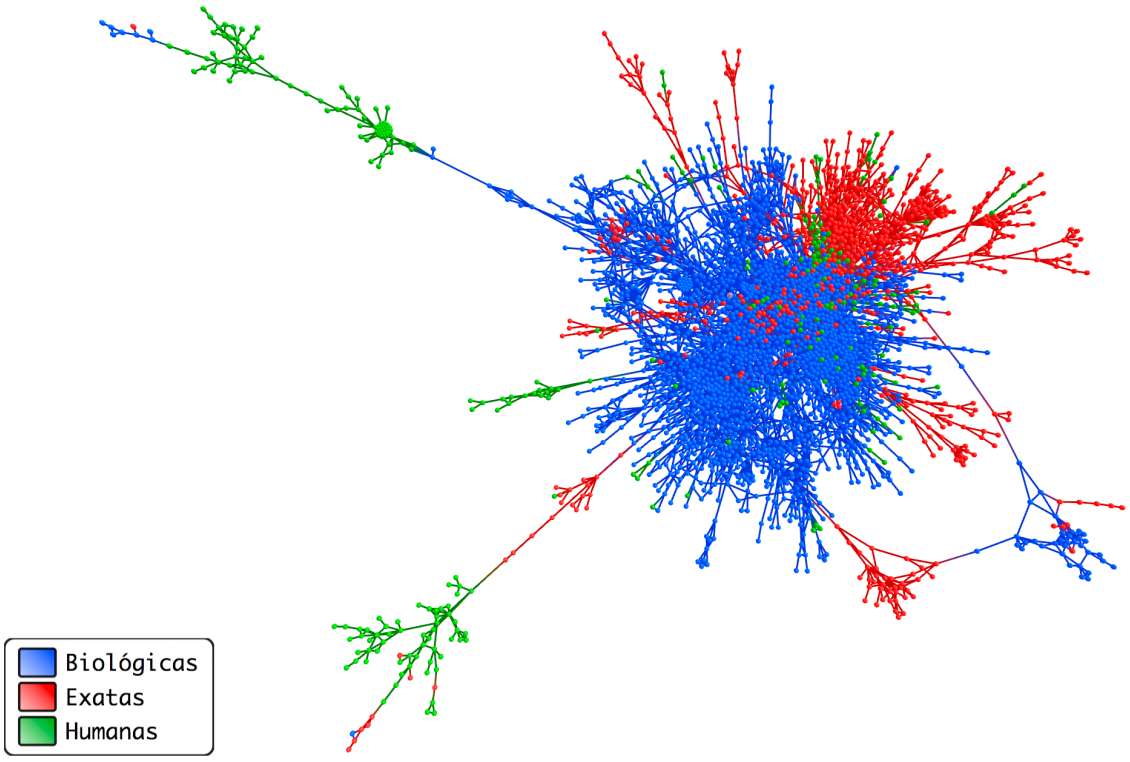


Figura 4.7 – Rede de colaboração da USP destacando as diferentes áreas do conhecimento, na imagem nomeadas exatas, humanas e biológicas.

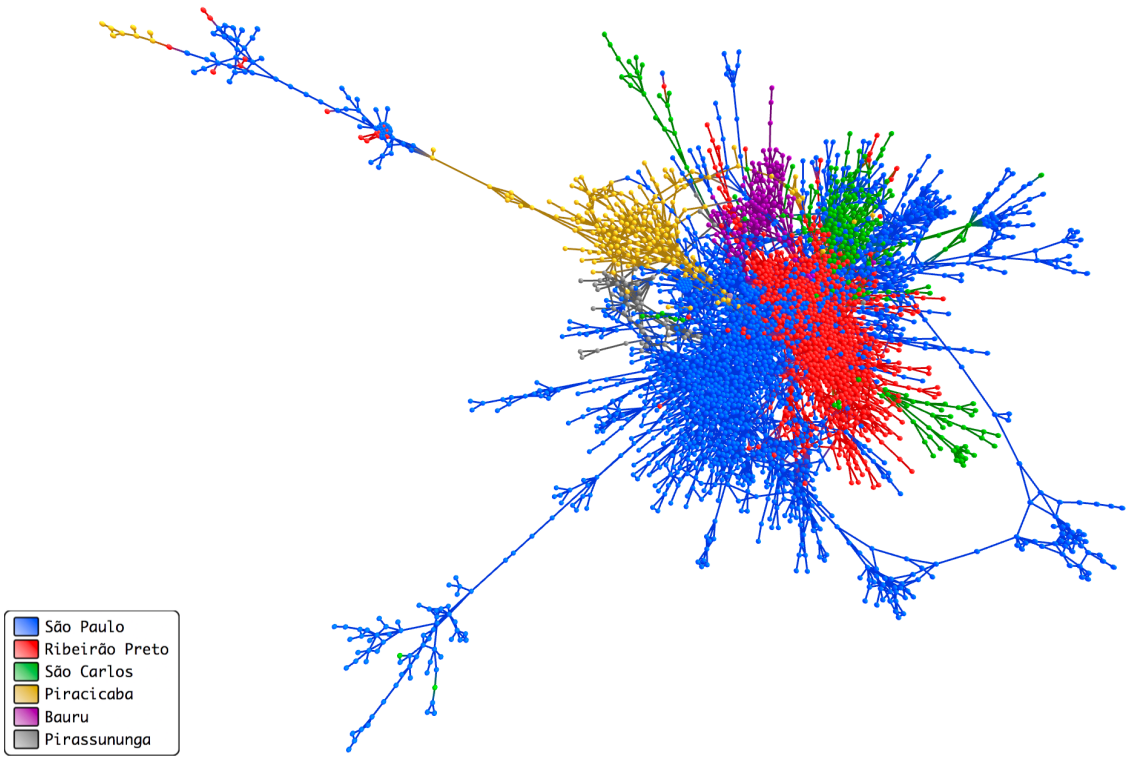
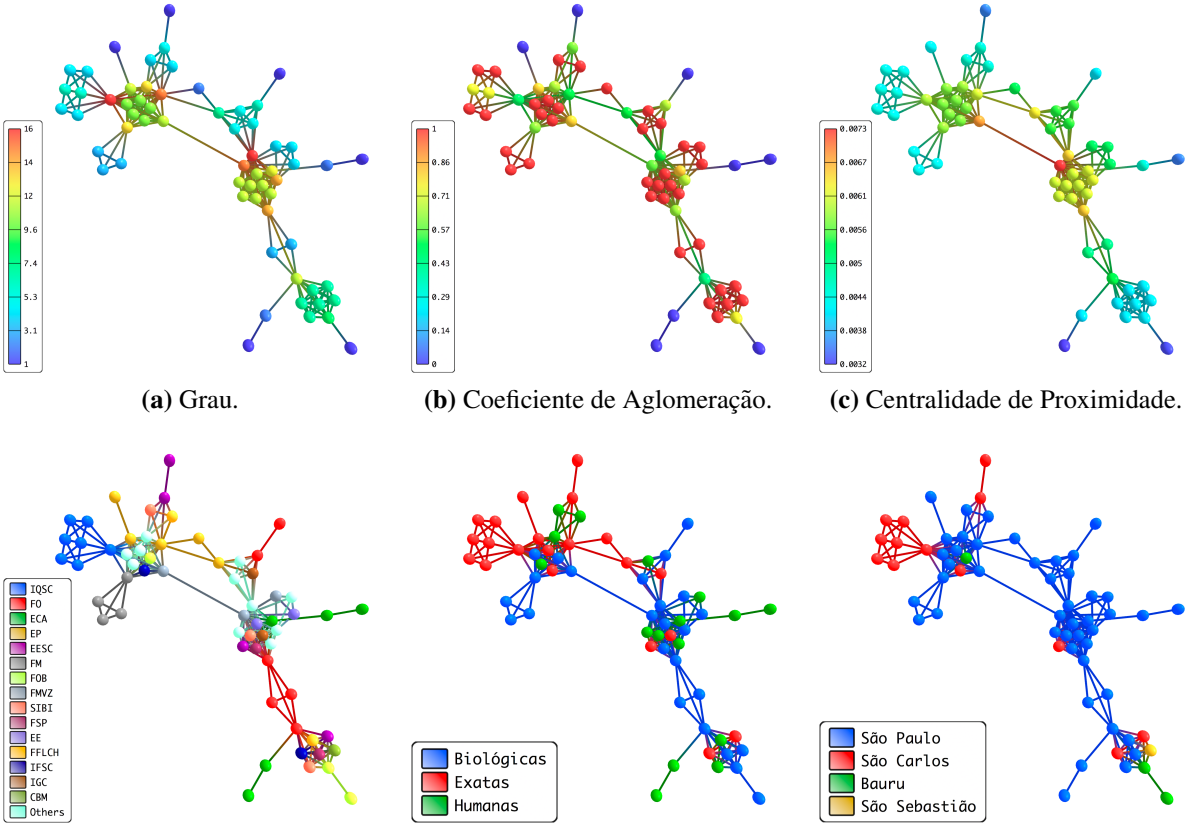


Figura 4.8 – Projeção 2D da rede de colaboração da USP destacando as cidades correspondentes às unidades de cada pesquisador.

As figuras 4.6, 4.7 e 4.8 apresentam a projeção bidimensional da rede destacando as diferentes categorias dos vértices, respectivamente, unidade, área do conhecimento e cidade. Primeiramente, nota-se a correspondência com as porcentagens das categorias obtidas na subseção anterior, e a maioria das classes se estruturam em comunidades relativamente bem definidas.

Pode-se inferir algumas características da interdisciplinaridade, apenas observando a figura 4.6, como o fato da Faculdade de Medicina (FM) estar relativamente menos integrada ao centro quando comparada a Faculdade de Medicina de Ribeirão Preto (FMRP), que, apesar de seus pesquisadores colaborarem muito entre si, possui uma ampla camada de colaboração com diversas outras unidades.



(a) Grau. **(b)** Coeficiente de Aglomeração. **(c)** Centralidade de Proximidade.

(d) Unidades **(e)** Áreas do Conhecimento. **(f)** Cidades.

Figura 4.9 – Representação bidimensional do segundo maior componente da rede de colaboração da USP, com seus vértices associados a diferentes propriedades e grupos.

As projeções bidimensionais do segundo maior componente conectado podem ser vistas na figura 4.9. Cada gráfico apresenta uma propriedade obtida para os vértices ou as classes as quais eles pertencem. O segundo maior componente conectado possui três grandes aglomerados de vértices, entretanto, possuem pesquisadores de diferentes áreas do conhecimento, com mais de 15 unidades participantes, e, portanto, pode ser caracterizado como altamente interdisciplinar. Entretanto, curiosamente, não há participação de unidades de Ribeirão Preto e, dado o

caráter interdisciplianar dessa sub-rede, é estranho não estar conectada a nenhum colaborador do componente principal.

4.2 Redes de Teoremas da Wikipédia

A rede inicial de teoremas da Wikipédia, obtida pela metodologia descrita em 3.3.2, resultou em um digrafo com 860 vértices e 766 arestas. Entretanto, devido a natureza incompleta da rede, as arestas foram tomadas sem direção privilegiada e, como foram considerados apenas os componentes conectados de maior tamanho, a rede resultante possui 371 vértices e 502 arestas, isto é $\langle k \rangle = 2.7$. O segundo maior componente conectado obtido era irrelevante, com apenas 5 vértices.

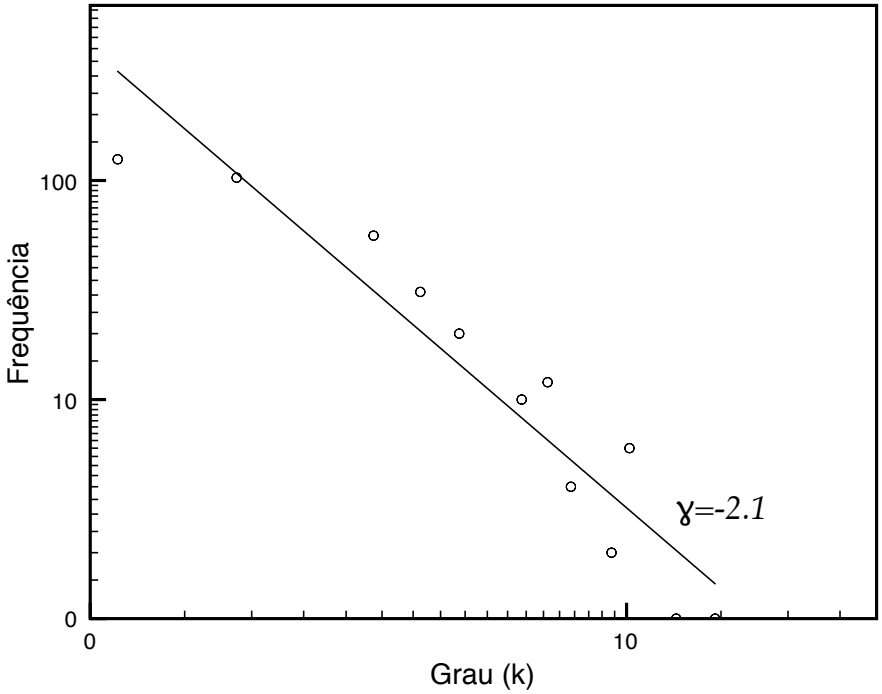


Figura 4.10 – Distribuição de grau da rede de teoremas da Wikipédia e a aproximação por lei de potência respectiva, com expoente $\gamma \simeq -2.1$.

A figura 4.10 apresenta a distribuição do grau dos vértices, revelando que a rede tem características de redes livres de escala, com expoente $\gamma \simeq -2.1$.

O maior componente conectado apresentou coeficiente de aglomeração médio $\langle Cc(i) \rangle = 0.16$ e mínimo caminho médio $l = 6.9$, revelando características de redes pequeno-mundo.

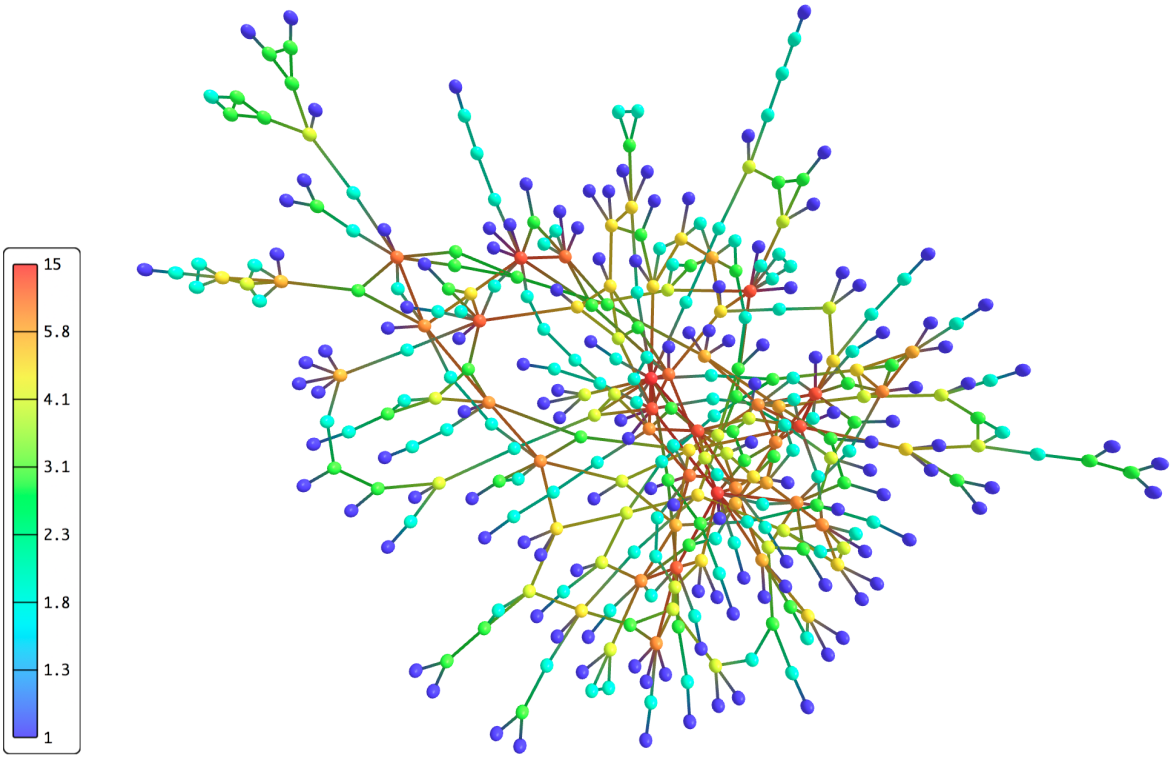


Figura 4.11 – Visualização da rede de teoremas da Wikipédia, com as cores representando o grau dos vértices como indicado pela legenda.

As visualizações geradas para a rede de teoremas, projetando-a em um plano, encontram-se nas figuras 4.11, 4.12, 4.13 e 4.14. Cada gráfico apresenta uma propriedade clássica obtida para os vértices da rede.

Considerando a figura 4.11 observa-se que a rede é composta por um grupo relativamente conectado no centro e por diversos ramos pequenos. Vértices com grau cima de 6 parecem estar uniformemente espalhados pela rede. Outra característica importante é a existência de muitos arcos que originam ramos, este fenômeno pode ser compreendido pela existência de teoremas originados dois campos de estudos da matemática diferentes. Devido a rede não ser completamente planar, esta última propriedade só pode ser observada com mais clareza através da visualização 3D interativa, disponível em:
`http://cyvision.if.sc.usp.br/TheoremNetwork/`.

A figura 4.12 indica o coeficiente de aglomeração dos vértices, mostrando que há poucos grupos muito conectados entre eles. A centralidade de proximidade, na figura 4.13, localizou o centro da rede composto pelos teoremas fundamentais da matemática, teorema fundamental do cálculo, teorema fundamental da aritmética, teorema fundamental da algebra e decomposição de Helmholtz, também conhecido como teorema fundamental do cálculo vetorial.

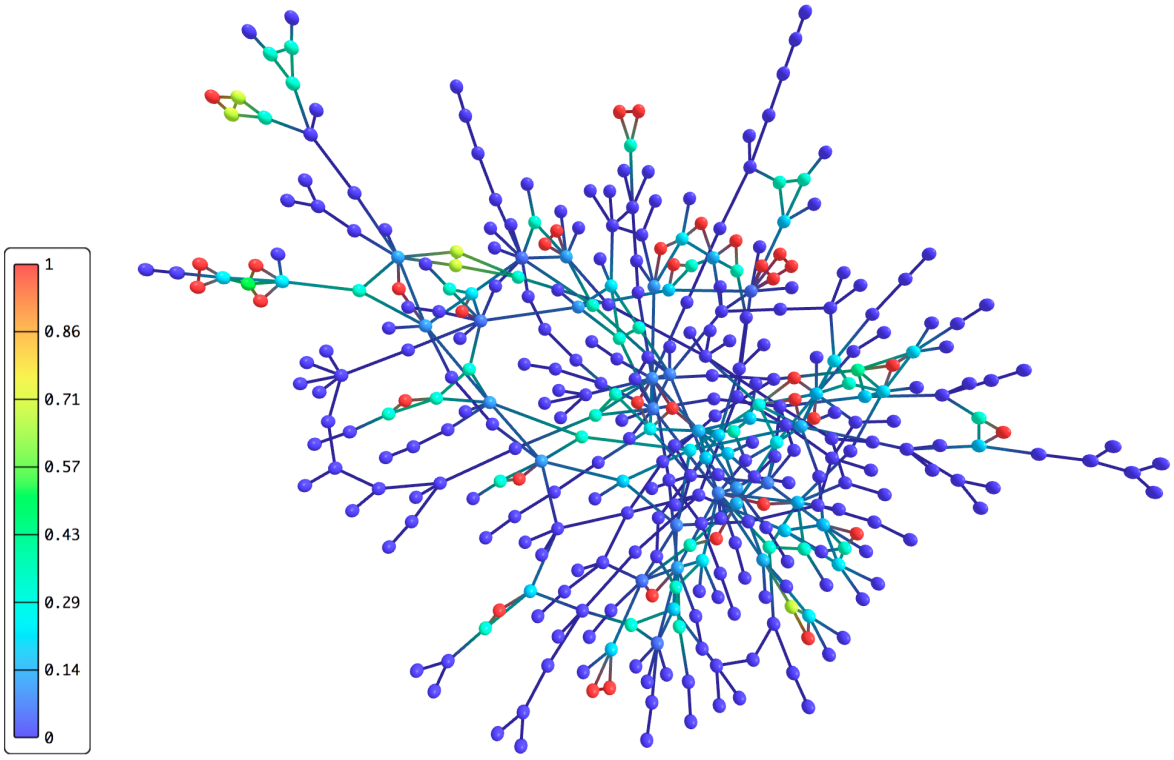


Figura 4.12 – Rede de teoremas da Wikipédia indicando o coeficiente de aglomeração.

A figura 4.14 apresenta o número de visitas ao artigo da Wikipédia correspondente cada vértice da rede durante o período do ano de 2008. Nota-se que a popularidade dos artigos não está relacionado a nenhuma das medidas topológicas obtidas para a rede.

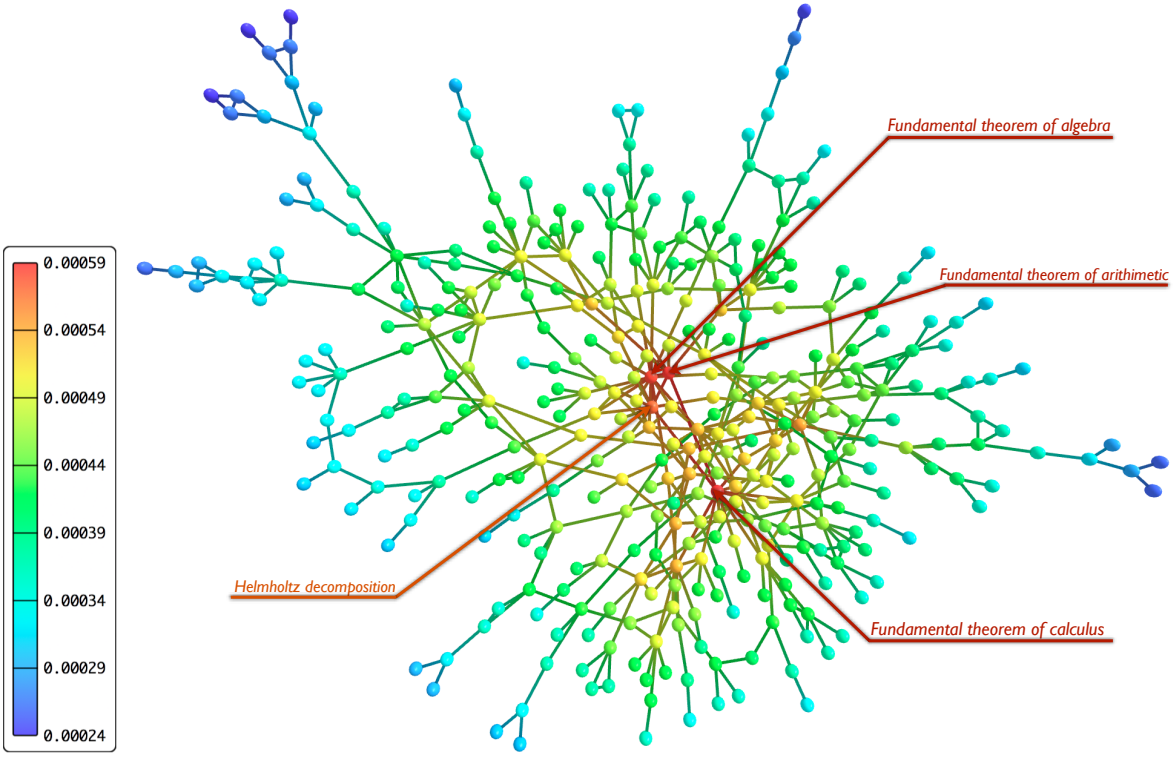


Figura 4.13 – Centralidade de proximidade calculada para cada vértice da rede de teoremas. Os teoremas com maior valor de centralidade estão em destaque.

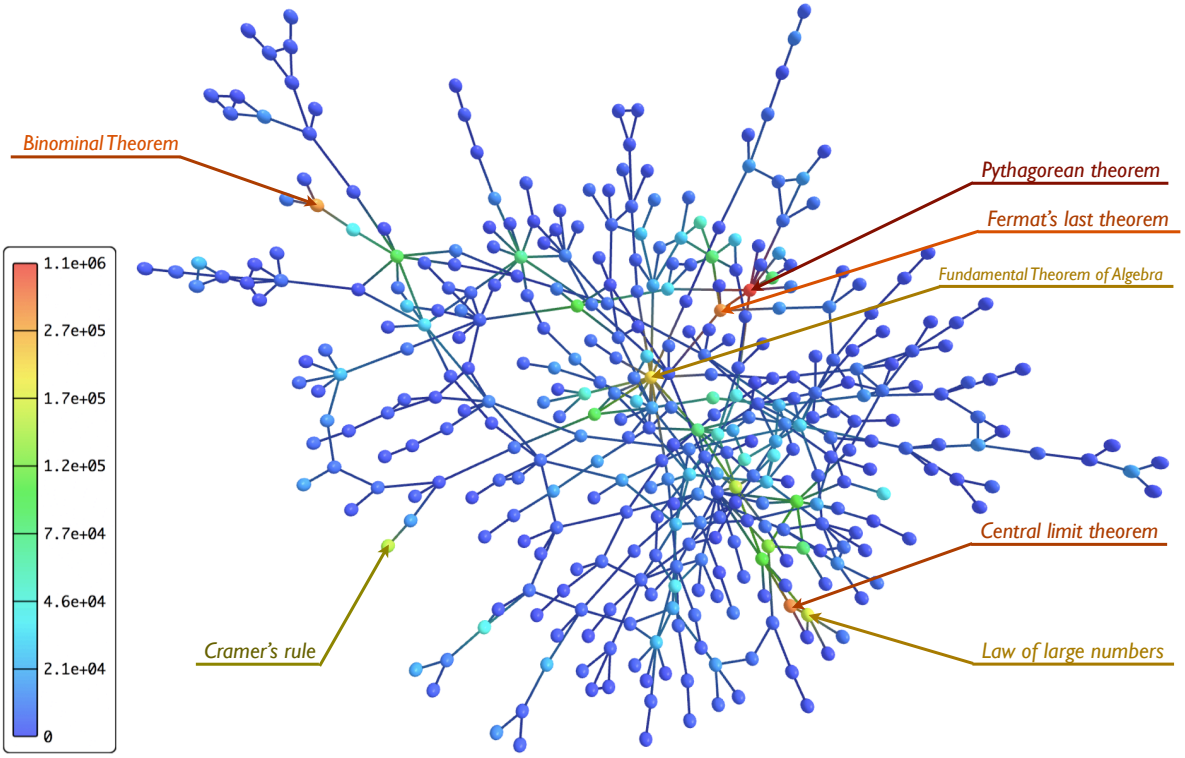


Figura 4.14 – Número de acessos aos artigos correspondentes aos vértices da rede de teoremas da Wikipédia para o ano de 2008. Os teoremas mais acessados estão em destaque.

4.3 Caracterização Concêntrica de Redes Complexas

Esta seção apresenta os principais resultados da metodologia de caracterização de redes complexas pelas propriedades concêntricas, descrita em 3.4. Inicialmente, os gráficos das distribuições das propriedades concêntricas ao longo dos níveis são apresentados para as redes complexas geradas pelos modelos teóricos, assim como as respectivas descrições e interpretações dos resultados. As distribuições obtidas paras as redes reais consideradas são apresentadas em diferentes subseções, discutindo e comparando os resultados com os modelos teóricos. Todas figuras de distribuições apresentadas nas próximas subseções apresentam as respectivas curvas de distribuição média considerando todos os vértices da rede considerada, acompanhadas pelas barras de erro representando o desvio padrão.

Os resultados da caracterização dos vértices é apresentado para a rede de colaboração da USP através de diagramas de pizza e dendrograma característico da aglomeração hierárquica. A seção concluí com a investigação de algumas características observadas das propriedades concêntricas e apresentando os resultados das projeções obtidas por PCA.

4.3.1 Distribuição das propriedades concêntricas para os modelos teóricos

Nesta sub-seção são apresentados os resultados das distribuições das propriedades concêntricas obtidos para os modelos teóricos de redes complexas.

Modelo de rede aleatória Erdös-Rényry, ER

A figura 4.15 apresenta as distribuições das propriedades concêntricas ao longo dos níveis concêntricos, obtidas para uma rede aleatória (ER) com 10 mil vértices e $\langle k \rangle \simeq 10$. As curvas obtidas apresentam nível concêntrico máximo 7, que equivale ao diametro da rede, ou máxima geodésica. Com exceção do grau entre-nívéis, as distribuições apresentam uma região inicial crescente, seguida por uma região de planalto para o coeficiente de aglomeração ou por um pico para as demais. O perfil dos gráficos termina com uma região decrescente devido a característica finita da rede.

As propriedades concêntricas: número de nós, número de arestas, grau concêntrico e grau

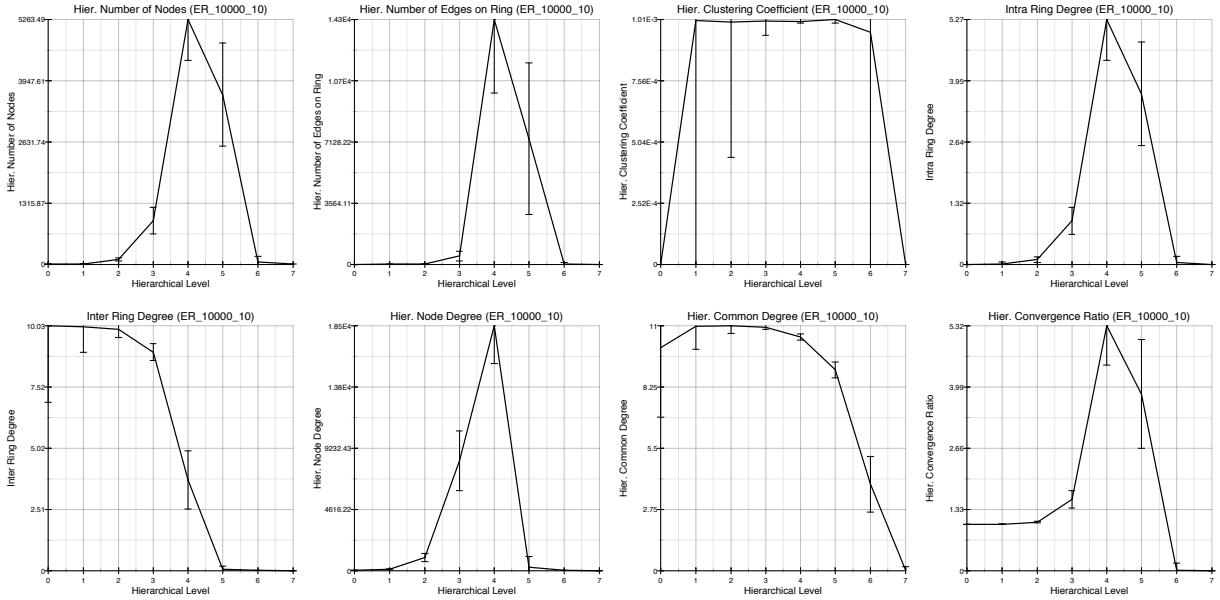


Figura 4.15 – Distribuição das propriedades concêntricas para uma rede aleatória com 10 mil vértices e grau médio $\langle k \rangle \simeq 10$. As curvas são apresentadas pela média dos valores correspondentes a cada nívei concêntrico assim como o respectivo desvio padrão, representado pelas barras de erro.

intra-níveis; apresentaram curvas de distribuição semelhantes para a rede ER, mostrando que, em média, tanto os vértices como as arestas estão distribuídas, em sua maioria, a uma distância 4 a partir de qualquer vértice. Entretanto, a parcela de conexões entre vértices de diferentes níveis concêntricos, em média, decrescem muito mais rapidamente quando comparado ao grau concêntrico. Este comportamento pode ser observado pela distribuição do grau entre-níveis, que diminui conforme o aumento do número de nós concêntricos. É importante notar que as propriedades de grau entre-níveis, grau intra-níveis e grau comum são normalizadas de acordo com o número de vértices em cada nível concêntrico.

O coeficiente de aglomeração concêntrico obtido apresentou uma distribuição praticamente constante com relação aos níveis concêntricos. Entretanto, o valor máximo é da ordem de 0.001, corroborando com o valor característico do coeficiente de aglomeração tradicional de redes aleatórias.

Modelo de rede livre de escala Barabási-Albert, BA

A figura 4.16 apresenta as distribuições das propriedades concêntricas para uma rede do modelo Barabási-Albert com 10 mil vértices e $\langle k \rangle \simeq 10$. Algumas curvas de distribuições são muito semelhantes àquelas obtidas para o modelo ER, como o número de nós, número de arestas, grau concêntrico, grau entre-níveis e taxa de convergência. Entretanto, as curvas obtidas para o grau entre níveis, grau comum e coeficiente de aglomeração, resultaram em curvas menos

suaves, principalmente na região do primeiro nível concêntrico.

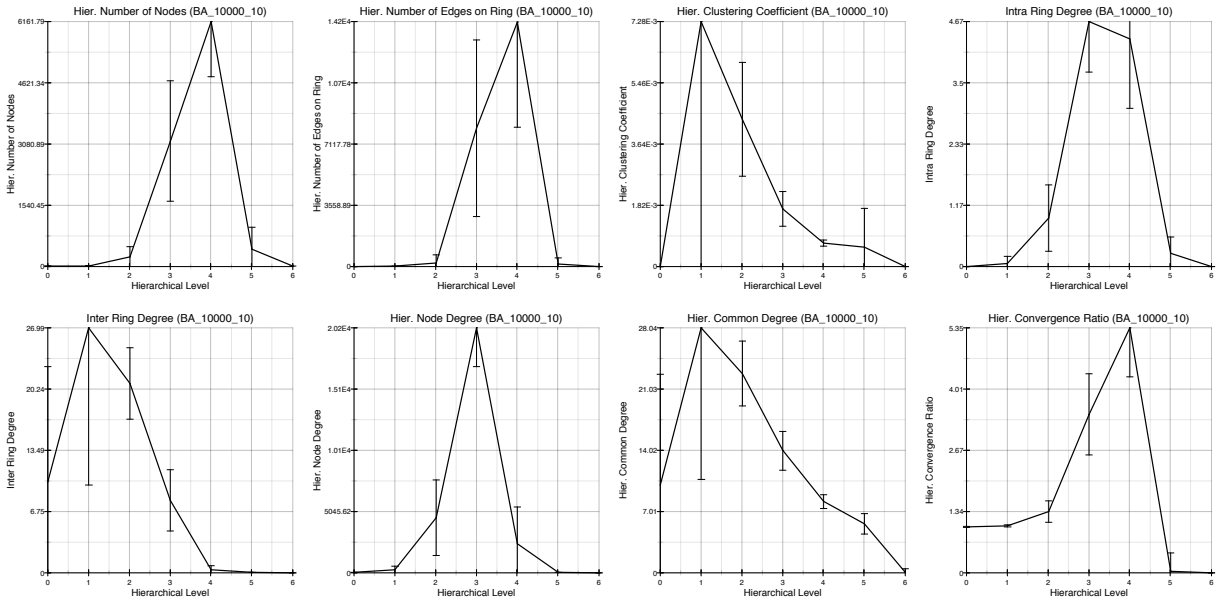


Figura 4.16 – Distribuição das propriedades concêntricas para uma rede livre de escala do modelo Barabási-Albert. A rede possui 10 mil vértices e grau médio $\langle k \rangle \simeq 10$.

O coeficiente de aglomeração concêntrico, assim como para a rede aleatória, é baixo, entretanto, para o primeiro anel apresenta um valor médio consideravalmente mais alto do que para os outros. Este comportamento é uma consequencia direta da presença de hubs na rede, que pertencem aos primeiros níveis concêntricos, isto é, em geral, distam dos vértices por poucos passos, resultando no saturamento de vértices conectados entre si dentro dos anéis mais distantes.

A existência dos hubs também pode ser verificada pelo súbito aumento do grau comum e grau entre-níveis, observado para os valores da distribuição no primeiro anel concêntrico, mostrando que os hubs são acessados rapidamente, pertencendo, em média, aos primeiros níveis concêntricos.

Modelos de redes regulares

As distribuições das propriedades concêntricas também foram obtidas para duas redes regulares, uma com efeitos de borda, apresentada na figura 4.17 e outra sem bordas, na figura 4.18, ambas dispostas como uma grade de 100×100, totalizando 10000 vértices e com $\langle k \rangle \simeq 8$.

As curvas das distribuições para a rede regular com efeito de bordas são suaves, com as propriedades de grau e número de nós caracterizadas por um pico largo, similar àqueles obtidos das

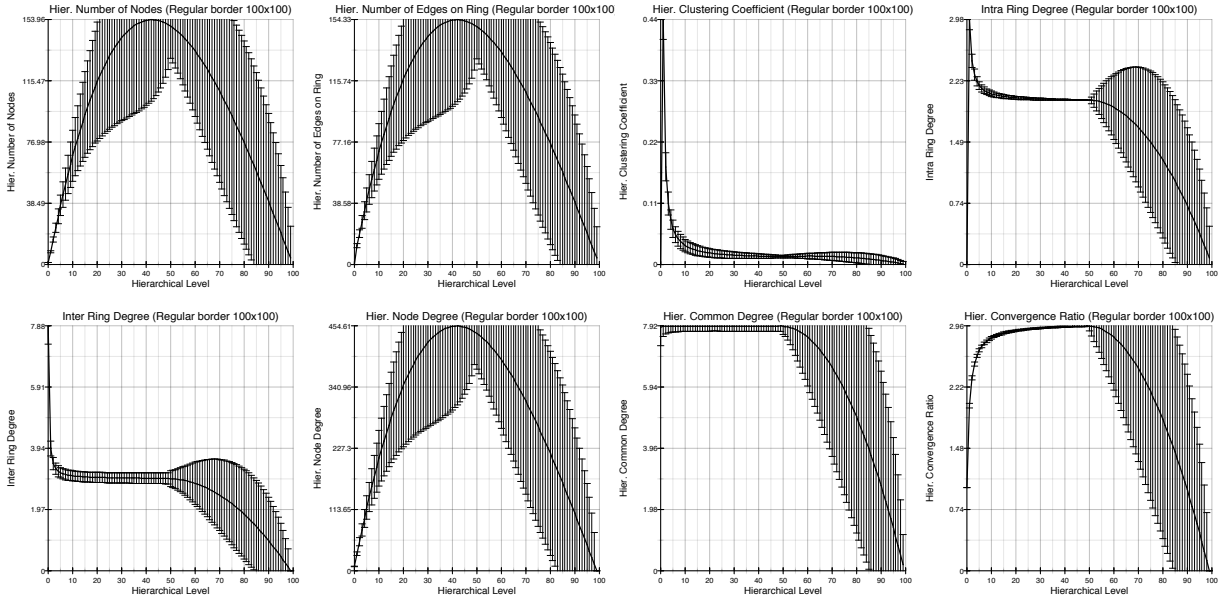


Figura 4.17 – Curvas das distribuições das propriedades concêntricas para uma rede regular bidimensional com borda quadrada de tamanho 100×100.

redes ER e BA, entretanto, apresentou alto valor de variação ao longo dos níveis concêntricos. As outras propriedades também apresentaram similaridades com os modelos ER e BA, como o decaimento dos valores das medidas para os níveis concêntricos mais distantes, propriedade característica de redes finitas.

