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

Guilherme de Guzzi Bagnato

**Análise estrutural de redes complexas modulares por
meio de caminhadas auto-excludentes**

São Carlos

2018

Guilherme de Guzzi Bagnato

**Análise estrutural de redes complexas modulares por
meio de caminhadas auto-excludentes**

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

Área de concentração: Física Aplicada
Opção: Física Computacional

Orientador: Prof. Dr. Gonzalo Travieso

Versão original

**São Carlos
2018**

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

Bagnato, Guilherme de Guzzi

Análise estrutural de redes complexas modulares por meio de caminhadas auto-excludentes / Guilherme de Guzzi Bagnato; orientador Gonzalo Travieso -- São Carlos, 2018. 131 p.

Tese (Doutorado - Programa de Pós-Graduação em Física Aplicada Computacional) -- Instituto de Física de São Carlos, Universidade de São Paulo, 2018.

1. Caminhada aleatória. 2. Redes Complexas. 3. Comunidades. I. Travieso, Gonzalo, orient. II. Título.

FOLHA DE APROVAÇÃO

Guilherme de Guzzi Bagnato

Tese apresentada ao Instituto de Física de São Carlos da Universidade de São Paulo para obtenção do título de Doutor em Ciências. Área de Concentração: Física Aplicada - Opção: Física Computacional.

Aprovado(a) em: 27/04/2018

Comissão Julgadora

Dr(a). Gonzalo Travieso

Instituição: (IFSC/USP)

Dr(a). César Henrique Comin

Instituição: (UFSCar/São Carlos)

Dr(a). Luciano da Fontoura Costa

Instituição: (IFSC/USP)

Dr(a). Alexandre Souto Martinez

Instituição: (FFCLRP/USP)

Dr(a). Alneu de Andrade Lopes

Instituição: (ICMC/USP)

À minha eterna filha Livia Ciol Bagnato

AGRADECIMENTOS

A toda minha família, em especial: Aos meus pais, Fernando e Lislane, pelo amor incondicional e eterno incentivo aos estudos, aos meus irmãos, Gabriel e Gabriela, pelos amorosos atritos que só os irmão sabem resolver, à minha avó Therezinha, por ser tão presente em minha vida.

A minha esposa Heloísa, pelo amor, respeito, confiança e força que me dá todos dias. Sem ela esta tese pareceria um grande rascunho, obrigado pela revisão sintática e ortográfica.

Ao professor Gonzalo por me dar a oportunidade de desenvolver este trabalho, estando sempre disposto a me dar conselho e sugerir melhorias.

Ao companheiro de sala e de trabalho José Ricardo, pelas horas de café e discussões científicas e não científicas que ajudaram muito durante todo o período do doutorado.

Aos amigos Thomas e César por despertar meu interesse em redes complexas e sempre estarem dispostos a sanar minhas dúvidas e discutir sobre os mais diversos assuntos.

Aos amigos de graduação, Tiago, Paulo, Lucas, Fernando e Denis, pelas frequentes discussões que não falam de nada e que não levam a lugar nenhum, mas que são indispensáveis.

A todos os meus professores, por despertarem meu interesse pela academia e, consequentemente, me incentivarem durante toda a pós-graduação.

À Universidade de São Paulo e ao Instituto de Física de São Carlos, por me darem o privilégio de estudar em um dos melhores centros acadêmicos do mundo.

À CAPES, pelo apoio financeiro.

“Tenha coragem de seguir o seu próprio coração e a sua intuição. Eles de alguma maneira já sabem o que você realmente quer se tornar. Todo o resto é secundário.”

Steve Jobs

RESUMO

BAGNATO, G. G. **Análise estrutural de redes complexas modulares por meio de caminhadas auto-excludentes**. 2018. 131p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2018.

O avanço das pesquisas em redes complexas proporcionou desenvolvimentos significativos para a compreensão de sistemas complexos. Uma rede complexa é modelada matematicamente por meio de um grafo, onde cada vértice representa uma unidade dinâmica e suas interações são simbolizadas por um conjunto de arestas. Para se determinar propriedades estruturais desse sistema, caminhadas aleatórias tem-se mostrado muito úteis pois dependem apenas de informações locais (vértices vizinhos). Entre elas, destaca-se o passeio auto-excludente (SAW) que possui a restrição de não visitar um vértice que já foi alcançado, ou seja, apresenta memória do caminho percorrido. Por este motivo o SAW tem apresentado melhores resultados do que caminhantes sem restrição, na exploração da rede. Entretanto, por não se tratar de um processo Markoviano ele apresenta grande complexidade analítica, tornando indispensável o uso de simulações computacionais para melhor compreensão de sua dinâmica em diferentes topologias. Mesmo com as dificuldades analíticas, o SAW se tornou uma ferramenta promissora na identificação de estruturas de comunidades. Apesar de sua importância, detecção de comunidades permanece um problema em aberto devido à alta complexidade computacional associada ao problema de otimização, além da falta de uma definição formal do significado de comunidade. Neste trabalho, propomos um método de detecção de comunidades baseado em SAW para extrair uma estrutura de comunidades da rede otimizando o parâmetro modularidade. Combinamos características extraídas desta dinâmica com a análise de componentes principais para posteriormente classificar os vértices em grupos por meio da clusterização hierárquica aglomerativa. Para avaliar a performance deste novo algoritmo, comparamos os resultados com outras quatro técnicas populares: Girvan-Newman, Fastgreedy, Walktrap e Infomap, aplicados em dois tipos de redes sintéticas e nove redes reais diversificadas e bem conhecidas. Para os benchmarks, esta nova técnica produziu resultados satisfatórios em diferentes combinações de parâmetros, como tamanho de rede, distribuição de grau e número de comunidades. Já para as redes reais, obtivemos valores de modularidade superior aos métodos tradicionais, indicando uma distribuição de grupos mais adequada à realidade. Feito isso, generalizamos o algoritmo para redes ponderadas e digrafos, além de incorporar metadados à estrutura topológica a fim de melhorar a classificação em grupos.

Palavras-chave: Caminhada aleatória. Redes complexas. Comunidades.

ABSTRACT

BAGNATO, G. G **Structural analysis of modular complex networks through self avoiding walk**. 2018. 131p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2018.

The progress in complex networks research has provided significant understanding of complex systems. A complex network is mathematically modeled by a graph, where each vertex represents a dynamic unit and its interactions are symbolized by groups of edges. To determine the system structural properties, random walks have shown to be a useful tool since they depend only on local information (neighboring vertices). Among them, the self-avoiding walk (SAW) stands out for not visiting vertices that have already been reached, meaning it can record the path that has been travelled. For this reason, SAW has shown better results when compared to non-restricted walkers network exploration methods. However, as SAW is not a Markovian process, it has a great analytical complexity and needs computational simulations to improve its dynamics in different topologies. Even with the analytical complexity, SAW has become a promising tool to identify the community structure. Despite its significance, detecting communities remains an unsolved problem due to its high computational complexity associated to optimization issues and the lack of a formal definition of communities. In this work, we propose a method to identify communities based on SAW to extract community structure of a network through optimization of the modularity score. Combining technical features of this dynamic with principal components analyses, we classify the vertices in groups by using hierarchical agglomerative clustering. To evaluate the performance of this new algorithm, we compare the results with four other popular techniques: Girvan-Newman, Fastgreedy, Walktrap and Infomap, applying the algorithm in two types of synthetic networks and nine different and well known real ones. For the benchmarks, this new technique shows satisfactory results for different combination of parameters as network size, degree distribution and number of communities. As for real networks, our data shows better modularity values when compared to traditional methods, indicating a group distribution most suitable to reality. Furthermore, the algorithm was adapted for general weighted networks and digraphs in addition to metadata incorporated to topological structure, in order to improve the results of groups classifications.

Keywords: Random walks. Complex networks. Communities

LISTA DE ILUSTRAÇÕES

Figura 1 – Estrutura da tese representada por uma grafo. As caixas em verde representam os capítulos, que se ligam com as respectivas seções, indicadas pelas caixas amarelas. Arestas mais grossas (verde) mostram a sequência dos capítulos, arestas médias (vermelhas) ligam as seções aos seus capítulos e as conexões mais finas indicam que ambas as partes do trabalho estão relacionadas.	37
Figura 2 – Exemplos de redes (a) simples, (b) com peso (representado na largura da aresta) e (c) direcionada (indicada pelas setas).	40
Figura 3 – (a) Rede simples com $N = 4$ e $E = 4$ com as seguintes conectividades ($k_1 = 1, k_2 = 3, k_3 = 2, k_4 = 2$). Já (b) mostra a distribuição de graus da respectiva rede.	41
Figura 4 – (a) Rede simples contendo dois caminhos mínimos entre os vértice i, j (cores verde e amarela). (b) Digrafo mostrando um único caminho mínimo entre o vértice origem i e o vértice destino j , evidenciando que, diferentemente do caso anterior, não existe geodésica entre j e i	42
Figura 5 – (a) Exemplo de uma rede simples com dois componentes conectados, destacados pelas regiões em cinza, com o componente gigante contendo seis vértices. (b) Rede direcionada ilustrando a diferença entre componentes fracamente e fortemente conectados. Neste caso, o primeiro tipo de componente engloba todos os vértices da rede enquanto o segundo é dividido em três partes, destacadas em cinza.	43
Figura 6 – (a) Exemplo de um grafo aleatório de Erdős e Rényi. (b) Distribuição de conexões para uma rede com $N = 50000$, $p = 0.0002$ e, consequentemente, $\langle k \rangle = 10$	45
Figura 7 – (a) Exemplo de uma rede livre de escala gera pelo modelo BA. (b) Distribuição das conexões para uma rede BA com $N = 1000000$ e $m = 4$	46
Figura 8 – Um grafo simples com quatro comunidades indicadas pelas cores vermelha, verde, azul e amarela.	47
Figura 9 – <i>Benchmark</i> de Girvan e Newman (BGM). As três redes correspondem a realizações do modelo para (a) $z_{out} = 2$, (b) $z_{out} = 7$ e (c) $z_{out} = z_{in} = 8$. No último caso, os quatro grupos são muito difusos uma vez que estamos no limite da definição de comunidades.	53
Figura 10 – Exemplo de rede gerada pelo modelo LFR com 185 vértices divididos em nove partições bem definidas dado que as distribuições de grau e de tamanho das comunidades seguem uma lei de escala com $\gamma = 2$ e $\beta = 1$, respectivamente. Além disso, $\mu = 0, 1$	55