Outra propriedade interessante, observada para os gráficos da rede regular com efeito de bordas, são os valores dos desvios ao longo dos níveis concêntricos, que apresenta comportamentos diferentes para as metades das curvas, isto é, para a região formada pelos níveis menores que 50, e pela região de níveis maiores que 50. Esta característica parece estar relacionada à geometria da borda.

O modelo regular sem efeito de bordas apresentou curvas de distribuição lineares para as propriedades de grau e número de nós, revelando o caráter infinito da rede, devido a sua topologia cíclica. A rede não apresentou variação dos valores, com desvio padrão zero para todos os pontos, consequência do fato de que todos os vértices dessa rede são degenerados, sendo impossível distingui-los uns dos outros, logo não importa qual vértice é tomado como referência, as propriedades concêntricas devem ser iguais.

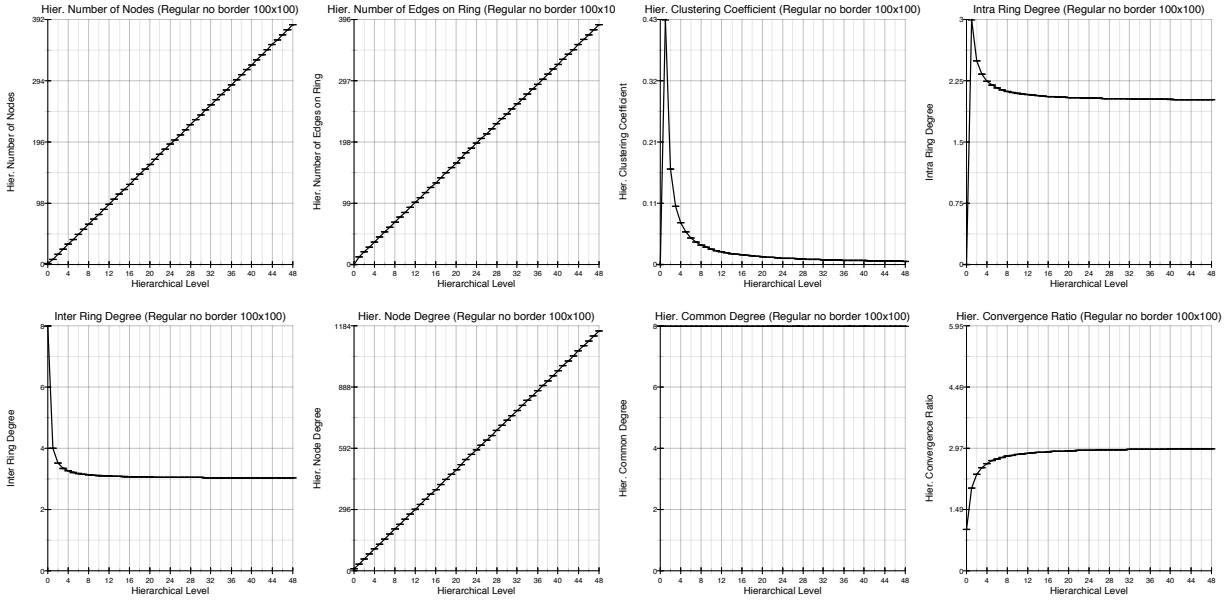


Figura 4.18 – Curvas das distribuições das propriedades concêntricas para uma rede regular bidimensional sem efeitos de bordas.

Modelo de rede geográfica

As distribuições das propriedades concêntricas, obtidas para uma rede geográfica com 10000 vértices e $\langle k \rangle \simeq 10$, é apresentada na figura 4.20. As curvas das propriedades concêntricas simples: grau concêntrico, número de vértices e número de arestas; apresentam um pico largo e são relativamente suaves, assemelhando-se àquelas obtidas para as redes regulares com borda.

O coeficiente de aglomeração concêntrico para a rede geográfica, apresentou dois picos característicos, um para os níveis concêntricos mais próximos e outro para os mais distântes. A região do segundo pico apresenta alto valor do desvio padrão, indicando que pode haver vértices que apresentam e outros que não apresentam o segundo pico. O aparecimento desse pico caracteriza redes do tipo geográficas, e um estudo mais detalhado de seu significado pode ser visto na sub-seção 4.3.4.

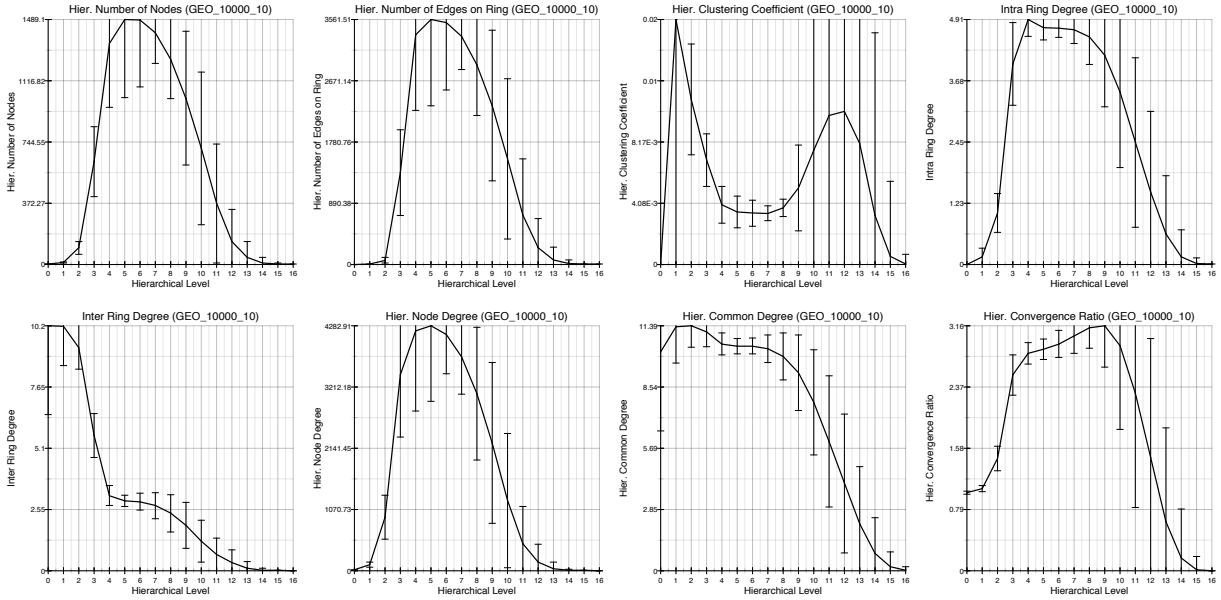


Figura 4.19 – Curvas das distribuições das propriedades concêntricas para uma rede geográfica com 10 mil vértices e $\langle k \rangle \simeq 10$.

Modelo de rede Watts-Strogatz (WS)

A figura 4.20 apresenta as distribuições das propriedades concêntricas para uma rede Watts-Strogatz, gerada com probabilidade de reconexão de arestas $p = 0.04$ e 10000 vértices com $\langle k \rangle \simeq 10$.

As distribuições das propriedades concêntricas simples, apresentaram curvas semelhantes àquelas obtidas para os modelos ER, BA e geográfico, caracterizadas por um pico próximo na região dos níveis concêntricos intermediarios.

Assim como para a rede geográfica, o coeficiente de aglomeração concêntrico apresentou um segundo pico na região dos últimos níveis concêntricos, entretanto com em menor intensidade comparada ao valor máximo, que apresenta alta aglomeração, $\langle Cc_1 \rangle \simeq 0.6$.

O comportamento da curva do grau comum foi semelhante àquela obtida para a rede regular sem bordas, exceto pelo rápido decréscimo de seu valor nós últimos 3 níveis concêntricos, o que indica que as redes Watts-Strogatz apresentam uma borda muito fina de vértices com poucas arestas.

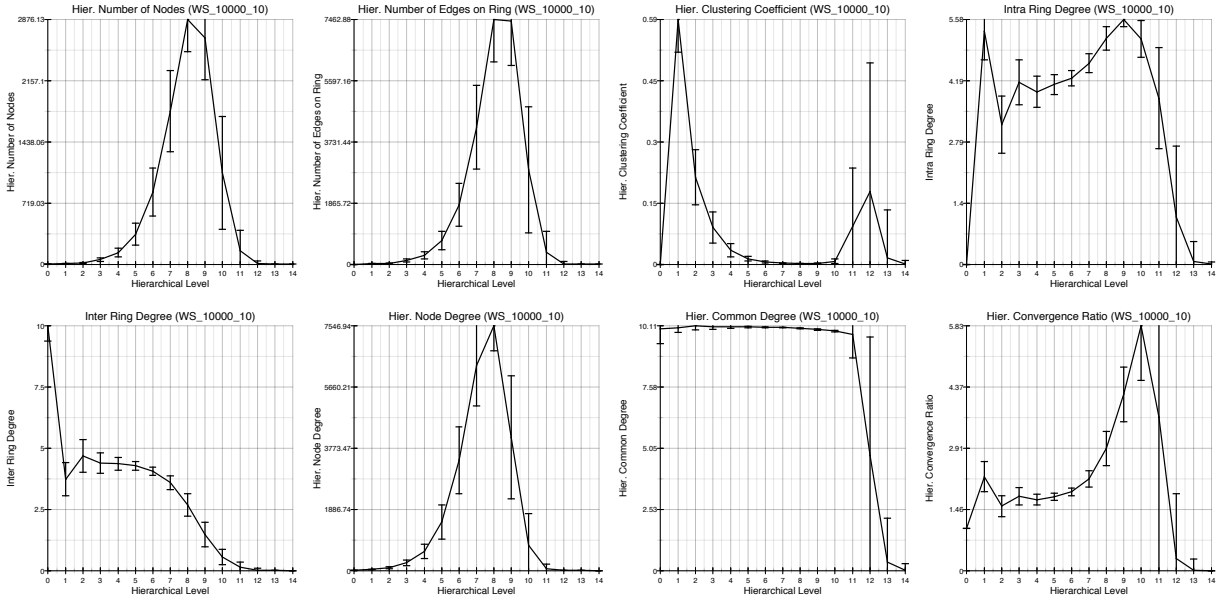


Figura 4.20 – Curvas das distribuições das propriedades concêntricas para uma rede Watts-Strogatz de 10 mil vétices e $\langle k \rangle \simeq 10$, com probabilidade de religação de arestas $p = 0.04$.

4.3.2 Distribuição das propriedades concêntricas para as redes reais

Nesta seção são apresentados e discutidos os principais resultados obtidos para as distribuições das propriedades concêntricas aplicadas às redes reais descritas no capítulo anterior.

Rede de Colaboração da USP

As curvas de distribuições das propriedades concêntricas obtida para a rede de colaboração da USP encontram-se na figura 4.21. As distribuições também foram determinadas para um conjunto de redes teóricas de número de vértices e grau médio semelhantes à rede de colaboração da USP, e se encontram nas figuras 4.22, 4.23, 4.24 e 4.25, respectivamente, para os modelos ER, BA, geográfico e WS. Devido a natureza dos métodos de geração de redes BA e WS, apenas redes com grau médio par podem ser criadas, de modo que foram construídas redes com o valor par mais próximo do grau médio da rede em questão.

Todas as propriedades apresentaram alto valor de desvio padrão, revelando que há grande variação de seus valores ao longo dos vértices da rede. Similarmente às curvas obtidas para os modelos de redes teóricas, o grau concêntrico, número de vértices e número de arestas da rede de colaboração são semelhantes entre si, entretanto apresentam um decaimento muito mais

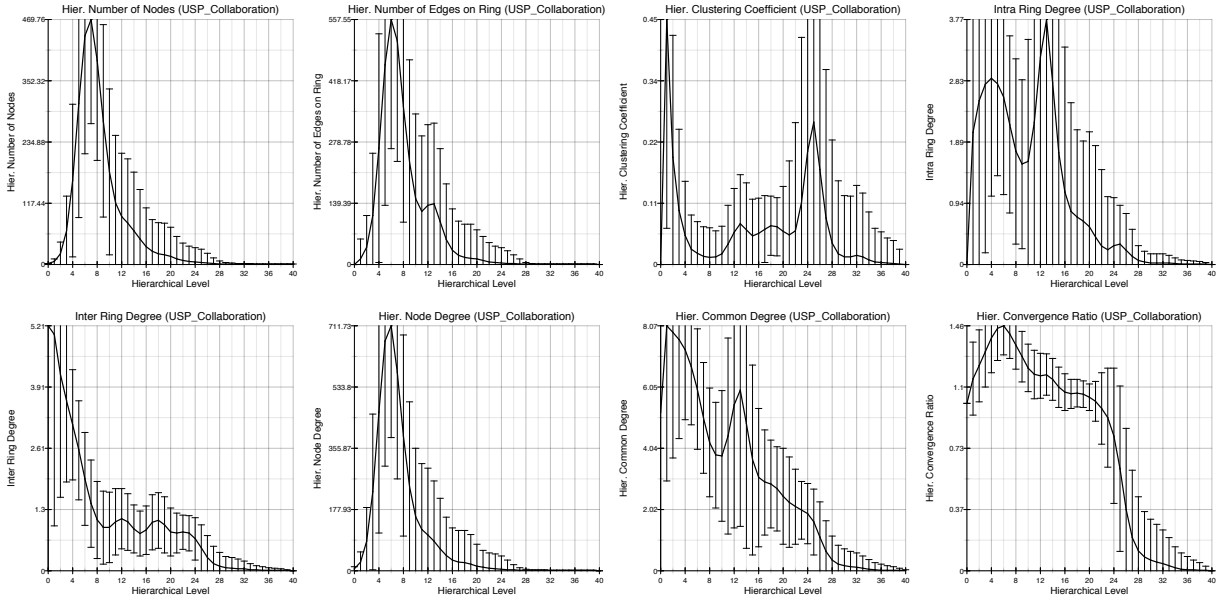


Figura 4.21 – Distribuição das propriedades concêntricas obtidas para a rede de colaboração da USP com 2864 vértices e $\langle k \rangle \simeq 5$.

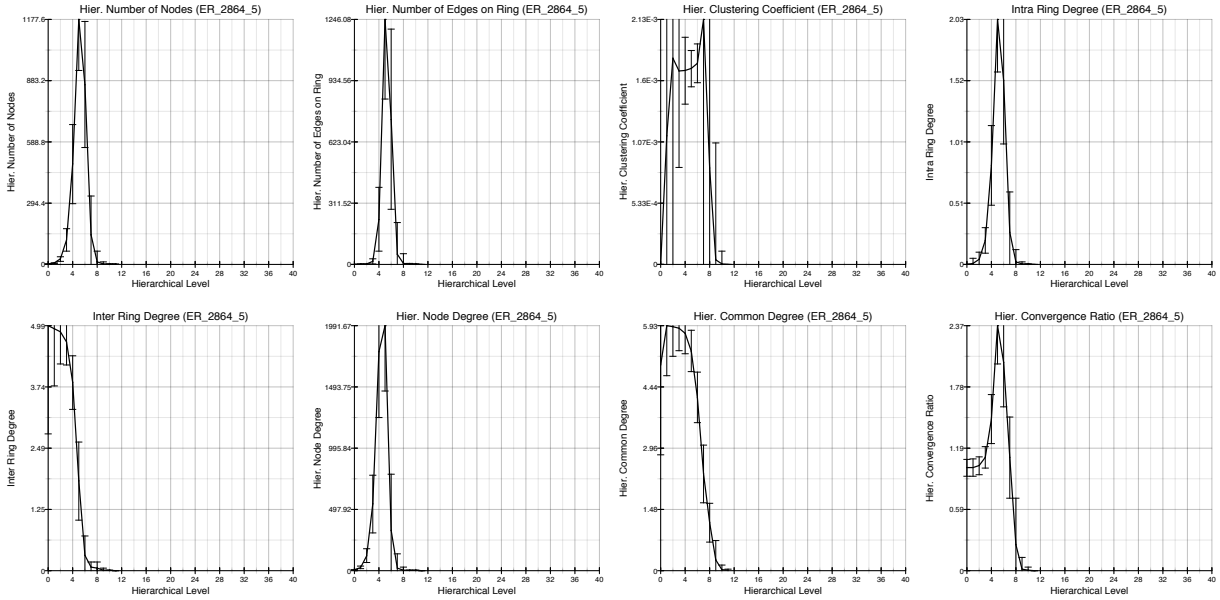


Figura 4.22 – Distribuição das propriedades concêntricas para a rede ER comparável à rede de colaboração da USP.

lento após o terceiro nível concêntrico, estendendo-se até os níveis próximos ao 30^o.

A curva de distribuição para o coeficiente de aglomeração apresentou um pico na região dos últimos níveis concêntricos, característica presente no modelo geográfico e em menor intensidade no modelo WS. O desvio padrão nessa região também foi bem elevado, indicando que a presença desse pico pode não ocorrer para uma parte dos vértices. Os resultados da caracterização dos vértices pelo coeficiente de aglomeração concêntrico são encontrados na

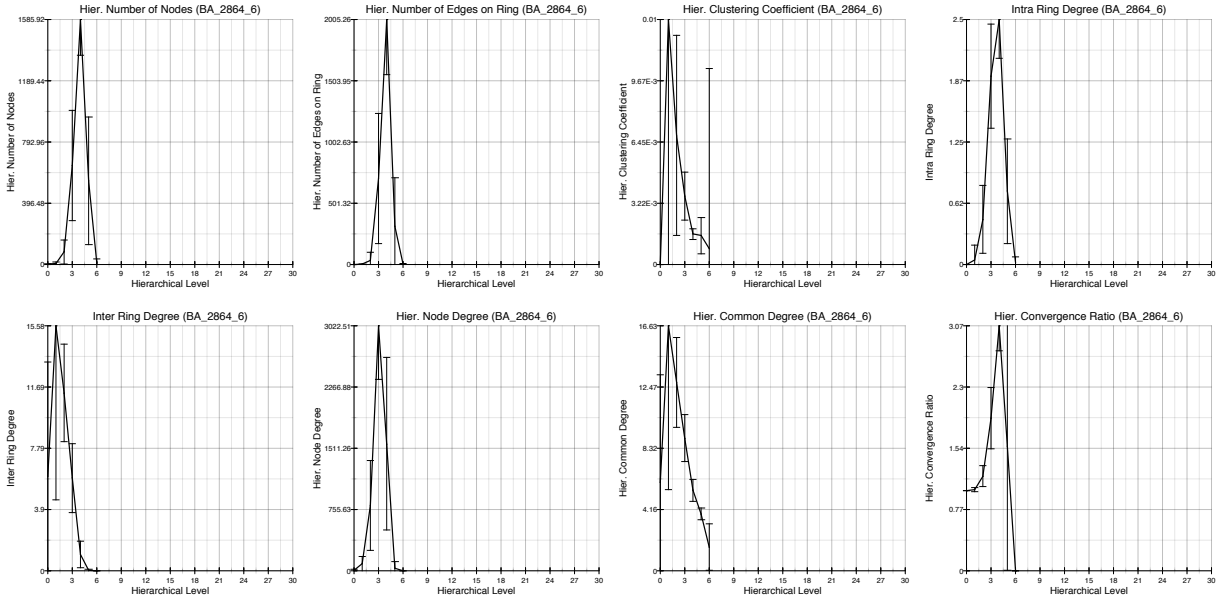


Figura 4.23 – Distribuição das propriedades concêntricas para a rede BA comparável à rede de colaboração da USP

subseção 4.3.3.

O grau entre-níveis apresentou uma distribuição composta por uma combinação das curvas obtidas para o modelo ER, caracterizada por um pico estreito e decaimento rápido nos primeiros níveis, e para o modelo geográfico e WS, caracterizadas por uma região de valor constante seguida por um decaimento lento.

A propriedade taxa de convergência apresentou-se como a mais diferente entres os outros modelos teóricos e a rede de colaboração. Sua distribuição apresentou um pico largo com o centro deslocado para a esquerda. Isto é uma consequência do fato que diferentemente dos modelos BA, os hubs são acessados gradualmente ao longo dos níveis concêntricos, isto significa que os hubs nem sempre estão conectados a outros hubs.

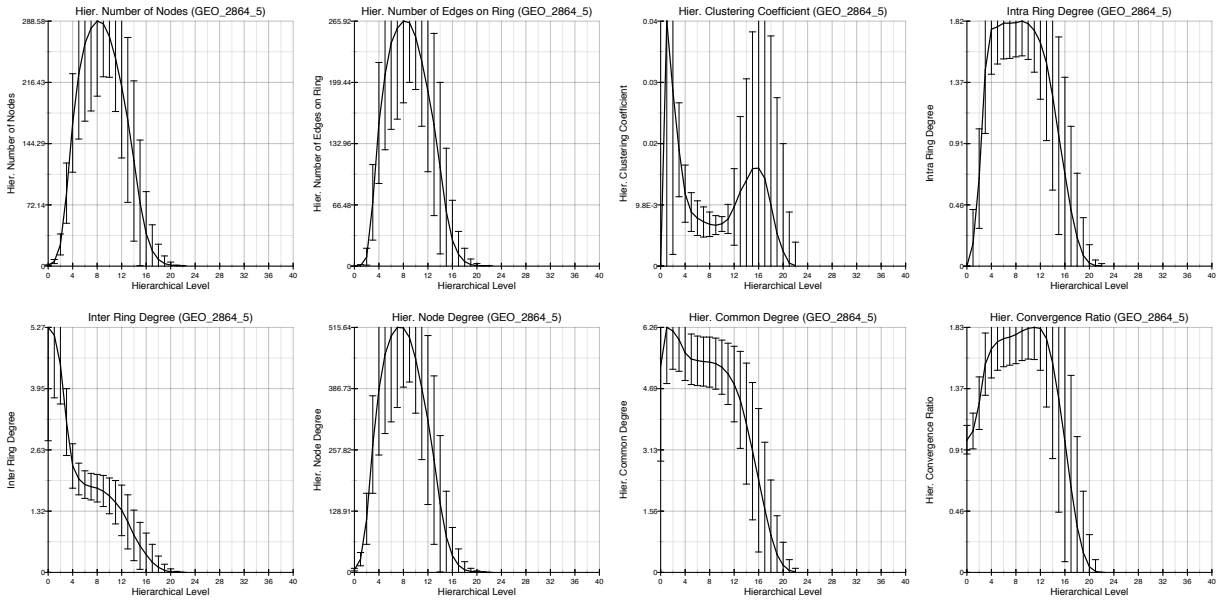


Figura 4.24 – Distribuição das propriedades concêntricas para a rede geográfca comparável à rede de colaboração da USP.

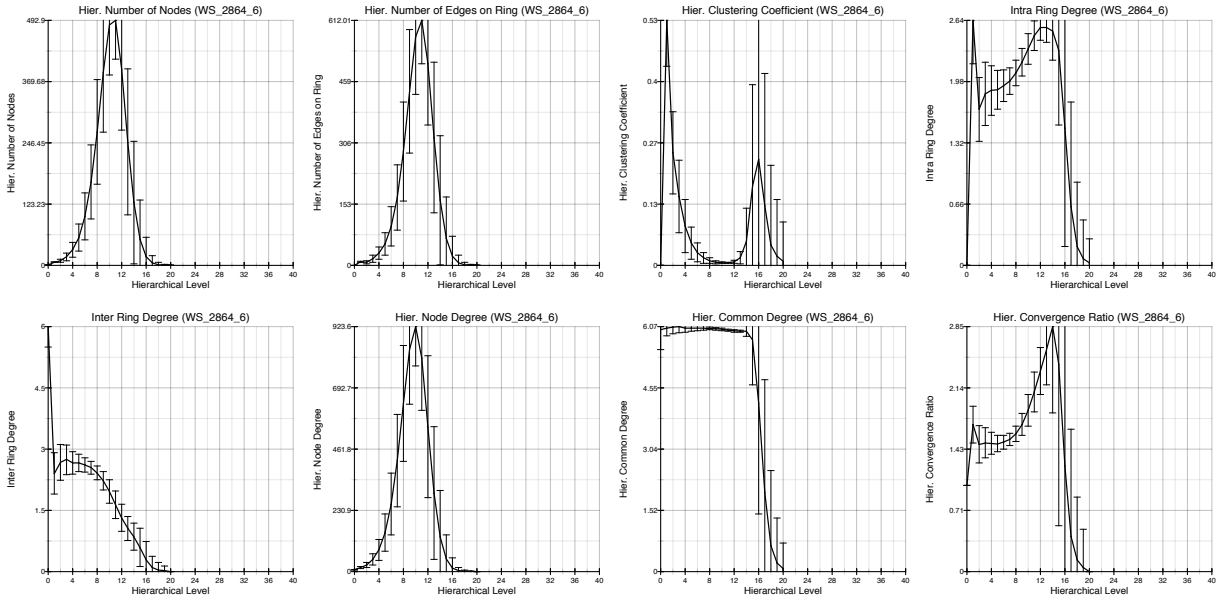


Figura 4.25 – Distribuição das propriedades concêntricas para a rede WS comparável à rede de colaboração da USP.

Rede de teoremas da Wikipédia

As curvas das distribuições obtidas para a rede de teoremas da Wikipédia podem ser vistas na figura 4.26. As curvas para as redes geradas seguindo os modelos teóricos comparáveis à rede de teoremas, encontram-se nas figuras 4.27, 4.28, 4.29 e 4.30, respectivamente obtidas para os modelos: ER, BA, geográfico e WS.

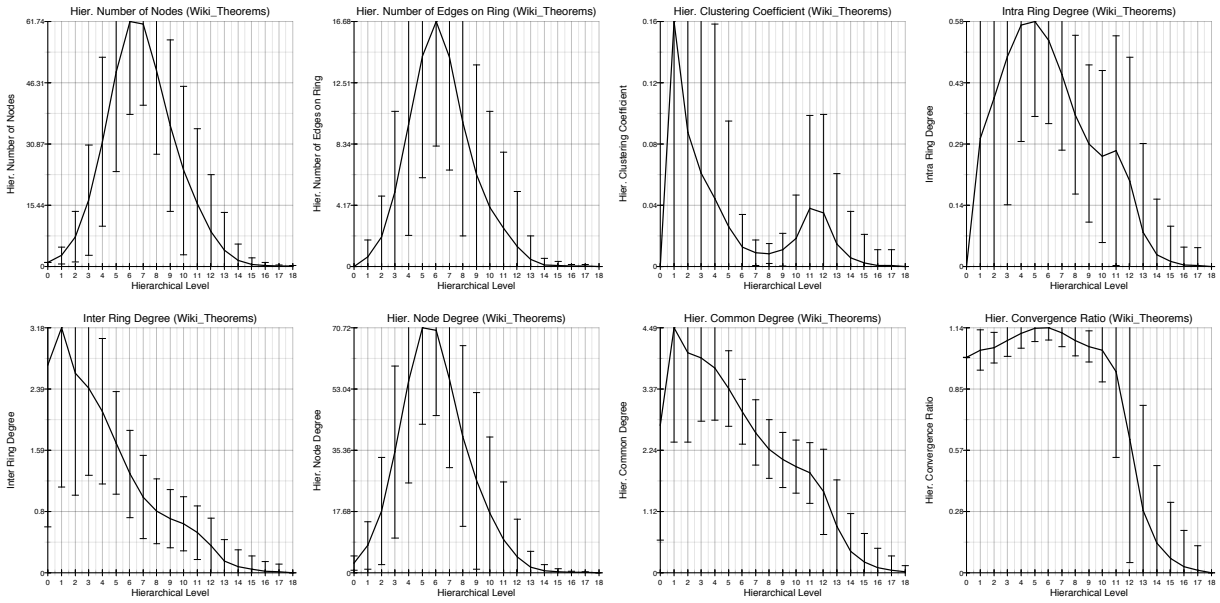


Figura 4.26 – Distribuição das propriedades concêntricas obtidas para a rede de teoremas da Wikipédia com 371 vértices e $\langle k \rangle \simeq 2.7$.

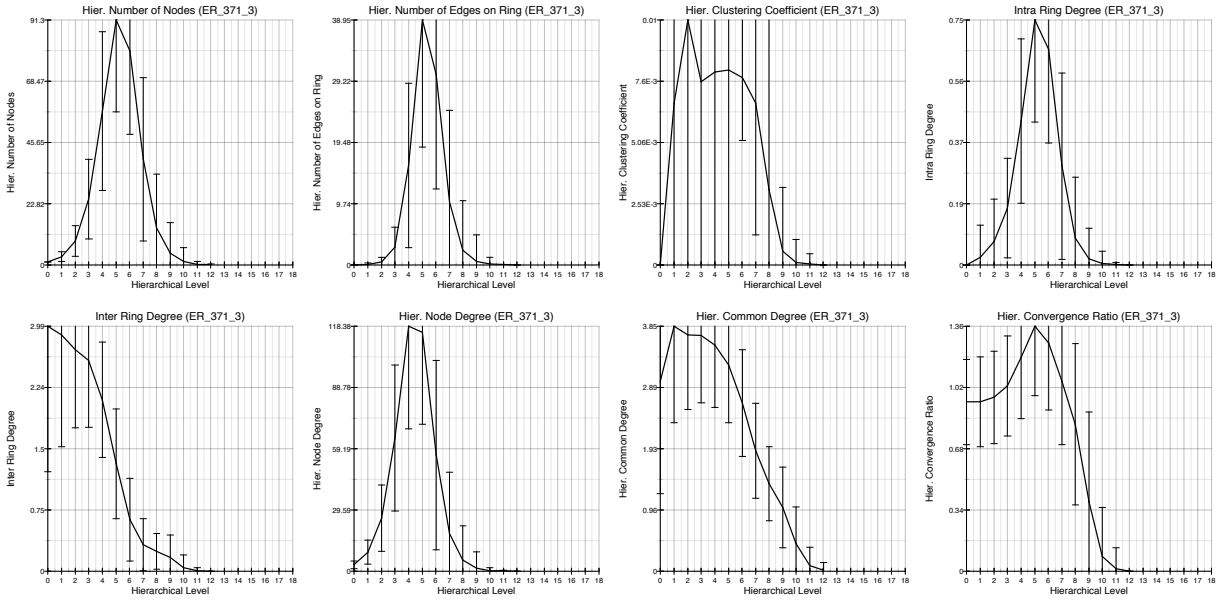


Figura 4.27 – Distribuição das propriedades concêntricas para a rede ER comparável à rede de teoremas da Wikipédia.

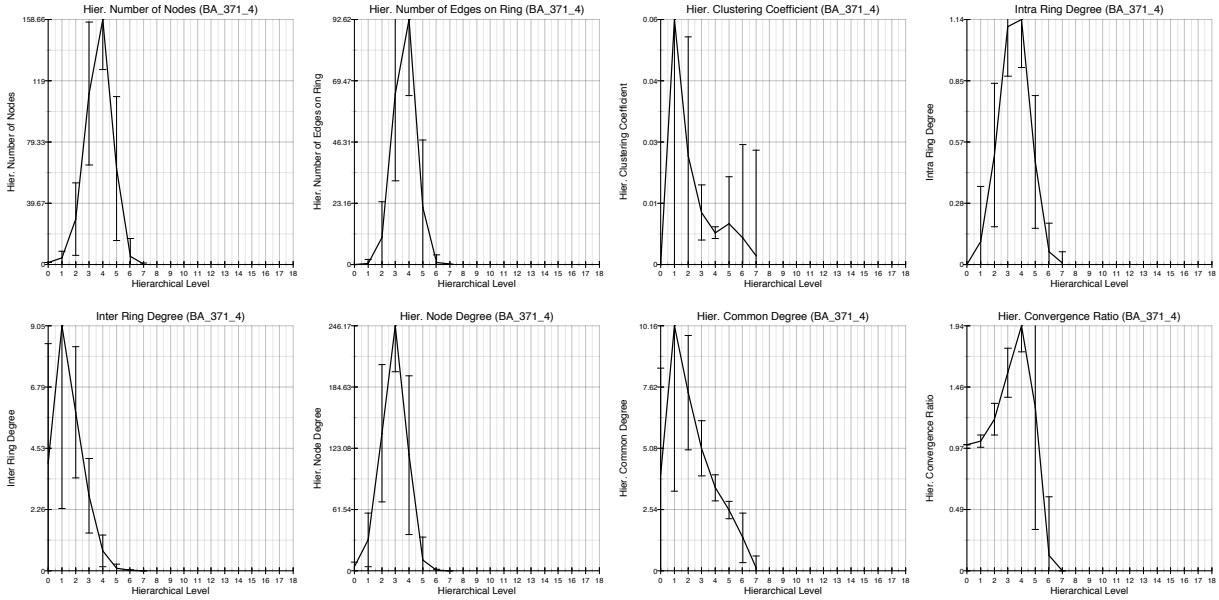


Figura 4.28 – Distribuição das propriedades concêntricas para a rede BA comparável à rede de teoremas da Wikipédia.

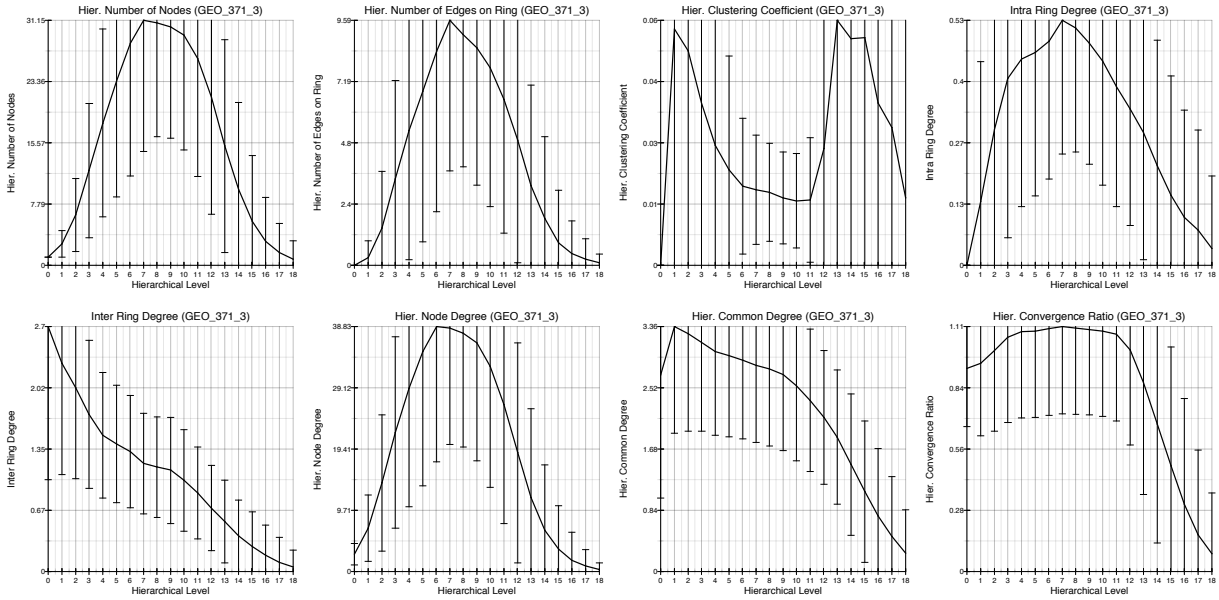


Figura 4.29 – Distribuição das propriedades concêntricas para a rede geográfca comparável à rede de teoremas da Wikipédia.

As curvas das propriedades concêntricas simples obtidas para a rede de teoremas da Wikipédia apresentaram um pico com largura entre àquelas obtidas para as redes BA e ER e e as obtidas para os modelos geográfico e WS. O que indica que a distribuição dos vértices ao longo dos níveis concêntricos é mais espalhada do que as obtidas para os modelos ER e BA, entretanto não tão espalhadas quanto os modelos geográficos e WS.

A distribuição do coeficiente de aglomeração concêntrico apresentou um pico com alta

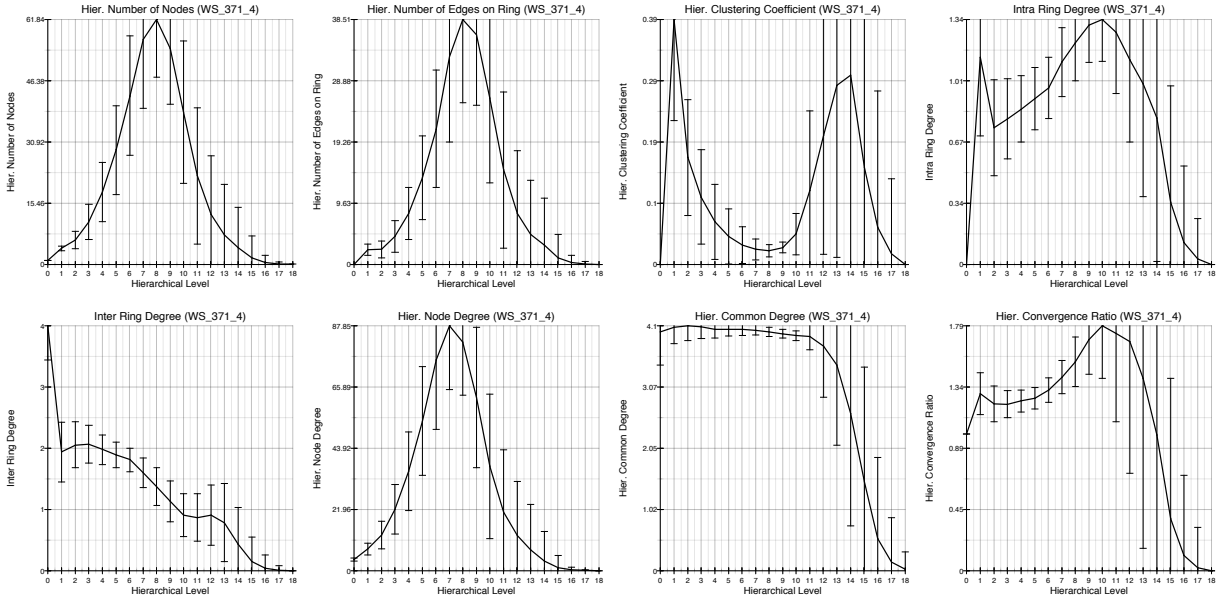


Figura 4.30 – Distribuição das propriedades concêntricas para a rede WS comparável à rede de teoremas da Wikipédia.

variação na região dos últimos níveis concêntricos, característico das redes Wattz-Strogatz e geográficas.