Figura 11 – Exemplo do passeio auto-excludente em uma rede simples. (a) Ele se inicia no vértice 3 e, no primeiro passo, atinge o nó 4. (b) Uma vez em 4, o caminhante só pode ir para o vértice 2, pois 3 já foi visitado. (c) Partindo de 2, o caminhante atinge obrigatoriamente o nó 1 e consequentemente termina o passeio dado que todos os vértices já foram visitados.	60
Figura 12 – Comparação entre as caminhadas RW no estado estacionário (linhas contínuas) e SAW (pontos com linhas tracejada) para redes tipo (a) ER e (b) BA para diferentes graus médios e mesma quantidade de vértices, $N = 5000$. O gráfico interno da figura (b) corresponde a um aumento da região inicial tal que $0 \leq k \leq 230$ e os valores indicados na escala representam o ponto de cruzamento das curvas vermelha, azul e verde respectivamente.	61
Figura 13 – Distribuições do número de passos para caminhantes realizando o SAW em três modelos distintos de redes: BA, ER e LFR para três tamanhos diferentes $N = 5000, 10000$ e 20000 e $\langle k \rangle$ definido para 10, 20 e 30. Para obter uma boa estatística utilizamos $N = 10^5$ para cada uma das curvas. Ademais, no modelo com comunidades usamos os parâmetros adicionais: $\beta = 2, \gamma = 3, \mu = 0,3$ e os tamanhos das comunidades variando entre 1% e 10% do tamanho da rede.	63
Figura 14 – Frações média e máxima de vértices alcançados em função do tamanho da rede. Novamente analisamos os resultados aplicados aos modelos ER, BA e LFR para $\langle k \rangle = 10, 20, 30$	64
Figura 15 – Classificação de um par de vértices de acordo com $p_{i,j}$ e $\langle l_{i,j} \rangle$. Os vértices na cor verde indicam a origem dos caminhantes enquanto os amarelos o destino. (a) TIPO A: Situação na qual a fração de caminhantes e a quantidade média de passos é elevada. (b) TIPO B: Este grafo representa poucos caminhantes atingindo o destino em poucos passos. (c) TIPO C: Nesta rede temos uma baixa fração de caminhantes percorrendo altas distâncias. (d) TIPO D: Finalmente, muitos caminhantes que saíram do vértice verde atingem o amarelo em poucos passos.	65
Figura 16 – Gráfico de espalhamento construído a partir das medidas definidas nas expressões (3.1) e (3.2) e aplicadas nos modelos BA, ER e LFR, todos com $\langle k \rangle = 8$ e $N = 500$. No último, os parâmetros adicionais foram escolhidos tal que $\beta = 2, \gamma = 3, \mu = 0,5$ e o tamanho das comunidades variando entre 10 e 50. Aqui, utilizamos os quatro padrões de classificação entre vértices definidos na figura 15 para distinguir esses três tipos de rede.	66

- Figura 17 – Distribuição dos valores de $f_{i,j}$ para os modelos (a) ER, (b) BA e (c) LFR. Os parâmetros utilizados aqui foram: $N = 500$, $\langle k \rangle = 8$ e no caso com comunidades, $\beta = 2$, $\gamma = 3$, $\mu = 0,5$ com as partições variando entre $[10, 50]$ vértices cada. Os gráficos internos de (a) e (b) correspondem a ampliação da região inicial das respectivas distribuições. 68
- Figura 18 – Norma (equação 3.5) da diferença de $\mathbf{F}$ em função do número médio de caminhantes por vértice para um intervalo de $\Delta M = 100$ aplicado a redes do modelo LFR com tamanhos 100, 500 e 1000. Para melhor visualização, o eixo das ordenadas está em escala logarítmica. 69
- Figura 19 – (a) Grafo com uma estrutura de comunidades bem definida com dez vértices em cada uma das cinco partições (estes estão numerados sequencialmente para favorecer a visualização dos padrões em $f_{i,j}$). (b) Imagem representando a matriz das probabilidades de transições, como todos os valores são altos não é possível distinguir claramente os cinco grupos. (c) Matriz com os valores de $\langle l_{i,j} \rangle$, aqui fica evidente que os caminhantes realizam trajetórias menores dentro das comunidades e maiores entre elas. (d) Figura indicando a matriz $\mathbf{F}$, as sub-matrizes com valores mais elevados indicam os vértices de uma mesma partição. Os elementos da diagonal são determinados através da expressão 3.6. . 71
- Figura 20 – Fração de vértices classificados corretamente para os diferentes algoritmos de detecção de comunidades submetidos ao *benchmark* de Girvan-Newman. Cada ponto do gráfico é uma média de 100 grafos e as barras de erro correspondem ao desvio padrão. Os métodos Wt e SAW apresentaram os melhores resultados, perdendo precisão em torno de $z_{out} = 7,5$. Por outro lado o IMap, único algoritmo que não minimiza a modularidade, começou a perder qualidade na classificação para $z_{out} = 6$ 75
- Figura 21 – Comparação de cinco algoritmos de detecção de comunidades para redes geradas por meio do *benchmark* LFR para diferentes valores de β e γ com $N = 500$ e $\langle k \rangle = 16$. O tamanho das comunidades variam entre 10 e 50 para cada uma das 100 realizações por μ com seus respectivos desvios padrões. O método do SAW, assim como no BGN, também produziu bons resultados para este conjunto de testes. 76
- Figura 22 – (a) Dendrograma resultante do nosso método de detecção de comunidade para rede do clube de karatê Zachary. As cores azul e vermelho correspondem à divisão da rede após o rompimento do clube, com exceção do vértice 2. (b) Estrutura de comunidade que apresentou o maior valor de modularidade ($Q = 0,4197$). As áreas coloridas são referência à estrutura de mesma cor mostrada no dendrograma. 77

Figura 23 – Distribuição de comunidades para a rede dos Miseráveis. As cores representam a melhor divisão encontrada pelo nosso algoritmo, que alcançou $Q = 0,5467$ com 7 partições. As áreas limitadas pelas linhas trastejadas é a distribuição resultante do método GN e indicam 11 grupos com $Q = 0,5380$	78
Figura 24 – Rede da liga universitária de futebol americano para a temporada do ano 2000. Os vértices são coloridos de acordo com a detecção de comunidades do SAW e e as formas geométricas representam as respectivas conferências. O mapa dos Estados Unidos mostra a influência da localização geográfica na nossa classificação (universidades geograficamente próximas disputam mais partidas entre si). Esta distribuição de comunidades atingiu $Q = 0,6044$ com 10 divisões.	80
Figura 25 – Resultados comparativos da qualidade da detecção de comunidades por meio dos métodos Wt, IMap, GN e SAW em redes artificiais geradas pelo modelo LFR com peso a medida que variamos os parâmetros β e γ . Cada ponto corresponde a média de 50 redes com $N = 500$, $\langle k \rangle = 16$, $\tau = 1,5$, $0,3 \leq \mu = \mu_w \leq 0,8$ e o tamanho das comunidades variando entre 10 e 50. As barras de erros foram determinadas através do desvio padrão.	82
Figura 26 – Detecção de comunidades via SAW na rede do romance Os Miseráveis em sua versão ponderada, onde a força das arestas indicam a frequência com que dois atores contracenam. Diferentemente do caso sem peso, quando obtivemos sete grupos, aqui encontramos seis comunidades. As principais diferenças se deram nos vértices indicados pelas elipses. . . .	84
Figura 27 – Resultado da detecção de comunidades obtida pelo método do SAW na rede ponderada da Guerra dos Tronos. Foram identificadas sete comunidades que representam a influência geopolítica das quatro principais famílias de Westeros (Stark, Lannister, Baratheon e Targaryen) pela disputa do poder. A esquerda temos o mapa onde se passa o romance com os círculos coloridos indicando a região geográfica de cada comunidade. . . .	86
Figura 28 – Divisão da rede de ações em comunidades por meio do algoritmo do SAW ponderado, as cores indicam as partições e as formas geométricas o setor econômico de cada empresa, com o peso das arestas representando a correlação entre duas companhias.	87
Figura 29 – Resultados da aplicação dos métodos GN, IMap e SAW em redes artificiais geradas pelo modelo LFR direcionado para diferentes valores de β e γ a medida que variamos μ em redes com $N = 500$ e $\langle k \rangle = 16$. Cada ponto corresponde a média de 100 redes, o respectivo desvio padrão foi tão baixo que não pôde ser observado nos gráficos.	90

Figura 30 – Rede direcionada dos Blogs de política, onde a direção representa quem citou quem na <i>Web</i> . A divisão em duas comunidades está de acordo com o alinhamento político das páginas referenciadas, em vermelho temos a predominância de <i>sites</i> com ideologia de esquerda, enquanto que em verde, vemos os vértices de tendência mais conservadora, ficando evidente que blogs de mesmo alinhamento político costumam citar outros que concordam com seus ideais.	91
Figura 31 – Exemplo da distribuição aleatória de metadados em uma rede com $N = 40$ dividida em 5 comunidades de mesmo tamanho para (a) $p = 0,4$ (b) $p = 0,6$ (c) $p = 0,8$. As cores indicam as partições e as formas geométricas os respectivos atributos.	94
Figura 32 – Gráfico mostrando a relação linear entre o parâmetro p e o coeficiente de correlação de Pearson. Para cada valor de p foram realizadas 100 amostragens. A rede base para a produção da distribuição de ψ foi gerada pelo modelo LFR com $N = 500$, $\beta = 1$, $\gamma = 3$ e $\mu = 0,5$	95
Figura 33 – (a) Modularidade em função da ponderação artificial entre vértices com mesmo metadado (ζ) à medida que aumentamos a probabilidade p . (b) Gráfico representando a NMI para a classificação de comunidades à medida que variamos (ζ) para $p \in \{0,4; 0,6; 0,8\}$. O valor ideal que encontramos para esta grandeza foi $\zeta = 3$, pois mantém a modularidade elevada e aumenta de forma significativa a informação mútua normalizada. Apesar do baixo número amostral (5 redes por ponto), refletido nas elevadas barras de erros, as curvas não se sobrepõe, deixando o resultado satisfatório.	95
Figura 34 – Comparação da influência dos metadados em redes LFR a medida que aumentamos o parâmetro de mistura (μ) para rede sem metadados (SAW) e com correlação entre os atributos adicionais e as comunidades ($p \in \{0,4; 0,6; 0,8\}$). Cada ponto corresponde a uma média de 100 redes com $N = 500$, $\beta = 1$ e $\gamma = 3$, onde as barras de erros indicam o respectivo desvio padrão.	96
Figura 35 – Exemplo de três métodos de união de grupos usado na hierarquização aglomerativa. (a) <i>Single linkage</i> : união de dois membros de acordo com a menor distância. (b) <i>Complete linkage</i> : Combina dois dados pela maior distância. (c) <i>Average linkage</i> : A junção de duas partições é feita conforme o menor valor médio.	122
Figura 36 – Exemplo do resultado da hierarquização aglomerativa (figura à esquerda) por meio de um dendrograma (imagem à direita). As cores indicam os níveis de agrupamento a medida que se sobe na árvore.	123

Figura 37 – Ilustração do método de construção de uma rede de mercado financeiro a partir de três ativos. Por meio da série temporal do preço, determina-se o retorno diário e posteriormente as conexões entre as ações, ponderadas pela correlação destes retornos.	126
Figura 38 – Distribuição estatística do coeficiente de Pearson (eq. B.1) para todos os pares de ações do índice S&P500 em um período de 10 anos. A média e o desvio padrão deste sistema são respectivamente 0.37 e 0.12.	127

LISTA DE TABELAS

Tabela 1	– Lista de algoritmos utilizados neste trabalho. A primeira coluna indica o nome do método de detecção de comunidades; a segunda, a sigla de identificação do algoritmo; a terceira, a complexidade computacional para redes esparsas e as duas últimas, para qual tipo de grafo a técnica se aplica.	52
Tabela 2	– Resumo de todas as redes utilizadas neste trabalho, indicando o número total de vértices e arestas na maior componente fracamente conectada, sua identificação com relação ao peso nas arestas e se possui direção. .	58
Tabela 3	– Comparação de diferentes métodos e métricas a fim de escolher o par mais adequado para a clusterização hierárquica aglomerativa por meio do SAW. Apesar da modularidade ser estatisticamente igual para as redes do modelo LFR (aproximadamente $0,6393 \pm 0.0005$), a métrica de Bray-Curtis e o método do <i>average linkage</i> (destacados em negrito) obtiveram os melhores resultados em todas as sete redes reais e, conseqüentemente, foram adotados como medidas padrão do nosso algoritmo de detecção de comunidades.	73
Tabela 4	– Comparação da modularidade e do número de partições para os diferentes algoritmos de detecção de comunidades aplicados em algumas redes reais.	81
Tabela 5	– Comparação da otimização da modularidade em três redes reais ponderadas submetidas aos métodos GN, Wt, IMap e SAW. Em todos os casos testados nosso algoritmo foi o que apresentou o maior valor de modularidade.	89
Tabela 6	– Comparação entre as modularidades encontradas por três diferentes métodos (GN, IMap e SAW) nas redes direcionadas dos Blogs de política e do verme <i>C. elegans</i> . Mais uma vez nosso algoritmo se mostrou uma boa alternativa a estas técnicas mais tradicionais.	91
Tabela 7	– Comparação do resultado da detecção de comunidades por meio do SAW entre quatro redes reais com e sem a presença de metadados. Como era de se esperar, observamos uma queda na modularidade devido ao número fixo de comunidades (número de metadados distintos) e um aumento significativo da NMI, revelando uma boa identificação com sistema real. Aqui, assim como no modelo LFR, utilizamos $\zeta = 3$	97
Tabela 8	– <i>Softwares</i> e pacotes de terceiros utilizados para geração de <i>scripts</i> de análises de dados e simulações, construção de gráficos e ilustrações de redes complexas.	131

LISTA DE ABREVIATURAS E SIGLAS

BA	Grafo aleatório de Barabási e Albert
BGN	<i>Benchmark</i> de Girvan e Newman
ER	Grafo aleatório de Erdős e Rényi
FG	Algoritmo <i>Fast Greedy</i>
GN	Algoritmo de Girvan e Newman
IMap	Algoritmo Infomap
LFR	<i>Benchmark</i> de Lancichinetti, Fortunato e Radicchi
NMI	Informação Mútua Normalizada (<i>Normalized Mutual Information</i>)
PCA	Análise de Componentes Principais (<i>Principal Component Analysis</i>)
RW	Passeio Aleatório (<i>Random Walk</i>)
SAW	Passeio Aleatório Auto-excludente (<i>Self Avoiding Walk</i>)
S&P500	<i>Standard & Poor's 500</i>
Wt	Algoritmo <i>Walktrap</i>

LISTA DE SÍMBOLOS

$\mathcal{V}$	Conjunto de vértices
$\mathcal{E}$	Conjunto de arestas
$\mathbf{A}$	Matriz de Adjacência
$\mathbf{W}$	Matriz de Adjacência ponderada
$\mathbf{D}$	Matriz diagonal de graus
$\mathcal{L}$	Matriz Laplaciana
$a_{i,j}$	Elemento da linha i e coluna j da matriz de adjacência
$w_{i,j}$	Elemento da linha i e coluna j da matriz de adjacência ponderada
w_i	Caminhante aleatório partindo do vértice i
N	Número de vértices de uma rede
E	Número de arestas de uma rede
k_i	Grau do vértice i
$\langle k \rangle$	Grau médio do vértice i
s_i	Força do vértice i
s_i^{in}	Força do vértice i em uma comunidade
k_i^{in}	Grau de entrada do vértice i
k_i^{out}	Grau de saída do vértice i
k_i^{total}	Grau total do vértice i
P_k	Distribuição de graus
$d_{i,j}$	Caminho mínimo entre os vértices i e j (geodésica)
$\langle d \rangle$	Média dos menores caminhos mínimos entre todos os pares de vértices
b_i	<i>Betweenness Centrality</i>
b_e	<i>Edge Betweenness Centrality</i>
γ	Expoente da distribuição de graus de uma lei de potências

Q	Modularidade de uma rede simples
Q_p	Modularidade de uma rede ponderada
Q_d	Modularidade da uma rede com direção
$P_{i,j}$	Termo de modelo nulo
l	Número de comunidade em um grafo
g	Número de vértices em uma comunidade
z_{in}	Grau médio dos vértices dentro de uma comunidade
z_{out}	Grau médio dos vértices entre comunidades
β	Expoente da distribuição de comunidades para o <i>benchmark</i> LFR
τ	Expoente da distribuição de pesos para o <i>benchmark</i> LFR ponderado
μ	Parâmetro de mistura do <i>benchmark</i> LFR
μ_w	Parâmetro de mistura ponderado do <i>benchmark</i> LFR
$\mathcal{P}_i(t)$	Probabilidade de encontrar um caminhante aleatório no vértice i no instante t
$\mathcal{P}_i^\infty$	Probabilidade estacionária do caminhante aleatório
M_i	Número total de caminhantes partindo do vértice i
$m_{i,j}$	Número de caminhantes que saíram de i e passaram por j antes de parar
$l_{i,j}^w$	Quantidade de passos necessária para o k -ésimo caminhante w_i alcançar j partindo de i
$p_{i,j}$	Probabilidade de transição entre os vértice i e j no modelo SAW
$\langle l_{i,j} \rangle$	Número médio de passos entre os vértice i e j no modelo SAW
$f_{i,j}$	Razão entre $p_{i,j}$ e $\langle l_{i,j} \rangle$
$\mathbf{F}$	Representação matricial de $f_{i,j}$
$\mathbf{P}$	Base composta pelas componentes principais mais relevantes de $\mathbf{F}$
$\tilde{\mathbf{F}}$	$\mathbf{F}$ projetado em $\mathbf{P}$
$ \mathbf{F} $	Norma de Frobenius para $\mathbf{F}$

ψ	Metadado de uma rede complexa
ζ	Peso artificial atribuído a uma aresta dado um metadado
d_{AB}^E	Distância Euclideana
d_{AB}^M	Distância de Manhattan
d_{AB}^{Ch}	Distância de Chebyshev
d_{AB}^{Cn}	Distância de Canberra
d_{AB}^{BC}	Distância de Bray-Curtis
$\rho_{i,j}$	Coeficiente de correlação de Pearson entre dois conjuntos de dados i e j
Y_i	Retorno do ativo i
ν	Parâmetro de corte para filtrar as arestas na rede de ações
η	Nível de significância estatística

SUMÁRIO

1	INTRODUÇÃO	31
1.1	Motivação e Objetivos	35
1.2	Organização de Capítulos	36
2	CONCEITOS BÁSICOS	39
2.1	Fundamentos em redes complexas	39
2.1.1	Noções Básicas	39
2.1.1.1	Grau e Distribuição de Grau	39
2.1.1.2	Trajetórias e Distâncias	41
2.1.1.3	Componentes Conectados	43
2.1.1.4	Medidas de Centralidade	44
2.1.2	Modelos de redes	44
2.1.2.1	Modelo Erdős-Rényi	44
2.1.2.2	Modelo Barabási-Albert	45
2.1.3	Comunidades	47
2.1.3.1	Algoritmo de Girvan e Newman	49
2.1.3.2	Algoritmo <i>Fast Greedy</i>	50
2.1.3.3	Algoritmo <i>Walktrap</i>	51
2.1.3.4	Algoritmo <i>Infomap</i>	51
2.1.3.5	<i>Benchmarks</i>	52
2.1.4	Redes reais	56
2.2	Passeios aleatórios em redes complexas	58
2.2.1	Passeio aleatório	58
2.2.2	Passeio aleatório auto-excludente	59
3	RESULTADOS	61
3.1	Caracterização do modelo auto-excludente	61
3.2	Deteccção de comunidades através do passeio auto-excludente	70
3.2.1	Descrição do Método	71
3.2.2	Redes simples	75
3.2.3	Redes com peso	82
3.2.4	Redes com direcção	89
3.2.5	Redes com metadados	92
4	CONCLUSÕES E PERSPECTIVAS	99

REFERÊNCIAS	107
APÊNDICES	119
APÊNDICE A – CLUSTERIZAÇÃO HIERÁRQUICA	121
APÊNDICE B – REDE DE AÇÕES	125
ANEXOS	129
ANEXO A – INFORMES COMPUTACIONAIS	131

1 INTRODUÇÃO

O estudo de redes complexas origina-se em uma das grandes áreas da matemática discreta conhecida como teoria dos grafos, que se iniciou em 1736 quando Leonhard Euler publicou a solução do problema das pontes de Königsberg.¹ O problema consistia em encontrar um caminho que atravessasse cada uma das sete pontes da cidade prussiana de Königsberg apenas uma vez. A proposta de Euler foi considerar cada ponte uma conexão (aresta) entre diferentes regiões da cidade, que foram transformadas em pontos (vértices). A conclusão obtida por ele foi que cada vértice do percurso deveria possuir um número par de arestas, ou seja, pelo menos uma rota de chegada e outra de saída, com exceção dos vértices inicial e final. Feito isso, Euler generalizou esta solução para qualquer grafo, ficando conhecida como caminhos de Euler: todas as arestas de um dado grafo devem ser percorridas apenas uma vez. Euler provou que a única dependência destes caminhos é a quantidade de ligações de cada vértice (conectividade). A partir disto, outros matemáticos se interessaram por problemas semelhantes, proporcionando o desenvolvimento da área com avanços práticos em problemas de otimização como, por exemplo, os problemas do cacheiro viajante² e das quatro cores.³

Entretanto, os maiores avanços da teoria dos grafos ocorreram apenas no século XX quando matemáticos passaram a estudar topologias mais complexas através de métodos estatísticos para determinar características estruturais de grande escala. Na década de 1920, a teoria dos grafos começou a ser utilizada para representar variados tipos de sistemas na grande área das ciências sociais como, por exemplo, comunicação entre membros de um grupo, negociações entre países e até mesmo transações econômicas corporativas. Já na década de 1950, os matemáticos Paul Erdős e Alfréd Rényi, baseados nos trabalhos de Solomonoff e Rapoport,⁴ estudaram a fundo os grafos aleatórios⁵ e demonstraram formalmente muitas propriedades estruturais solúveis no limite de muitos vértices. Os avanços alcançados pelos matemáticos húngaros Erdős e Rényi⁶⁻⁸ marcaram o início da teoria moderna dos grafos,^{9,10} mas apesar de sua importância, este modelo não é capaz de reproduzir a maioria das propriedades presentes nas redes do mundo real.