Tanto o grau entre-níveis quanto o grau comum apresentaram distribuições compostas por uma mistura das curvas características dos modelos ER, BA e geográfico; revelando o caráter híbrido da rede.

A distribuição obtida para a propriedade de taxa de convergência resultou em uma curva constituída de um pico largo com rápido decaimento para os últimos níveis, revelando, assim como para a rede de colaboração da USP e para o modelo geográfico, que os hubs da rede de teoremas são acessados gradualmente ao longo dos níveis concêntricos.

Rede de aeroportos dos EUA

A figura 4.31 apresenta as curvas das distribuições das propriedades concêntricas obtidas para a rede de aeroportos dos EUA, enquanto as figuras 4.32, 4.33, 4.34 e 4.35; as àquelas obtidas para os modelos teóricos de mesmo número de vértices e $\langle k \rangle$.

Diferentemente da rede de teoremas e de colaboração, o nível concêntrico máximo obtido para a rede de aeroportos, isto é, seu diâmetro topológico, é semelhante àqueles obtidos para o

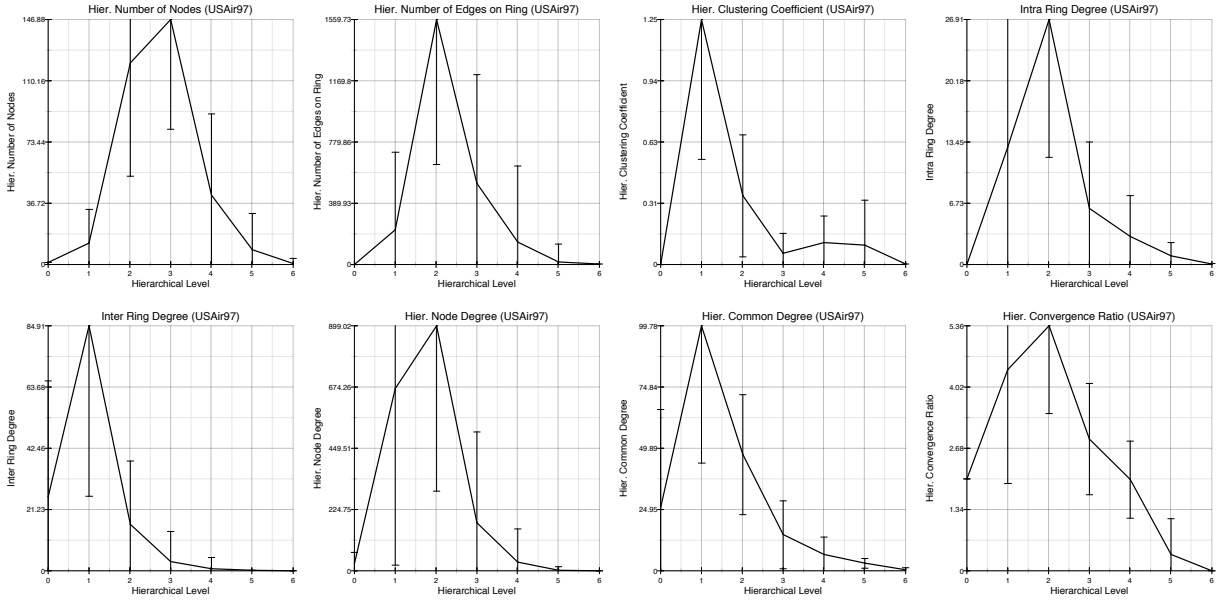


Figura 4.31 – Distribuição das propriedades concêntricas obtidas para a rede de aeroportos de EUA com 332 vértices e $\langle k \rangle \simeq 6$.

modelo BA e ER, com valor 6.

As propriedades concêntricas básicas obtidas para a rede de aeroportos resultaram em curvas em distribuições semelhantes, analogamente aos modelos e as outras redes reais estudadas, caracterizadas por um pico central. Entretanto, para essa rede, a largura do pico é semelhante às obtidas para os modelos ER e BA.

O coeficiente de aglomeração concêntrico apresentou uma distribuição caracterizada por um pico, seguido de um decaimento e por fim uma pequena elevação, semelhante à curva obtida para a distribuição considerando a rede BA comparável.

A propriedade de grau entre níveis resultou em uma distribuição composta por uma curva apresentando um pico, semelhante àquela obtida para o modelo BA, entretanto mais estreito, revelando que o acesso aos vértices de maior conectividade é mais rápido na rede de aeroportos. Os resultados obtidos para a taxa de convergência concêntrica corroboram com esse fato, apresentando um pico deslocado para a esquerda quando comparado aos obtidos para os modelos teóricos.

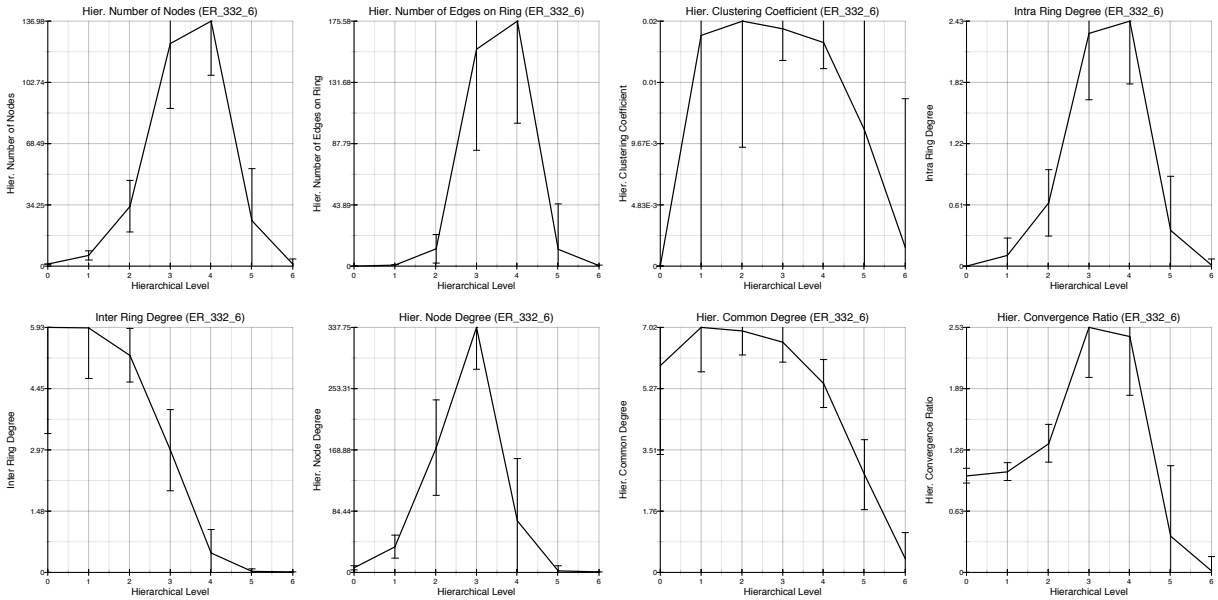


Figura 4.32 – Distribuição das propriedades concêntricas para a rede ER comparável à rede de aeroportos.

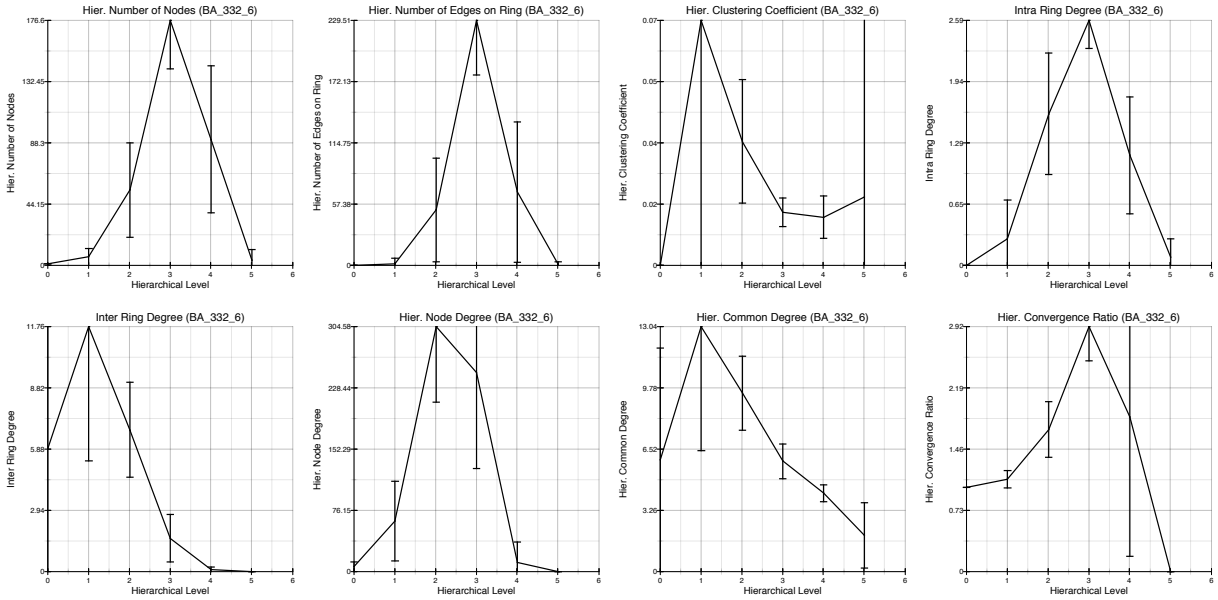


Figura 4.33 – Distribuição das propriedades concêntricas para a rede BA comparável à rede de aeroportos.

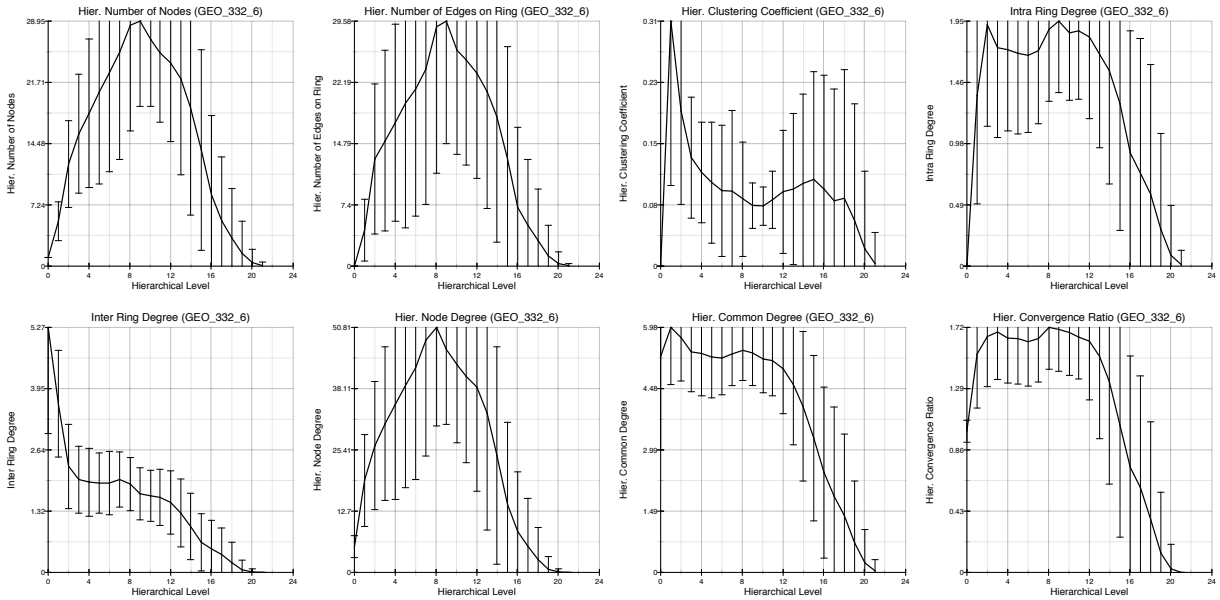


Figura 4.34 – Distribuição das propriedades concêntricas para a rede geográfca comparável à rede de aeroportos.

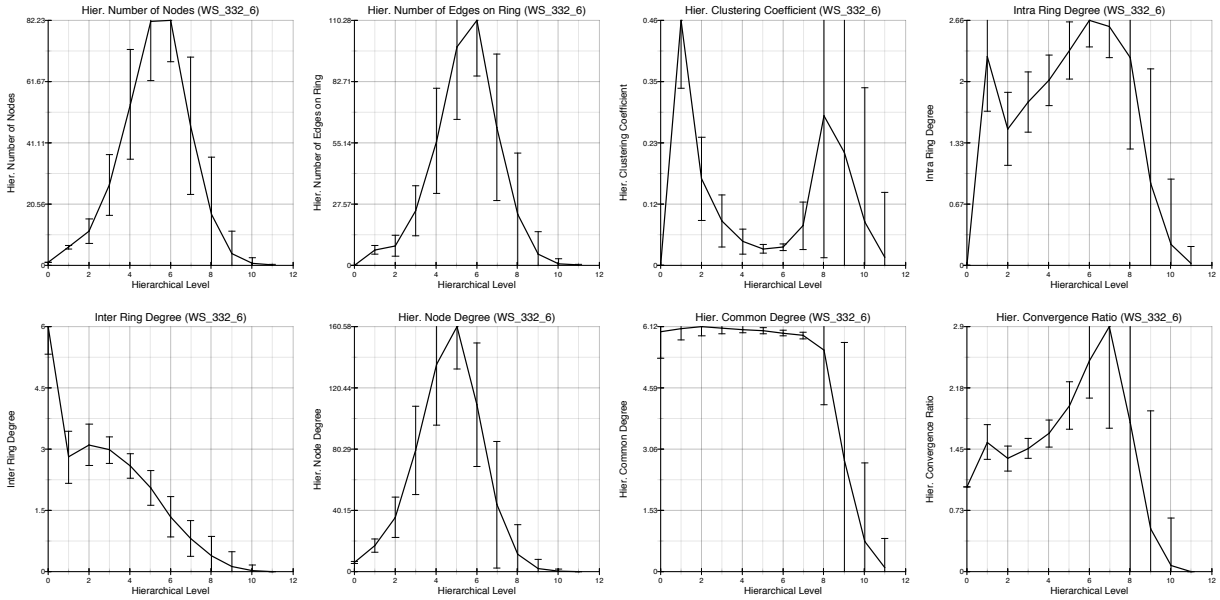


Figura 4.35 – Distribuição das propriedades concêntricas para a rede WS comparável à rede de aeroportos.

Rede de associação de palavras de Edinburgh

As curvas das propriedades concêntricas obtidas para a rede de palavras de Edinburgh podem ser vistas na figura 4.36. As figuras 4.37, 4.38, 4.39 e 4.40; apresentam as propriedades obtidas para os modelos teóricos de mesmo número de vértices e $\langle k \rangle$.

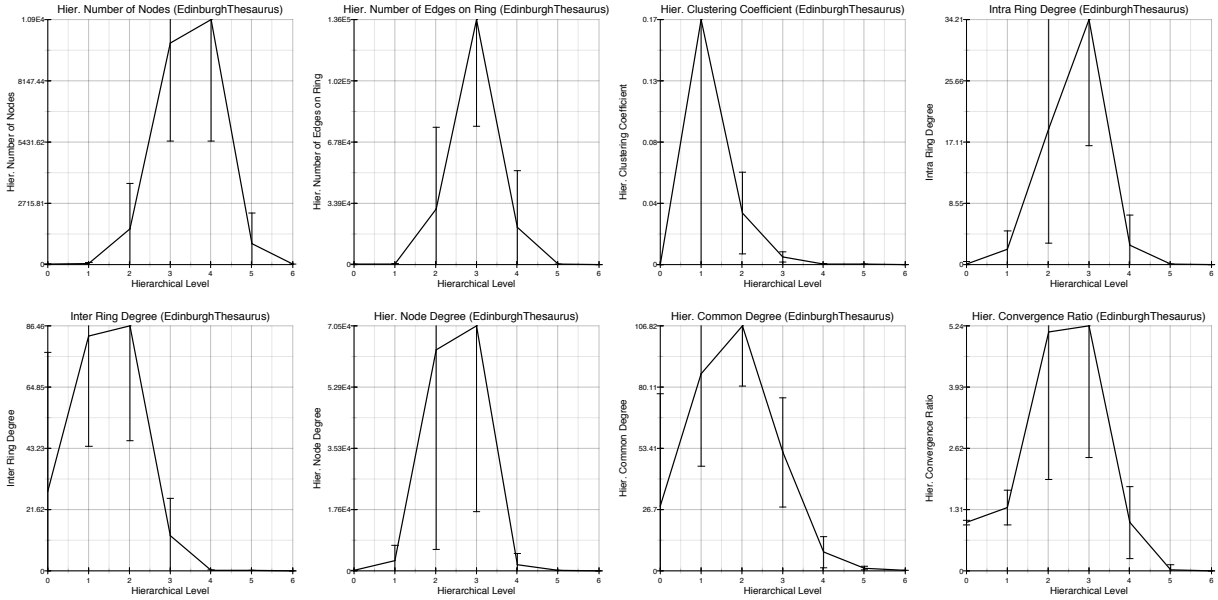


Figura 4.36 – Distribuição das propriedades concêntricas obtidas para a rede de associação de palavras de edinburgh com 23219 vértices e $\langle k \rangle \simeq 28$.

As propriedades concêntricas básicas não se diferenciaram dos outros modelos, apresentando uma curva com o pico característico muito semelhante àquele obtido para a rede BA.

As propriedades de coeficiente de aglomeração concêntrico, grau entre-níveis e grau comum e taxa de convêrgencia também apresentaram propriedades semelhantes às obtidas para a rede BA, no entanto com curvas de picos mais largos e deslocadas por até um nível.

A rede de associação de palavras é muito semelhante topologicamente à redes geradas pelo modelo BA em contrate com as outras redes reais estudadas, que apresentavam comportamento concêntrico composto por uma mistura de vários modelos.

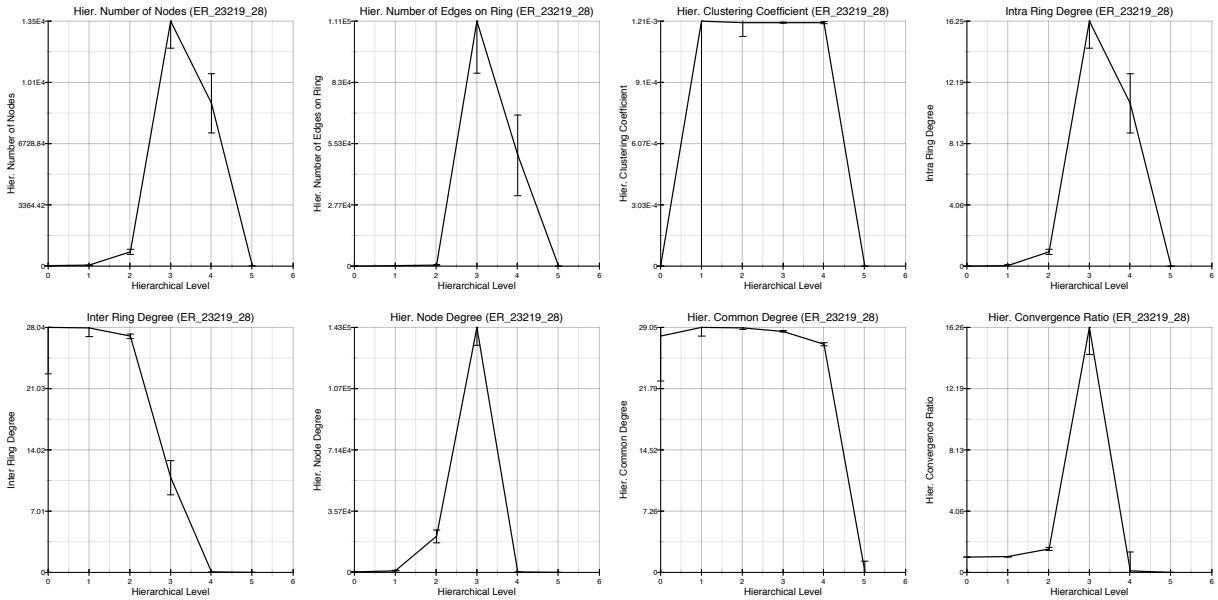


Figura 4.37 – Distribuição das propriedades concêntricas para a rede ER comparável à rede de associação de palavras.

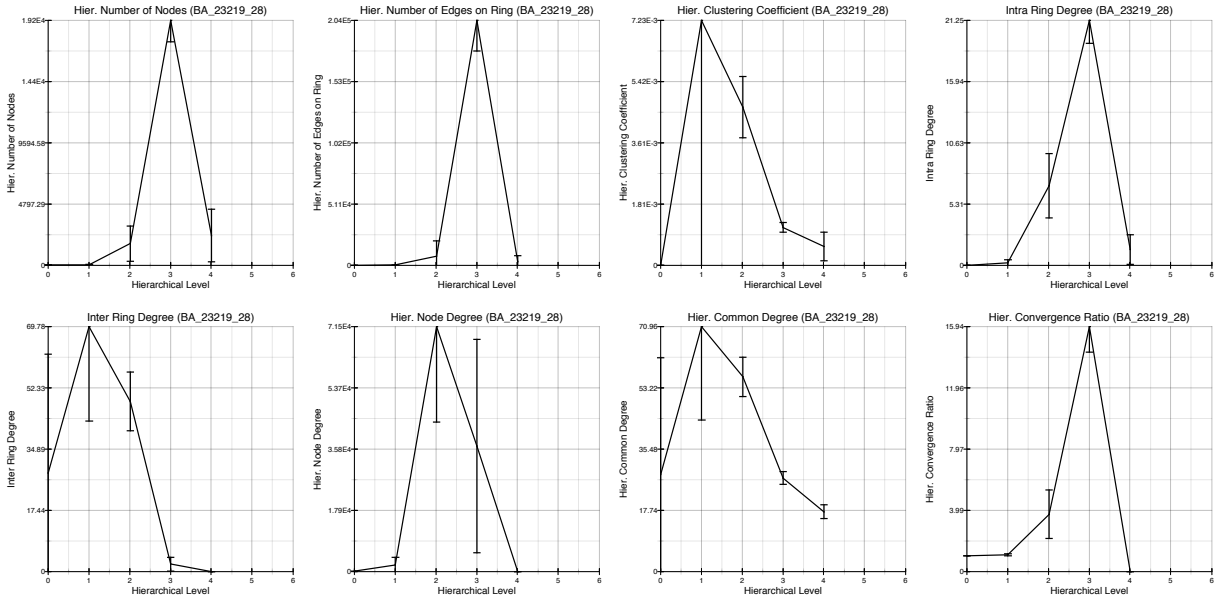


Figura 4.38 – Distribuição das propriedades concêntricas para a rede BA comparável à rede de associação de palavras.

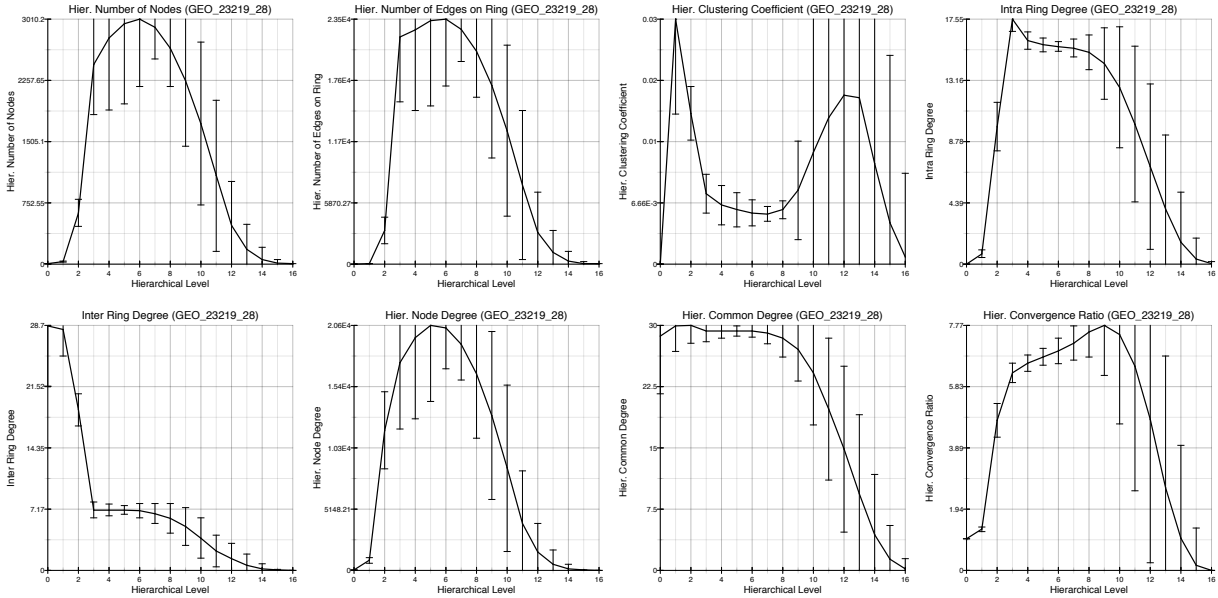


Figura 4.39 – Distribuição das propriedades concêntricas para a rede geográfca comparável à rede de associação de palavras.

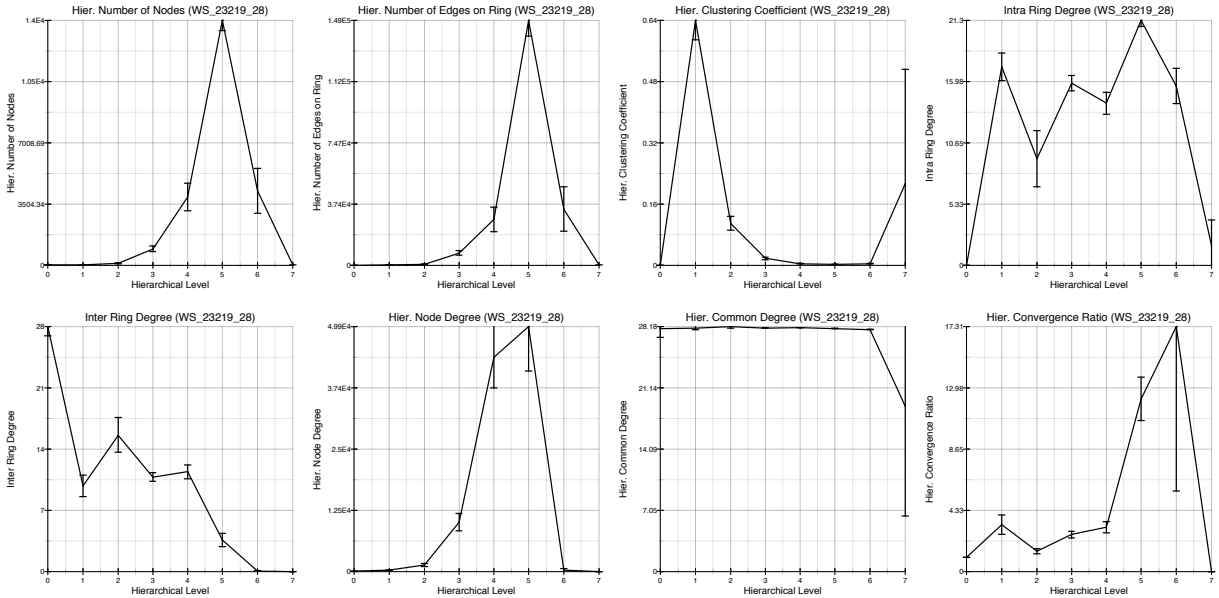


Figura 4.40 – Distribuição das propriedades concêntricas para a rede WS comparável à rede de associação de palavras.

Rede de proteínas, Yeast.

As curvas das distribuições obtidas para a rede de interação de proteínas, Yeast, podem ser vistas na figura 4.41. As curvas para as redes geradas seguindo os modelos teóricos, encontram-se nas figuras 4.42, 4.43, 4.44 e 4.45.

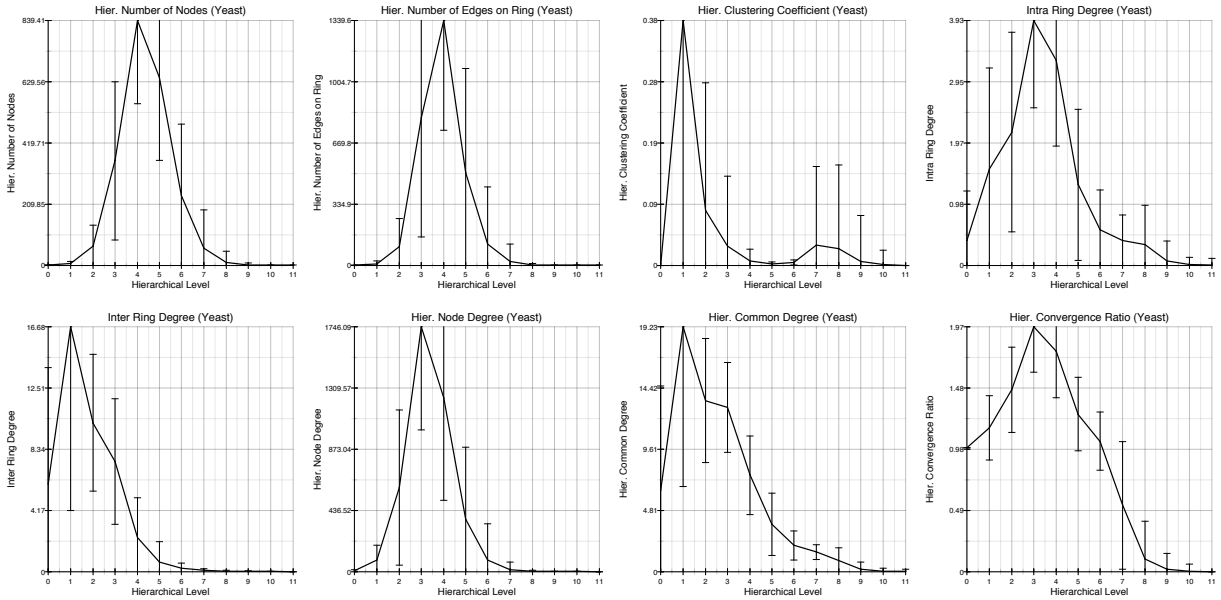


Figura 4.41 – Distribuição das propriedades concêntricas obtidas para a rede de proteínas com 2224 vértices e $\langle k \rangle \simeq 6$.

Assim como as outras redes reais, a rede de proteínas apresenta as propriedades básicas semelhantes, constituídas por um pico na região dos níveis centrais. Para esta rede, a largura do pico foi semelhante à obtida para os modelos ER e BA.

O coeficiente de aglomeração apresentou uma pequena elevação na região dos últimos níveis concêntricos, entretanto com alta variação, assemelhando-se à curva obtida para a rede BA. As outras propriedades concêntricas também apresentaram semelhanças com a rede gerada pelo modelo BA, no entanto a taxa de convergência apresentou uma curva com pico sutilmente mais largo e deslocado para a esquerda, revelando que o acesso aos primeiros hubs é mais rápido, mas com acesso aos outros gradualmente, quando comparado ao modelo BA.

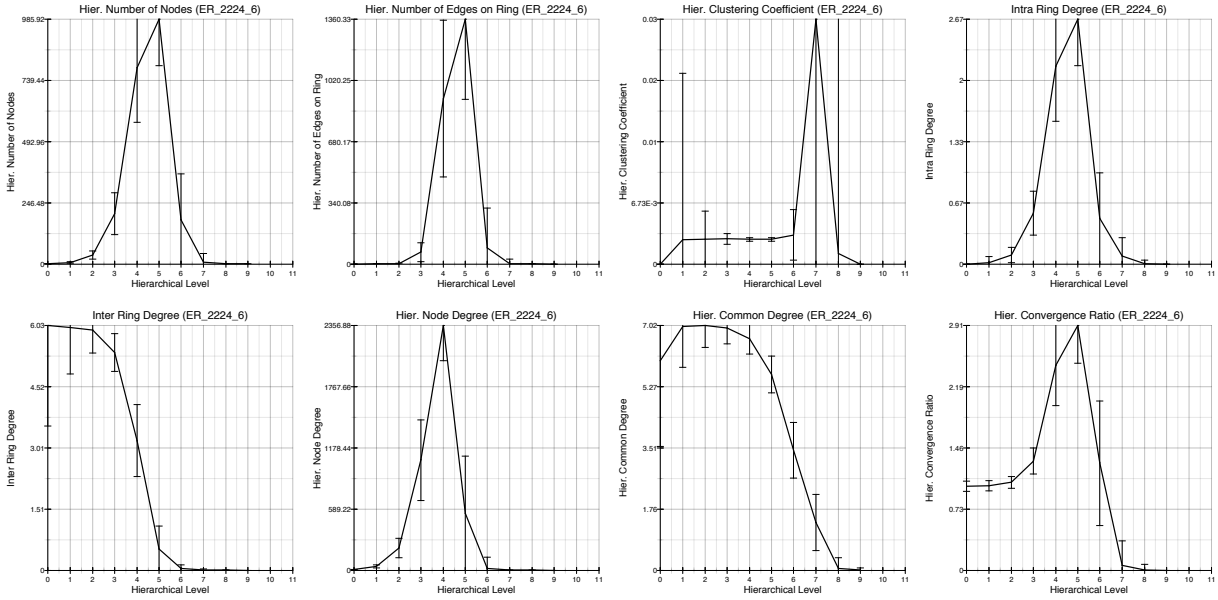


Figura 4.42 – Distribuição das propriedades concêntricas para a rede ER comparável à rede de proteínas.

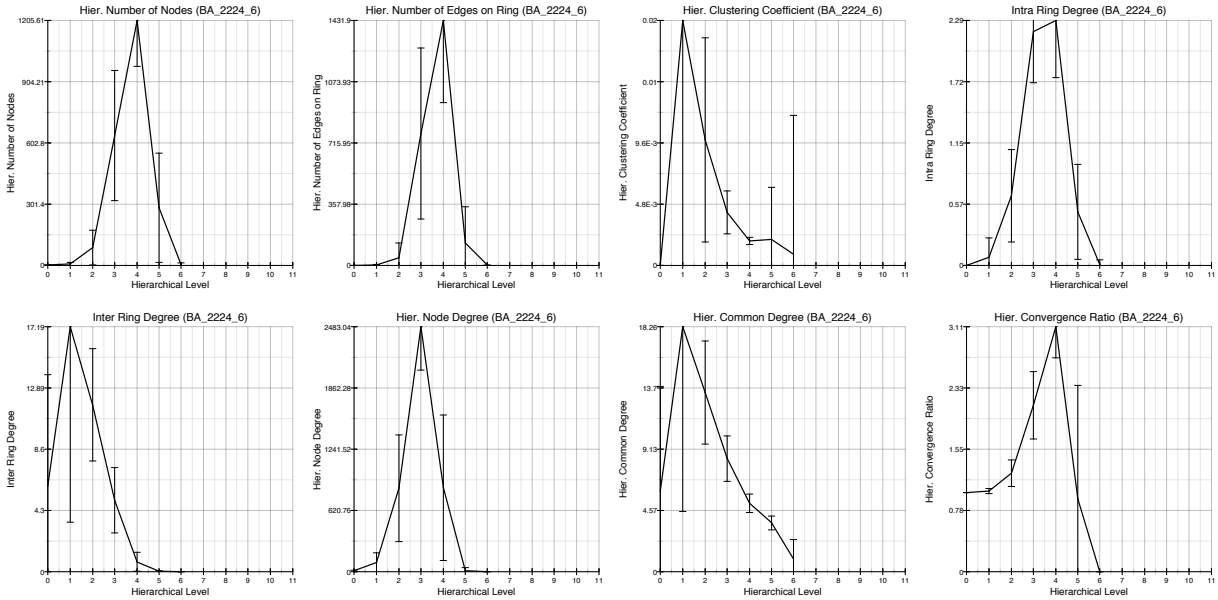


Figura 4.43 – Distribuição das propriedades concêntricas para a rede BA comparável à rede de proteínas.

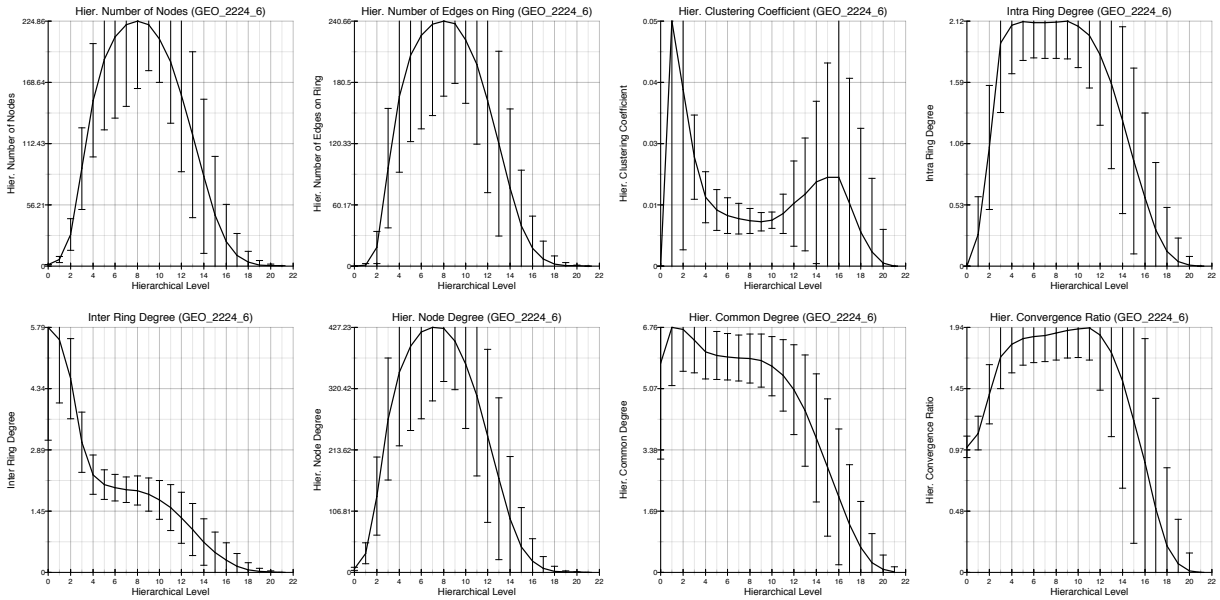


Figura 4.44 – Distribuição das propriedades concêntricas para a rede geográfca comparável à rede de proteínas.