Nas últimas décadas, com o avanço da tecnologia de informação e o grande aumento do poder computacional, o interesse no estudo de grafos cresceu e originou um novo movimento de pesquisa, conhecido como redes complexas. Estas são caracterizadas por sistemas com estruturas irregulares, complexas e que evoluem dinamicamente com o tempo, viabilizando a modelagem de pequenos grafos até sistemas com milhares ou até milhões de vértices, como por exemplo a *World Wide Web*¹¹ e a Internet.¹² Este avanço possibilitou estudos estatísticos mais precisos com redes reais, com grande destaque para os trabalhos de Albert-László Barabási e Reka Albert sobre distribuições livre de escala¹³⁻¹⁵ e para o

modelo de pequeno mundo proposto por Duncan J. Watts e Steven H. Strogatz.¹⁶

Neste último, publicado em 1998, os autores mostraram que uma rede inicialmente regular (todos os vértices apresentam a mesma conectividade) submetida a trocas aleatórias de um pequeno número de ligações reduz significativamente as distâncias médias entre os componentes. Este efeito conhecido como pequeno-mundo foi primeiramente observado pelo psicólogo Stanley Milgram,¹⁷ que convidou alguns participantes a enviarem cartas a um destinatário final, porém elas poderiam ser encaminhadas apenas por pessoas conhecidas. Ao final do experimento Milgram observou que um terço das cartas chegaram ao destino final com uma média de seis passos. Este fenômeno ficou conhecido pelo termo *seis graus de separação* e, atualmente, é detectado em vários outros sistemas reais, como sistemas de comunicação, redes sociais, sistemas biológicos, rotas aéreas e cadeias alimentares.

Com o estudo crescente e intrinsecamente interdisciplinar, a análise comparativa das diferentes redes produziu novos conceitos e medidas para caracterizar a topologia das redes reais como, por exemplo, a importância de cada vértice no sistema, a presença de correlação nas ligações e principalmente a distribuição de graus P_k , definida como a fração de vértices em um grafo que possui conectividade k . Muitos casos reais apresentam P_k no padrão de uma distribuição lei de potências $P_k \sim k^{-\gamma}$ (livre de escala), como proposto por Barabási e Albert, onde o expoente γ , em geral, varia entre 2 e 3 dependendo da rede real estudada. Este resultado contrasta com o modelo de grafo aleatório inicialmente desenvolvido por Erdős e Rényi, que produz uma distribuição de Poisson com k 's mais bem definidos e não muito comuns na natureza.

Além dessas medidas topológicas, destaca-se a identificação de estruturas de comunidades, fenômeno que ocorre na rede quando existe uma alta concentração de ligações dentro de certos grupos de vértices e baixa concentração entre estes grupos,¹⁸ que na prática indica o compartilhamento de alguma propriedade comum entre membros de um mesmo grupo. Determinar estas estruturas é de grande importância pois pode nos ajudar a compreender como a rede é organizada, permite a classificação de vértices de forma a respeitar a estrutura dos grupos, possibilita a compreensão das funcionalidades dos elementos da rede, além de determinar níveis hierárquicos entre os membros. Comunidades estão presentes em vários tipos de redes reais. Em biologia, nas redes de interação de proteínas, por exemplo, as comunidades reúnem as proteínas que desempenham a mesma função dentro da célula.¹⁹⁻²¹ Já no caso do *World Wide Web*, elas correspondem às páginas que apresentam conteúdo semelhante ou relacionado.^{22,23} Nos últimos anos, *sites* de interação social como *Myspace* (myspace.com), *Facebook* (www.facebook.com) e *Twitter* (twitter.com) se tornaram muito populares e foram objeto de estudo na detecção de comunidades em redes sociais.^{24,25}

O objetivo principal da detecção de comunidades em grafos é identificar estes conjuntos de vértices utilizando apenas informações topológicas. Este tradicional problema

de grande aplicação prática já foi explorado de várias formas em muitos campos da ciência. As primeiras análises sobre estruturas de comunidades são atribuídas a Weiss e Jacobson,²⁶ que procuravam grupos de trabalho dentro de uma agência governamental. Porém, pesquisas sobre comunidades começaram mesmo antes de Weiss e Jacobson. Já em 1927, Stuart Rice procurou grupos de pessoas em pequenos órgãos políticos, com base na semelhança de seus padrões de votação.²⁷ Duas décadas depois, George Homans mostrou que grupos sociais poderiam ser revelados reorganizando adequadamente linhas e colunas de matrizes que descrevem as relações sociais, até que elas tomem a forma de um bloco diagonal.²⁸

Em 2002, Girvan e Newman propuseram um novo algoritmo para encontrar esses grupos visando a identificação das ligações entre comunidades e sua sucessiva remoção que leva ao isolamento progressivo das partições.^{29,30} Estas ligações entre comunidades são detectadas de acordo com uma medida conhecida como centralidade de intermediação de arestas (*edge betweenness centrality*) que expressa a importância de cada aresta em processos onde existe transmissão de “sinais” através do grafo seguindo os caminhos mínimos (menor sequência de arestas entre dois vértices). Além disso, os autores definiram uma medida que quantifica a qualidade da divisão encontrada utilizando apenas o padrão de ligações do sistema. Esta medida foi chamada de modularidade e basicamente compara a densidade de arestas dentro de uma comunidade com o número esperado se os vértices fossem conectados aleatoriamente, já que em grafos aleatórios não se espera encontrar nenhuma estrutura de comunidade. A partir deste trabalho ocorreu um aumento no interesse pelo desenvolvimento de métodos de detecção de comunidades, em particular pelos físicos que através de processos dinâmicos como modelo de *spin*,³¹ percolação³² e sincronização³³ elaboraram novos algoritmos e conceitos relacionados ao estudo dessas partições.^{34–36}

Embora muito utilizado na detecção de comunidades, os processos dinâmicos em redes complexas, de forma geral, tem atraído a atenção de pesquisadores desde as últimas décadas com simulações de propagação de doenças³⁷ e opiniões,³⁸ fluxos de informações,³⁹ falhas⁴⁰ e autômatos celulares,⁴¹ por exemplo. A análise temporal destes processos em sistemas de larga escala por meio de modelos com regras dinâmicas microscópicas (interação apenas entre cada vértice e seus vizinhos mais próximos) levam a padrões de auto organização que estão muitas vezes associadas a criticalidades^{42,43} e transições de fases (mudança repentina de alguma propriedade do sistema).^{44,45}

Outro aspecto que também desperta interesse devido à importância em sistemas reais é a relação entre topologia e processos dinâmicos.^{46–49} O estudo deste relacionamento pode ser usado para verificar propriedades estruturais que influenciam comportamentos dinâmicos, permitindo antecipar e, conseqüentemente, prevenir efeitos indesejados no sistema. A combinação topologia-dinâmica também pode ser aplicada para otimizar certas

propriedades temporais de uma rede com a modificação progressiva da estrutura. Além disso, se a dependência entre este par for previamente conhecida, o sistema pode ser definido para apresentar as propriedades dinâmicas desejáveis. Um processo temporal relativamente simples e flexível muito popular em diversas áreas da ciência que possibilita o estudo topológico de redes através da dinâmica microscópica é o passeio aleatório (*random walk*).

Modelos de caminhadas aleatórias são muito utilizadas em uma grande variedade de processos estocásticos, conjunto de variáveis aleatórias usadas para estudar a evolução temporal de um sistema.^{50,51} Em 1900, Louis Bachelier em sua tese de doutorado propôs o uso da dinâmica aleatória como um modelo fundamental para séries temporais financeiras,⁵² ideia que influenciou muitas décadas depois a teoria moderna das finanças. Alguns anos depois, 1905, Karl Pearson utilizou o *random walk* para descrever de forma simples uma infestação de mosquitos em uma floresta através de uma distribuição estatística.⁵³ No mesmo ano, Albert Einstein elaborou o movimento Browniano para explicar o complicado movimento de uma partícula suspensa no ar, que foi modelado como um passeio aleatório guiado por colisões com um gás molecular.⁵⁴ Com o passar dos anos, a caminhada aleatória se mostrou muito útil em outras áreas, por exemplo, na biologia com o estudo de estruturas e dinâmicas de polímeros,^{55,56} na computação com algoritmos de buscas⁵⁷ e na matemática com fractais⁵⁸ e grafos regulares de diferentes dimensões espaciais.⁵⁹

No contexto de redes complexas, o passeio aleatório consiste em um caminhante visitando vértices consecutivos escolhidos aleatoriamente entre todos os vizinhos do vértice atual. Apesar de simples, a vantagem dele utilizar apenas informações locais da rede o torna adequado quando conhecemos poucas propriedades do sistema como um todo. Diversas dinâmicas de sistemas naturais e artificiais já foram bem descritas com este processo estocásticos: padrões de tráfego e mobilidade,^{60,61} fluxo de informações em redes sociais⁶² e enovelamento de proteínas.⁶³ Mais recentemente, foram desenvolvidos estudos centrados nas propriedades dessa caminhada em redes aleatórias, produzindo resultados analíticos para o tempo médio de primeira passagem entre dois vértices distintos,^{64,65} tempo médio de captura (tempo médio para alcançar um vértice específico partindo de qualquer ponto da rede),⁶⁶ número médio de vértices distintos visitados por um mesmo caminhante⁶⁷ e tempo médio de cobertura.⁶⁸ Além disso, também já foram exploradas leis de escala em redes de pequeno mundo,^{69,70} propriedades em redes direcionadas^{71–73} bem como o efeito de memória finita.^{74,75}

Muitos tipos alternativos de passeios aleatórios já foram propostos visando otimizar a análise topológica.^{76–80} Entre estas abordagens, o passeio aleatório auto-excludente, do inglês *self avoiding walk* (SAW), tem-se mostrado mais eficiente na exploração e navegação de diferentes estruturas do que a caminhada tradicional. Esta dinâmica já foi extensivamente explorada em redes regulares,^{81–85} proporcionando muito conhecimento sobre estruturas

e termodinâmica de polímeros.⁸⁶ Esse tipo de *random walk* tem a característica de não visitar um vértice que já foi alcançado.⁸⁷ Em cada passo, o caminhante escolhe o próximo destino aleatoriamente entre os vizinhos do nó atual, excluindo os vértices já visitados. O passeio termina quando todos os nós já foram alcançados ou quando não existe mais vizinhos acessíveis. Devido ao fato dos caminhantes guardarem todo o trajeto percorrido (memória infinita), a solução analítica não se torna trivial, embora alguns esforços teóricos proporcionem resultados interessantes em modelos de redes de pequeno mundo,⁸⁸ Erdős e Rényi⁸⁹ e topologias livre de escala.^{90,91} Contudo, justamente pela dificuldade analítica, sua aplicação em redes complexas do mundo real ainda é pouco explorada.