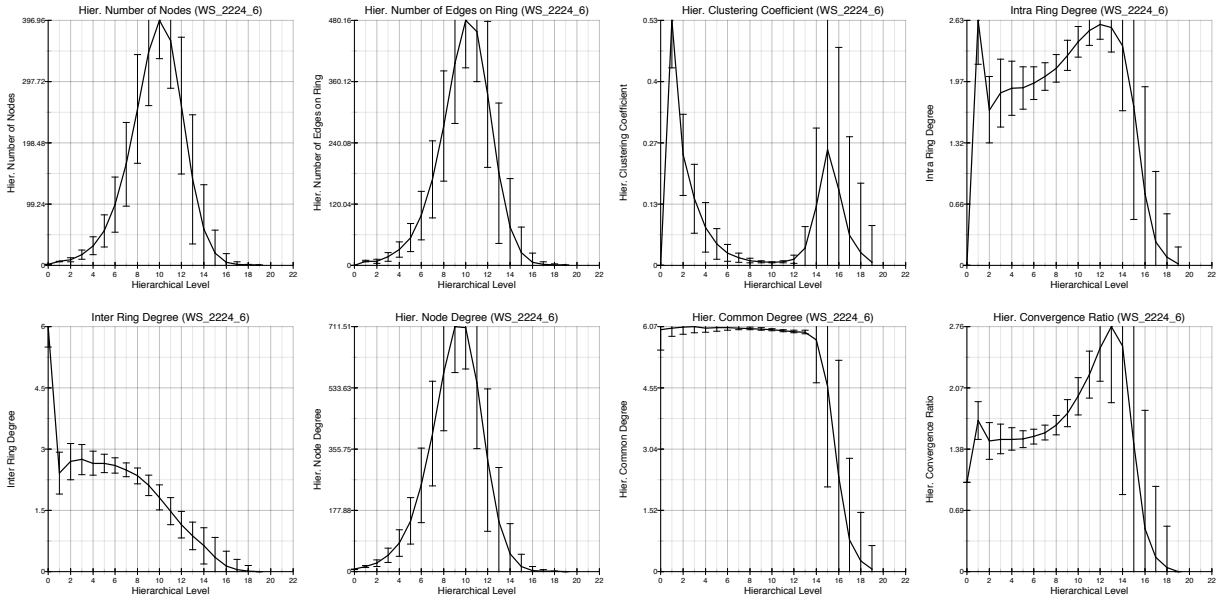


Figura 4.45 – Distribuição das propriedades concêntricas para a rede WS comparável à rede de proteínas.

Rede de alta tensão dos EUA.

Os gráficos das curvas das propriedades concêntricas obtidas para a rede de alta tensão dos EUA encontram-se na figura 4.46, àquelas obtidas para os modelos teóricos encontram-se nas figuras 4.47, 4.48, 4.49, 4.50.

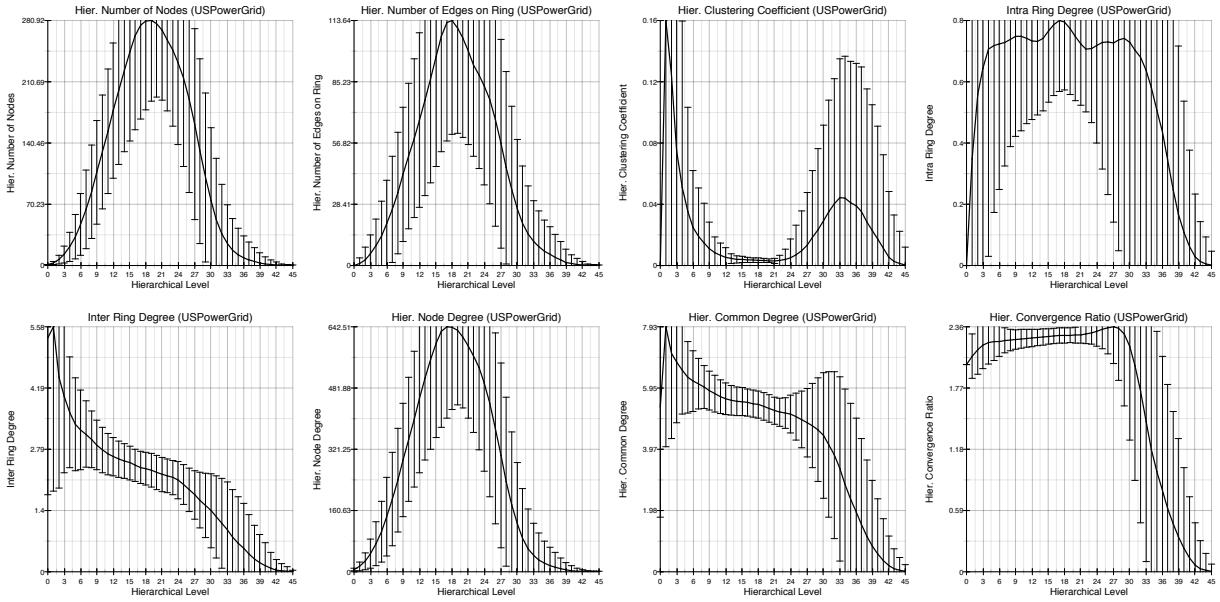


Figura 4.46 – Distribuição das propriedades concêntricas obtidas para a rede de alta tensão dos EUA com 4941 vértices e $\langle k \rangle \simeq 2.7$.

As distribuições das propriedades concêntricas apresentaram alto valor de desvio padrão ao longo dos níveis concêntricos. As propriedades básicas resultaram em distribuições caracterizadas por um pico largo, semelhante àquele obtido para a rede geográfica.

As curvas de distribuição para as demais propriedades concêntricas também apresentaram-se muito semelhantes àquelas obtidas para o modelo geográfico. De fato, era esperado que apresentasse características de redes geográficas devido a natureza de sua origem.

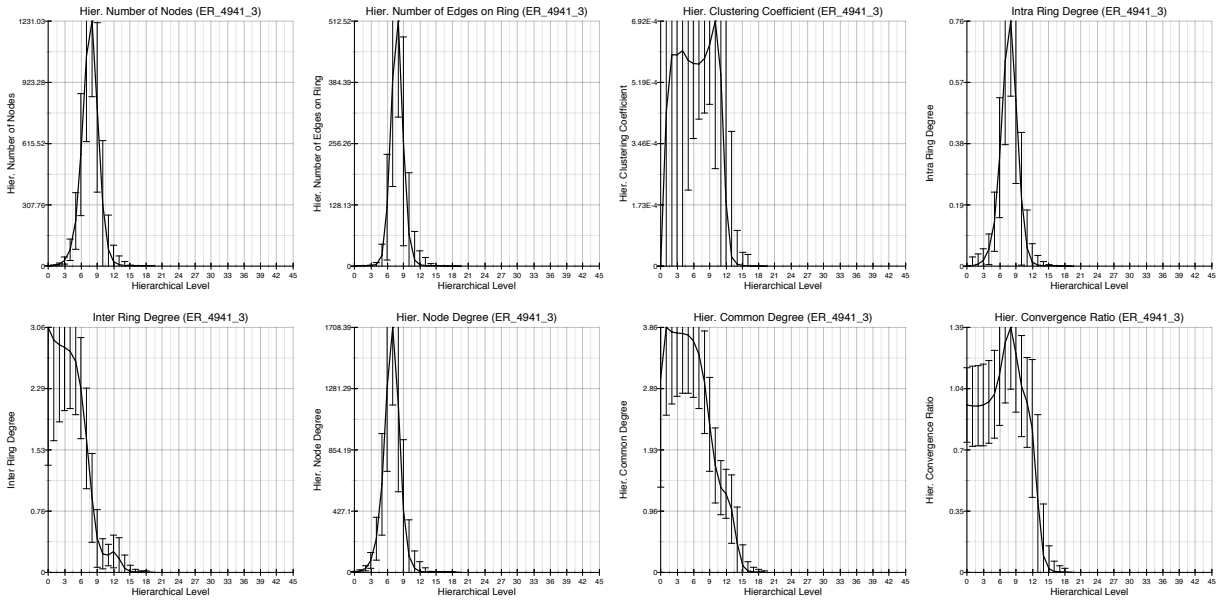


Figura 4.47 – Distribuição das propriedades concêntricas para a rede ER comparável à rede de alta tensão dos EUA.

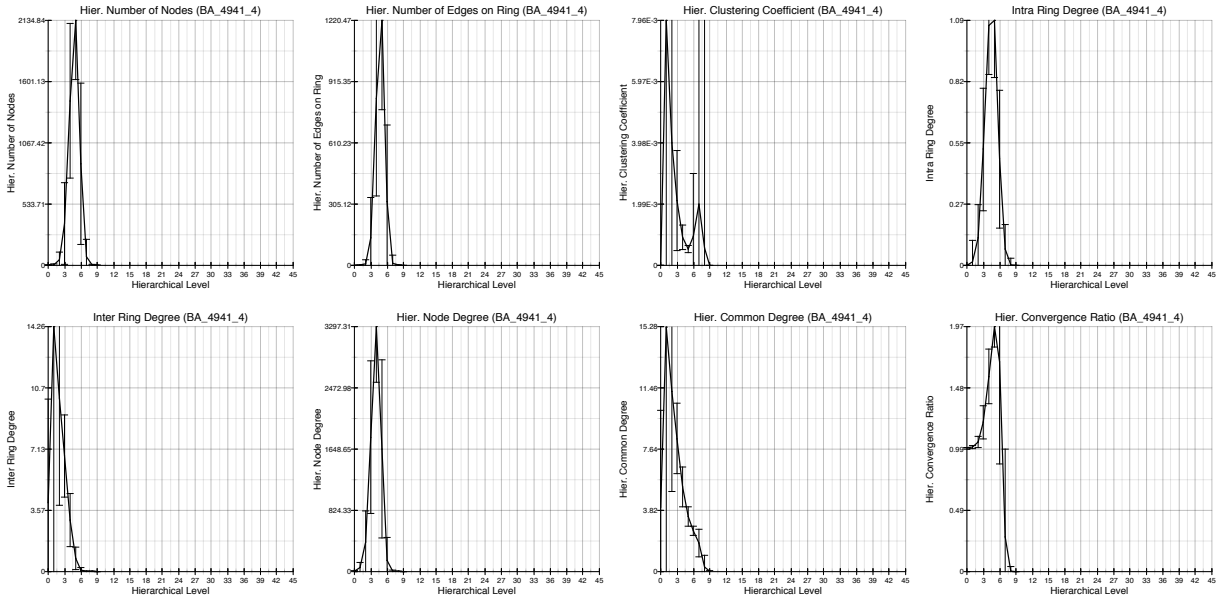


Figura 4.48 – Distribuição das propriedades concêntricas para a rede BA comparável à rede de alta tensão dos EUA.

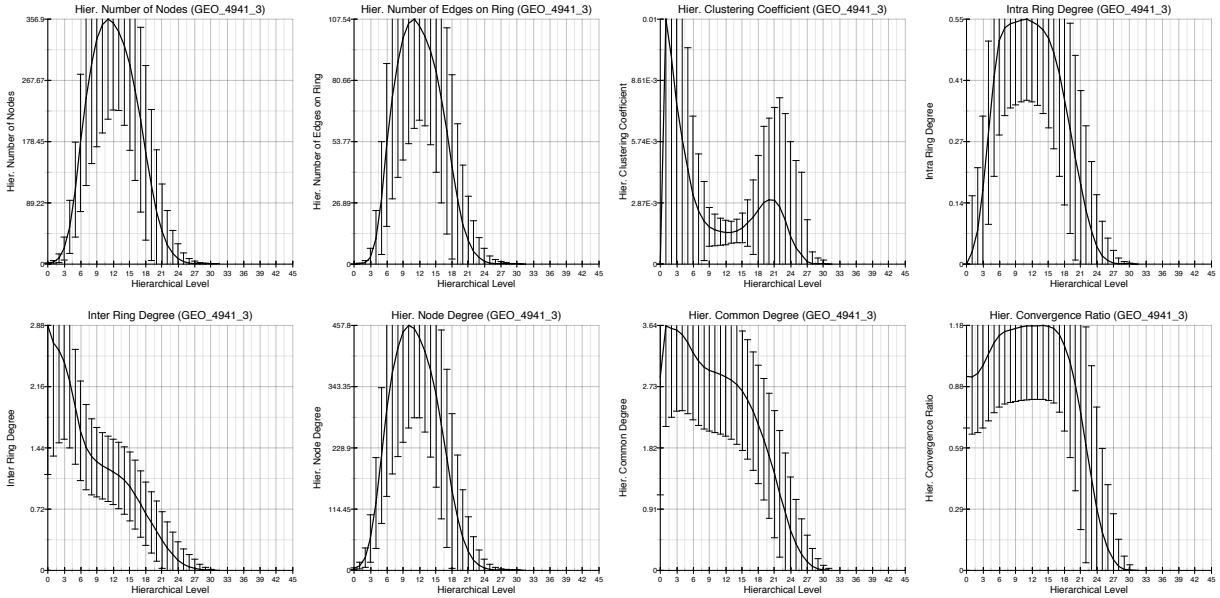


Figura 4.49 – Distribuição das propriedades concêntricas para a rede geográfca comparável à rede de alta tensão dos EUA.

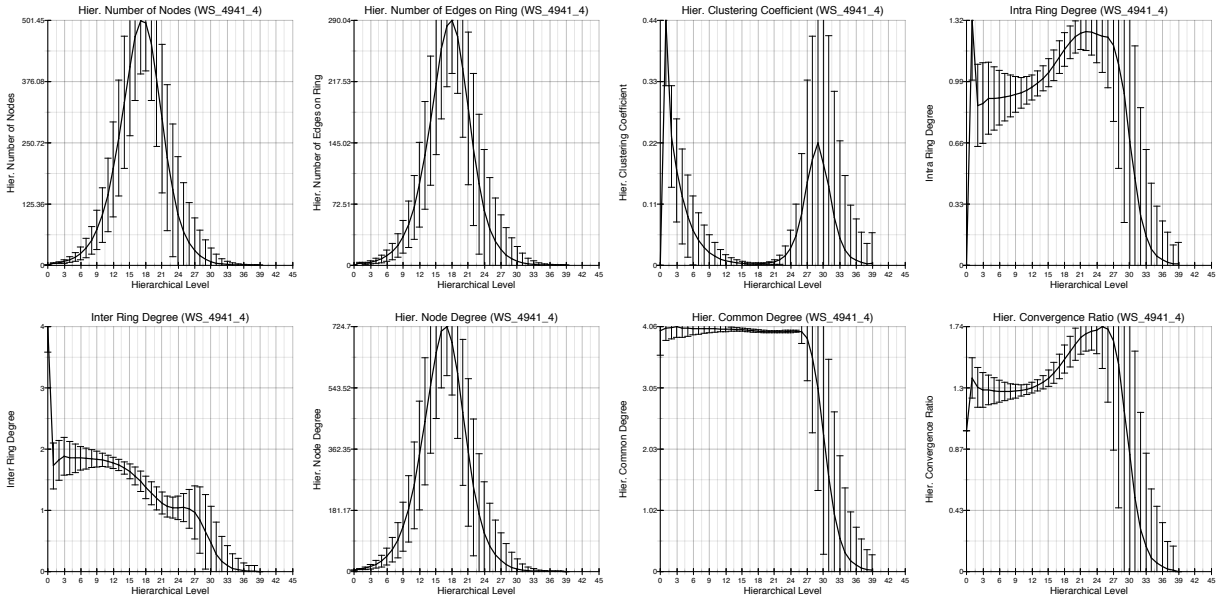


Figura 4.50 – Distribuição das propriedades concêntricas para a rede WS comparável à rede de alta tensão dos EUA.

Sub-rede da WWW dos resultados da busca ”California”.

As distribuições resultantes das propriedades concêntricas obtidas para a sub-rede da WWW encontram-se na figura 4.51. As curvas para as redes teóricas comparáveis à sub-rede da WWW, encontram-se nas figuras 4.52, 4.53, 4.54 e 4.55.

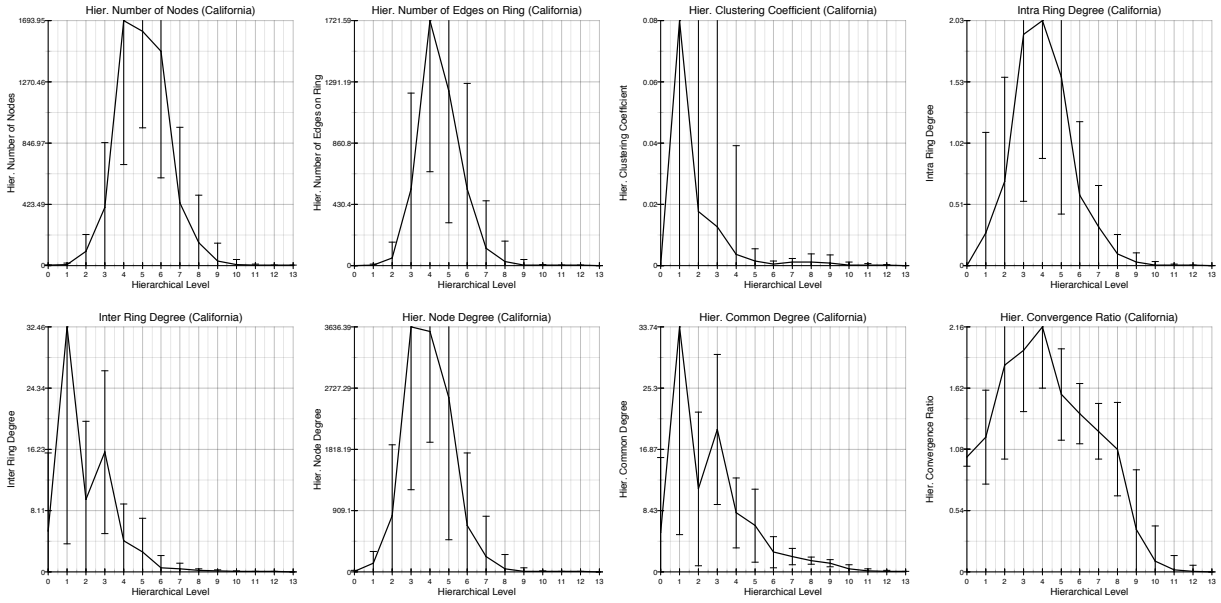


Figura 4.51 – Distribuição das propriedades concêntricas obtidas para a sub-rede da WWW com resultados da busca pelo termo ”California” com 5925 vértices e $\langle k \rangle \simeq 5.4$.

A rede da WWW apresentou curvas de distribuição das propriedades concêntricas básicas semelhantes entre si e, assim como para as outras redes, caracterizadas por um pico, entretanto deslocado na direção dos primeiros níveis concêntricos.

As outras propriedades apresentaram-se semelhantes àquelas obtidas para o modelo BA, com exceção para a taxa de convergência que resultou em uma curva caracterizada por um pico largo, indicando que vértices de maior conectividade são acessados gradualmente ao longo dos níveis concêntricos.

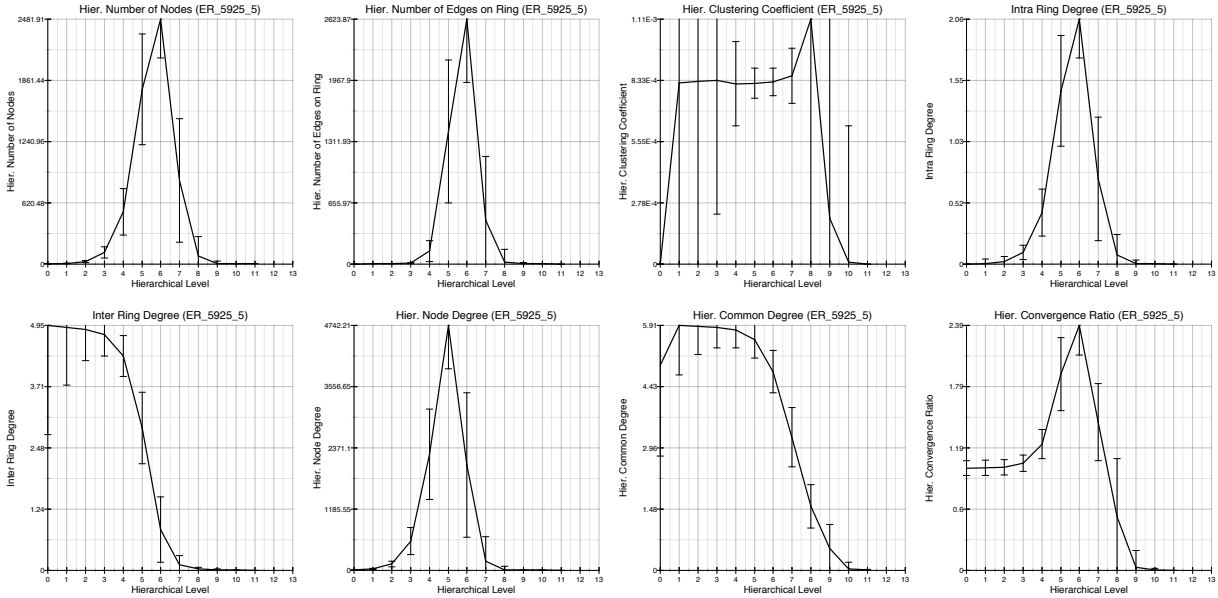


Figura 4.52 – Distribuição das propriedades concêntricas para a rede ER comparável à sub-rede da WWW.

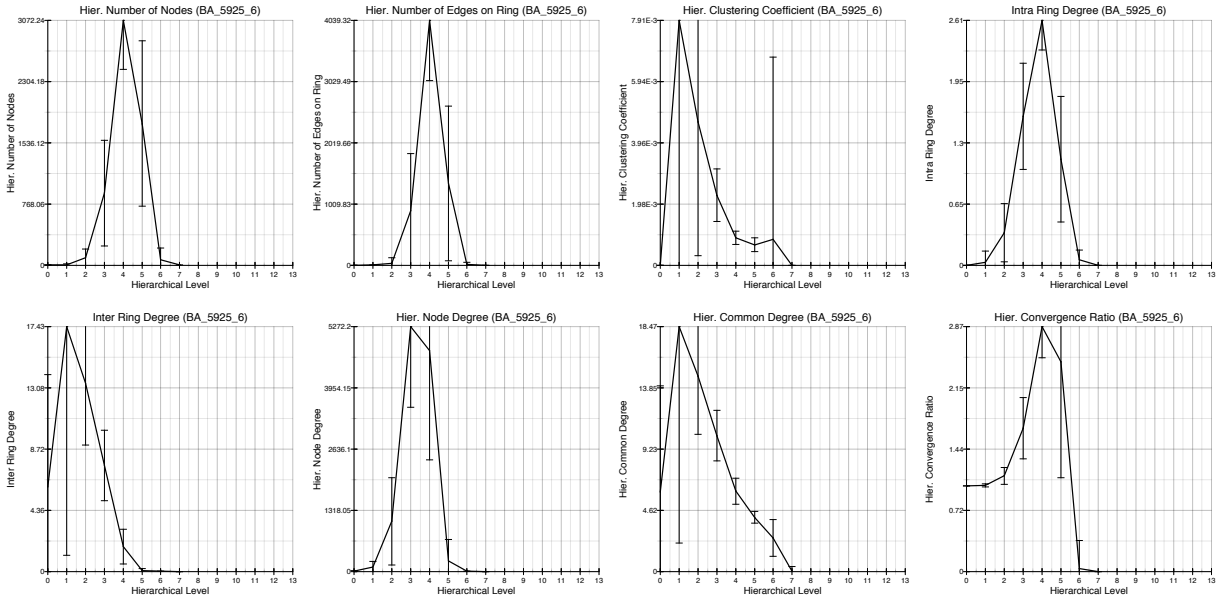


Figura 4.53 – Distribuição das propriedades concêntricas para a rede BA comparável à sub-rede da WWW.

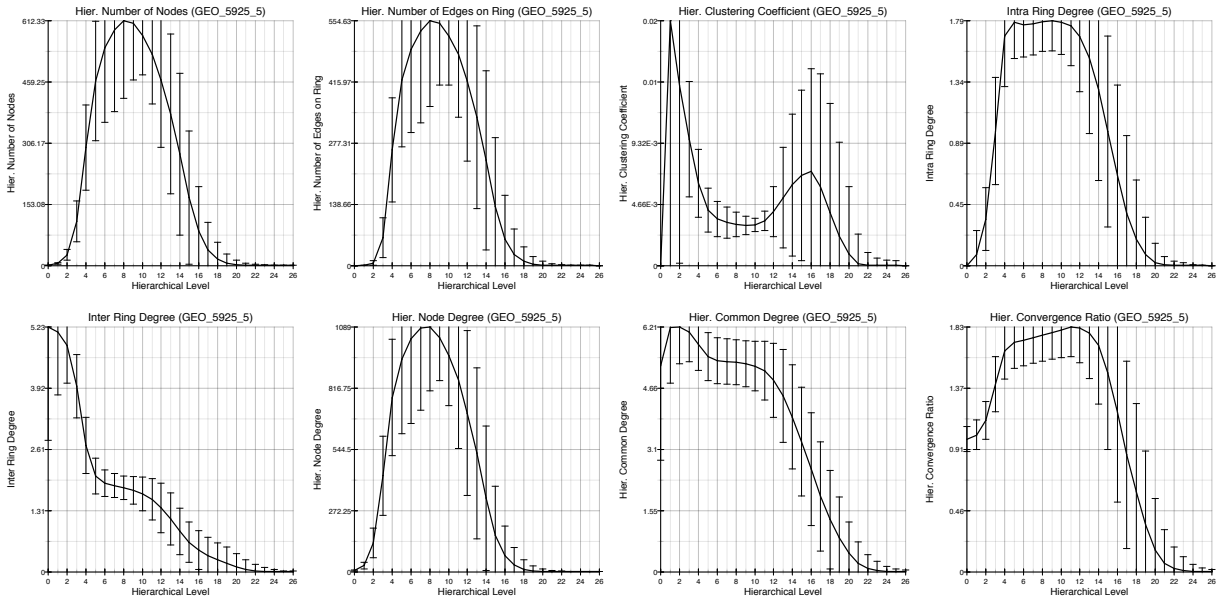


Figura 4.54 – Distribuição das propriedades concêntricas para a rede geográfca comparável à sub-rede da WWW.

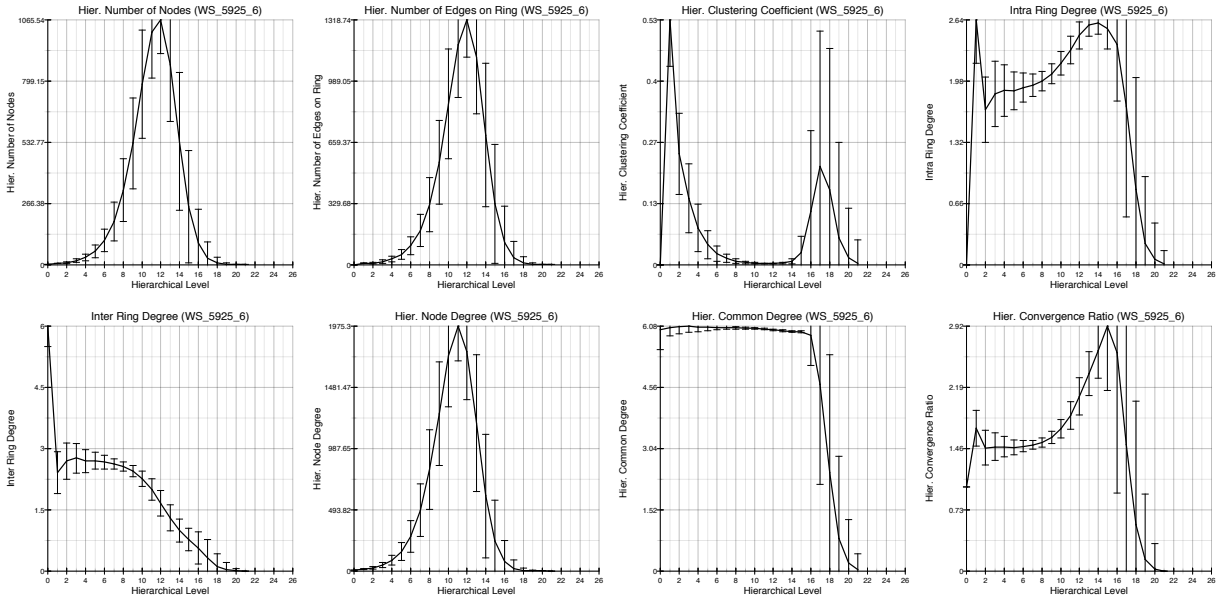


Figura 4.55 – Distribuição das propriedades concêntricas para a rede WS comparável à rede de sub-rede da WWW.

4.3.3 Caracterização dos vértices da rede de colaboração da USP

Os vértices do maior componente da rede de colaboração da USP foram caracterizados em termos de suas propriedades concêntricas através da metodologia de aglomeração hierárquica, obtendo um dendrograma dos grupos (ou aglomerados). A propriedade utilizada para essa análise foi o coeficiente de aglomeração concêntrico de cada vértice devido a alta variação de seus valores na distribuição, e, como métrica de distância, foi usado o coeficiente de correlação.

O dendrograma obtido para a rede encontra-se na figura 4.56, onde cada bifurcação indica uma subdivisão de aglomerados, com os valores de coeficiente de correlação aumentando na direção indicada. Com o objetivo de ilustrar a classificação dos vértices, o dendrograma foi limitado de modo que pudessem ser obtidas as 4 primeiras classes, como mostrado pelo corte pontilhado da figura. Cada bifurcação da árvore obtida foi associado a um grupo obtendo-se a taxinomia completa desses vértices considerando o coeficiente de aglomeração.

A figura 4.57 apresenta, em detalhes, as distribuições do coeficiente de aglomeração concêntrico obtidos apenas paras os vértices pertencentes a cada grupo considerado, nomeados assim como no dendrograma da figura 4.56. Para cada ramo também são indicados o número de vértices e a representação percentual dos vértices da rede.

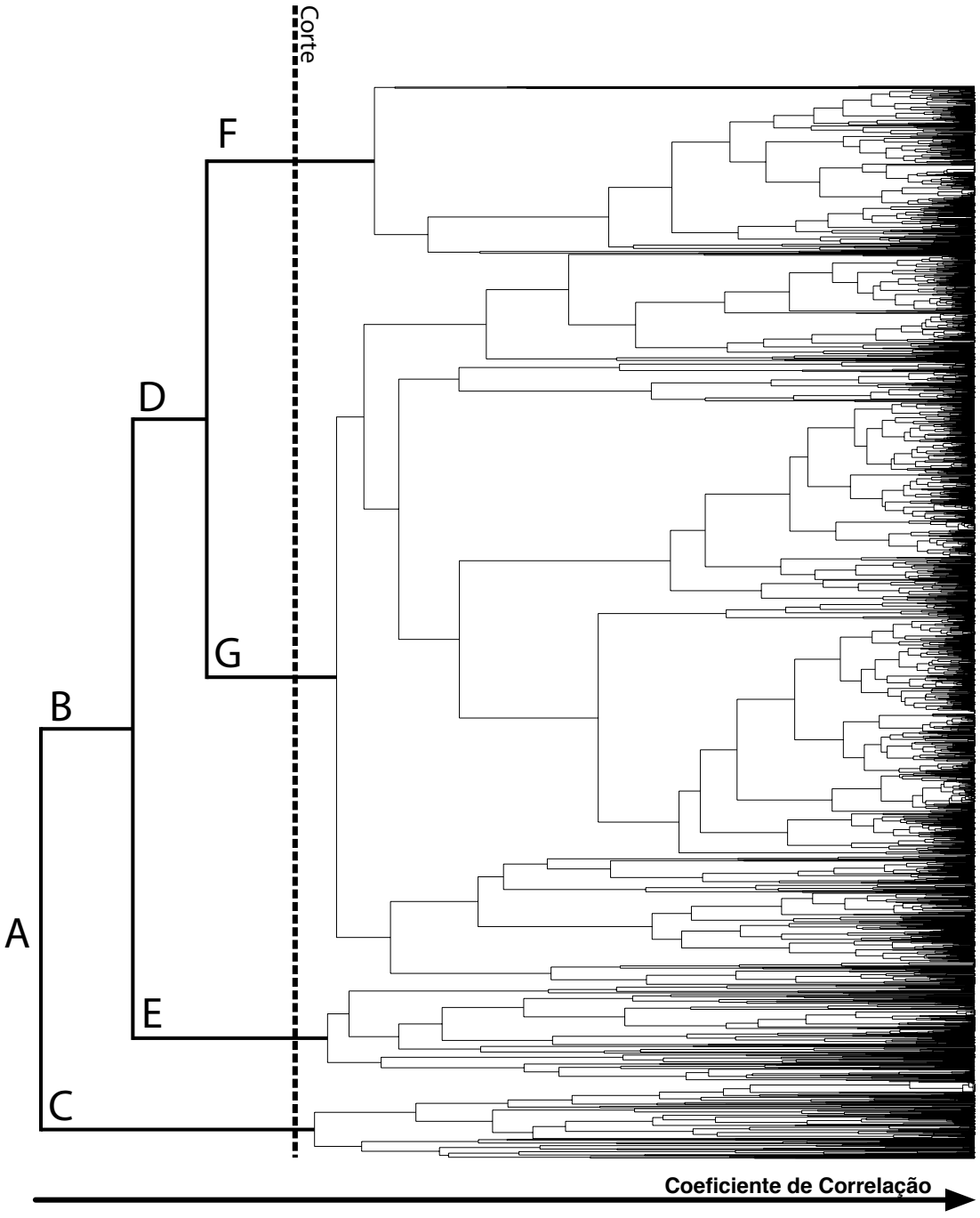


Figura 4.56 – dendrograma obtido pela aplicação do método de aglomeração hierárquica à rede de colaboração da USP. O corte e os 4 primeiros grupos são indicados, assim como a direção e sentido do aumento da medida de distância utilizada, coeficiente de correlação.

Verifica-se que a região das curvas de distribuição que mais se diferenciou entre os grupos foi aquela próxima dos últimos níveis concêntricos, a mesma que, na figura 4.21 da seção anterior, apresentou valores altos de desvio padrão.

A raiz da árvore, *A*, representa, a rede completa que, inicialmente, divide-se em dois grupos, sendo o grupo *B*, muito maior que o *C*, em termos de número de vértices. Diferentemente das distribuições obtidas para os outros grupos, os grupos *B* e *C* se distinguem pela curva de *B* apresentar um vale na região dos níveis concêntricos centrais, enquanto grupo *C* apresenta uma crista. O grupo *B* divide-se em dois outros grupos, *D* e *E*, com o grupo *E* caracterizado por diversas ondulações para os últimos níveis concêntricos, enquanto o grupo *D* permanece com o mesmo padrão que o grupo *B*. Por fim, o grupo *D* divide-se em dois ramos, *F* e *G*, com o grupo *G* caracterizado por uma elevação muito menor do que aquela obtida para o grupo *F* nos últimos níveis concêntricos.

As figuras 4.58, 4.60 e 4.62, apresentam, através de gráficos de pizza, a *representação percentual das categorias dos vértices*, respectivamente, por unidade, cidade e área do conhecimento para cada grupo considerado. Analogamente, as figuras 4.59, 4.61 e 4.63, apresentam a *distribuição percentual* de cada categoria ao longo dos grupos considerados.

A representação dos vértices do grupo *A* nas figuras de pizza, são as mesmas daquelas obtidas na seção 4.1 para o maior componente conectado. Observando os ramos *B* e *C*, verifica-se que o grupo *D* ainda representa grande parte da rede, com 93% dos vértices, entretanto, o grupo *C*, com apenas 7% da rede representa 50% de todos os vértices da área de humanas, constituído por 88% dos vértices da unidade FEA (Faculdade de Economia e Administração) e de vértices da EESC (Escola de Engenharia de São Carlos) e EP (Escola Politécnica).

Os ramos *D* e *E* apresentam considerável diferença quanto à distribuição dos vértices, com o grupo *E* caracterizado por 18% dos pesquisadores da área de exatas, representado por 41% daqueles da pertencentes à unidade EP, enquanto que o grupo *D* apresenta grande parte dos outros institutos e 84% dos vértices da rede. Dos quatro grupos de vértices obtidos para os últimos ramos da árvore considerada, aquele com maior quantidade de vértices é o grupo *G*, com 64% dos vértices e caracterizado, assim como a rede completa, pela maioria dos vértices pertencentes à área de biologia. Já o grupo F, apresentou representativa quantidade de vértices da área de biologia e exatas.

Unidades da cidade de São Paulo estão mais distribuídos pelos grupos do que outras unidades, como Pirassununga e Ribeirão Preto. Enquanto o grupo G apresenta 84% dos vértices de Ribeirão Preto e 94% daqueles de Pirassununga, apresenta apenas 62% dos vértices de São Paulo.

Estes resultados corroboram aqueles obtidos pela visualização da rede, apresentados na seção 4.1. Os unidades da área de biologia tendem a estar mais próximos do centro do maior grupo, enquanto que humanas e exatas tendem a estar na região das bordas.

Uma visualização da rede indicando os vértices de cada grupo obtido foi gerada e é apresentada na figura 4.64. Como esperado, o maior grupo, *G*, representa grande parte do núcleo da rede, englobando boa parte da área de biológicas, como mostrado na figura 4.7. Nota-se que os grupos menores correspondem a vértices cada vez mais distantes do centro, com o grupo *C*, representando a parte mais externa da rede e, como consequência, grande parte dos vértices da área de humanas. Vértices pertencentes ao grupo *E* estão situados na região na região de borda imediata ao grupo *G*, correspondendo a alguns vértices da área de exatas. O grupo *F* apresenta alguns vértices do aglomerado central, mas principalmente aqueles mais externos, também representando alguns vértices da área de exatas.

O corte mais profundo do dendrograma resultará em um maior número de classes, que podem caracterizar melhor o núcleo central, entretanto tal análise foge do escopo desse trabalho que tem como objetivo apenas ilustrar a metodologia.

Com a finalidade de comparação com as medidas concêntricas, foram obtidas outras métricas de caracterização dos vértices, como as medidas de centralidade. Entretanto, as distribuições para as diferentes categorias de vértices praticamente se sobrepõem quando normalizadas, resultando em curvas como a da figura 4.65, que apresenta a distribuição logarítmica da distribuição da centralidade de subgrafos obtida para a rede de colaboração, considerando as diferentes áreas do conhecimento. Apesar disso, é importante ressaltar que as métricas de centralidade são importantíssimas para a caracterização de vértices segundo uma hierarquia de importância na rede, quando se diz respeito ao número de caminhos fechados ou fluxo que atravessa um vértice.