1.1 Motivação e Objetivos

Apesar da interdisciplinaridade e das inúmeras e importantes aplicações práticas da identificação de estruturas modulares em redes complexas, a detecção de comunidades continua um problema em aberto. Muito se deve à falta de uma definição formal do significado de comunidade juntamente com a subjetividade das análises do particionamento, uma vez que, também não há uma definição geral do que se está procurando. Além disso, a busca por estes grupos exige, de forma geral, uma alta complexidade computacional associada ao problema de otimização. Contudo, muitos métodos heurísticos que visam a otimização da modularidade tem mostrado uma boa correspondência com características relevantes de vários sistemas reais.

O passeio aleatório auto excludente se mostra vantajoso e de alta eficácia na busca e exploração de estruturas de redes complexas, logo utilizamos estes benefícios para implementar um novo algoritmo de detecção de comunidades baseado na otimização da modularidade. Além disso, uma vantagem de utilizar a dinâmica de *random walks* em redes é sua flexibilidade em diferentes estruturas como, por exemplo, em grafos ponderados (quando as arestas apresentam valores que indicam alguma propriedade do sistema) e grafos com direções ou digrafos (quando as arestas possuem orientação, partem de um vértice i e atingem um vértice j). Assim, no caso particular do SAW, em redes ponderadas o caminhante deixa de escolher os vizinhos acessíveis uniformemente e passa a decidir seu próximo destino aleatório de acordo com o valor presente nas arestas e, para digrafos, o caminhante começa a respeitar as direções preferenciais das ligações. Desta forma, com pequenas alterações na dinâmica do SAW é possível generalizar nosso método de detecção de comunidades para redes ponderadas e digrafos.

A maioria das pesquisas nesta área trata as redes como objetos puramente topológicos, um conjunto estático de vértices e suas interações. No entanto, a maioria das fontes de dados de onde surgem as redes, são acompanhados por características adicionais que descrevem propriedades gerais dos vértices, como idade, gênero e opção sexual. Estes atributos que enriquecem o sistema são conhecidos como metadados e podem ser incluídos

no grafo para melhorar as análises e o entendimento da estrutura da rede. Assim sendo, outro objetivo de estudos foi combinar as informações estruturais com os metadados visando aperfeiçoar a classificação dos vértices em grupos por meio do novo método pré estabelecido.

Portanto, o presente trabalho é dedicado à investigação de estruturas topológicas modulares em redes complexas por meio da dinâmica do passeio aleatório auto-excludente. Em particular, nesta tese buscamos conciliar as vantagens da dinâmica do SAW visando enriquecer a análise tanto quantitativa quanto qualitativa de redes reais de diferentes categorias. Para isso, definimos uma medida que concentra toda informação topológica da rede obtida por meio deste processo dinâmico, e a partir dela, elaboramos um método que visa otimizar a medida de modularidade para determinar a melhor distribuição de comunidades. A partir disso, generalizamos o algoritmo desenvolvido para identificar grupos também em redes ponderadas e digrafos, além de combinar propriedades estruturais com informações provindas de metadados para aprimorar a divisão em partições.

1.2 Organização de Capítulos

Após a introdução, a estrutura do trabalho prossegue com o conteúdo relacionado à formulação teórica, o qual aborda os conceitos básicos de redes complexas e a dinâmica do passeio aleatório neste tipo de sistema. A seguir, apresentamos os procedimentos metodológicos e os resultados obtidos. Por fim são apresentadas a conclusão, as perspectivas de trabalhos futuros e as considerações finais. Na figura 1, temos a organização da estrutura do trabalho no formato de um grafo. Os capítulos são representados pelas caixas verdes e as seções pelas caixas vermelhas. A sequência dos capítulos são marcadas por arestas mais grossas e cada seção está ligada ao seu capítulo por conexões com espessura intermediária. As ligações mais finas indicam que as duas partes do trabalho estão relacionadas, mesmo não estando na sequência.

No capítulo 2, introduziremos conceitos teóricos sobre redes complexas que são usualmente tratados em livros-texto. Inicialmente descreveremos as definições básicas de grafos, como representa-los matematicamente, algumas características como tipos de componentes, trajetórias e distâncias. Feito isso, discutimos brevemente sobre medidas de centralidade e abordaremos as principais propriedades de dois modelos de redes aleatórias: o grafo estocástico de Erdős e Rényi e o modelo livre de escala de Barabási e Albert. Posteriormente, exploraremos os principais conceitos sobre comunidades, discutiremos quatro algoritmos clássicos de detecção de grupos em redes diversificadas além de apresentar dois tradicionais *benchmarks* para comparação e validação das diferentes técnicas de particionamento. Na sequência, serão descritas as nove redes reais utilizadas neste trabalho. Por fim, desenvolvemos matematicamente dois tipos de passeios aleatórios em redes complexas: o sem restrição e o auto-excludente.

Figura 1: Estrutura da tese representada por um grafo. As caixas em verde representam os capítulos, que se ligam com as respectivas seções, indicadas pelas caixas amarelas. Arestas mais grossas (verde) mostram a sequência dos capítulos, arestas médias (vermelhas) ligam as seções aos seus capítulos e as conexões mais finas indicam que ambas as partes do trabalho estão relacionadas.

Fonte: Elaborada pelo autor.

A seguir, capítulo 3, discutiremos os resultados analíticos provenientes da dinâmica aleatória auto-excludente. Para a caracterização inicial deste passeio, utilizamos os dois modelos teóricos de redes descrito anteriormente visando compará-lo ao do caminhante sem restrição em um ambiente onde os parâmetros são controlados e alguns efeitos conhecidos. Após esta comparação, levantamos algumas informações relevantes do SAW como distribuição do número de passos e fração de vértices alcançados. A partir dessa análise preliminar, descreveremos nosso algoritmo de detecção de comunidades baseado no passeio aleatório auto-excludente. Em seguida, realizaremos sua validação quando submetido aos dois *benchmarks* e comparado aos outros quatro métodos para finalmente mostrarmos os resultados em redes reais com estruturas topológicas variadas.

Por fim, o capítulo 6 é destinado à apresentação das conclusões dos trabalhos

realizados, considerações finais e perspectivas para trabalhos futuros decorrente da nossa contribuição científica.

2 CONCEITOS BÁSICOS

Neste capítulo introduziremos os fundamentos básicos sobre redes complexas presentes em livros-texto.^{92,93} Também discutiremos os conceitos de comunidades,^{94–96} assim como métodos tradicionais de identificação de partições e diferentes formas de testá-los utilizando dois tipos de *benchmarks*. Ao término desta sessão, discutiremos a caminhada aleatória com e sem restrição aplicada a redes complexas.

2.1 Fundamentos em redes complexas

2.1.1 Noções Básicas

Uma rede complexa é uma estrutura topológica representada matematicamente por um grafo. Um grafo $G(\mathcal{V}, \mathcal{E})$ consiste em um conjunto de vértices $\mathcal{V}$ (ou nós) relacionados por meio de um conjunto de arestas $\mathcal{E}$ (ligações ou conexões), tal que $\mathcal{V} = (v_1, v_2, v_3, \dots, v_N)$ com N o número total de nós e $\mathcal{E} = \{e_{i,j} = (v_i, v_j) \mid v_i, v_j \in \mathcal{V}\}$.

Existem várias formas de se representar uma rede, a mais conveniente para realização de cálculos é através de uma matriz de adjacência $\mathbf{A}$, onde cada elemento $a_{i,j}$ ($i, j = 1, \dots, N$) é dado pelo par linha i coluna j . Para um rede simples, temos que $a_{i,j}$ recebe o valor 1 caso exista uma ligação entre os vértices i e j , e zero caso não, como mostrado na figura 2(a). Redes também podem apresentar arestas com valores numéricos, indicando o peso ou a força da conexão. Neste caso, os elementos da matriz de adjacência são dados por $w(e_{i,j}) = w_{i,j}$, como representado na figura 2(b). Diferentemente das duas representações anteriores, outra estrutura bastante usual é o digrafo, grafo no qual as conexões apresentam direção, ou seja, uma dada orientação, saindo do vértice i e atingindo o vértice j , por exemplo. Esta situação é ilustrada na rede da figura 2(c).

Além destas três estruturas básicas existem outras representações possíveis como presença de autoconexões, aresta que ligam um vértice com ele mesmo, implicando em uma diagonal diferente de zero em $\mathbf{A}$, ligações repetidas entre um mesmo par de vértices conhecida como múltiplas arestas e até mesmo redes bipartidas, onde os vértices podem ser divididos em dois conjuntos distintos no qual cada vértice do primeiro conjunto está associado a outro do segundo conjunto.

2.1.1.1 Grau e Distribuição de Grau

A medida mais simples de um vértice é conhecida como grau ou conectividade (k). Ela consiste no número de arestas incidente em um dado vértice i e, para uma rede

Figura 2: Exemplos de redes (a) simples, (b) com peso (representado na largura da aresta) e (c) direcionada (indicada pelas setas).

Fonte: Elaborada pelo autor.

simples, é definida em termos da matriz de adjacência como

$$k_i = \sum_{j=1}^N a_{i,j}, \quad (2.1)$$

ou seja, é basicamente a soma de todos os elementos da linha correspondente. Este conceito é facilmente estendido para redes ponderadas, onde a força de um vértice (s) é definida como a soma dos pesos das arestas,

$$s_i = \sum_{j=1}^N w_{i,j}, \quad (2.2)$$

com w sendo o peso da respectiva aresta $e_{i,j}$.