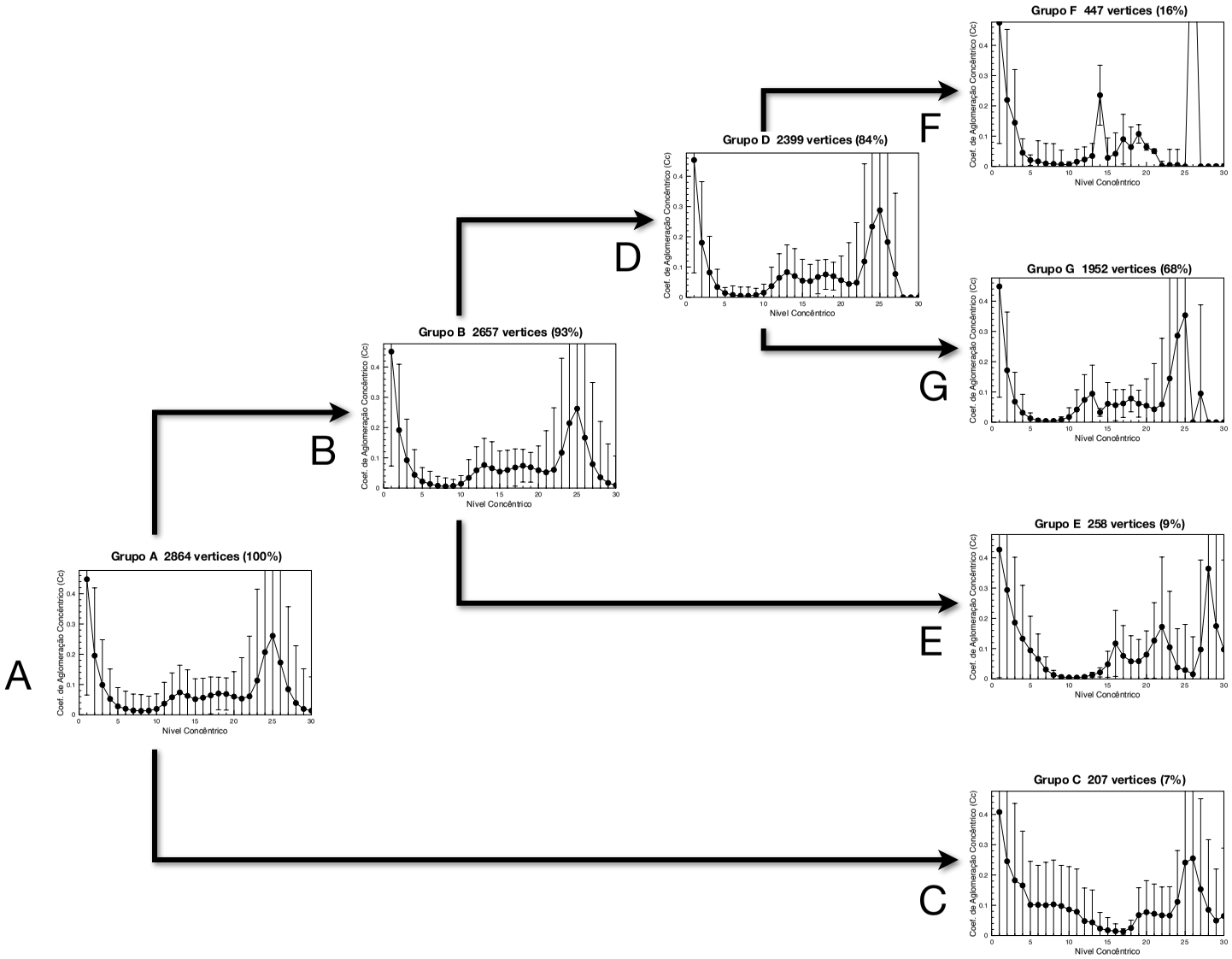


Figura 4.57 – Distribuições do coeficiente de aglomeração concêntrico obtido para cada grupo determinado pela análise de aglomeração hierárquica para os vértices da rede de colaboração da USP.

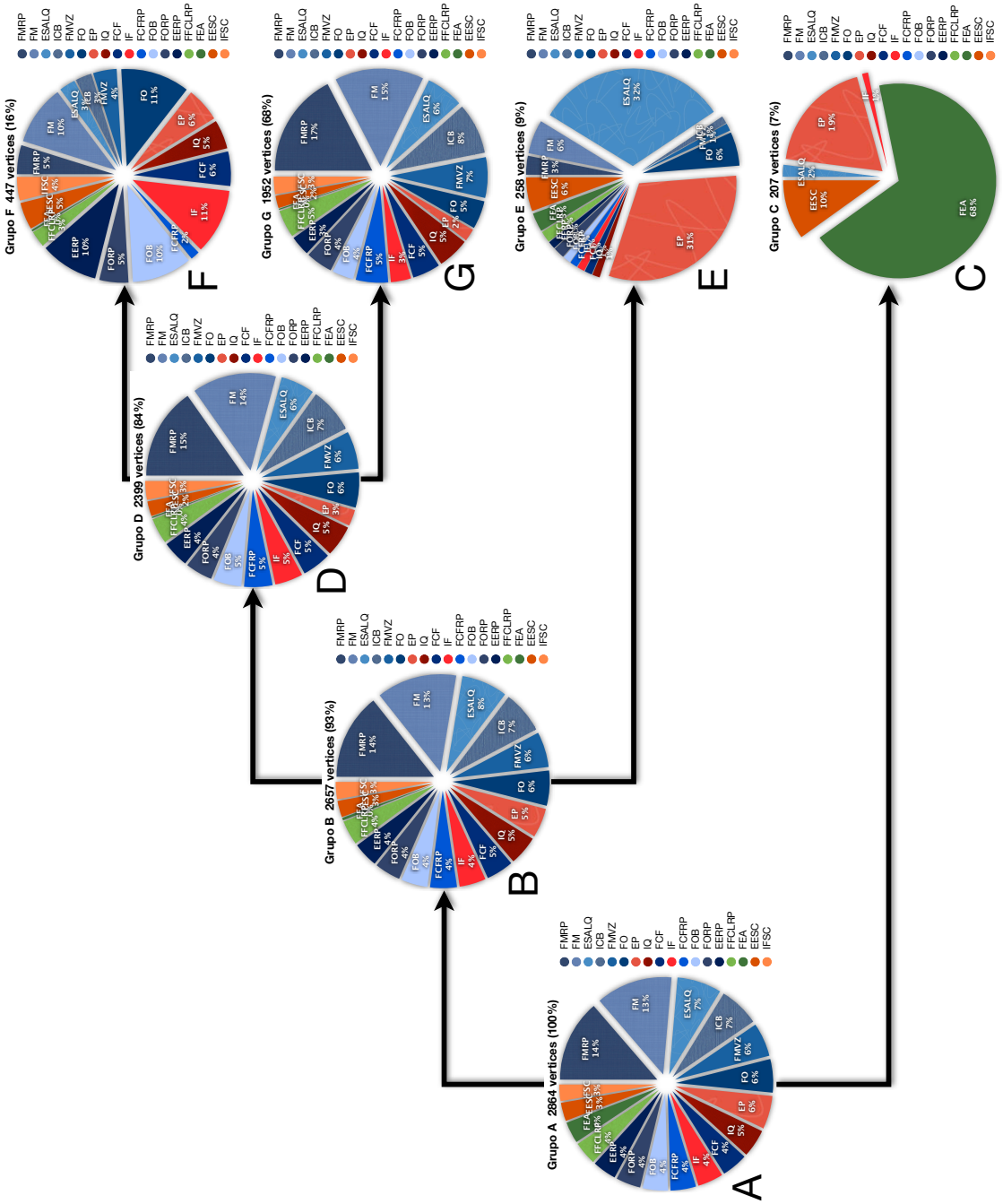


Figura 4.58 – Representações percentuais dos vértices das categorias de *unidades* em cada grupo obtido pela análise de aglomeração hierárquica para os vértices da rede de colaboração da USP.

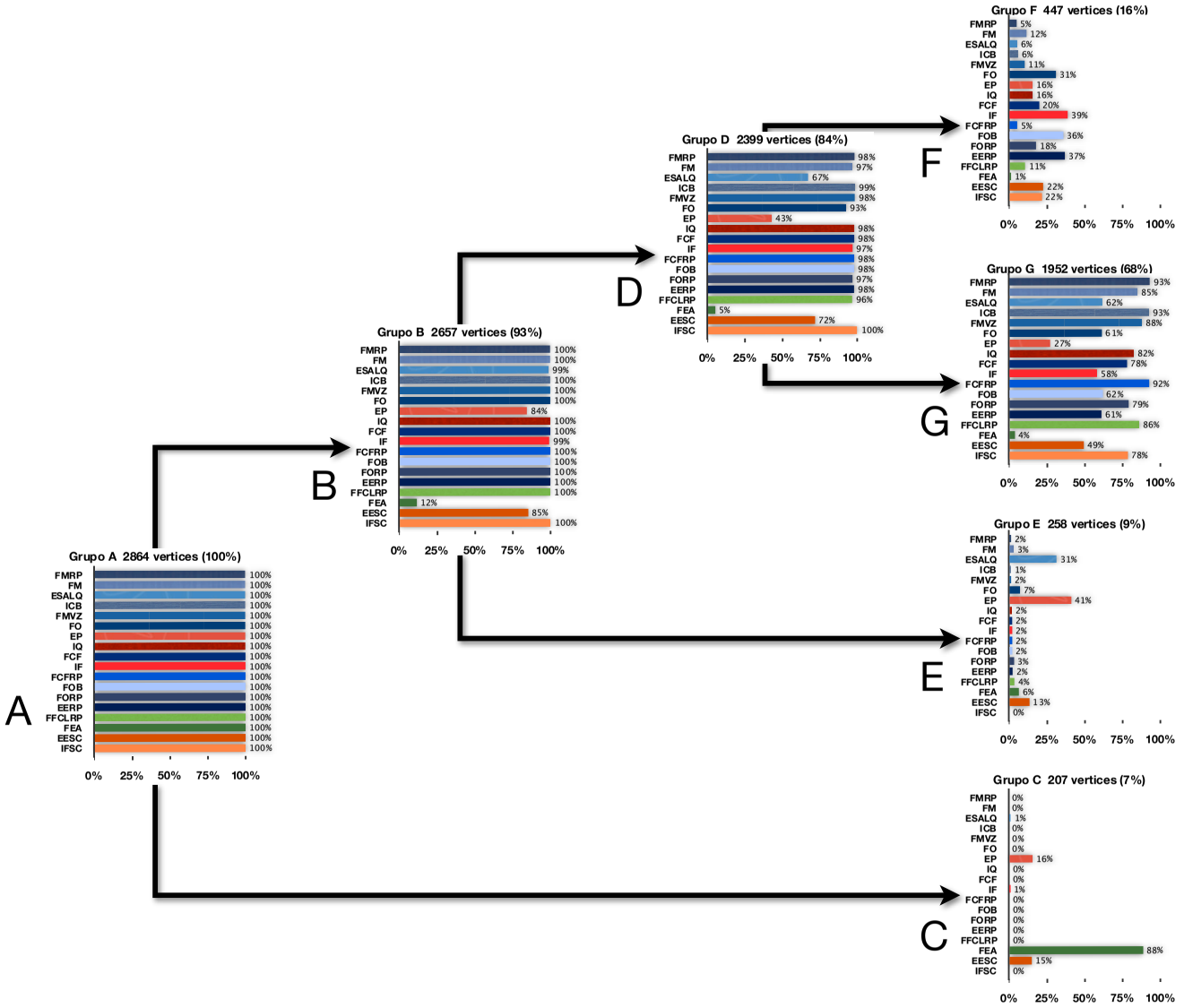


Figura 4.59 – Distribuição dos vértices de cada *unidade* para cada grupo obtido pela análise de aglomeração hierárquica para os vértices da rede de colaboração da USP.

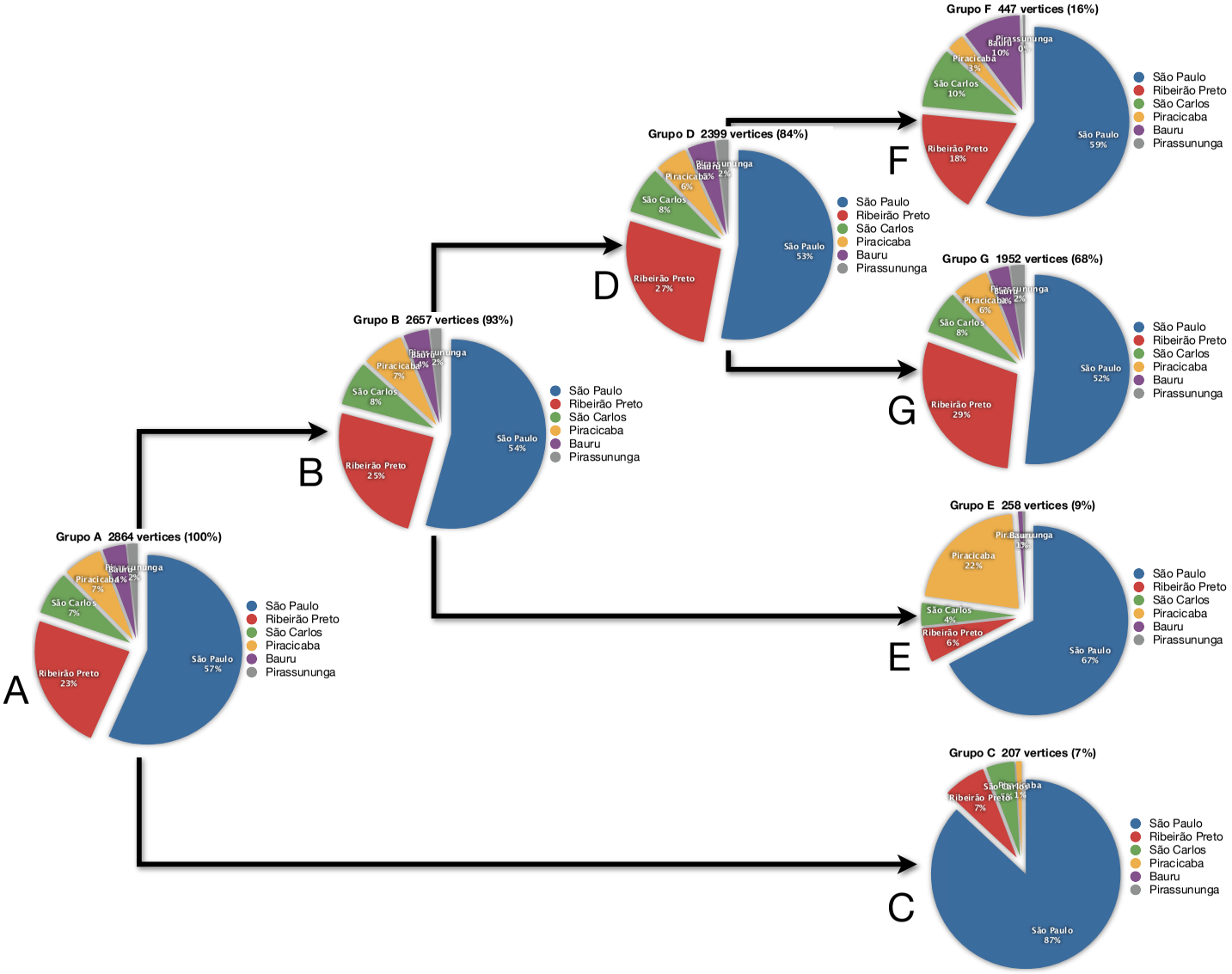


Figura 4.60 – Representações percentuais dos vértices das categorias de *cidades* em cada grupo obtido pela análise de aglomeração hierárquica para os vértices da rede de colaboração da USP.

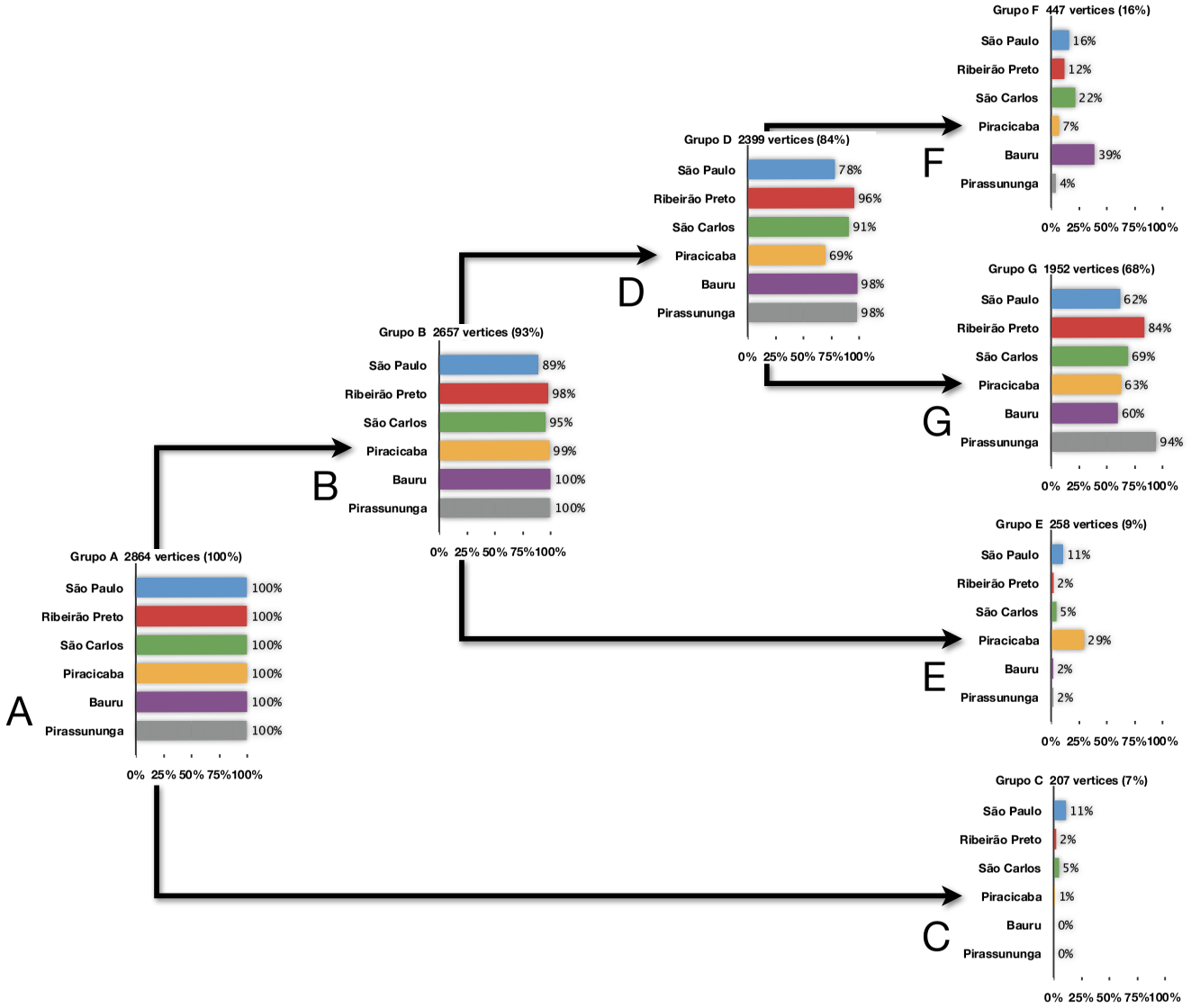


Figura 4.61 – Distribuição dos vértices de cada *cidade* para cada grupo obtido pela análise de aglomeração hierárquica para os vértices da rede de colaboração da USP.

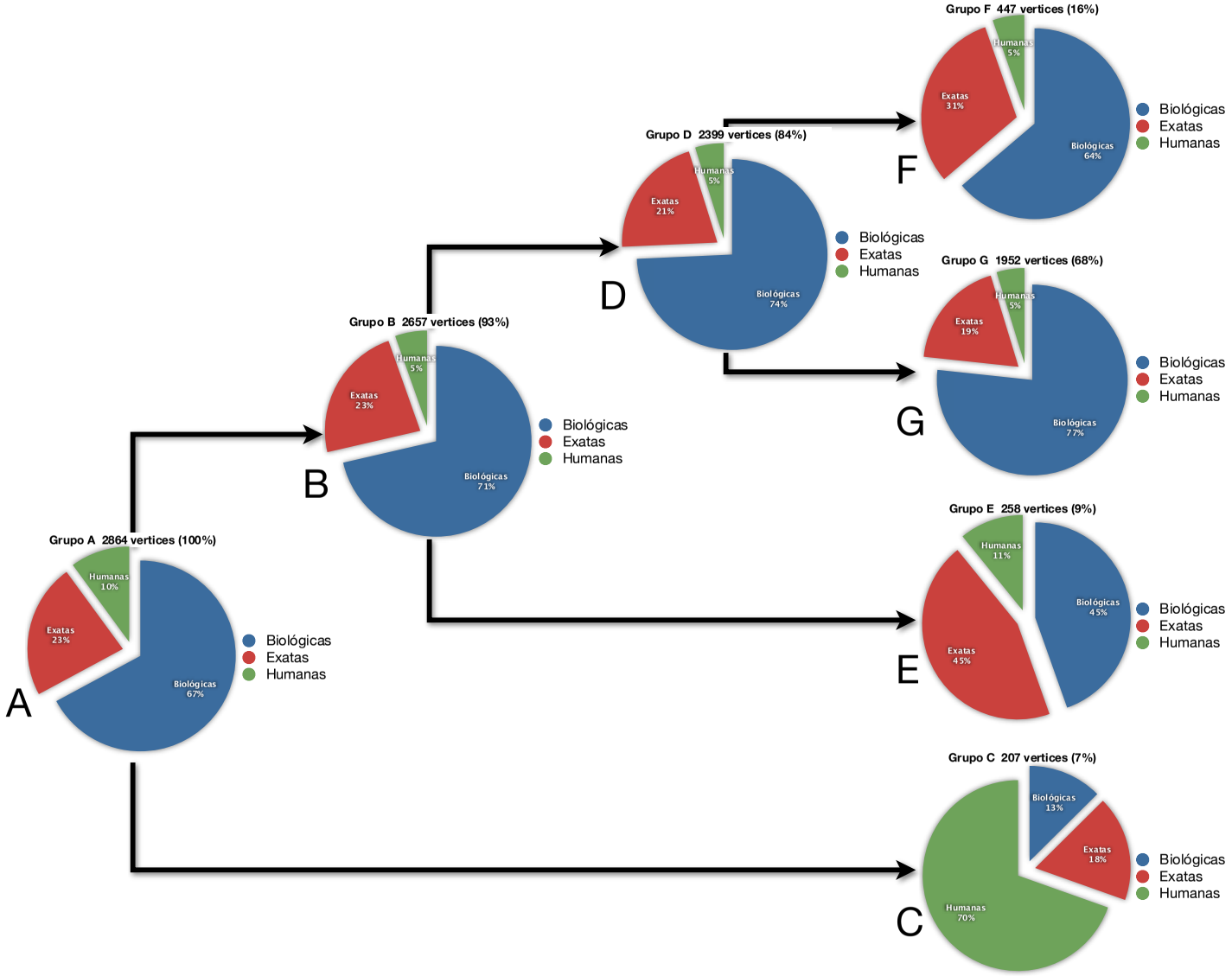


Figura 4.62 – Representações percentuais dos vértices das categorias de *áreas do conhecimento* em cada grupo obtido pela análise de aglomeração hierárquica para os vértices da rede de colaboração da USP.

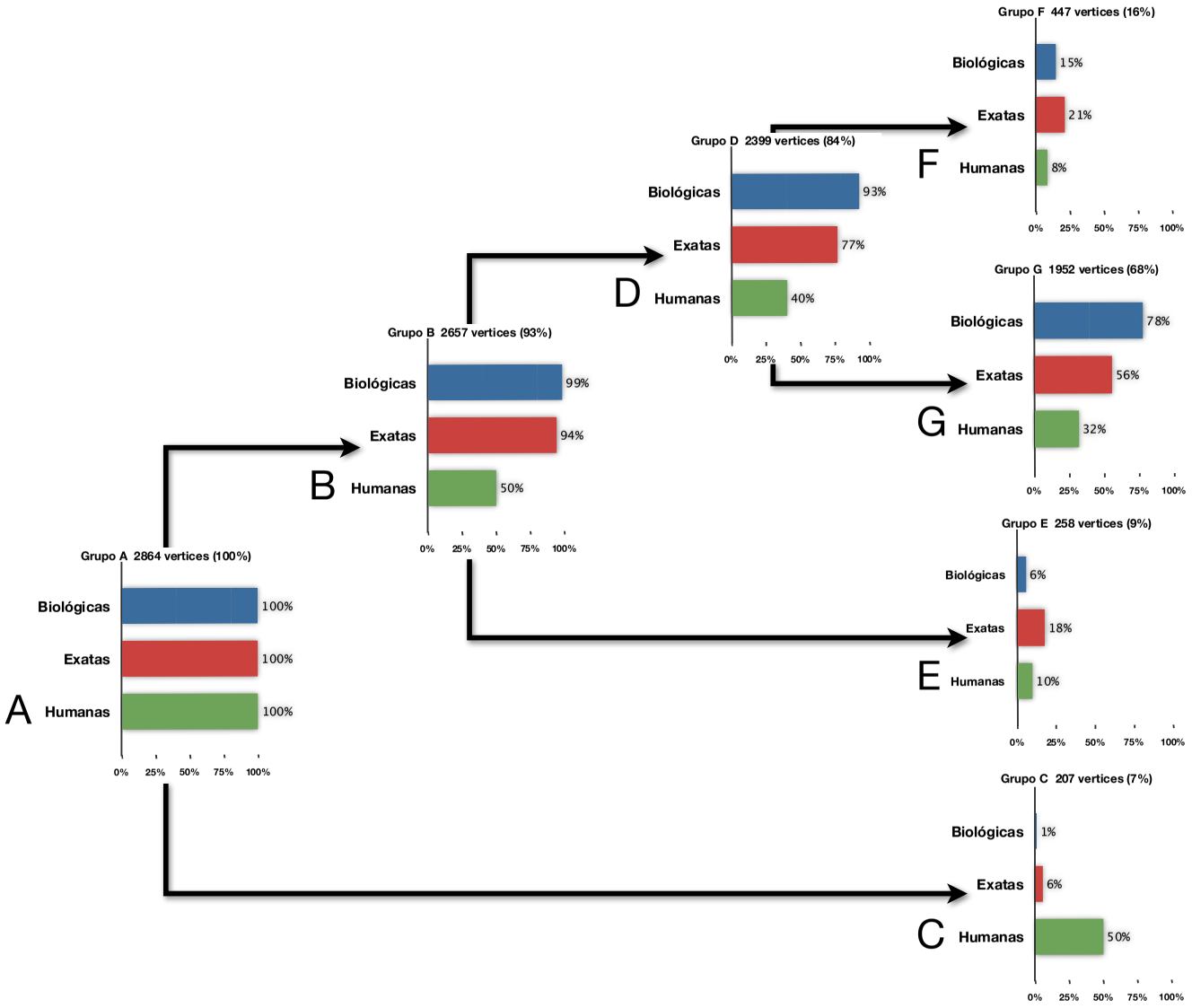


Figura 4.63 – Distribuição dos vértices de cada *área do conhecimento* para cada grupo obtido pela análise de aglomeração hierárquica para os vértices da rede de colaboração da USP.

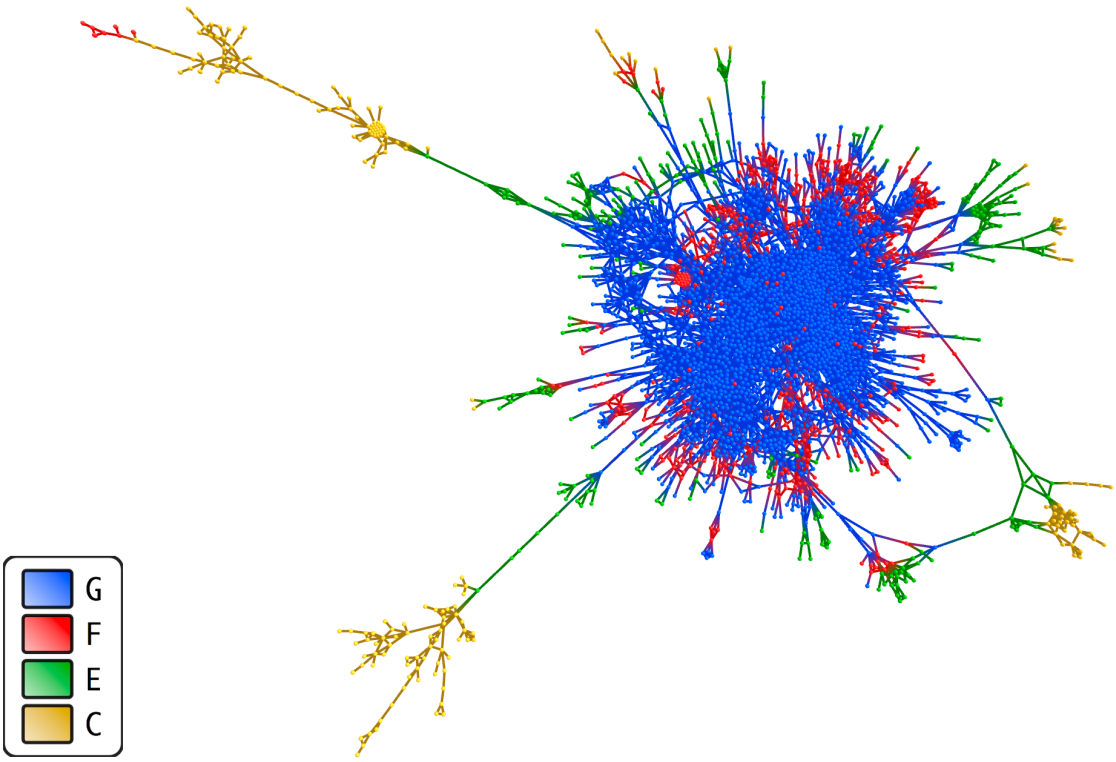


Figura 4.64 – Rede de colaboração da USP indicando os grupos obtidos pela análise de aglomeração hierárquica aplicada ao coeficiente de aglomeração concêntrico.

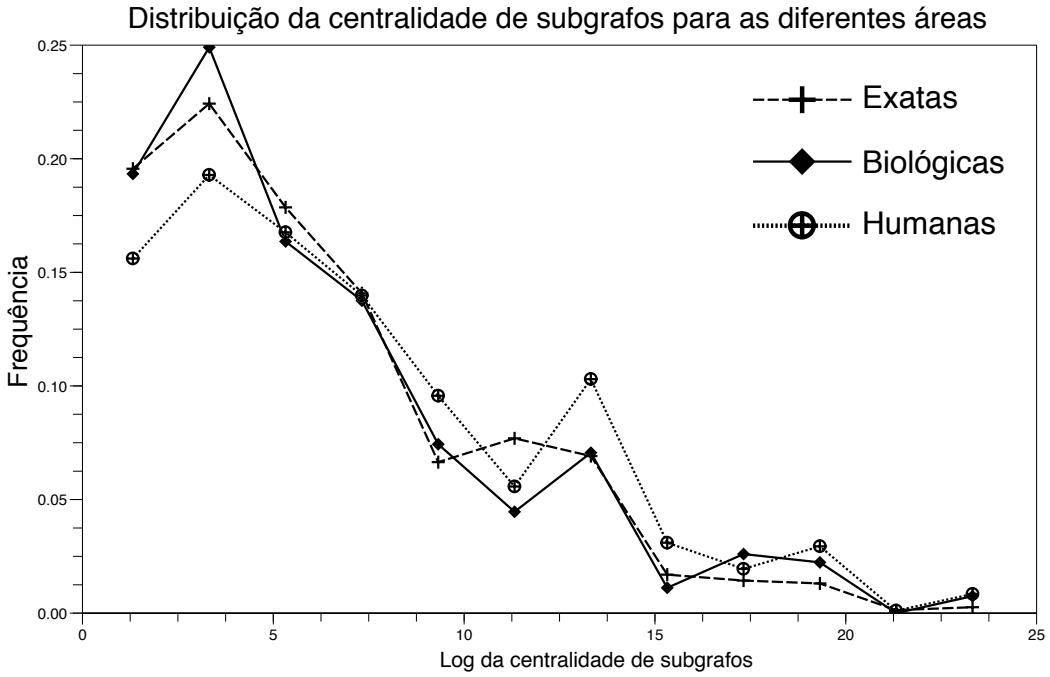


Figura 4.65 – Distribuição da centralidade de subgrafos obtida para a rede de colaboração da USP, considerando os vértices das diferentes áreas do conhecimento.

4.3.4 Análise empírica do centro de distribuição do coeficiente de aglomeração concêntrico.

Devido a alta variância nos valores da propriedade coeficiente de aglomeração concêntrico obtido para a maioria das redes, e de sua importância para caracterizar os aglomerados de vértices, é interessante investigar mais a fundo as características dessa propriedade.

Foi verificado, na subseção anterior e pelas figuras das distribuições do coeficiente de aglomeração, que a característica que mais varia entre os vértices é o aparecimento ou não do pico na região dos últimos níveis concêntricos. Pode-se então, definir uma propriedade que represente o nível concêntrico central com relação à distribuição dos valores do coeficiente de aglomeração concêntrico. O centro de distribuição normalizado do coeficiente de aglomeração concêntrico, $\bar{N}_{Cc}(i)$, de um vértice i, é definido então como:

$$\bar{N}_{Cc}(i) = \frac{\sum_{d=0}^{d_{\max}} d\ Cc_d(i)}{\sum_{d=0}^{d_{\max}} Cc_d(i)} \tag{4.1}$$

onde $d_{\max}$ é o nível concêntrico máximo para o vértice i.

Espera-se que vértices com valores do centro de distribuição altos estejam distantes dos centros bem conectados, enquanto aqueles com baixos valores estão próximos ou pertencem a um grupo bem conectado.

A propriedade foi determinada para os vértices de algumas das redes estudadas anteriormente, e, através do software de visualização, foram geradas imagens das redes com as cores associadas ao centro de distribuição do coeficiente de aglomeração.

A figura 4.66 apresenta os resultados da medida de centro de distribuição obtidos para a rede de colaboração da USP. Como foi visto anteriormente, em 4.1.2, os vértices com coeficiente de aglomeração altos estão bem espalhados pela rede, assim como pode ser visto para o centro de distribuição. Entretanto, nota-se que os vértices mais distantes do aglomerado central apresentam valores baixos (em vermelho no gráfico) da medida.

A figura 4.67 apresenta uma visualização dos resultados obtidos para a medida de centro de distribuição do coeficiente de aglomeração concêntrico considerando os vértices da rede de teoremas da Wikipédia. Como visto 4.2, os vértices com altos valores do coeficiente de aglomeração são poucos e encontram-se espalhados pela rede, quanto aos valores do centro de distribuição, vértices próximos àqueles centros tem valores baixos, enquanto àqueles mais distântes, valores bem mais altos.

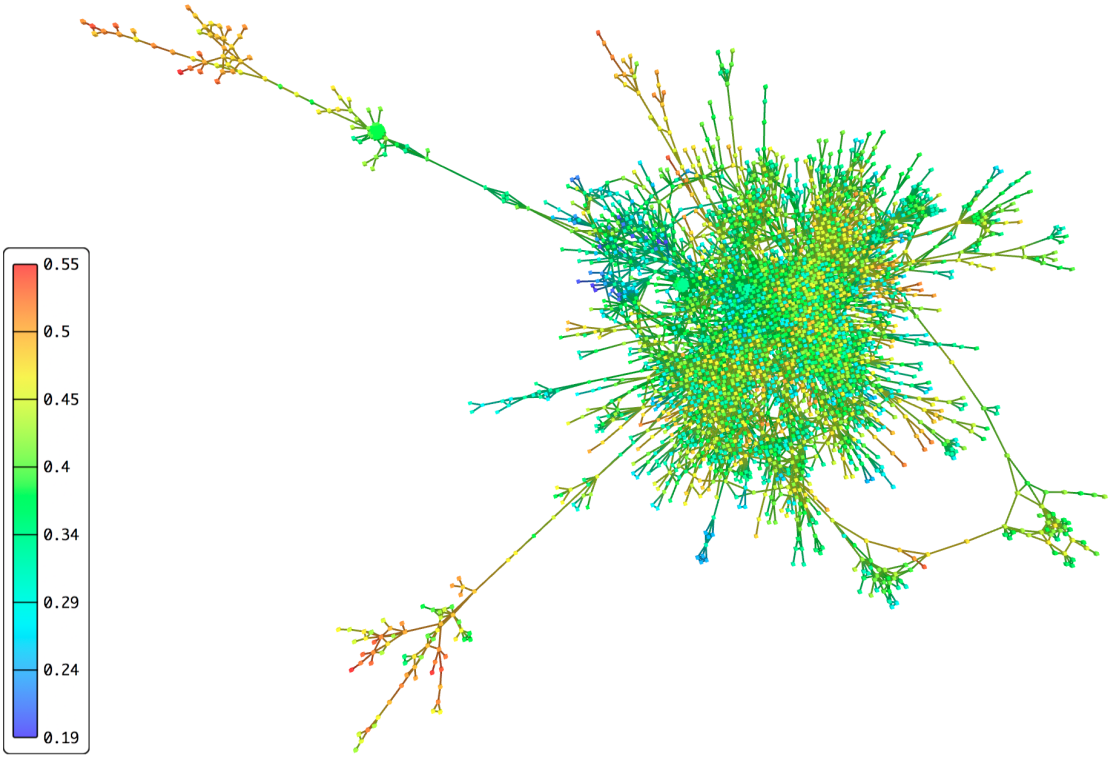


Figura 4.66 – Centro de distribuição obtido para os vértices da rede de colaboração da USP. É importante notar que os vértices com maiores valores, em vermelho, indicam que estão longe dos grandes centros, enquanto os vértices com valores baixos, em azul, são aqueles mais próximos aos centros.

Os valores da medida também foram obtidos para uma rede regular, apresentada na figura 4.68, neste caso verifica-se que, apesar dos vértices que não pertencem às bordas possuírem mesmo valor de coeficiente de aglomeração tradicional devido a simetria e pela propriedade considerar apenas a primeira vizinhança de vértices; o centro de distribuição apresentou variação ao longo de todos os vértices da rede, indicando que os aqueles pertencentes ao centro da rede apresentam apresentam baixos valores. As pontas também apresentaram baixos valores da medida, isto devido ao fato de que nessas bordas, o coeficiente de aglomeração é maior, já que há menos vértices vizinhos.

O centro de distribuição dos vértices, obtidos para a rede de alta tensão dos EUA, mostrados na figura 4.69, apresenta melhor as características referentes a redes caracterizadas como geográficas. Verifica-se que há um gradiente dos valores no sentido da esquerda, mais conectada, para a direita, menos conectada. Essa característica deve estar presente para todas as redes geográficas que possui alguma assimetria com relação ao coeficiente de aglomeração. No caso de uma rede geográfica gerada pelo modelo teórico, a assimetria ocorre para a região central da rede, que é muito mais aglomerada do que as bordas, como mostra a figura 4.70.

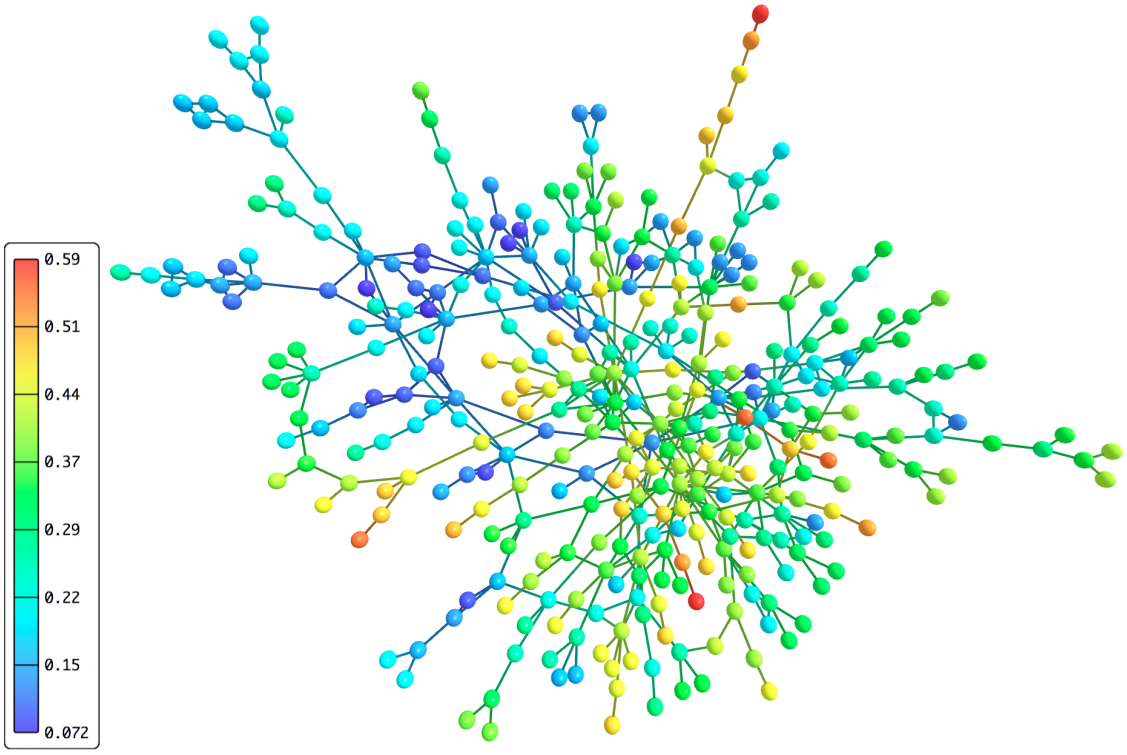


Figura 4.67 – Centro de distribuição obtido para os vértices da rede de teoremas da Wikipédia.

A metodologia é efetiva mesmo para redes compostas por um aglomerado central muito bem conectado, como é o caso da subrede da WWW, California, como mostrado na figura 4.71. Para essa rede, os grupos mais conectados estão presentes no núcleo fechado da rede, apresentando baixos valores da medida. A partir do núcleo partem diversos ramos de vértices, que, em grande maioria, não estão conectados entre si, ou apresentam estruturas de árvores (sem aglomeração), estes se destacam pelo alto valor do centro de distribuição, enquanto que há o aparecimento de alguns ramos muito conectados entre si, que se destacam pela coloração azul no gráfico, isto é, baixos valores da medida.

O centro de distribuição da métrica concêntrica pode ser usada, se devidamente formalizada, como complemento para o coeficiente de aglomeração, pois apresenta variação gradual ao longo da topologia dos vértices, isto é, vértices vizinhos entre si, devem apresentar valores semelhantes; podendo ser usado como métrica para separação de aglomerados, em contraste com a medida clássica que é descontinua para vértices próximos e dependem exclusivamente da topologia local dos vértices.

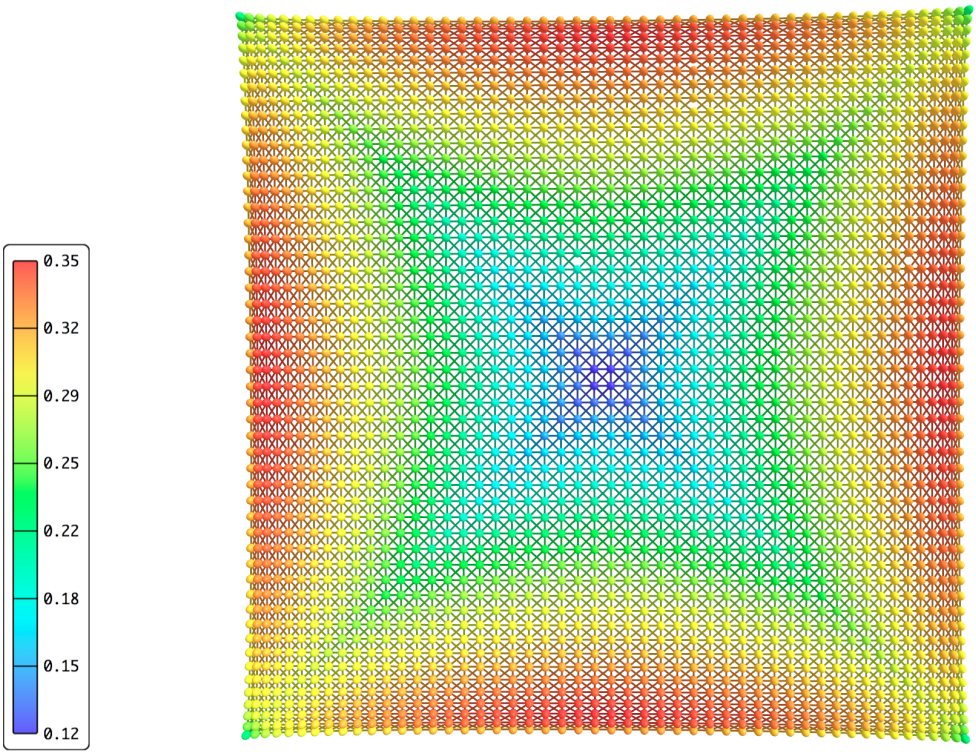


Figura 4.68 – Centro de distribuição obtido para os vértices para uma rede Regular com bordas.

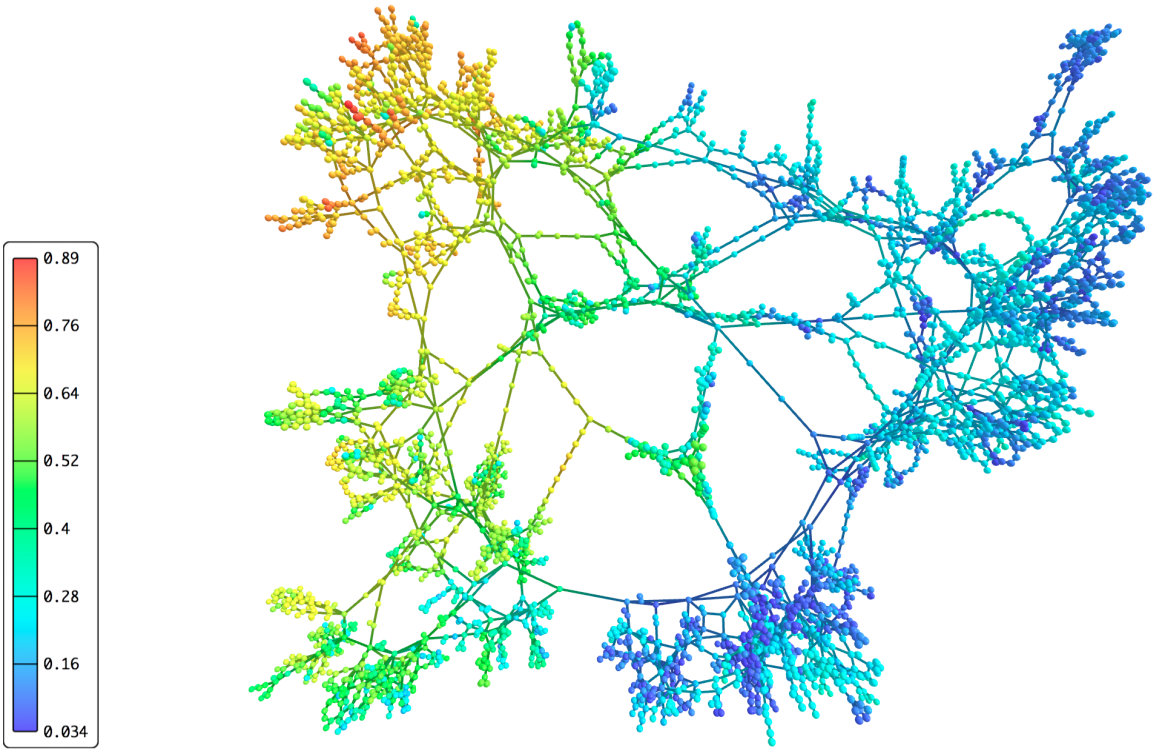


Figura 4.69 – Centro de distribuição obtido para os vértices da rede de alta tensão dos EUA.

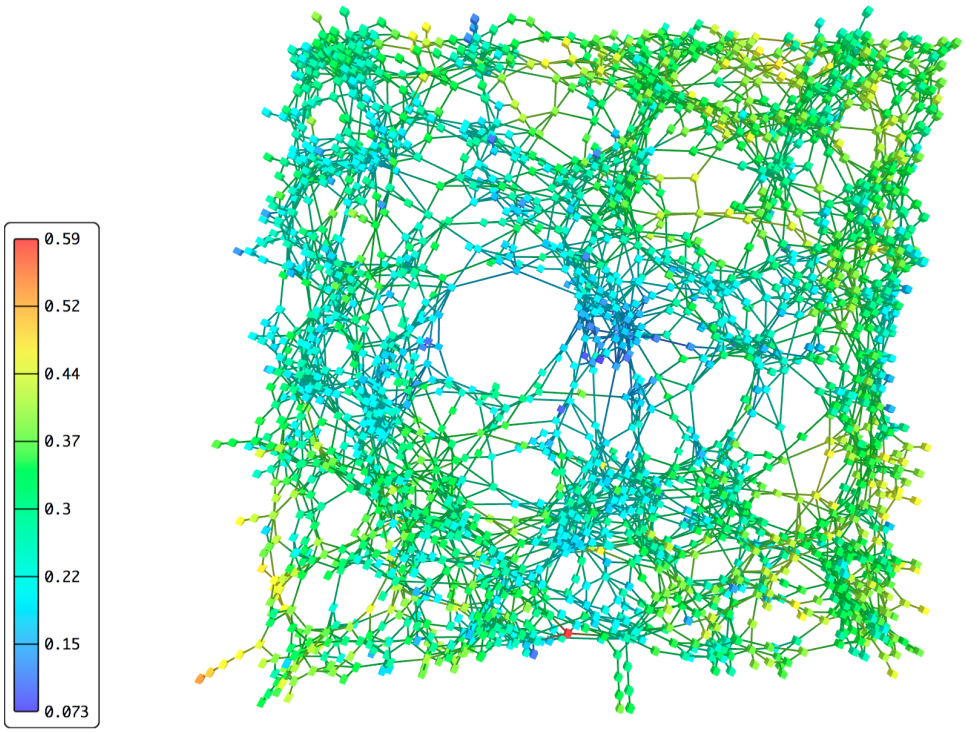


Figura 4.70 – Centro de distribuição obtido para os vértices de uma rede geográfica gerada pelo modelo teórico.

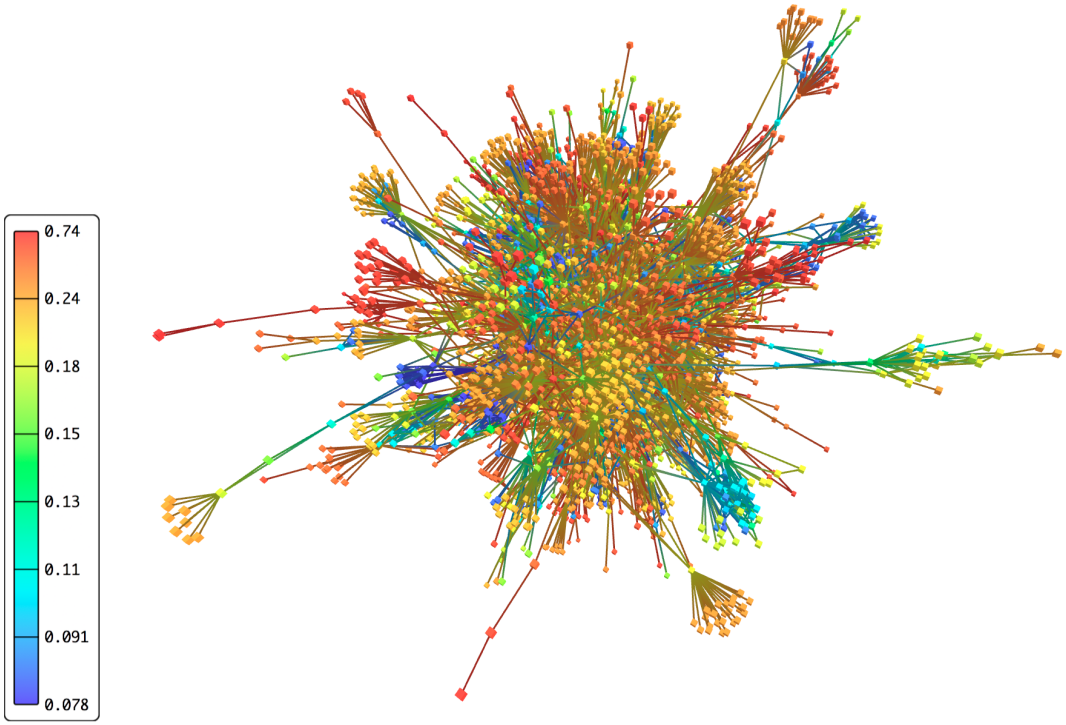


Figura 4.71 – Centro de distribuição obtido para os vértices da subrede da WWW, California.

4.4 PCA aplicado a redes complexas

A metodologia PCA foi aplicada a algumas redes estudadas anteriormente, com o objetivo de ilustrar a possibilidade de classificação de redes, de acordo com um conjunto de modelos teóricos através de diferentes medidas, assim como determinar, empiricamente, a relevância das propriedades concêntricas.

Inicialmente, foi criado, para cada rede estudada, um conjunto de diversas redes geradas, baseando-se nos quatro modelos teóricos apresentados: ER, BA, geográfico (GEO) e WS. Estas redes pertencem aos grupos de treinamento e foram geradas de modo a apresentarem valores de grau médio e número de vértices semelhante semelhantes aos da respectiva rede estudada.

As propriedades usadas para este estudo abrangem tanto as propriedades concêntricas e centralidades, quanto as métricas tradicionais obtidas em média para cada rede estudada.

Para a análise da rede de alta tensão dos EUA, foi usada uma combinação das diversas propriedades a seguir: grau médio, diâmetro, coeficiente de aglomeração médio, mínimo caminho médio, centralidade de intermediação e as propriedades concêntricas obtidas para o segundo anel concêntrico (nível 2), o grau concêntrico, coeficiente de aglomeração concêntrico e taxa de convergência.

A figura 4.72 apresenta a projeção em três dimensões obtida pela análise canônica do conjunto de redes para a classificação da rede de alta tensão dos EUA. Verifica-se que os modelos estudados apresentam distribuições espaciais bem definidas e não se sobrepõem. Entretanto, curiosamente, a rede de alta tensão dos EUA foi posicionada relativamente distante dos outros vértices, sendo classificada, pela análise de máxima verossimilhança como uma rede geográfica. Neste caso, as medidas concêntricas foram muito representativas nos vetores de projeção, representando, juntas, quase $1/2$ do primeiro vetor de projeção (CAN1), e cerca de $1/4$ do segundo vetor (CAN2).

As redes de aeroportos dos EUA e de proteínas, Yeast, também foram estudadas pela metodologia, entretanto apenas considerando as medidas concêntricas. Todas as medidas concêntricas foram obtidas para os modelos equivalentes, considerando apenas as medidas ao longo dos 5 primeiros anéis, para a rede de aeroportos, e dos 10 primeiros para a rede de proteínas.

A figura 4.73 apresenta a projeção obtida por PCA da rede de aeroportos. Assim como a obtida para a rede de alta tensão dos EUA, os diferentes modelos estão distribuídos em regiões bem definidas e, não se sobrepondo, entretanto, neste caso não foi necessária a aplicação do

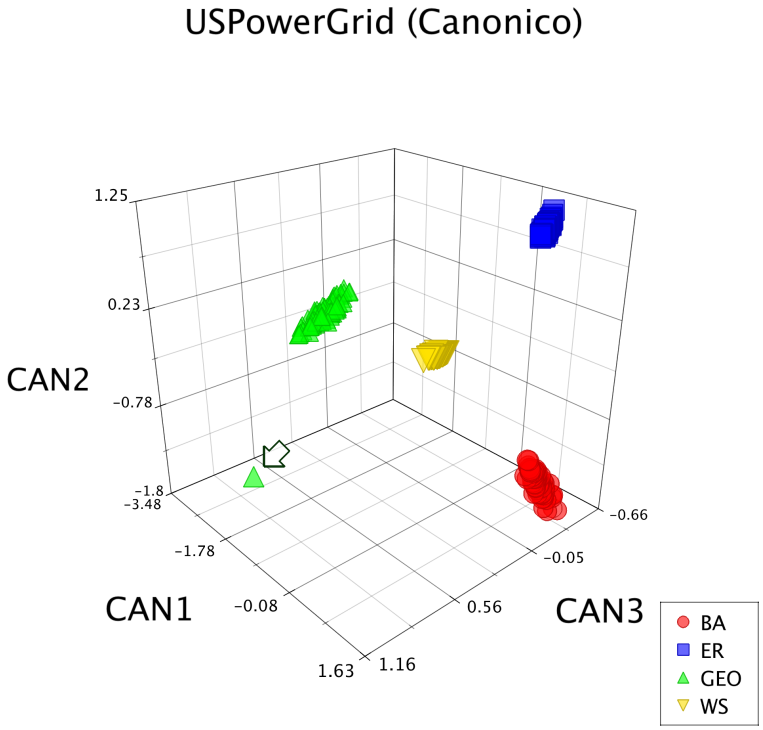


Figura 4.72 – Projeção 3D da análise canônica obtida para o conjunto de redes relacionadas à rede de alta-tensão dos EUA. A rede estudada foi classificada como geográfica pela análise de máxima verossimilhança e é indicada por uma seta.

método canônico já que a separação dos modelos foi satisfatória apenas por PCA. A rede de aeroportos foi classificada como BA pela análise de máxima verosimilhança. Dentre as propriedades concêntricas, aquelas que foram mais representativas para a projeção dos dados são o coeficiente de aglomeração e a taxa de convergência obtidas para o segundo nível concêntrico.

Os resultados da projeção obtida para a rede de proteínas estão apresentados na figura 4.74, assemelhando-se muito àqueles obtidos para a rede de aeroportos, classificando-a como uma rede BA, com as medidas mais representativas na projeção sendo o coeficiente de aglomeração concêntrico e a taxa de convergência.

Os resultados mostraram que as propriedades concêntricas podem ser bastante relevantes na classificação de redes complexas de acordo com modelos teóricos, entretanto novos estudos devem ser realizados para explorar essa capacidade e validar a metodologia.

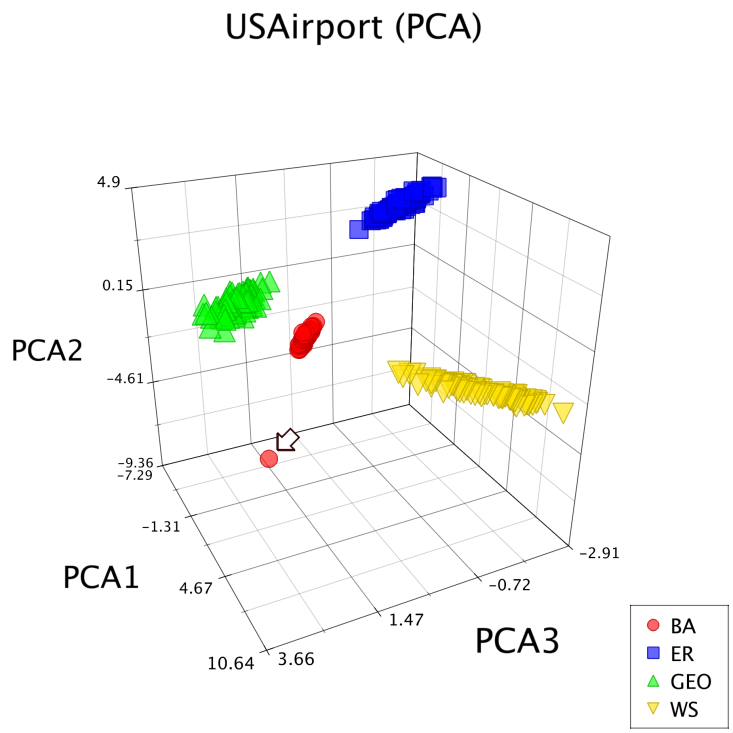


Figura 4.73 – Projeção 3D obtida através de PCA para o conjunto de redes relacionadas à rede de aeroportos dos EUA, classificada como BA e indicada pela seta.

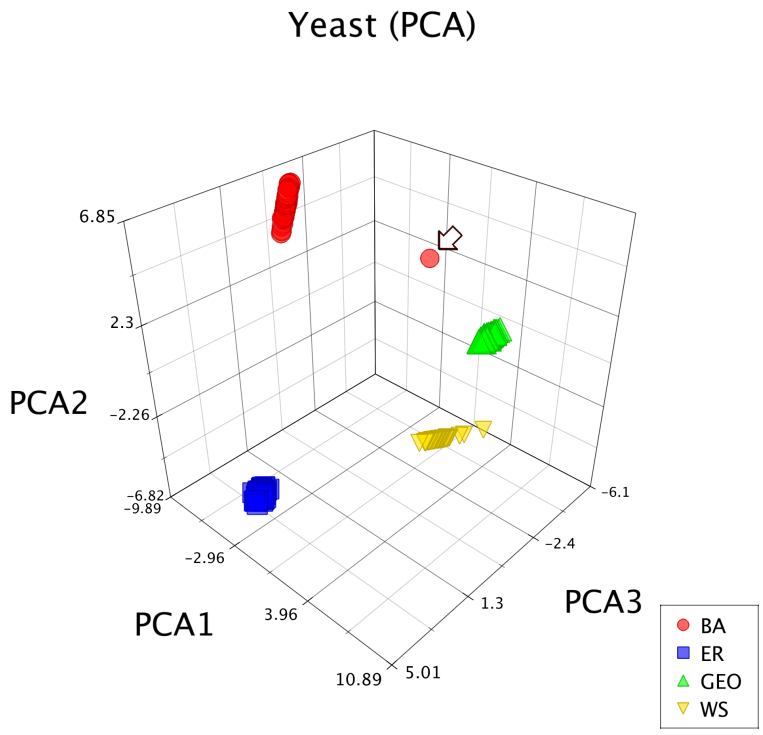


Figura 4.74 – Projeção 3D obtida por PCA para as redes relacionadas à rede de proteínas, Yeast, classificada como BA.

4.5 Modelagem de Aquisição de Conhecimento

A metodologia de simulação de aquisição de conhecimento, descrita na seção 3.5, foi aplicada à rede de teoremas da Wikipédia e aos modelos de redes teóricas considerados neste trabalho: ER, BA, geográfico e WS.

Cada simulação inicia dispondo, aleatoriamente, um certo número de agentes ao longo dos vértices de uma rede. As propriedades dos agentes, como memória e susceptibilidade a erros, são escolhidas e fixadas no inicio da simulação, assim como a rede de interação entre eles. Com o objetivo de simplificar a apresentação dos resultados, para cada simulação todos os agentes apresentam as mesmas propriedades. Em seguida, os agentes iniciam a navegação pela rede, de acordo com a heurística apresentada. A simulação termina quando ao menos 90% da rede é percorrida, com o desempenho quantificado pela quantidade de vértices únicos percorridos pelo número de passos que foram necessários.

Os resultados são apresentados através de gráficos, que mostram curvas de escalabilidade do desempenho de aquisição de conhecimento com o número de agentes, para diferentes redes e configurações de parâmetros da simulação. A abcissa de cada curva representa o número de agentes e a ordenada, o desempenho. Cada ponto da curva indica o *valor médio* do desempenho obtido, considerando diversas simulações com parâmetros iguais, para este trabalho, foram realizadas cerca de 5000 simulações para cada ponto dos gráficos.

As redes baseadas nos modelos teóricos, usadas nesta seção, são as mesmas que foram geradas para comparação das medidas concêntricas obtidas para a rede de teoremas da wikipédia, e portanto, possuem o grau médio semelhantes, $\langle k \rangle = 2.7$, assim como mesmo número de vértices, $N = 371$.

Cada figura a seguir apresenta uma legenda formatada de modo a indicar os diferentes parâmetros de simulação escolhidos para as curvas compostos da seguinte forma:
(*rede*)_m(*(Memória)*)_e(*Porcentagem de erro*)_(*Rede de interação, se houver*)
Por exemplo, uma legenda indicando uma curva intitulada *BA_m0100_e025_ra8* corresponde a uma curva obtida pela simulação em uma rede BA, com agentes de memória $M = 100$, susceptibilidade a erros $P_E = 25\%$ e rede de interação aleatória de grau 8.

A figura 4.75 apresenta uma curva de escalabilidade obtida pelas simulações na rede de teoremas da Wikipédia, com agentes de memória $m = 10$, nenhuma susceptibilidade a erros e não interagentes. Verifica-se que o desempenho aumenta linearmente com a adição de novos agentes, esse comportamento, assim como o aumento crescente do desvio padrão, se repete para

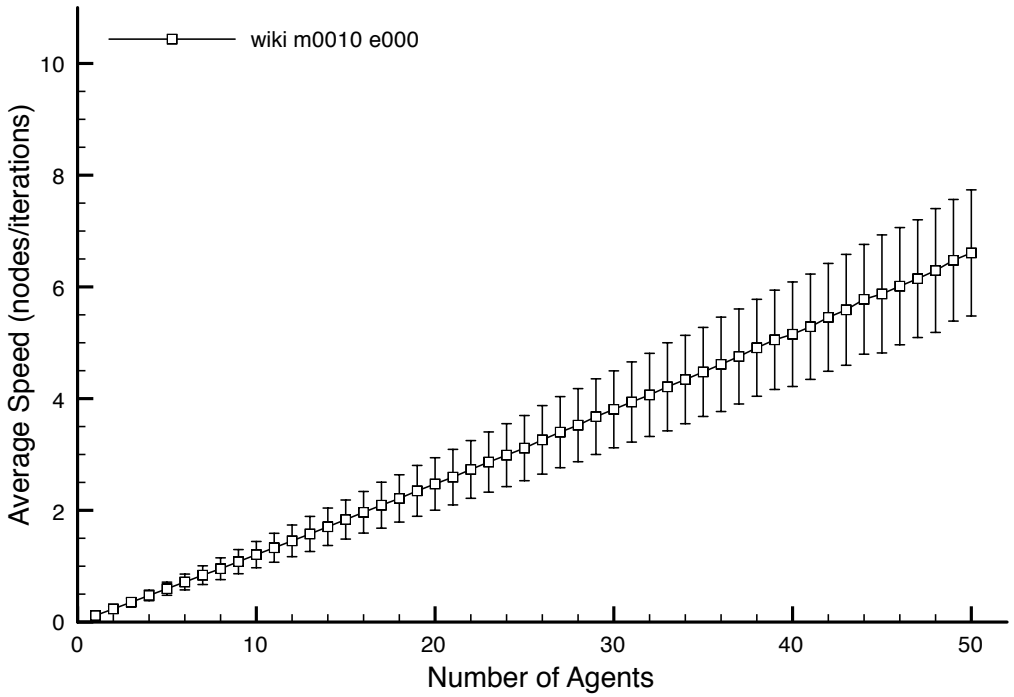


Figura 4.75 – Curva de escalabilidade obtida para a rede de teoremas da Wikipédia usando agentes com memória de até 10 vértices cada. O eixo vertical apresenta os valores de desempenho médio (*Average Speed*) enquanto o horizontal apresenta o número de agentes (*Number of Agents*) para cada simulação. A figura também mostra barras de erro que indicam o desvio padrão de cada conjunto de medidas.

a maioria dos modelos e configurações, com algumas poucas exceções.

As curvas obtidas para a rede de teoremas considerando agentes com diversos valores de memória e sem susceptibilidade a erros são apresentadas na figura 4.76. Todas as curvas apresentam comportamento crescente ao longo de todo o eixo considerado e, a partir de 20 agentes, são praticamente lineares. As curvas que representam simulações com valores altos de memória apresentaram melhor desempenho na primeira metade do gráfico, no entanto, também apresentaram pior desempenho na segunda metade. É interessante destacar que a curva que apresentou melhor desempenho para número de agentes maior do que 25 foi aquela que representa as simulações com agentes de memória $m = 10$. Apesar das simulações considerarem como redes de interação, apenas, redes aleatória de grau 8, o mesmo comportamento foi observado para outras redes de interação.

O comportamento das curvas considerando a variação da memória dos agentes indica que o desempenho não se altera muito com o aumento da memória, podendo até ser pior, no caso de grande quantidade de agentes, revelando que o desempenho da navegação dos agentes na rede de teoremas é localmente otimizado. Este resultado está ligado à característica da rede

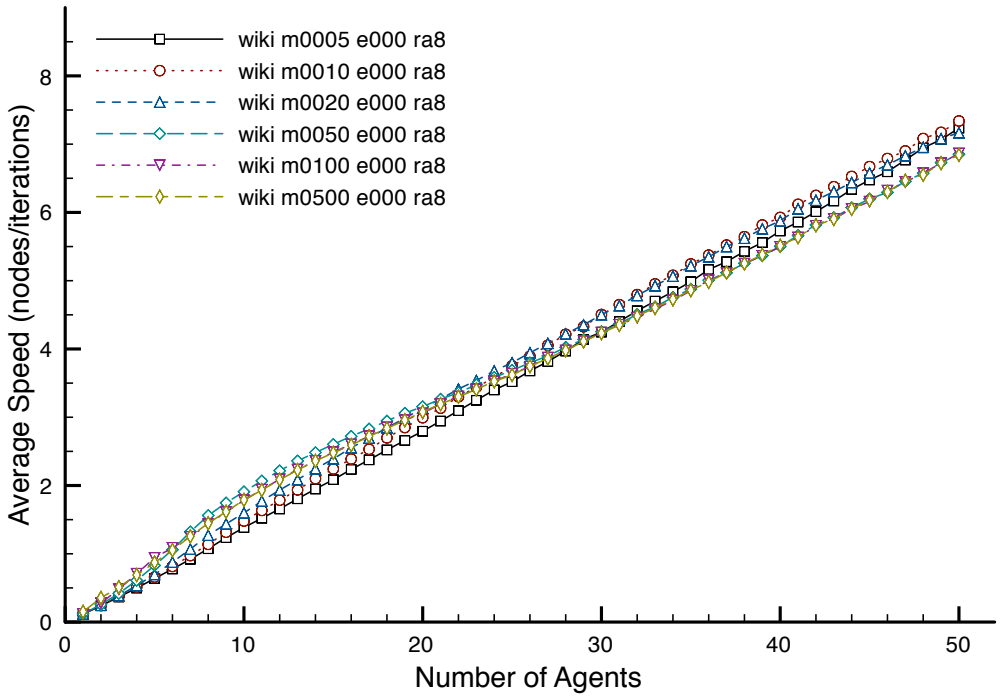


Figura 4.76 – Curvas da escalabilidade da rede de teoremas considerando agentes com diferentes valores de memória (de $m = 5$ a 500), sem susceptibilidade a erros e rede de interação aleatória de grau 8.

de teoremas possuir muitos arcos e arvores, assim como à heurística utilizada. Agentes com valores altos de memória tenderão a evitar vértices já visitados, necessitando caminharem cada vez mais longe de seu ponto de partida. Para muitos agentes, distribuídos uniformemente ao longo da rede, não é ideal que caminhem muito longe de seus pontos de partida, pois assim, em média, podem estar se sobrepondo, podendo diminuir o desempenho de dois agentes.

Quando os agentes estão sujeitos a erros, aqueles com alto valor de memória apresentam desempenho pior, quando comparados às mesmas simulações sem efeito de erros, como mostra a figura 4.77. Agentes com memória tendem a guardar os vértices errados por mais tempo, diminuindo consideravelmente o desempenho de aquisição de conhecimento.

A figura 4.78 apresenta as curvas de escalabilidade resultantes das simulações para os diferentes modelos teóricos de redes complexas considerando, para cada rede, três parâmetros de memória dos agentes. As simulações resultaram em curvas de desempenho crescentes com o número de agentes, e, tornam-se praticamente lineares com grande número de agentes. Com exceção daquelas obtidas para o modelo Watts-Strogatz, as curvas obtidas paras os modelos teóricos, com maior quantidade de memória, apresentaram os melhores desempenhos quando há poucos agentes, e desempenho similar ou menor que as outras curvas quando há muitos

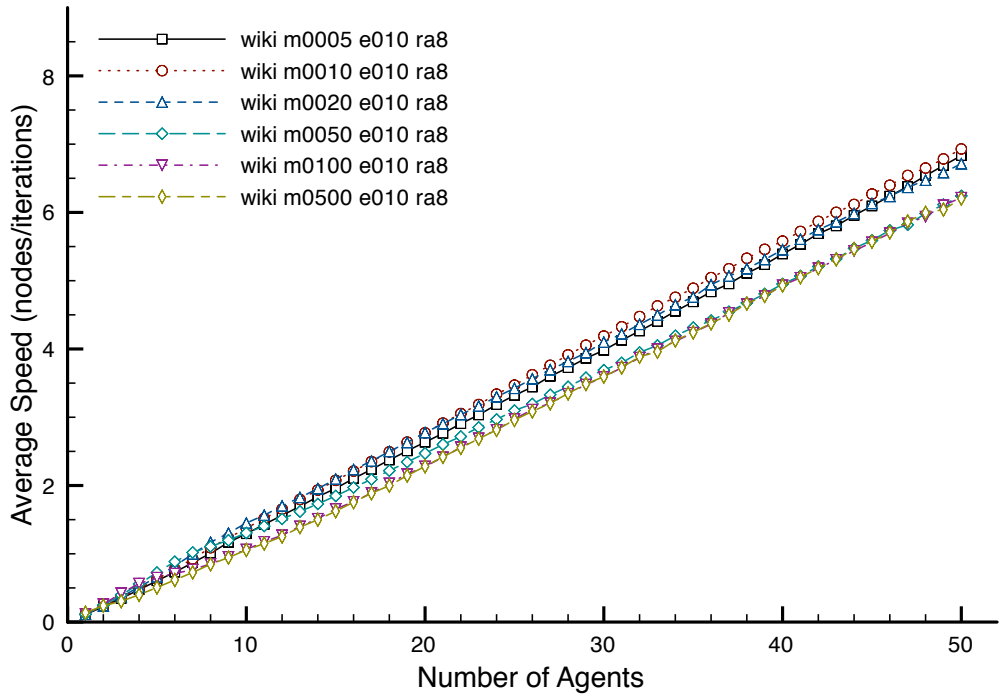


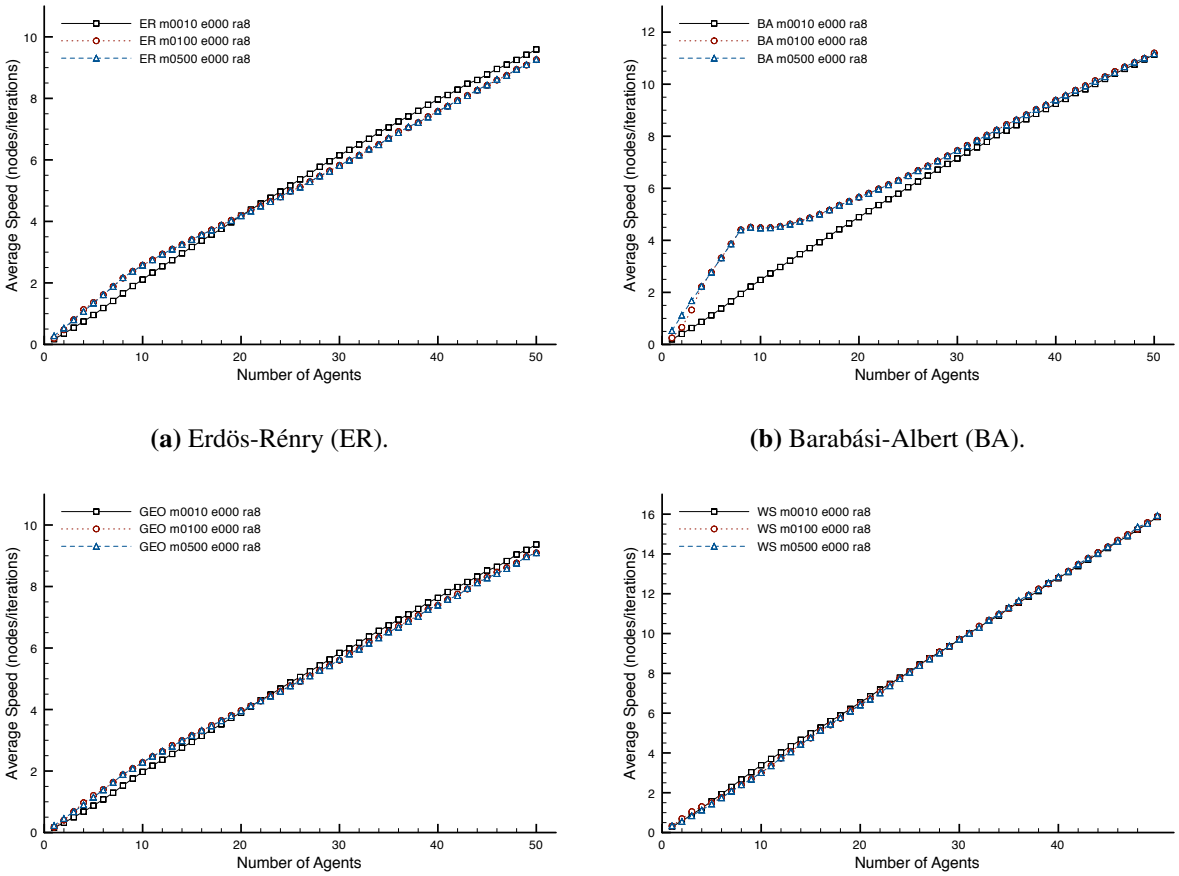
Figura 4.77 – Curvas de escalabilidade da rede de teoremas da Wikipédia considerando agentes com diferentes valores de memória e 10% de susceptibilidade a erros.

agentes.

O fato do modelo WS apresentar resultados opostos aos obtidos para os outros modelos pode ser explicado pela existência de arcos bem conectados, de modo que, mesmo para um pequeno número de agentes com alta quantidade de memória, os caminhos compostos por esses arcos devem ser percorridos várias vezes para que seus vértices sejam visitados. As curvas obtidas para o modelo BA são as que mais se diferenciam entre si, com aquelas que representam as simulações de agentes com memória $m = 10$ apresentando um desempenho muito menor do que as de memória $m = 100$ e $m = 500$ para uma quantidade pequena de agentes.