Já para redes direcionadas cada vértice possui um grau de entrada (k_i^{in}), número de ligações que chegam no vértice i , e um grau de saída (k_i^{out}), quantidade de conexões que saem do vértice i . Matematicamente, essas medidas são dadas por

$$k_i^{in} = \sum_{j=1}^N a_{j,i}, \quad (2.3)$$

$$k_i^{out} = \sum_{j=1}^N a_{i,j}, \quad (2.4)$$

que resulta no grau total do vértice i ,

$$k_i^{total} = k_i^{in} + k_i^{out}. \quad (2.5)$$

Repare que a diferença dessas equações em relação ao grau de uma rede simples (2.1) se dá justamente pela falta de simétrica no caso do digrafo.

A partir dessas medida, podemos determinar a distribuição de graus de uma rede, P_k , definida como a fração de vértices que possuem grau k ,

$$P_k = \frac{1}{N} \sum_{i=1}^N \delta_{k_i,k} \quad (2.6)$$

Figura 3: (a) Rede simples com $N = 4$ e $E = 4$ com as seguintes conectividades ($k_1 = 1, k_2 = 3, k_3 = 2, k_4 = 2$). Já (b) mostra a distribuição de graus da respectiva rede.

Fonte: Elaborada pelo autor.

com $\delta_{i,j}$, valendo 1 se $i = j$ e zero caso contrário. Consequentemente, os momentos estatísticos desta distribuição são dados por

$$\langle k^n \rangle = \sum_k k^n P_k. \quad (2.7)$$

Para $n = 1$ temos o grau médio de G e para $n = 2$ obtemos uma grandeza relacionada ao desvio padrão que indica a flutuação de P_k . A figura 3 ilustra a distribuição de graus para uma rede simples, $G(\{1, 2, 3, 4\}, \{(1, 2), (2, 3), (2, 4), (3, 4)\})$. Neste exemplo, temos $P(1) = P(3) = 1/4$ uma vez que $k_1 = 1$ e $k_2 = 3$ e $P(2) = 1/2$ pois $k_3 = k_4 = 2$. Como P_k é uma distribuição de probabilidades, temos, por construção, $P(1) + P(2) + P(3) = 1$.

A partir de (2.7), temos que para qualquer rede simples o grau médio pode ser escrito como

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^N k_i = \frac{2E}{N}, \quad (2.8)$$

e para redes direcionadas,

$$\langle k^{in} \rangle = \frac{1}{N} \sum_{i=1}^N k_i^{in} = \frac{E}{N} \quad (2.9)$$

$$\langle k^{out} \rangle = \frac{1}{N} \sum_{i=1}^N k_i^{out} = \frac{E}{N} \quad (2.10)$$

2.1.1.2 Trajetórias e Distâncias

Além da conectividade, outro conceito importante, base de muitas medidas em grafos, são as trajetórias (caminho) e as distâncias. Caminho é definido como uma sequência de vértices diferentes tal que cada vértice está conectado ao próximo por uma aresta. O comprimento do caminho (distância) é dado pelo número de ligações entre os nós que compõem a trajetória.

Figura 4: (a) Rede simples contendo dois caminhos mínimos entre os vértice i, j (cores verde e amarela). (b) Dagrafo mostrando um único caminho mínimo entre o vértice origem i e o vértice destino j , evidenciando que, diferentemente do caso anterior, não existe geodésica entre j e i .

Fonte: Elaborada pelo autor.

Em geral, existem várias distâncias possíveis entre um par de vértices, contudo a mais importante dentre estas é a menor, também conhecida como geodésica, distância topológica ou simplesmente distância entre os vértices i e j , denotado por $d_{i,j}$. Caso não exista caminho entre i e j definimos $d_{i,j} = \infty$. No caso de redes sem direção a distância entre i e j é a mesma que a distância entre j e i , $d_{i,j} = d_{j,i}$. Para redes com direção, frequentemente $d_{i,j} \neq d_{j,i}$ e a existência de um caminho de i para j não implica que exista o inverso. Já para redes com peso, esta informação extra pode ser adicionada na geodésica e interpretada de acordo com a necessidade, podendo indicar, por exemplo, uma distância física, um custo financeiro ou até mesmo uma unidade de tempo. Computacionalmente, existem diversos algoritmos para determinar as distâncias mínimas, cada um indicado para um certo tipo de grafo. Entre eles destacamos, a busca em largura (*Breadth Firsh Search*),⁹⁷ o algoritmo de Bellman–Ford,⁹⁸ o algoritmo de Dijkstra,⁹⁹ o Algoritmo de Johnson¹⁰⁰ e o algoritmo de Floyd–Warshall.¹⁰¹

Na figura 4, temos um exemplo da geodésica para uma rede simples e uma rede com direção. No primeiro caso a distância mínima entre os vértices i e j é degenerada em dois caminhos (cores verde e amarela) contendo três arestas cada. Vale ressaltar que, nesta situação não definimos a origem e o destino dos trajetos, uma vez que a geodésica é a mesma de i para j e de j para i . Por outro lado, na rede com direção apresentada em (b), existe apenas um caminho mínimo ligando o vértice origem i e o vértice destino j (destacado na cor verde) e não há caminho possível de j para i .

Através da geodésica podemos definir uma medida de separação média entre dois vértices na rede conhecida como a média dos menores caminhos mínimos entre todos os pares de vértices,

$$\langle d \rangle = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij}. \quad (2.11)$$

Figura 5: (a) Exemplo de uma rede simples com dois componentes conectados, destacados pelas regiões em cinza, com o componente gigante contendo seis vértices. (b) Rede direcionada ilustrando a diferença entre componentes fracamente e fortemente conectados. Neste caso, o primeiro tipo de componente engloba todos os vértices da rede enquanto o segundo é dividido em três partes, destacadas em cinza.

Fonte: Elaborada pelo autor.

Como esta média é diretamente proporcional ao caminho mínimo, temos um problema evidente: caso os vértices i e j não sejam conectados, $\langle d \rangle$ divergirá. Portanto, esta medida é válida quando existe pelo menos um caminho entre todos os pares de nós da rede, ou seja, apenas um componente conectado, como veremos a seguir.

2.1.1.3 Componentes Conectados

Uma rede sem direção é dita conectada se todos os pares de vértices possuem um caminho entre eles e desconectada quando existir pelo menos um par i, j com $d_{i,j} = \infty$. Assim, um subgrafo onde todos os vértices possuem ao menos um caminho entre eles é chamado de componente conectado. Uma rede pode ter vários componentes, sendo chamado de maior componente conectado ou componente gigante aquele que contém o maior número de nós. Isso é ilustrado na figura 5(a), que mostra uma rede simples contendo dois componentes conectados (regiões em cinza), com o componente gigante possuindo seis vértices.

Redes direcionadas apresentam dois tipos de componentes. O componente fracamente conectado, situação na qual a direção das ligações não tem relevância para a determinação do caminho, ou seja, análogo a uma rede sem direção, e o componente fortemente conectado, caso onde as trajetórias são dadas respeitando a direção das conexões. A figura 5(b) apresenta uma rede direcionada que possui um componente fracamente conectado contendo todos os vértices ou três componentes fortemente conectados demarcados pelas regiões em cinza.

2.1.1.4 Medidas de Centralidade

Uma característica muito importante e popular no campo de redes complexas é identificar a importância de um nó, ou o quão central ele é na estrutura geral do sistema. As medidas que buscam quantificar esta característica são conhecidas como centralidade e, como exemplo, destacamos as centralidades de proximidade (*closeness*),¹⁰² grau e autovetor (*eigenvector*),¹⁰³ Katz¹⁰⁴ e Page-Rank.¹⁰⁵

Contudo, a medida de centralidade mais relevante para este trabalho é a centralidade de intermediação ou *betweenness centrality*.¹⁰⁶ Ela é definida para cada vértice i da rede e basicamente mensura a quantidade de geodésicas que passam por ele. Mais precisamente, a intermediação, b_i , é definida como

$$b_i = \sum_{j \neq k} \frac{\sigma_{j,k}^i}{\sigma_{j,k}}, \quad (2.12)$$

onde $\sigma_{j,k}^i$ é o número de caminhos mínimos entre os vértices j e k que obrigatoriamente passam por i , e $\sigma_{j,k}$ é o número total de caminhos mínimos que conectam j e k . Portanto quanto maior o número de caminhos mínimos que passam por i , mais importante ele é em relação à propagação da informação. Este conceito de intermediação também pode ser estendido para as arestas, e o *edge betweenness centrality* (b_e) passa a ser definido como o número de caminhos mínimos entre um par de vértices que passam pela aresta e ,³⁰

$$b_e = \sum_{j \neq k} \frac{\sigma_{j,k}^e}{\sigma_{j,k}}, \quad (2.13)$$

Como o cálculo da intermediação, tanto b_i com b_e , depende apenas da forma como a distância mínima é calculada, ele se aplica tanto para redes simples como para redes direcionadas, com ou sem peso.

2.1.2 Modelos de redes

Uma das principais motivações para o estudo de redes complexas é a possibilidade de representar e compreender propriedades e fenômenos encontrados na natureza. Assim, os modelos de redes visam reproduzir características observadas em sistemas reais explorando a falta de regularidade por meio da aleatoriedade. Apesar desses modelos não corresponderem totalmente a estes fenômenos naturais, a grande vantagem que apresentam é o controle sistemático dos parâmetros simulados, permitindo análises estatísticas e consequente caracterização das redes reais por meio de uma boa aproximação. A seguir, apresentaremos dois modelos utilizados neste trabalho.

2.1.2.1 Modelo Erdős-Rényi

No final da década de 1950, os matemáticos húngaros Paul Erdős e Alfréd Rényi introduziram um modelo de grafo baseado em ligações aleatórias que ficou conhecido como

Figura 6: (a) Exemplo de um grafo aleatório de Erdős e Rényi. (b) Distribuição de conexões para uma rede com $N = 50000$, $p = 0.0002$ e, consequentemente, $\langle k \rangle = 10$.