A figura 4.79 apresenta as curvas de simulações semelhantes às aplicadas para obter aquelas da figura 4.78, entretanto com probabilidade de erros $P_E = 10\%$. As curvas são semelhantes àquelas obtidas para as simulações sem probabilidade de erros, entretanto, para simulações de agentes com grande quantidade de memória, o desempenho foi relativamente menor. Assim como para a rede de teoremas, este comportamento está diretamente ligado ao fato de que agentes com grande quantidade de memória tendem a guardar por mais tempo os erros, evitando que novos vértices sejam visitados.

É interessante entender como o desempenho de aquisição de conhecimento é influenciado pela rede de interação dos agentes. A figura 4.80 apresenta esses resultados através de curvas de



(a) Erdös-Rényry (ER).

(b) Barabási-Albert (BA).

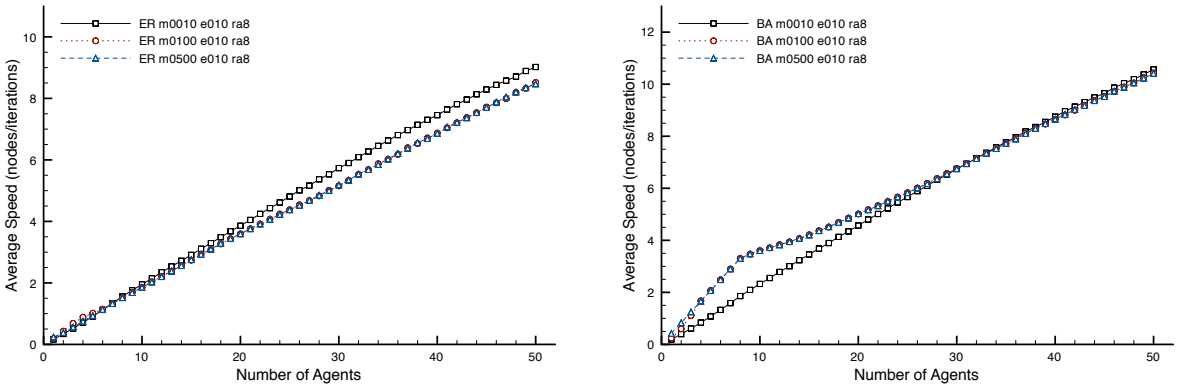
(c) Geográfica (GEO).

(d) Watts-Strogatz (WS).

Figura 4.78 – Curvas do desempenho contra o número de agentes para as simulações nos modelos teóricos de redes complexas, considerando diferentes valores de memória e sem susceptibilidade à erros.

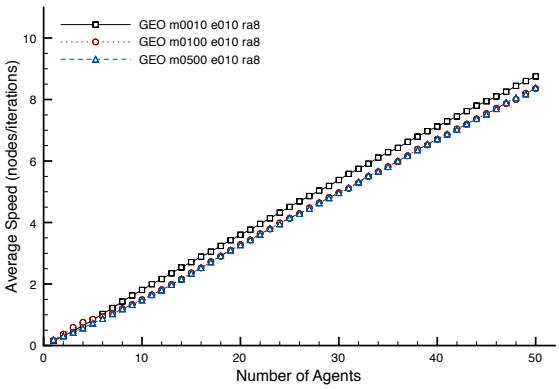
desempenho da rede de teoremas, considerando diferentes redes de interação. Primeiramente, verifica-se que não há diferença significativa de desempenho entre as simulações obtidas com redes de interação baseadas nos modelos de redes BA e ER para grau médio 4 ou 8, no entanto, simulações com rede de interação ER de grau 16 apresentaram desempenho superior às equivalentes para uma rede BA.

Acredita-se que este comportamento está relacionado com o melhor gerenciamento de informação para a rede aleatória, onde o compartilhamento é realizado localmente, não sobrecarregando nenhum vértice em especial. Em contraste, uma rede de interação baseada em um modelo livre de escala tenderá a apresentar hubs, e por consequência, grande parte da informação compartilhada viajará por caminhos que contêm esses vértices. Por possuírem memória limitada, não conseguem guardar efetivamente todo o conhecimento para transmiti-lo. Também é importante notar que as simulações sem rede de interação apresentaram os piores

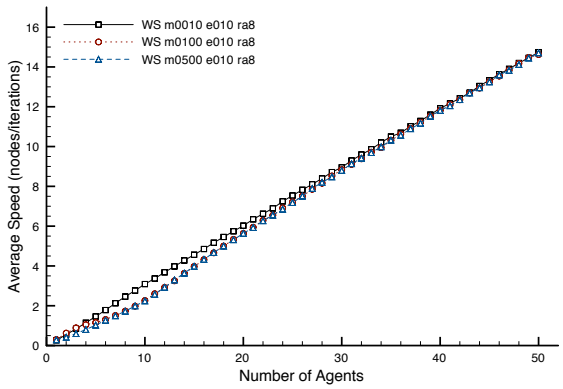


(a) Erdös-Rényry (ER).

(b) Barabási-Albert (BA).



(c) Geográfica (GEO).



(d) Watts-Strogatz (WS).

Figura 4.79 – Curvas do desempenho contra o número de agentes para as simulações nos modelos teóricos de redes complexas, considerando diferentes valores de memória com probabilidade de erros $P_E = 10\%$.

desempenhos considerando qualquer quantidade de agentes.

Assim como as curvas da figura 4.80, a figura 4.81 apresenta os resultados de desempenho da aquisição de conhecimento na rede de teoremas, considerando diferentes redes de interação, entretanto com probabilidade dos agentes à erros de $P_E = 10\%$. Os resultados são semelhantes àqueles obtidos na primeira figura, entretanto, há maior queda de desempenho para redes de interação que apresentavam melhor desempenho quando não havia susceptibilidade a erros. Este resultado pode ser explicado pelo fato redes de melhor desempenho em comunicação, propagarão tanto as informações corretas, quando as incorretas, levando a um balanceamento do ganho de desempenho pelo compartilhamento contra a propagação de erros dos agentes.

Para ilustrar melhor a queda de desempenho devido a probabilidade de erros para a rede de teoremas da Wikipédia, foram realizadas simulações fixando todos os parâmetros com exceção da taxa de susceptibilidade a erros. Os resultados podem ser vistos na figura 4.82. Como

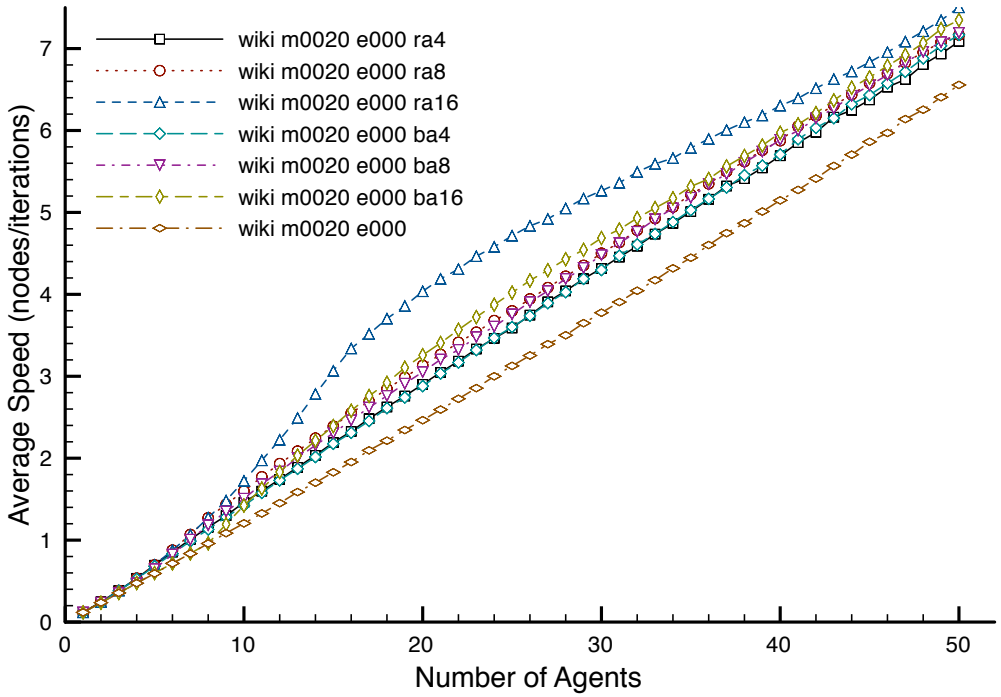


Figura 4.80 – Curvas de escalabilidade da rede de teoremas da Wikipédia considerando agentes de memória 20, não sujeitos a erros e diversas redes de interação entre eles.

esperado, há considerável perda de desempenho com o aumento da taxa de erros dos agentes. As simulações para os modelos teóricos também apresentaram comportamento semelhante.

Para comparar o desempenho de aquisição de conhecimento entre as diferentes redes consideradas, foram realizadas simulações para agentes de memória $m = 10$ e, inicialmente, sem probabilidade de cometerem erros. As curvas obtidas para cada rede são apresentadas da figura 4.83. A curva que apresentou o melhor desempenho ao longo do número de agentes foi aquela obtida para a rede do modelo WS, com coeficiente linear $\alpha_{WS} \simeq 0.31$, seguida pela rede BA, com $\alpha_{BA} \simeq 0.22$. As redes baseadas no modelo geográfico e aleatório apresentaram desempenho semelhantes, com $\alpha_{ER} \simeq \alpha_{GEO} \simeq 0.19$. A rede de teoremas da Wikipédia foi aquela que apresentou o pior desempenho dentre os modelos estudados, com coeficiente linear $\alpha_{WIKI} \simeq 0.15$, menos do que a metade do desempenho obtido para a rede WS.

Com o objetivo de compreender melhor as possíveis causas das variações de desempenho obtidas para as diferentes redes de conhecimento, foram determinadas as médias das frequências com que cada vértice dessas redes era visitado por um agente, chamada de *frequência de acesso*. As figuras 4.84, 4.85, 4.86, 4.86, 4.87 e 4.88 apresentam as frequências de acesso para cada vértice, respectivamente, das redes de teoremas, ER, BA, geográfica e WS.

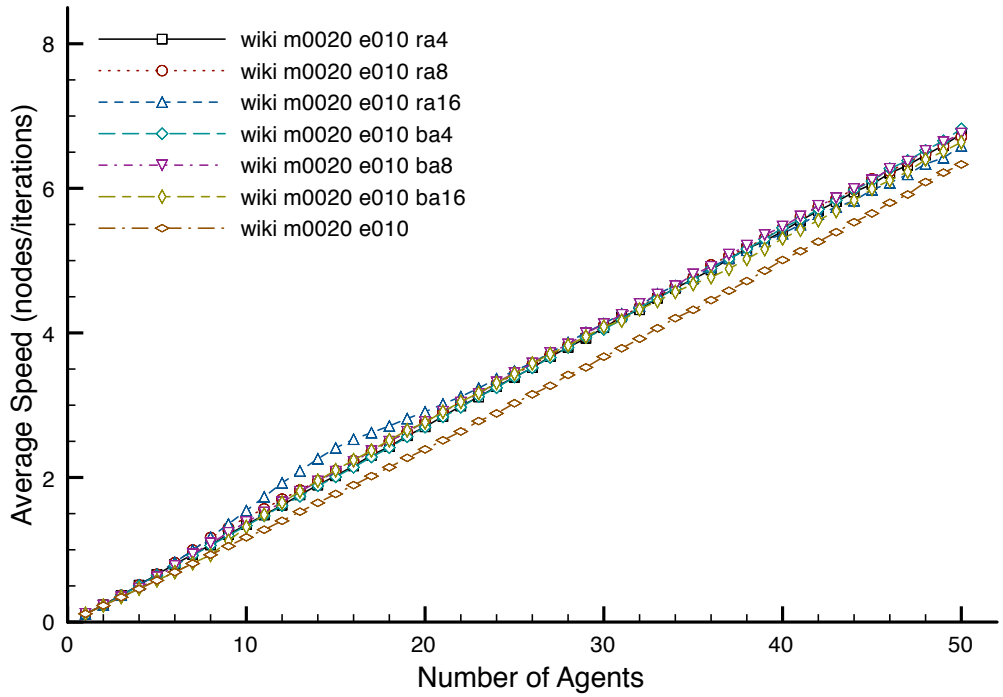


Figura 4.81 – Curvas de escalabilidade da rede de teoremas considerando agentes de memória 20, 10% de chance de susceptibilidade a erros e diversas redes de interação entre eles.

Para a rede de teoremas, observa-se que há dois vértices que destacam-se por serem muito mais visitados que os outros, estes dois correspondem aos teoremas fundamentais do cálculo e da álgebra. A frequência de acesso e o grau dos vértices parecem estar relacionados, ao menos para esta rede, como mostra o gráfico de correlação da figura 4.89a, com alto coeficiente de correlação, 0.91. Apesar da maior tendência em visitar os vértices de alta conectividade, os agentes circulam muito pelas bordas das redes, revelando que a existência de árvores nas bordas da rede é um dos principais fatores pelo baixo desempenho de aquisição de conhecimento da rede de teoremas. Este comportamento também é observado para a rede geográfica e ER, já que estas também possuem, em menor quantidade, árvores nas bordas da rede.

Assim como obtido para a rede da wikipedia, as redes ER e geográfica apresentaram alta correlação do grau com a frequência de acesso dos vértices, com coeficiente de correlação 0.80 e 0.75, respectivamente.

A rede BA e WS, que apresentaram os melhores desempenhos de aquisição de conhecimento, não possuem árvores em suas bordas, sendo a WS caracterizada por arcos bem conectados e a BA por vértices conectados entre si, com ao menos grau 2. A rede BA foi aquela que apresentou a maior correlação do grau com a frequência de acesso, com coeficiente de correlação 0.95. Em constraste, a rede WS apresentou a menor correlação entre o grau e a frequência de acesso, com 0.50 de coeficiente de correlação.

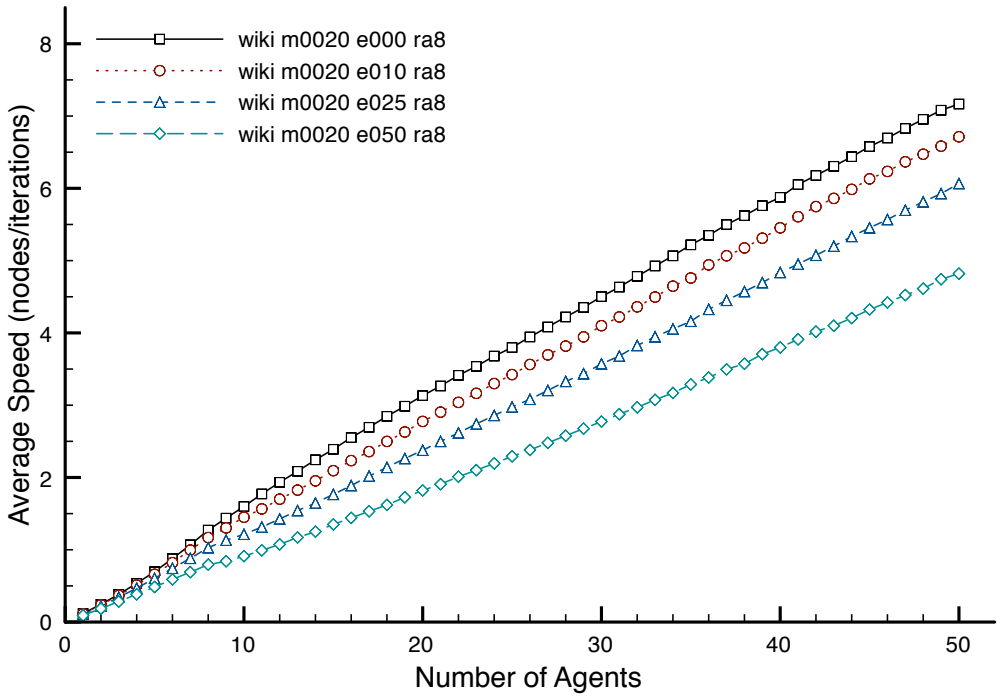


Figura 4.82 – Curvas de desempenho contra o número de agentes obtidas para a rede de teoremas da Wikipédia, considerando diversos valores de susceptibilidade dos agentes a erros.

Com a devida formalização da metodologia e dos resultados apresentados, a simulação de agentes também pode ser útil para a caracterização de redes complexas e até mesmo de seus vértices, através das medidas de desempenho, frequência de acesso, ou outras que podem se mostrar se tornarem relevantes no futuro.

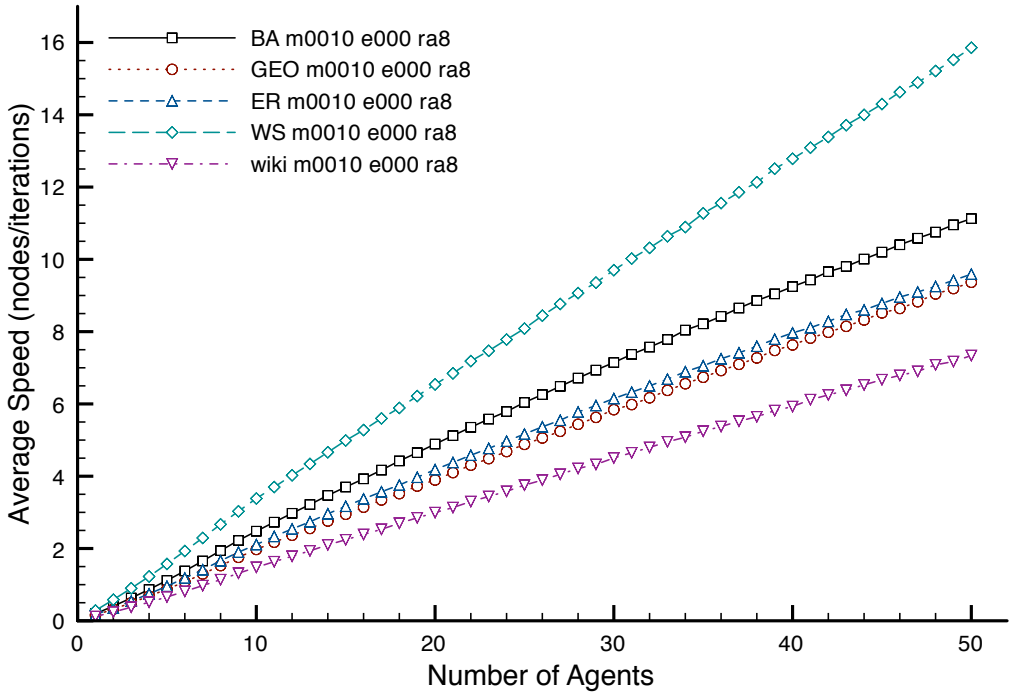


Figura 4.83 – Comparação das curvas de desempenho de aquisição de conhecimento obtidas para as redes de conhecimento estudadas.

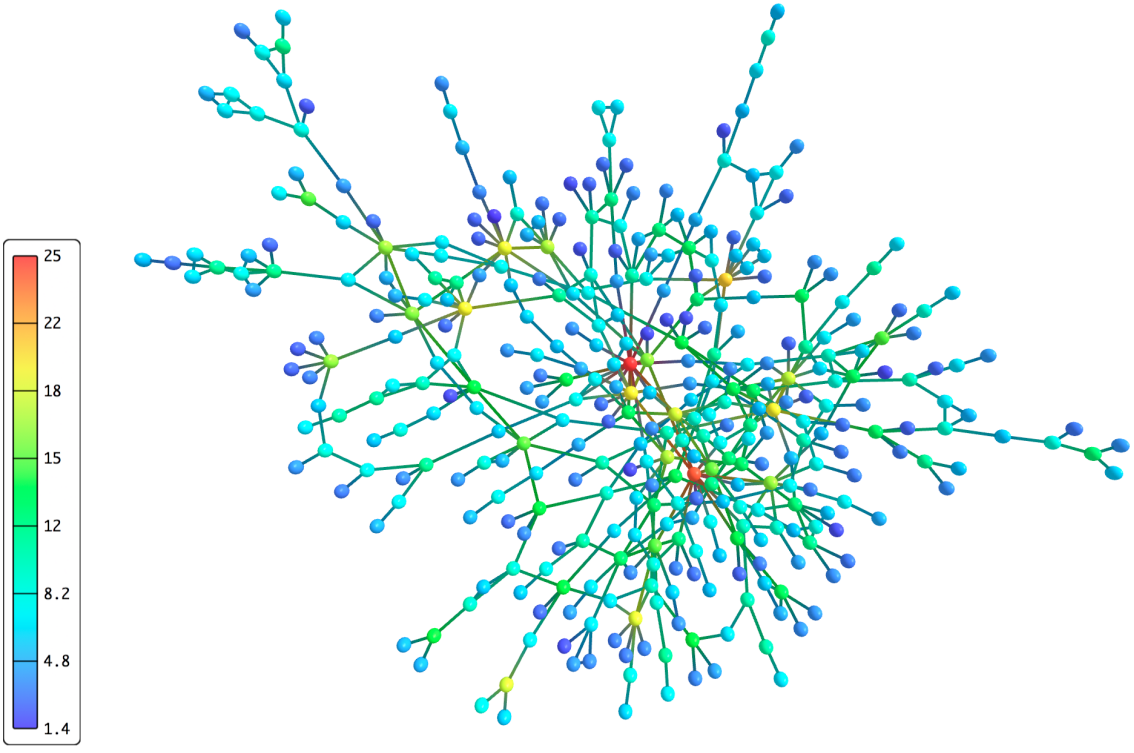


Figura 4.84 – Projeção 2D da rede de teoremas da Wikipédia apresentando, para cada vértice, a média da frequência com que ele é visitado pelos agentes. A simulação considerou 26 agentes de memória 10 sem susceptibilidade a erros.

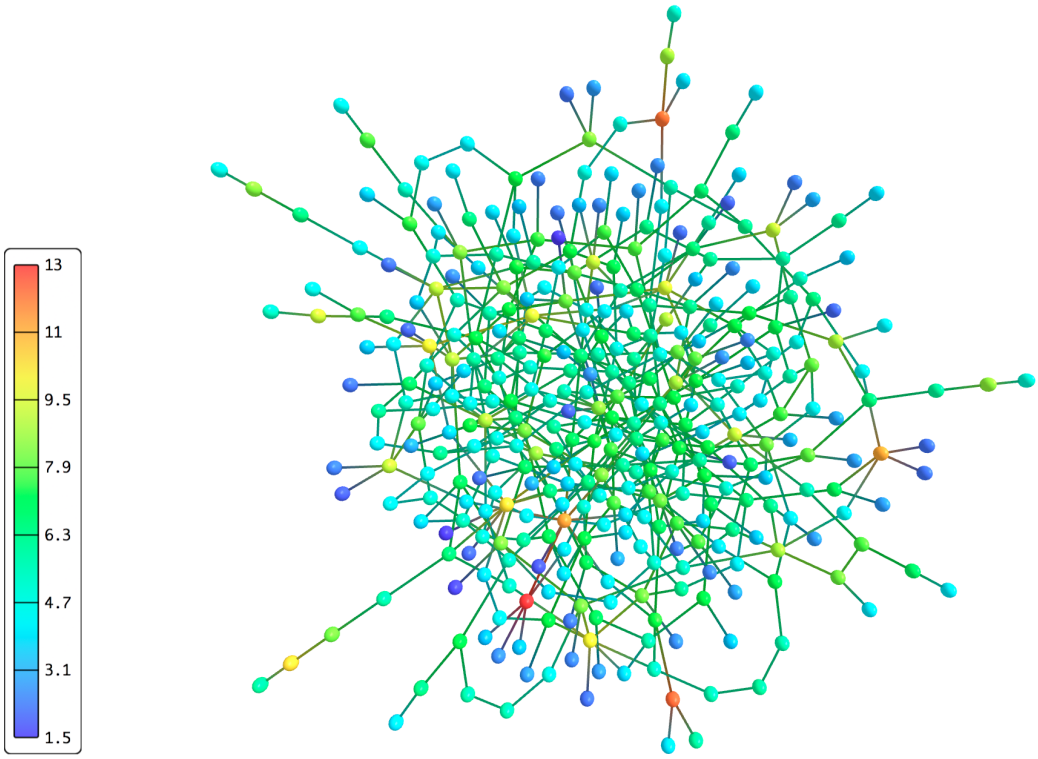


Figura 4.85 – Frequência de acesso obtidas para os vértices da rede ER, considerando 26 agentes de memória 10 sem susceptibilidade a erros.

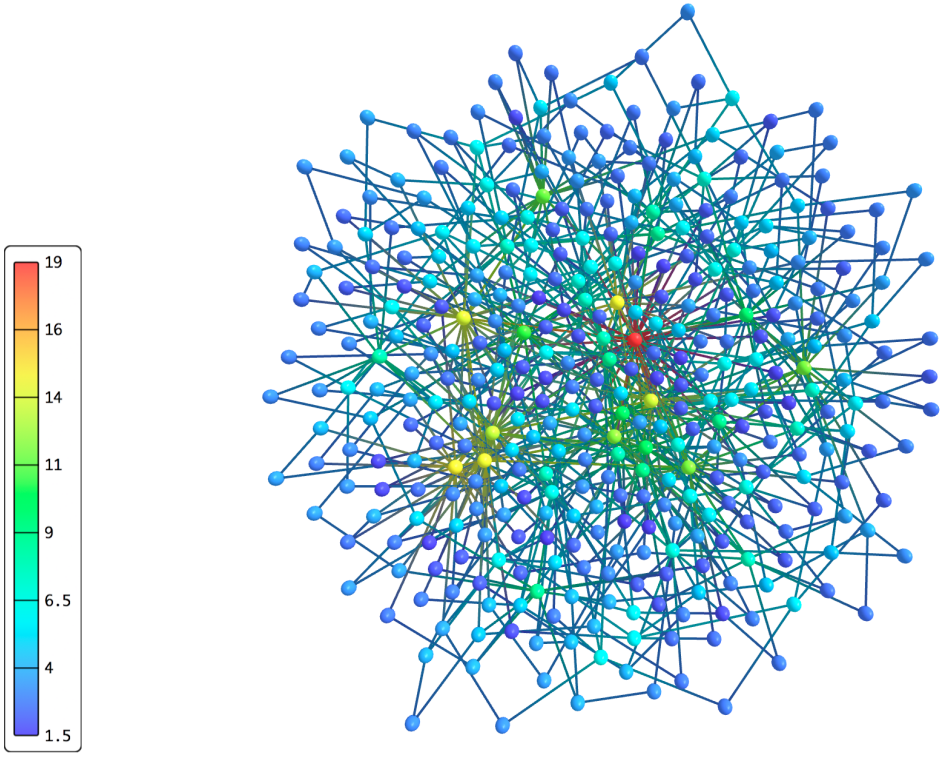


Figura 4.86 – Frequência de acesso obtidas para os vértices da rede BA, considerando 26 agentes de memória 10 sem susceptibilidade a erros.

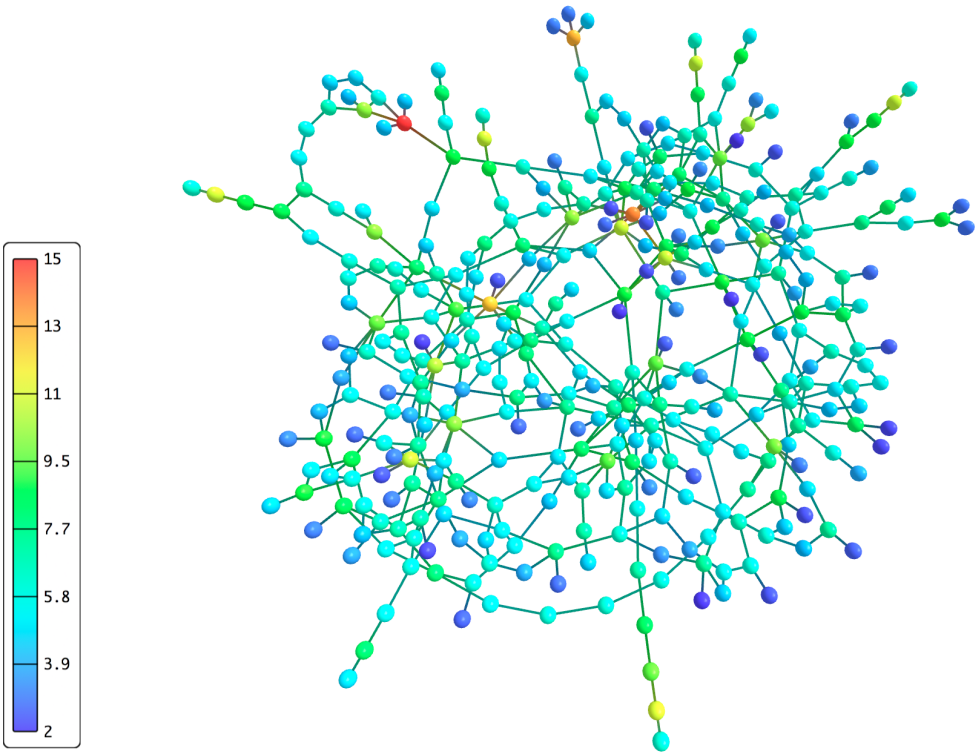


Figura 4.87 – Frequência de acesso obtida para os vértices da rede geográfica, considerando 26 agentes de memória 10 sem susceptibilidade a erros.

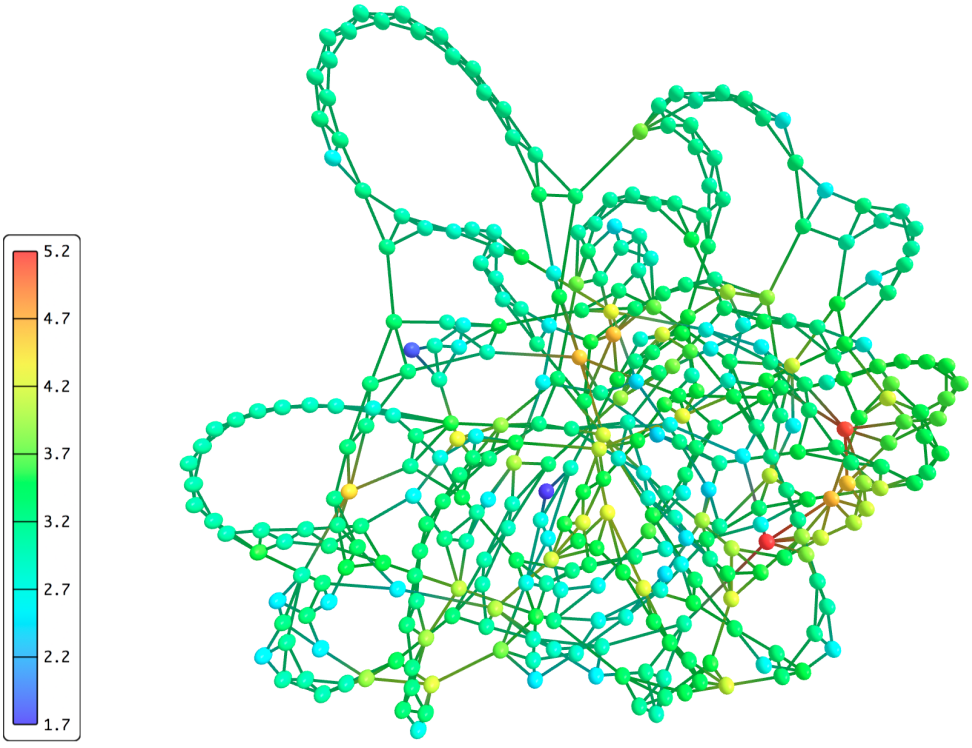
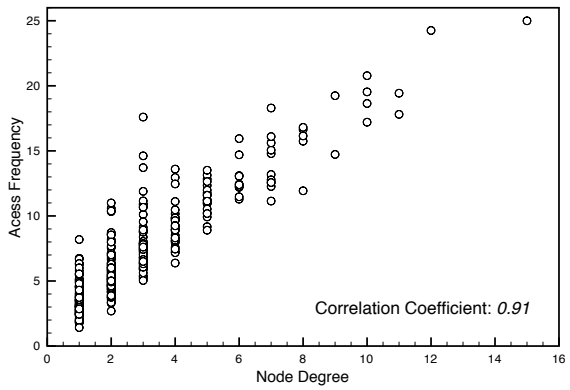
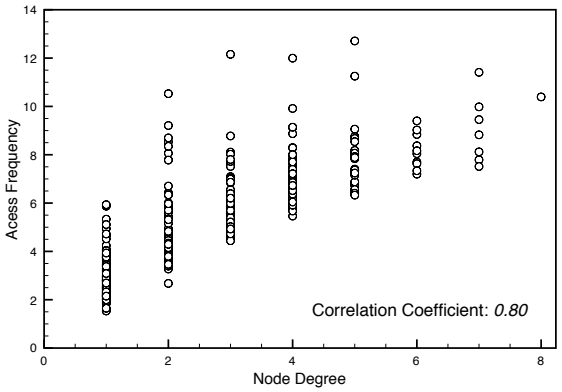


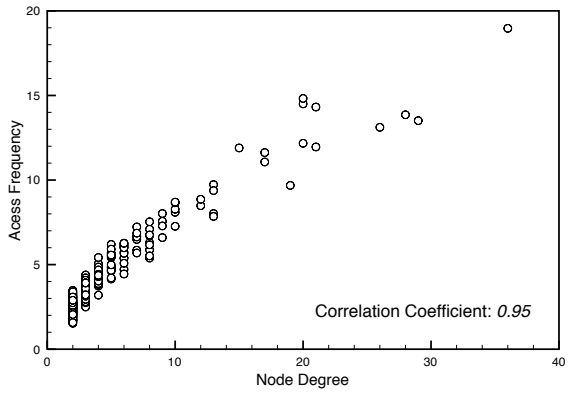
Figura 4.88 – Frequência de acesso obtidas para os vértices da rede WS, considerando 26 agentes de memória 10 sem susceptibilidade a erros.



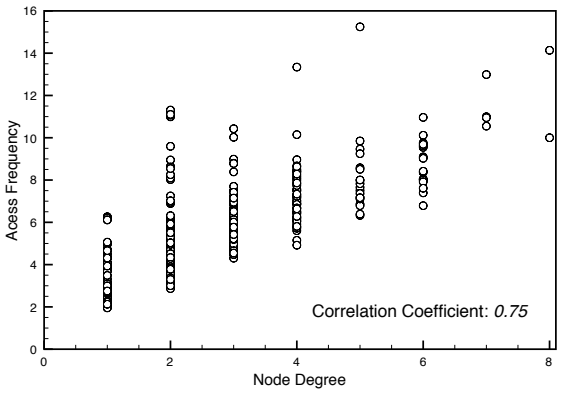
(a) Wikipédia (WIKI).



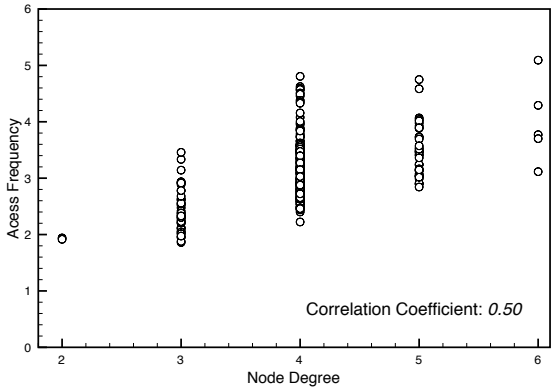
(b) Erdös-Rényry (ER).



(c) Barabási-Albert (BA).



(d) Geográfica (GEO).



(e) Watts-Strogatz (WS).

Figura 4.89 – Correlação da frequência de acesso (*Access Frequency*) com o grau (*Node Degree*) dos vértices para as redes consideradas.

5 *Conclusões*

As metodologias de caracterização de redes complexas ainda constituem um problema em aberto. Apesar do crescimento da abrangência de aplicações de redes complexas, novas metodologias, diferentes daquelas consideradas tradicionais, ainda são pouco desenvolvidas, e, muitas vezes, vistas com ceticismo pela comunidade científica. Infelizmente, alguns pesquisadores ainda têm uma visão um tanto dogmática de que redes complexas são suficientemente caracterizadas apenas por distribuições de grau e valores médios do coeficiente de aglomeração, e classificadas unicamente de acordo com os modelos teóricos tradicionais. Entretanto, os recentes desenvolvimentos na área de metodologias de caracterização de redes complexas vêm se mostrando muito promissores, revelando propriedades não observáveis por metodologias tradicionais.

Tendo em vista a motivação de estimular o uso e desenvolvimento de novas metodologias em redes complexas, este projeto de mestrado teve como objetivo estudar e caracterizar um conjunto diverso de redes através de metodologias recentemente formalizadas, em especial, pelo uso de métricas concêntricas, criadas pelo grupo de pesquisa ao qual este trabalho está vinculado.

O presente trabalho explorou as novas metodologias no contexto da modelagem de conhecimento, onde foram elaboradas duas redes complexas, uma rede de colaboração entre pesquisadores da Universidade de São Paulo, enquanto a outra, uma rede de conhecimento baseada nas conexões dos artigos de teoremas matemáticos encontrados na Wikipédia.

Ambas as redes foram devidamente caracterizadas pelas medidas concêntricas e tradicionais, assim como diversas outras redes reais e baseadas em modelos teóricos, entretanto com maior ênfase para a rede de colaboração da USP.

Um modelo de aquisição do conhecimento, baseado na caminhada de múltiplos agentes por redes semânticas, foi proposto e aplicado à rede de teoremas da Wikipédia, assim como para outras redes baseadas em modelos teóricos. A caracterização prévia dessas redes contribuiu significamente para o entendimento da dinâmica sugerida.

Durante a elaboração deste projeto também foram desenvolvidos diversos softwares relacionados ao cálculo e análise de novas métricas, assim como um para a visualização interativa de redes complexas. Este último foi essencial para a apresentação e compreensão das outras contribuições descritas neste trabalho.

A rede de colaboração de pesquisadores da USP foi obtida em colaboração com o Sistema Integrado de Bibliotecas da Universidade de São Paulo (SIBi-USP), considerando um banco de dados contendo trabalhos acadêmicos realizados nos anos de 2003 e 2004. O perfil estatístico da rede foi obtido em termos das grandes áreas do conhecimento, das cidades e unidades aos quais os pesquisadores estão vinculados, revelando que a área de conhecimento com maior participação da rede foi a de biológicas, compondo 67% do maior componente conectado. Ao considerar a presença dos vértices tanto no maior componente conectado quanto naqueles desconectados a ele, observou-se maior número de pesquisadores das áreas de exatas e humanas pertencentes aos componentes desconectados do que aqueles das área de humanas. O mesmo comportamento foi observado, para a cidade de São Paulo, revelando que tanto pesquisadores das áreas de exatas e humanas, assim como aqueles que pertencem a algumas unidades de São Paulo, apresentaram pesquisas pouco interdisciplinares ou fechadas em pequenos grupos de pesquisa.

A visualização da rede de colaboração da USP permitiu que fossem observados diversas propriedades interessantes de modo simples e efetivo, através da apresentação de projeções dos vértices coloridos de acordo com alguma métrica ou categoria. Foram observados grupos bem definidos para as diferentes áreas do conhecimento, assim como para as cidades e unidades da USP, revelando também as interfaces de colaboração entre diferentes grupos. Estes últimos resultados indicaram, por exemplo, que vértices da área de biológicas correspondem àqueles com maior interdisciplinariedade pois possuem maior interface de colaboração com diversos grupos de diferentes áreas do conhecimento. A mesma técnica de visualização foi aplicada em conjunto com as medidas tradicionais de redes complexas, revelando, através de uma escala de cores, que há uma distribuição uniforme de vértices com alto coeficiente de aglomeração ao longo da topologia da rede. Em contrapartida, a aplicação da metodologia tradicional através de distribuições e valores médios, pouco contribuiu com a caracterização da rede, revelando apenas que possui alto coeficiente de aglomeração médio e apresenta características de redes livres de escala assim como de pequeno mundo.

A mesma metodologia de visualização de propriedades, usada para a rede de colaboração da USP, foi aplicada à rede de teoremas da Wikipédia, revelando estruturas compostas por vértices que formam arcos ou árvores na região das bordas da rede, comportamento este que não está presente nos principais modelos teóricos comumente usados, assim como não podem

ser detectados por metodologias tradicionais. As métricas de centralidade dos vértices também forneceram informações relevantes sobre os vértices da rede, com aqueles considerados centrais, correspondendo notavelmente aos teoremas fundamentais da matemática.

As propriedades concêntricas também foram aplicadas a diferentes redes, incluindo aquelas elaboradas neste trabalho. Inicialmente, foram obtidos os valores médios de distribuições das propriedades concêntricas ao longo dos níveis concêntricos para os principais modelos teóricos. Resultados mostraram que cada modelo apresentou características únicas em suas curvas de distribuição, revelando tanto propriedades observadas pelas metodologias tradicionais, quando propriedades observadas somente pela nova metodologia.

A caracterização concêntrica de redes reais foi realizada comparando as curvas obtidas para os modelos teóricos com aqueles obtidos para as redes estudadas. Para traçar um perfil mais completo das propriedades concêntricas, diversas redes reais, de diferentes origens foram estudadas. Tanto a rede de colaboração da USP quanto a rede de teoremas da Wikipédia apresentaram comportamentos híbridos, compostos por características dos modelos BA, WS e geográfico, além disso também apresentaram curvas da propriedade taxa de convergência muito diferentes daquelas obtidas para qualquer outro modelo, revelando que as distribuições dos hubs, ao longo da estrutura topológica dessas redes é muito diferente daquelas observadas para os modelos BA. Enquanto a rede BA apresenta hubs muito conectados entre si, essas redes apresentaram uma distribuição uniforme dos hubs ao longo da rede. Tal resultado não seria observável se fosse usada apenas a metodologia tradicional da análise da distribuição do grau.

Com exceção da rede de associação de palavras, os comportamentos das curvas de propriedades concêntricas, observados para as outras redes reais, também apresentaram composições das características de diversos modelos teóricos. Também puderam ser obtidas informações sobre a distribuição dos vértices de alta conectividade ao longo da topologia das redes, representado, principalmente, pela taxa de convergência.

Apesar de todas as redes reais estudadas apresentarem características de redes livres de escala, o acesso aos hubs pôde ser caracterizado em termos da posição e largura do pico presente na taxa de convergência concêntrica. Quando o pico localiza-se na região dos primeiros níveis concêntricos, a rede deve apresentar vértices bem conectados a pequenas distâncias de qualquer vértice, enquanto que para um pico localizado em níveis concêntricos maiores, a rede apresenta hubs mais distantes dos vértices da rede. Já a largura do pico forneceu a informação da velocidade de acesso aos hubs, assim como da uniformidade da distribuição topológica dos mesmos ao longo da rede.

A caracterização de vértices pelas medidas concêntricas foi explorada considerando a rede

de colaboração da USP. Os vértices foram categorizados em grupos de acordo com a metodologia de aglomeração hierárquica aplicada à propriedade do coeficiente de aglomeração concêntrico. Os grupos obtidos apresentaram resultados relevantes para a segregação dos vértices em áreas do conhecimento e departamentos. A aplicação da visualização interativa também revelou que os grupos estavam relacionados com diferentes regiões topológicas da rede, com grupos menores representando as bordas, e o maior representando o centro da rede. Entretanto, como o objetivo desta parte do trabalho era ilustrar a metodologia, apenas 4 grupos foram obtidos, e, acredita-se que os resultados podem melhorar muito realizando o corte do dendrograma de modo a obter maior número de grupos.

Outro método explorado para a caracterização concêntrica de dos vértices em redes complexas foi a utilização de do centro de distribuição do coeficiente de aglomeração que foi aplicado a algumas redes reais e modelos teóricos. Apesar da análise superficial, a propriedade forneceu resultados promissores por revelar regiões topologicamente próximas a aglomerados de vértices, sendo muito efetiva para redes do tipo geográficas. Mesmo sendo promissora, esta métrica ainda não foi explorada eficientemente, de modo que novos estudos devem ser realizados para averiguar sua validade, e portanto, deve ser usada com cautela.

O estudo das novas métricas é finalizado pela análise da relevância das medidas concêntricas através da aplicação de PCA e da metodologia de variáveis canônicas. Os resultados obtidos para a análise canônica considerando diversas métricas, incluindo as concêntricas, da rede de alta tensão dos EUA concordaram com os esperados, classificando a rede como geográfica. Surpreendentemente, para este caso a propriedade mais relevante para a projeção foi coeficiente de aglomeração concêntrico de segundo nível. A metodologia PCA também foi aplicada às redes de aeroportos e de interação de proteínas, no entanto, considerando apenas as medidas concêntricas ao longo de alguns níveis concêntricos. Resultados mostraram que as propriedades mais importantes foram o coeficiente de aglomeração concêntrica e a taxa de convergência, principalmente aqueles obtidos para o segundo nível concêntrico.

Metodologias de caminhadas aleatórias de agentes foram estudas neste trabalho considerando o contexto da simulação de um modelo para a aquisição de conhecimento. Para isso foi desenvolvida uma metodologia baseada na caminhada de múltiplos agentes por uma rede complexa, com cada agente caracterizado por propriedades como memória e susceptibilidade a erros. As informações entre os agentes também podiam ser transmitidas entre eles através de redes de interação. As simulações foram consideradas para a rede de teoremas da Wikipédia, assim como para modelos teóricos.

Inicialmente o desempenho da aquisição do conhecimento foi determinado para diferentes

valores de memória dos agentes. Os resultados revelaram que o aumento da memória dos agentes não reflete em um aumento significativo do desempenho da aquisição de conhecimento, que, dependendo do número de agentes e da susceptibilidade a erros, pode resultar até mesmo em perda de desempenho.

A comparação dos desempenhos entre as diferentes redes de conhecimento, mostrou que a rede da Wikipédia apresentou o pior desempenho, com menos da metade do valor obtido para a rede WS, que foi a rede onde os agentes apresentaram a melhor performance. Acredita-se que o baixo desempenho obtido para a rede de teoremas seja devido a existência de arcos e árvores nas bordas da rede.

A análise da frequência de acesso, isto é a frequência com que os vértices eram visitados pelos agentes, apoiada pela visualização computacional ajudou a determinar por quais regiões os agentes tendem a caminhar mais, apresentando, com exceção da rede WS, alta correlação com o grau dos vértices. Em uma rede WS, os agentes tendem a trafegar muito pelos arcos e pelas regiões que os conectam. O uso da frequência de acesso, se devidamente formalizado e estudado, pode vir a se tornar uma métrica interessante para a determinação de estruturas de vértices dentro de redes complexas.

Em geral, esta dissertação apresentou a aplicação de diversas novas metodologias elaboradas recentemente, apesar de algumas terem sido estudadas de modo superficial, apresentaram resultados muito relevantes e promissores para a caracterização de redes complexas. Novos trabalhos nesta linha de pesquisa poderiam aplicar as metodologias descritas aqui a mais redes reais e modelos teóricos, ajudando na determinação experimental dos perfis concêntricos de redes complexas.

Há a necessidade da formalização analítica e estatística de algumas propriedades. Apesar da compreensão intuitiva dos resultados obtidos pela aplicação de algumas metodologias descritas, sua formalização analítica ou estatística pode ajudar a enriquecer a análise dos resultados, podendo, inclusive, revelar novas características.

As metodologias usadas para gerar as redes reais elaboradas neste trabalho também podem ser usadas no futuro para a criação de novas redes de colaboração ou semânticas. O método usado para gerar a rede de colaboração da USP, pode, por exemplo ser aplicado a outros bancos de dados, ou mesmo considerando trabalhos de outros períodos, de modo a obter um perfil cronológico dessas redes, assim como compará-la com redes de outras instituições.

O método de obtenção de redes semânticas a partir de artigos da Wikipédia pode ser aplicado para a criação de diversas outras redes semânticas considerando categorias de outras áreas do conhecimento. Devido às dificuldades encontradas em gerar redes semânticas automatica-

mente, a classe de redes obtidas pela metodologia descrita apresentam grande potencial para estudos e pesquisas relacionadas à modelagem do conhecimento, e, acredita-se que a aplicação tanto das metodologias descritas aqui, quanto outras, podem ajudar a descrever diferentes características das estruturas do conhecimento.

A aplicação dos métodos descritos aqui e de outra metodologias recentes a essas redes, poderá ajudar a descrever e comparar diferentes facetas do conhecimento.

O modelo de aquisição de conhecimento descrito neste trabalho também necessita de uma formalização elaborada e de maior quantidade de resultados para diversas outras redes. É importante notar que a metodologia de simulação de agentes não necessariamente deve ser aplicada a redes de conhecimento, podendo ajudar a descrever dinâmicas em sistemas de computação paralela como grids, ou contribuir futuramente para sistemas de aquisição de dados através de autômatos.

A utilização de dados reais relacionados a aquisição de conhecimento podem contribuir muito a este trabalho ajudando a validar o modelo e obter algumas conclusões efetivas sobre a otimização do desempenho. Um exemplo de contribuição seria a obtenção de dados sobre a escalabilidade das pesquisas de uma instituição com o número de pesquisadores, assim como a descrição da colaboração entre eles através de uma rede. Entretanto, estes dados não puderam ser encontrados ou extraídos até a conclusão deste trabalho.

Espera-se que com as novas metodologias aqui descritas, assim como os resultados apresentados, outros pesquisadores sejam estimulados a aplica-las em seus trabalhos, enriquecendo tanto qualidade das novas metodologias quanto os resultados dos pesquisadores.

Referências

1 NEWMAN, M. E. J. The structure and function of complex networks. *Siam Review*, v. 45, n. 2, p. 167–256, Jan 2003. Disponível em: <http://dx.doi.org/10.1137-/S003614450342480>. Acesso em: 10 nov. 2009.

2 ALBERT, R.; BARABASI, A.-L. Statistical mechanics of complex networks. *Reviews of Modern Physics*, v. 74, n. 1, p. 47–97, Jan 2002. Disponível em: <http://dx.doi.org/10-.1103/RevModPhys.74.47>. Acesso em: 10 nov. 2009.

3 STROGATZ, S. Exploring complex networks. *Nature*, v. 410, n. 6825, p. 268–276, Jan 2001. Disponível em: <http://dx.doi.org/10.1038/35065725>. Acesso em: 10 nov. 2009.

4 PASTOR-SATORRAS, R.; VESPIGNANI, A. Epidemic dynamics and endemic states in complex networks. *Physical Review E*, v. 63, n. 6, p. 066117, Jan 2001. Disponível em: <http://dx.doi.org/10.1103/PhysRevE.63-.066117>. Acesso em: 10 nov. 2009.

5 EPSTEIN, J. M. Modelling to contain pandemics. *Nature*, v. 460, n. 7256, p. 687, Aug 2009. Disponível em: <http://dx.doi.org/doi:10.1038/460687a>. Acesso em: 10 nov. 2009.

6 BARBOSA, V. C.; DONANGELO, R.; SOUZA, S. R. Emergence of scale-free networks from local connectivity and communication trade-offs. *Physical Review E*, v. 74, n. 1, p. 016113, Jan 2006. Disponível em: <http://dx.doi.org/10.1103/PhysRevE.74.016113>. Acesso em: 10 nov. 2009.

7 CRIADO, R; FLORES, J, GONZALEZ-VASCO, M.I.; PELLO, J. Choosing a leader on a complex network. *Journal of Computational and Applied Mathematics*, v. 204, n. 1, p. 10–17, Jan 2007. Disponível em: <http://dx.doi.org/10.1016/j.cam.2006.04.024>. Acesso em: 10 nov. 2009.

8 NEWMAN, M. E. J. Who is the best connected scientist? A study of scientific coauthorship networks.. In: BEN-NAIM,E.; FRAUENFELDER, H.; TOROCZKAI, Z. *Complex networks*. Berlin: Springer Berlin Heidelberg, 2004, p. 337–370. ISBN 3540223541. Disponível em: <httpa://www.santafe.edu/˜mark/papers/cnlspre.pdf>. Acesso em: 10 nov. 2009.

9 ASSAD, A. A. Leonhard Euler: a brief appreciation. *Networks*, v. 49, n. 3, p. 190–198, 2007. Disponível em: <http://dx.doi.org/10.1002/net.20158>. Acesso em: 10 nov. 2009.

10 EULER, L. Solutio problematis ad geometriam situs pertinentis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, v. 8, p. 128–140, 1736. Disponível em: <http://math.dartmouth.edu/~euler/docs/originals/E053.pdf>. Acesso em: 10 nov. 2009.

11 SANDIFER, C. E. The early mathematics of Leonhard Euler. In: ____ *Solution of a problem relating to the geometry of position*. Washington: MAA, 2007, p. 195–200. ISBN 0883855593.

12 FLORY, P. Molecular size distribution in three dimensional polymers. I. gelation. *Journal of the American Chemical Society*, v. 63, n. 11, p. 3083–3090, Jan 1941. Disponível em: <http://pubs.acs.org/doi/abs/10.1021/ja01856a061>. Acesso em: 10 nov. 2009.

13 RAPOPORT, A. Contribution to the theory of random and biased nets. *Bulletin of Mathematical Biophysics*, v. 19, n. 4, p. 257–277, Dec 1957. Disponível em: <http://www.springerlink.com/content/mn2u24251n05n45j/>. Acesso em: 10 nov. 2009.

14 ERDÖS, P.; RÉNYI, A. On random graphs. I. *Publicationes Mathematicae Debrecen*, v. 6, p. 290–297, 1959. Disponível em: <http://www2.renyi.hu/~p_erdos/1959-11.pdf>. Acesso em: 10 nov. 2009.

15 WATTS, D.; STROGATZ, S. Collective dynamics of 'small-world' networks. *Nature*, v. 393, n. 6684, p. 440–442, Jan 1998. Disponível em: <http://dx.doi.org/10.1038/30918>. Acesso em: 10 nov. 2009.

16 ALBERT, R.; JEONG, H.; BARABASI, A.-L. Internet - diameter of the world-wide web. *Nature*, v. 401, n. 6749, p. 130–131, Jan 1999. Disponível em: <http://dx.doi.org/10.1038-/43601>. Acesso em: 10 nov. 2009.

17 BARABASI, A.-L.; ALBERT, R.; JEONG, H. Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A*, v. 281, n. 1-4, p. 69–77, Jun 2000. Disponível em: <http://dx.doi.org/10.1016/S0378-4371(00)00018-2>. Acesso em: 10 nov. 2009.

18 BORNHOLDT, S.; SCHUSTER, H. G. Handbook of graphs and networks: from the genome to the internet. Weinheim: Wiley-JCH, 2003. ISBN 3527403361.

19 AMARAL, L.; OTTINO, J. Complex networks - augmenting the framework for the study of complex systems. *European Physical Journal B*, v. 38, n. 2, p. 147–162, Jan 2004.

Disponível em: <http://dx.doi.org/10.1140/epjb/e2004-00110-5>. Acesso em: 10 nov. 2009.

20 BARABASI, A.-L.; OLTVAI, Z. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, v. 5, n. 2, p. 101–U15, Jan 2004. Disponível em: <http://dx.doi.org/10.1038/nrg1272>. Acesso em: 10 nov. 2009.

21 SCOTT, J. Social network analysis: a handbook. London: SAGE, 2000. p. 208. ISBN 0803984804.

22 NEWMAN, M. E. J.; PARK, J. Why social networks are different from other types of networks. *Physical Review E*, v. 68, n. 3, p. 036122, Jan 2003. Disponível em: <http://dx.doi.org/10.1103/PhysRevE.68.036122>. Acesso em: 10 nov. 2009.

23 WASSERMAN, S.; FAUST, K. Social network analysis: methods and applications. Cambridge: Cambridge University Press, 1994. p. 825. ISBN 0521387078.

24 COSTA, L. F. SPORNS, O.; ANTIQUEIRA, L.; NUNES, M. H. V.;OLIVEIRA Jr, O. N. Correlations between structure and random walk dynamics in directed complex networks. *Applied Physics Letters*, v. 91, n. 5, p. 054107, Jan 2007. Disponível em: <http://dx.doi.org/10.1063/1.2766683>. Acesso em: 10 nov. 2009.

25 ROSVALL, M.; SNEPPEN, K. Self-assembly of information in networks. *Europhysics Letters*, v. 74, n. 6, p. 1109–1115, Jan 2006. Disponível em: <http://dx.doi.org/10.1209/epl-/i2006-10064-2>. Acesso em: 10 nov. 2009.

26 CATTUTO, C.; BARRAT, A.; BALDASSARRI, A.; SCHEHR, G.; LORETO, V. Collective dynamics of social annotation. *Proceedings of the National Academy of Sciences of the United States of America*, v. 106, n. 26, p. 10511–10515, Jan 2009. Disponível em: <http://dx.doi.org/10.1073/pnas.0901136106>. Acesso em: 10 nov. 2009.

27 ROCHA, L. E. C. *Redes acopladas: estrutura e dinâmica*, 2007, 123p. Dissertação (Mestrado) — Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2007.

28 WANG, P.; GONZALEZ, M. C.; HILDALGO, C. A.; BARABASI, A-L. Understanding the spreading patterns of mobile phone viruses. *Science*, v. 324, n. 5930, p. 1071–1076, Jan 2009. Disponível em: <http://dx.doi.org/10.1126/science.1167053>. Acesso em: 10 nov. 2009.

29 COSTA, L. F. The hierarchical backbone of complex networks. *Physical Review Letters*, v. 93, n. 9, p. 098702, Jan 2004. Disponível em: <http://dx.doi.org/10.1103/PhysRevLett-.93.098702>. Acesso em: 10 nov. 2009.

30 COSTA, L. F.; SILVA, F. N. Hierarchical characterization of complex networks. *Journal of Statistical Physics*, v. 125, n. 4, p. 845–876, Jan 2006. Disponível em: <http://dx.doi.org/10.1007/s10955-006-9130-y>. Acesso em: 10 nov. 2009.

31 CHENG, T.; WANG, H.; WANG, L. A. Study of Tacit Knowledge Transfer Based on Complex Networks Technology in Hierarchical Organizations. In: ZHOU, J. (ed.) *Complex sciences*. Berlin: Springer Berlin Heidelberg, 2009, p. 1485–1494. ISBN 978-3-642-02468-9.

32 BRACHMAN, R. What is-a is and isn't: An analysis of taxonomic links in semantic networks. *Computer*, v. 16, n. 10, p. 30 – 36, Oct 1983. Disponível em: <http://dx.doi.org-/10.1109/MC.1983.1654194>. Acesso em: 10 nov. 2009.

33 RAVASZ, E.; BARABASI, A.-L. Hierarchical organization in complex networks. *Physical Review E*, v. 67, n. 2, p. 026112, Feb 2003. Disponível em: <http://dx.doi.org/10.1103-/PhysRevE.67.026112>. Acesso em: 10 nov. 2009.

34 HELBIG, H. Knowledge representation and the semantics of natural language. Berlin: Springer Berlin Heidelberg, 2005. ISBN 3540244611.

35 JACKSON, P. C. Introduction to artificial intelligence. 2nd ed. New York: Dover Publications, 1985. ISBN 048624864X.

36 TADIC, B. *Exploring Complex Graphs by Random Walks*. Jan 2003. Disponível em: <http://arxiv.org/abs/cond-mat/0310014v1>. Acesso em: 10 nov. 2009.

37 COSTA, L. F. Learning about knowledge: a complex network approach. *Physical Review E*, v. 74, n. 2, p. 026103, Jan 2006. Disponível em: <http://dx.doi.org/10.1103/PhysRevE-.74.026103>. Acesso em: 10 nov. 2009.

38 SILVA, F. N.; COSTA, L. F. *Identifying complex networks by random walks*. Jan 2006. Disponível em: <http://arxiv.org/abs/physics/0612121v1>. Acesso em: 10 nov. 2009.

39 BATISTA, J. B.; COSTA, L. da F. *Knowledge acquisition by networks of interacting agents in the presence of observation errors*. Jan 2008. Disponível em: <http://arxiv.org/abs/0807-.2031v2>. Acesso em: 10 nov. 2009.

40 SAJJA, P. Multi-agent system for knowledge-based access to distributed databases. *Interdisciplinary Journal of Information, Knowledge, and Management*, v. 3, p. 1–9, 2008. Disponível em: <http://ijikm.org/Volume3/IJIKMv3p001-009Sajja106.pdf>. Acesso em: 10 nov. 2009.

41 HU, Y. Efficient and high quality force-directed graph drawing. *The Mathematica Journal*, v. 10, n. 1, p. 1–37, May 2006. Disponível em: <http://www.research.att.com/~yifanhu-/PUB/graph_draw_small.pdf>. Acesso em: 10 nov. 2009.

42 BRANK, J. *Drawing graphs using simulated annealing and gradient descent*. Disponível em: <http://kt.ijs.si/Dunja/SiKDD2004/Papers/JanezBrank-GraphDrawing.pdf>. Acesso em: 10 nov. 2009.

43 WALSHAW, C. A multilevel algorithm for force-directed graph drawing. In: ____ *Graph drawing*. Berlin: Springer Berlin Heidelberg, 2001. v. 1984/2001, p. 31–55. Disponível em: <http://dx.doi.org/10.1007/3-540-44541-2_17>. Acesso em: 10 nov. 2009.

44 FRISHMAN, Y.; TAL, A. Online dynamic graph drawing. *IEEE Transactions on Visualization and Computer Graphics*, v. 14, n. 4, p. 727–740, Jan 2008. Disponível em: <http://dx.doi.org/10.1109/TVCG.2008.11>. Acesso em: 10 nov. 2009.

45 KUMAR, G.; GARLAND, M. Visual exploration of complex time-varying graphs. *IEEE Transactions on Visualization and Computer Graphics*, v. 12, n. 5, p. 805–812, Jan 2006. Disponível em: <http://dx.doi.org/10.1109/TVCG.2006.193>. Acesso em: 10 nov. 2009.

46 COSTA, L. 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, Jan 2007. Disponível em: <http://dx.doi.org/10.1080-/00018730601170527>. Acesso em: 10 nov. 2009.

47 AMARAL, L. A. N.; SCALA, A.; BARTHELEMY, M.; STANLEY, H. E. Classes of small-world networks. *Proceedings of the National Academy of Sciences of the United States of America*, v. 97, n. 21, p. 11149–11152, Jan 2000. Disponível em: <http://dx.doi.org/10.1073/pnas.200327197>. Acesso em: 10 nov. 2009.

48 BARABASI, A.-L.; ALBERT, R. Emergence of scaling in random networks. *Science*, v. 286, n. 5439, p. 509–512, Jan 1999. Disponível em: <http://dx.doi.org/10.1126/science-.286.5439.509>. Acesso em: 10 nov. 2009.

49 HAYASHI, Y. *A Review of Recent Studies of Geographical Scale-Free Networks*. Dec 2005. Disponível em: <http://arxiv.org/abs/physics/0512011v2>. Acesso em: 10 nov. 2009.

50 NEWMAN, M. E. J.; LEICHT, E. A. Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences of the United States of America*, v. 104, n. 23, p. 9564–9569, Jan 2007. Disponível em: <http://dx.doi.org/10.1073/pnas-.0610537104>. Acesso em: 10 nov. 2009.

51 NEWMAN, M. E. J. The mathematics of networks. *The new palgrave encyclopedia of economics*, v. 45, p. 167–256, Jan 2007. Disponível em: <http://www-personal.umich.edu-/~mejn/papers/palgrave.pdf>. Acesso em: 10 nov. 2009.

52 GODSIL, C. D.; ROYLE, G. Algebraic graph theory. Berlin: Springer Berlin Heidelberg, 2001. (Graduate texts in mathematics, v. 207). ISBN 0387952411.

53 SABIDUSSI, G. Centrality index of a graph. *Psychometrika*, v. 31, n. 4, p. 581–581, Jan 1966. Disponível em: <http://dx.doi.org/10.1007/BF02289527>. Acesso em: 10 nov. 2009.

54 BRANDES, U. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, v. 25, n. 2, p. 163–177, Jan 2001. Disponível em: <http://www.inf.uni-konstanz-.de/algo/publications/b-fabc-01.pdf>. Acesso em: 10 nov. 2009.

55 FREEMAN, L. C. Set of measures of centrality based on betweenness. *Sociometry*, v. 40, n. 1, p. 35–41, Jan 1977. Disponível em: <http://dx.doi.org/10.2307/3033543>. Acesso em: 10 nov. 2009.

56 ESTRADA, E.; RODRIGUEZ-VELAZQUEZ, J. Subgraph centrality in complex networks. *Physical Review E*, v. 71, n. 5, p. 056103, Jan 2005. Disponível em: <http://dx.doi.org/10.1103/PhysRevE.71.056103>. Acesso em: 10 nov. 2009.

57 DUDA, R. O.; HART, P. E.; STORK, D. G. Pattern classification. Weinheim: Wiley-JCH, 2001. ISBN 0471056693.

58 COSTA, L. F.; CESAR JR, R. M. Shape classification and analysis: Theory and practice. 2nd ed. Boca Raton: CRC Press, 2009. (*Image Processing Series*, v. 10). ISBN 0849379296.

59 CAMPBELL, N.; ATCHLEY, W. The geometry of canonical variate analysis. *Systematic Zoology*, v. 30, n. 3, p. 268–280, Sep 1981. Disponível em: <http://dx.doi.org/10.2307-/2413249>. Acesso em: 10 nov. 2009.

60 FISHER, R. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, v. 7, p. 179–188, Jan 1936. Disponível em: <http://hdl.handle.net/2440/15227>. Acesso em: 10 nov. 2009.

61 BATAGELJ, V.; MRVAR, A. Pajek - Analysis and visualization of large networks. In: WALSHAW, C. *Graph drawing*. Berlin: Springer Berlin Heidelberg, 2001, p. 8–11. (*Lecture Notes in Computer Science*. v. 2265). Disponível em: <http://dx.doi.org/10.1007-/3-540-45848-4_54>. Acesso em: 10 nov. 2009.

62 HOON, M.; IMOTO, S.; NOLAN, J.; MIYANO, S. Open source clustering software. *Bioinformatics*, v. 20, n. 9, p. 1453–1454, Jan 2004. Disponível em: <http://dx.doi.org/10-.1093/bioinformatics/bth078>. Acesso em: 10 nov. 2009.

63 HUA, Y.; XIANG, Y.; CHEN, T.; ABED-MERAIM, K.; MIAO, Y. A new look at the power method for fast subspace tracking. *Digital Signal Processing*, v. 9, n. 4, p. 297–314, Jan 1999. Disponível em: <http://dx.doi.org/10.1006/dspr.1999.0348>. Acesso em: 10 nov. 2009.

64 LARSEN, R. J.; MARX, M. L. An introduction to mathematical statistics and its applications. 3rd ed. Englewood Cliffs: Prentice Hall, 2000. ISBN 0139223037.

65 NEWMAN, M. E. J. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, v. 101, Suppl 1, p. 5200–5205, 2004. Disponível em: <http://dx.doi.org/10.1073/pnas-.0307545100>. Acesso em: 10 nov. 2009.

66 NEWMAN, M. E. J. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical Review E*, v. 64, n. 1, p. 016132, Jan 2001. Disponível em: <http://dx.doi.org/10.1103/PhysRevE.64.016132>. Acesso em: 10 nov. 2009.

67 NEWMAN, M. E. J. Scientific collaboration networks. i. network construction and fundamental results. *Physical Review E*, v. 64, n. 1, p. 016131, Jan 2001. Disponível em: <http://dx.doi.org/10.1103/PhysRevE.64.016131>. Acesso em: 10 nov. 2009.

68 NEWMAN, M. E. J. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, v. 98, n. 2, p. 404–409, Jan 2001. Disponível em: <http://dx.doi.org/10.1073/pnas.021544898>. Acesso em: 10 nov. 2009.

69 CARDILLO, A.; SCELLATO, S.; LATORA, V. A topological analysis of scientific coauthorship networks. *Physica A*, v. 372, n. 2, p. 333–339, 2006. Disponível em: <http://dx.doi.org/10.1016/j.physa.2006.08.059>. Acesso em: 10 nov. 2009.

70 COSTA, L. F.; TOGNETTI, M. A. R.; SILVA, F. N. Concentric characterization and classification of complex network nodes: Application to an institutional collaboration

network. *Physica A*, v. 387, n. 24, p. 6201–6214, Jan 2008. Disponível em: <http://dx.doi-.org/10.1016/j.physa.2008.06.034>. Acesso em: 10 nov. 2009.

71 COSTA, L. F. *On the Dynamics of the $h-$index in Complex Networks with Coexisting Communities*. Jan 2006. Disponível em: <http://arxiv.org/abs/physics/0609116v1>. Acesso em: 10 nov. 2009.

72 VAZQUEZ, A. *Statistics of citation networks*. Jan 2001. Disponível em: <http://arxiv.org-/abs/cond-mat/0105031v1>. Acesso em: 10 nov. 2009.

73 BILKE, S.; PETERSON, C. *Topological Properties of Citation and Metabolic Networks*. Jan 2001. Disponível em: <http://arxiv.org/abs/cond-mat/0103361v1>. Acesso em: 10 nov. 2009.

74 WU, W.; HAN, Y.; LI, D. The topology and motif analysis of journal citation networks. In: INTERNATIONAL CONFERENCE ON COMPUTER SCIENCE AND SOFTWARE ENGINEERING, 2008, Washington. *Proceedings...* Washington: IEEE Computer Society, 2008. p. 287–293. ISBN 978-0-7695-3336-0.

75 BATAGELJ, V.; MRVAR, A. *Pajek Datasets*. Disponível em: <http://vlado.fmf.uni-lj.si-/pub/networks/data/>. Acesso em: 10 nov. 2009.

76 KISS, G.; ARMSTRONG, C.; MILROY, R.; AND PIPER, J. An associative thesaurus of English and its computer analysis. In: AITKEN, A. J.; Bailey, R. W.; HAMILTON-SMITH, N. (ed.) *The Computer and Literary Studies*. Edinburgh: University of Edinburgh Press, 1973. ISBN 0852242328.

77 BU, D. et al. Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Research*, v. 31, n. 9, p. 2443–50, May 2003. Disponível em: <http://dx.doi.org/10.1093/nar/gkg340>. Acesso em: 10 nov. 2009.

78 KLEINBERG, J. Authoritative sources in a hyperlinked environment. *Journal of the Association Computing Machinery*, v. 46, n. 5, Sep 1999. Disponível em: <http://dx.doi-.org/10.1145/324133.324140>. Acesso em: 10 nov. 2009.

79 KAMADA, T.; KAWAI, S. An algorithm for drawing general undirected graphs. *Information Processing Letters*, v. 31, n. 1, p. 7 – 15, Sep 1989. Disponível em: <http://dx.doi.org/10.1016/0020-0190(89)90102-6>. Acesso em: 10 nov. 2009.

80 FRUCHTERMAN, T. M.; REINGOLD, E. M. Graph drawing by force-directed placement. *Software-Practice and Experience*, v. 21, n. 11, p. 1129 – 1164, Nov 1991. Disponível em: <http://dx.doi.org/10.1002/spe.4380211102 >. Acesso em: 10 nov. 2009.

81 JACKSON, J.D. Classical Electrodynamics. 7th ed. New York: Wiley-JCH, 1962.

APÊNDICE A - Algoritmo de visualização.

Este apêndice descreve sucintamente o algoritmo e os principais conceitos necessários para o posicionamento de vértices das visualizações de redes complexas elaboradas no software Network 3D.

Diferentes metodologias são empregadas na visualização de redes complexas, algumas delas são herdadas dos métodos de visualização de grafos, outras tentam se aproveitar de estruturas como hubs em redes reais para aprimorar o procedimento.

Em geral, o problema de se obter uma visualização para uma rede se baseia em determinar os vetores de posição (2D ou 3D) referentes a cada um dos vértices da rede complexa, isto é, mapear, para cada vértice uma posição no espaço. Dependendo do objetivo, diferentes metodologias podem ser empregadas. Outras propriedades como cor, forma, tamanho, etc; podem ser atribuídos aos vértices ou arestas, tanto usando informações extras, quanto utilizando a própria estrutura topológica.

O que se busca, em geral, são métodos automatizados que com poucos parâmetros gerem uma representação gráfica agradável ao usuário final, tanto no sentido estético quanto na utilidade como ferramenta. No entanto, principalmente com o objetivo de divulgação, o processo de disposição dos vértices e arestas em uma figura pode ser realizada manualmente por um artista ou pelo próprio cientista.

O algoritmo aqui apresentado baseia-se em uma metodologia de disposição dos vértices pelo uso de métodos dirigidos por forças (79, 80). Métodos dirigidos por forças utilizam uma analogia física onde cada vértice é representado por a uma partícula carregada e cada aresta por uma interação de forças entre elas; a técnica, então, baseia-se em encontrar a disposição dos vértices que apresente os menores valores de energia para o sistema. Essa metodologia é uma das mais usadas, pois não necessita de nenhuma informação extra sobre a rede, tendo em vista que apenas sua topologia será responsável pelo conjunto de posições.

Diferentes modelos de forças podem ser usados, como molas, atração e repulsão eletromagnéticas, forças do tipo Van der Waals, ou forças derivadas de potenciais com diversos mínimos locais. Já a determinação dos mínimos de energia também podem ser determinadas de diversas maneiras, como por exemplo através técnicas de *simulated annealing* e funções

de esfriamento.

Apesar de haverem técnicas extremamente sofisticadas para o cálculo eficiente do posicionamento dos vértices, o algoritmo descrito neste trabalho, mesmo não sendo tão eficiente é muito mais simples.

O algoritmo baseia-se largamente no método de Fruchtermen-Riengold (80), que abrange tanto forças repulsivas quanto forças atrativas. Esta técnica considera uma força atrativa entre dois vértices quando estes estão ligados por uma aresta, e, uma força repulsiva provida de todos os outros vértices. O sistema é análogo ao problema físico de N partículas carregadas, com algumas ligadas entre si por molas que seguem uma lei de força quadrática. Enquanto a força de atração aumenta linearmente com a distância, as forças de repulsão decaem com o quadrado da distância, de acordo com a equação:

$$\vec{F}_{(a)j} = \sum_{(i,j)\in\mathscr{E}} a_{ij}(\vec{R}_i - \vec{R}_j)^2 \hat{r}_{ij} \quad (1)$$

$$\vec{F}_{(r)j} = \sum_{i\in\mathscr{V}} \frac{-b_{ij}}{(\vec{R}_i - \vec{R}_j)^2} \hat{r}_{ij} \quad (2)$$

Nesse modelo a distância preferencial $d_{ij} = |\vec{R}_{i(ideal)} - \vec{R}_{j(ideal)}|$ será dada quando as forças são nulas, isto é $d_{ij} = \left(\frac{b_{ij}}{a_{ij}}\right)^{\frac{1}{4}}$.

A busca pela energia mínima é feita através de simulações reais do sistema, através da integração das equações diferencias do movimento físico dos vértices. Diferentemente das metodologias usuais, a busca pelo mínimo é realizada através da adição de uma força viscosa ao sistema, β, de modo que a cada instante de tempo da simulação o sistema perde energia.

Como o objetivo das simulações é chegar aos estados que correspondem a menor energia, e não exatamente a simulação correta das partículas, certos formalismos podem ser relaxados.

As equações diferenciais de movimento semelhantes às da física (só que na forma adimensional):

$$\frac{d\vec{R}_j}{dt} = \begin{cases} a_{ij}(\vec{R}_i - \vec{R}_j)^2 \hat{r}_{ij} + \frac{-b_{ij}}{(\vec{R}_i - \vec{R}_j)^2} \hat{r}_{ij}, \text{se } \{(i,j)\in\mathscr{E}\} \\ \frac{-b_{ij}}{(\vec{R}_i - \vec{R}_j)^2} \hat{r}_{ij}, \text{outros casos} \end{cases} \quad (3)$$

A resolução pode ser realizada de forma numérica, por exemplo usando a integração de

Euler:

Algoritmo .0.1: FRUCHTERMEN-RIENGOLD($\vec{R}_{Inicial}, \beta$)

$\vec{R} \leftarrow \vec{R}_{Inicial}$
$\vec{V} \leftarrow \vec{0}$
// Posições e velocidades recebem condições iniciais.

repetir
para cada $\{i : v_i \in V\}$
faça
$F_{\text{total}} \leftarrow \vec{F}_{(a)i} + \vec{F}_{(r)i}$
$V_i \leftarrow V_i + (F_{\text{total}} - \beta V_i)\delta t$
$\vec{R}_i \leftarrow \vec{R}_i + V_i \delta t$
// Realiza a integração numérica.
até satisfeito

É importante notar que o cálculo das forças atrativas de cada vértice é de complexidade $O(E)$ e o cálculo das forças repulstivas, $O(N)$, com N sendo o número de vértices e E o número de arestas da rede. Portanto a ordem do algoritmo completo é $O(N(N+E))$, entretanto, dividindo o algoritmo em duas partes, uma para o cálculo das forças atrativas e outra para as forças repulsivas, pode-se obter um desempenho melhor para as forças atrativas, inteirando sobre as arestas, resultando em uma complexidade $O(N^2+E)$. Em geral, $E \ll N^2$, portanto uma complexidade $O(N^2)$ é o mínimo que pode-se obter através de otimizações simples.

Uma forma de reduzir o custo computacional do algoritmo é através da metodologia de expansão multipolar (41). A expansão multipolar baseia-se em um conceito físico usado, por exemplo, em eletromagnetismo para obter aproximações dos campos elétricos de distribuições genéricas de cargas (81). Através dessa metodologia foi possível reduzir o custo computacional do cálculo da força repulsiva de $O(N^2)$ para $O(NLogN)$, tornando viável a visualização de redes com grande número de vértices.