Fonte: Elaborada pelo autor.

grafos aleatórios de Erdős e Rényi, ou modelo ER.^{7,8} A construção deste modelo se dá iniciando um conjunto de N vértices desconectados e, a cada passo, dois vértices deste conjunto são escolhidos e conectados dada uma probabilidade fixa p . Quando todos os pares já tiverem sido percorrido o processo termina e a rede ER está pronta. Com isso, a probabilidade de um vértice possuir grau k segue uma distribuição binomial com grau médio $\langle k \rangle = (N - 1)p$ constante, como mostra a figura 6. No limite de $N \rightarrow \infty$, $p \rightarrow 0$ e $\langle k \rangle$ constante a distribuição de grau tende a uma distribuição de Poisson,

$$P_k = \frac{\langle k \rangle^k}{k!} e^{-\langle k \rangle}. \quad (2.14)$$

Outra característica importante desse modelo é o pequeno caminho mínimo médio, proporcional ao logaritmo do tamanho da rede,

$$\langle d \rangle \sim \ln N / \ln \langle k \rangle. \quad (2.15)$$

Esta característica é conhecida como efeito do pequeno mundo e está presente em muitas redes reais.¹⁶

Contudo redes tipo ER, devido à probabilidade fixa p , apresentam alta homogeneidade, ou seja, o grau da maioria dos vértices da rede é muito similar. Além disso, ele não reproduz a maioria das propriedades presentes nas redes observadas na natureza como correlação de grau, estruturas de comunidades e topologia livre de escala.¹⁰⁷

2.1.2.2 Modelo Barabási-Albert

Em 1999, Albert-László Barabási e Réka Albert propuseram um modelo de formação de redes com o intuito de explicar a topologia livre de escala presente em muitas redes reais como, por exemplo, a *World Wide Web*¹² e a *Internet*,¹³ redes de proteínas,¹⁰⁸ citações,¹⁰⁹ metabólicas¹¹⁰ e sociais.¹¹¹ A principal característica da topologia livre de escala é a presença de uma função tipo lei de potências na distribuição de conectividades, $P_k \propto k^{-\gamma}$.

Figura 7: (a) Exemplo de uma rede livre de escala gera pelo modelo BA. (b) Distribuição das conexões para uma rede BA com $N = 1000000$ e $m = 4$.

Fonte: Elaborada pelo autor.

Este modelo ficou conhecido como modelo BA e foi um dos responsáveis pela popularização da área de redes complexas.¹³

O modelo BA, diferentemente do modelo ER, consiste em aumentar o tamanho da rede adicionando novos vértices com uma chance maior de se ligarem aos nós mais densamente conectados. Assim, o processo de formação pode ser dividido em duas etapas,

1. *Crescimento*: Inicia-se o sistema com uma configuração aleatória de vértices e arestas, a cada passo do algoritmo um novo vértice com m arestas é introduzido no grafo e conectado com os outros presentes.
2. *Ligação Preferencial*: A probabilidade de um vértice j ser escolhido para se conectar ao novo nó é diretamente proporcional ao seu grau, ou seja,

$$P_{v_j}(n+1) = \frac{k_j}{\sum_{v_i \in \mathcal{V}_n} k_i}, \quad (2.16)$$

sendo n o tempo e o número de vértices adicionados à rede.

Como cada vértice adicionado possui m arestas, o grau médio da rede será $\langle k \rangle = 2m$. Já a distribuição de graus resultará em uma lei de potência de expoente fixo, $\gamma = 3$, ou, $P_k \propto k^{-3}$. Apesar deste modelo gerar apenas uma particular lei de potências, o grande resultado aqui é mostrar que a ligação preferencial proporcional ao grau produz redes com topologia livre de escalas. A figura 7 mostra um exemplo de rede gerado por esse modelo assim como o P_k para $N = 1000000$ e $m = 4$, resultando em um gráfico linear com o uso da escala logarítmica em ambos os eixos.

Uma consequência da ligação preferencial que pode ser observado na rede 7(a) é a formação de alguns *hubs*, vértices com um número de conexões muito superior à média, uma vez que nós com mais ligações aumentam continuamente sua conectividade. Outra

Figura 8: Um grafo simples com quatro comunidades indicadas pelas cores vermelha, verde, azul e amarela.

Fonte: Elaborada pelo autor.

característica deste modelo é o surgimento do efeito *small world*, mas diferentemente das redes ER, a média dos menores caminhos possui a seguinte dependência para $m > 1$,¹¹²

$$\langle d \rangle \sim \frac{\ln N}{\ln \ln N}, \quad (2.17)$$

ou seja, $\langle d \rangle$ tem um crescimento muito atenuado com a quantidade total de nós, diferentemente da equação 2.15. Esta característica também é conhecida como efeito *ultra-small world*.¹¹³

Além do modelo Barabási e Albert, muitos outros já foram propostos visando maior flexibilidade no expoente γ , como por exemplo, os desenvolvidos por Dorogovtsev¹¹⁴ e Krapivky.¹¹⁵

2.1.3 Comunidades

Muitas redes de sistemas reais apresentam naturalmente estruturas topológicas modulares, ou seja, as conexões entre os vértices não acontece de forma homogênea, elas se concentram mais em certos grupo de nós do que em outros. Estas subestruturas são conhecidas como comunidades e são formadas por vértices que possuem alguma relação de semelhança. Em redes sociais, comunidades são observadas de diversas formas como contatos sexuais,¹¹⁶ idade,¹¹⁷ grupos étnicos,¹¹⁸ relações de amizade¹¹⁹ e organizações criminosas.¹²⁰ Já em redes biológicas, proteínas se agrupam de acordo com sua função molecular^{121,122} e redes de genes podem relacionar grupo de genes a doenças como o câncer, por exemplo.^{123,124} A figura 8 ilustra uma rede simples com quatro comunidades bem definidas, representadas pelas cores vermelha, verde, azul e amarela.

Subdividir um grafo em partições, apesar de parecer intuitivo, não é um processo simples, pois se trata de um problema *NP-completo*¹²⁵ onde geralmente não se sabe o número exato de partições que compõem a rede. Além disso, o conceito de comunidades em redes complexas não é bem definido, entretanto certas condições são geralmente empregadas. O grafo precisa ser esparso (o número de arestas deve ser da mesma ordem do número de vértices) e o número de conexões entre os membros de uma comunidade deve ser maior do que com o restante da rede, que matematicamente corresponde a

$$\frac{2m_{N_c}}{N_c(N_c - 1)} > \frac{2m}{N(N - 1)} > \frac{m_{N_c N}}{N_c(N - N_c)}, \quad (2.18)$$

com N_c o número de vértices presente na comunidade c , m_{N_c} a quantidade de ligações entre os nós deste grupo, N e m o número total de vértices e arestas respectivamente e finalmente $m_{N_c N}$ as arestas entre o grupo c e os $N - N_c$ vértices restantes da rede. Outro critério frequentemente utilizado é a similaridade de vértices, ou seja, vértices de um mesmo grupo apresentam perspectivas semelhantes da rede por possuírem padrões de conexão parecidos.

Para quantificar a qualidade de uma dada estrutura de comunidades, a medida mais amplamente utilizada é a modularidade.³⁰ Se c_i e c_j são as comunidades dos vértices i e j , a expressão geral da modularidade pode ser escrita como,

$$Q = \frac{1}{2m} \sum_{i,j} (A_{i,j} - P_{i,j}) \delta(c_i, c_j), \quad (2.19)$$

onde $\mathbf{A}$ é a matriz de adjacência, m o número total de arestas da rede, $P_{i,j}$ o termo do *modelo nulo* e $\delta(c_i, c_j)$, a função delta de Kronecker. O termo $P_{i,j}$ representa a matriz de adjacência para um *ensemble* de redes aleatórias geradas com as características do grafo original. Portanto, a modularidade compara a densidade de aresta em um dado subgrafo e a densidade esperada para este conjunto caso os vértices fossem ligados aleatoriamente. Como espera-se que redes aleatórias não sejam modulares, esta comparação quantifica a presença de comunidades à medida que a estrutura da rede diverge do esperado, consequentemente, quanto maior o Q , mais bem definida é a estrutura modular.

Para redes simples, temos $P_{i,j} = k_i k_j / 2m$, que corresponde ao número esperado de arestas entre i e j caso as conexões da rede fossem rearranjadas de forma a preservar o grau de todos os vértices do sistema. Isto produz a forma clássica da modularidade

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{i,j} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j). \quad (2.20)$$

Alterando a escolha do termo de modelo nulo, outros aspectos podem ser incorporados na estrutura da rede, como peso¹²⁶ e direção.¹²⁷ No caso onde as ligações possuem pesos, temos $P_{i,j} = s_i s_j / 2s$ e a modularidade dada por

$$Q_p = \frac{1}{2s} \sum_{i,j} \left(W_{i,j} - \frac{s_i s_j}{2s} \right) \delta(c_i, c_j), \quad (2.21)$$

onde $W_{i,j}$ representa a matriz de adjacência com peso, s_i e s_j a força dos vértices i e j (equação 2.2) e $2s = \sum_{i,j} w_{i,j}$, a força total. Já no caso com direção, ficamos com

$$Q_d = \frac{1}{m} \sum_{i,j} \left(A_{i,j} - \frac{k_i^{out} k_j^{in}}{m} \right) \delta(c_i, c_j), \quad (2.22)$$

sendo k_i^{out} e k_j^{in} os graus de saída e entrada respectivamente (equações 2.4 e 2.3). Nesta expressão, vale observar a ausência do fator 2 no denominador quando comparado com os resultados de Q e Q_p .

Um grande número de métodos de detecção de comunidades visam encontrar a distribuição de partições que maximiza a modularidade. Contudo esta abordagem apresenta um limite de resolução. O modelo nulo assume que cada nó pode se ligar com qualquer outro da rede, entretanto este comportamento implícito do modelo não é razoável para redes muito grandes, onde o alcance das conexões de um vértice se restringe em geral a uma pequena vizinhança. Com isso, se a rede for suficientemente grande, a modularidade fica incapaz de resolver pequenas comunidades, mesmo quando elas são topologicamente bem definidas.^{128–130} Uma abordagem utilizada para resolver esta limitação é adicionar um parâmetro $\gamma > 0$ junto ao termo do modelo nulo em 2.20, controlando a importância relativa entre os nós dentro das comunidades e o modelo nulo. Este método é conhecido como *multiresolution*³¹ uma vez que manipulando γ é possível abranger diversas escalas de partições. Entretanto, ele também apresenta problemas intrínsecos, não produzindo resultados satisfatórios.¹³¹

Os algoritmos de identificação de comunidades podem ser agrupados de acordo com a metodologia empregada. Entre elas destacam-se os métodos (i) espectrais, baseados na análise de autovetores de matrizes derivadas da matriz de adjacência¹³²; (ii) inferência estatística, cuja abordagem principal é ajustar o conjunto original de dados através de modelos estocásticos de redes^{133,134} e (iii) otimização da modularidade, que tenta obter a melhor divisão maximizando o parâmetro modularidade. Um grande número de algoritmos de detecção aplicam esta técnica que, apesar das limitações da modularidade, apresenta bons resultados quando comparado às divisões naturais do sistema, como será visto a seguir.

2.1.3.1 Algoritmo de Girvan e Newman

O algoritmo mais popular para identificar estruturas modulares em redes complexas foi proposto por Girvan e Newman no início dos anos 2000^{18,30} (GN). Eles propuseram um modelo divisivo (apêndice A) baseado na medida de centralidade de arestas *edge betweenness* (expressão 2.13). Esta técnica é usada para identificar comunidades, pois quanto maior o número de caminhos mínimos que passam por uma aresta, mais importante ela é em relação à propagação da informação, ou seja, arestas que ligam comunidades

distintas tendem a apresentar altos valores de b_e uma vez que os caminhos mínimos entre vértices de partições diferentes obrigatoriamente passam por estas arestas. Os passos desse algoritmo são dados por:

1. Calcular o *edge betweenness* de cada aresta;
2. Remover a aresta de maior centralidade;
3. Recalcular o *edge betweenness*;
4. Voltar para o passo 2 até a rede conter N componentes.

O resultado deste método é uma distribuição hierárquica normalmente representada por um dendrograma e a melhor distribuição de comunidades é a que apresenta maior modularidade para um determinado corte da árvore. Apesar desta técnica apresentar bons resultados, ela possui uma complexidade de $\mathcal{O}(NE^2)$ no caso geral e $\mathcal{O}(N^3)$ para grafos esparsos, uma vez que o *betweenness* para todas as arestas é calculado em $\mathcal{O}(NE)$ ($\mathcal{O}(N^2)$ para grafos esparsos) e para cada aresta a centralidade é recalculada. Portanto é um método custoso computacionalmente e não recomendado para redes muito grandes (dezenas de milhares de vértices). Contudo, além das redes simples, ele também pode ser estendido facilmente para redes com peso e com direção alterando o método de cálculo dos caminhos mínimos.

2.1.3.2 Algoritmo *Fast Greedy*

Devido à limitação do algoritmo de Girvan e Newman para grandes redes, Newman inicialmente propôs outra abordagem para otimização da modularidade¹³⁵ e posteriormente Clauset o aprimorou alterando a forma do cálculo da maximização de Q , resultando em um processo mais rápido para identificação das comunidades¹³⁶ (FG).

Este é um algoritmo de clusterização hierárquica aglomerativa onde inicialmente tem-se n *clusters* contendo apenas um vértice e nenhuma aresta. Para cada iteração, uma aresta conhecida é introduzida aos vértices desconectados de forma a maximizar a modularidade e diminuir o número de comunidades. Esta otimização é feita calculando

$$\Delta Q_{r,s} = 2(e_{r,s} - a_r a_s), \quad (2.23)$$

com $e_{r,s}$ a fração de arestas entre vértices das comunidades r e s e $a_{r(s)}$ a fração de arestas ligadas aos vértices da comunidade $r(s)$. Feito isso, junta-se as comunidades que apresentam maior $\Delta Q_{r,s}$. Todas as arestas são inseridas no grafo obedecendo o mesmo princípio: quando todas as arestas originais já foram inseridas no grafo, o algoritmo termina. A partir do dendrograma encontrado, determina-se a distribuição mais adequada para o corte que apresentar o maior valor da modularidade.

Como redes que apresentam estruturas modulares são esparsas, o algoritmo é otimizado utilizando as estruturas de dados adequadas, como uma matriz esparsa contendo $\Delta Q_{i,j}$ e um *max-heap* contendo o maior elemento de cada linha de $\Delta Q_{i,j}$. A complexidade do algoritmo é $\mathcal{O}(dE \log N)$, onde d é a profundidade do dendrograma no caso geral, e $\mathcal{O}(N \log^2 N)$ para grafos esparsos, uma vez que $E \sim N$ e $d \sim \log N$. Entretanto, este tipo de implementação tende a formar grandes comunidades e frequentemente prejudicar o valor da modularidade, ou seja, é um método rápido porém não muito preciso. Além disso, este caso não se estende a redes ponderadas e digrafos.

2.1.3.3 Algoritmo *Walktrap*

Este algoritmo, desenvolvido por Pons e Latapy, utiliza o conceito de *random walks* para definir uma medida de similaridade entre vértices e desenvolver uma estrutura hierárquica aglomerativa baseada também na otimização da modularidade¹³⁷ (Wt). A similaridade é calculada a partir da probabilidade do caminhante sair do vértice i e chegar em um vértice j em um tempo t , $P_{i,j}(t)$. O número de passos deve ser grande o bastante para explorar de forma significativa o grafo mas não pode ser muito longo para evitar o estado estacionário. Assim, dois vértices i e j pertencem à mesma comunidade se $P_{i,j}(t)$ é alto e se ambos possuem perspectivas semelhantes da rede. Desta forma, a medida de distância utilizada neste processo de hierarquização é dada por,

$$d_{i,j} = \sqrt{\sum_{z=1}^n \frac{(P_{i,z}(t) - P_{j,z}(t))^2}{k_z}}. \quad (2.24)$$

Para juntar duas comunidades, utiliza-se o método de Ward, que consiste na minimização da variância entre os elementos de cada comunidade como critério de escolha.¹³⁸ Assim como nos casos anteriores, a distribuição ideal é definida pela altura do dendrograma que corresponde ao maior valor da modularidade.

O tempo de execução desta implementação é $\mathcal{O}(dN^2)$ no pior caso e para redes reais (esparsas) a profundidade do dendrograma é pequena ($d \sim \log N$), então $\mathcal{O}(N^2 \log N)$. Assim como o *Fast Greedy*, esse algoritmo, apesar de rápido, não apresenta altos valores de modularidade para as divisões em redes reais e também não se aplica aos casos direcionados.

2.1.3.4 Algoritmo *Infomap*

Assim como o Wt, este algoritmo proposto por Roswall,¹³⁹ aplica a dinâmica de *random walks* para determinar comunidades analisando o fluxo de informação que percorre a rede. A ideia central aqui é codificar a trajetória dos caminhantes por meio do algoritmo de Huffman¹⁴⁰ e posteriormente minimizar a função objetiva (*map equation*¹⁴¹) resultante desta codificação, por esse motivo este método recebe o nome de *Infomap* (IMap).

Baseada na sequência de vértices alcançados pelos caminhantes em um percurso muito longo, cada vértice recebe inicialmente um código binário único, onde vértices com

Tabela 1: Lista de algoritmos utilizados neste trabalho. A primeira coluna indica o nome do método de detecção de comunidades; a segunda, a sigla de identificação do algoritmo; a terceira, a complexidade computacional para redes esparsas e as duas últimas, para qual tipo de grafo a técnica se aplica.

Método	Sigla	Complexidade	Ponderada	Direcionada
Girvan e Newman	GN	$\mathcal{O}(N^3)$	Sim	Sim
<i>Fast Greedy</i>	FG	$\mathcal{O}(N \log^2 N)$	Não	Não
<i>Walktrap</i>	Wt	$\mathcal{O}(N^2 \log N)$	Sim	Não
<i>Infomap</i>	IMap	$\mathcal{O}(N)$	Sim	Sim

Fonte: Elaborada pelo autor.

maiores probabilidades de serem visitados deverão ter nomes maiores do que os poucos visitados. Além disso, espera-se que num passeio completo, os vértices pertencentes à mesma partição devam aparecer em sequência por um longo período, já que o caminhante tende a permanecer no grupo. Para comprimir este código, o IMap utiliza uma descrição em dois níveis: no primeiro são mapeadas as comunidades, que terão um código próprio de identificação e de saída, já no segundo nível novos códigos são atribuídos aos membros de cada comunidade. Assim, dada a sequência binária correspondente ao trajeto, o primeiro código indica de qual partição o caminhante inicia o percurso, seguido dos códigos que representam os vértices visitados dentro da comunidade inicial até que o código de saída é identificado e assim por diante.

Uma vez que as possíveis partições e os caminhos foram devidamente codificados, utiliza-se a *map equation* para encontrar a comunidade ideal. Ela calcula o número total de *bits* necessários para codificar uma partição, ou seja, os códigos de identificação e saída do módulo, assim como o código dos vértices presentes nesta estrutura. Esta equação consiste em dois termos, expressados pela entropia de Shannon¹⁴² do passeio dentro e entre os grupos. A melhor partição é aquela que minimiza o número total de *bits*.

Além das redes simples, este método pode ser aplicado a grafos ponderados e digrafos. Neste último, a dinâmica do passeio aleatório é modificada para introduzir uma probabilidade de teleportação como a utilizada na centralidade *PageRank*, garantindo que o estado estacionário seja alcançado. Outras abordagens do Imap também já foram desenvolvidas, como detecção de comunidades com estruturas hierárquicas,¹⁴³ sobreposição¹⁴⁴ e caminhantes com memória recente (a probabilidade de transição depende dos últimos passos).¹⁴⁵

2.1.3.5 Benchmarks

A maior parte das redes reais possui características não triviais que influenciam na distribuição de comunidades e, portanto, não são ideais para comparação de desempenho e eficiência dos diferentes métodos de detecção. Portanto, para se testar um novo modelo,

Figura 9: *Benchmark* de Girvan e Newman (BGM). As três redes correspondem a realizações do modelo para (a) $z_{out} = 2$, (b) $z_{out} = 7$ e (c) $z_{out} = z_{in} = 8$. No último caso, os quatro grupos são muito difusos uma vez que estamos no limite da definição de comunidades.

Fonte: Elaborada pelo autor.

normalmente utiliza-se um *benchmark* que gera grafos com diferentes configurações de *clusters* com parâmetros bem controlados.

Um dos *benchmark* mais tradicionais é baseado no modelo *planted l -partition*.¹⁴⁶ Este gera grafos com $N = gl$ vértices em l partições com g membros. Vértices pertencentes ao mesmo grupo são conectados aleatoriamente como no modelo de Erdős-Rényi com probabilidade p_{in} , enquanto que vértices de grupos diferentes são ligados com probabilidade p_{out} . Assim, o grau médio dessas redes será $\langle k \rangle = (g-1)p_{in} + g(l-1)p_{out}$. Quando $p_{in} > p_{out}$ a densidade de conexões dentro das comunidades é maior do que a densidade de arestas entre os *cluster*, portanto a rede apresenta estrutura de comunidades. Girvan e Newman consideraram um caso especial deste modelo¹⁸ (BGM), fixaram $l = 4$, $g = 32$ (portanto $N = 128$) e $\langle k \rangle = 16$. Na prática, para construção dessas redes utiliza-se como parâmetros o grau médio de cada comunidade, $z_{in} = (g-1)p_{in} = 31p_{in}$, e o grau médio de ligações entre elas, $z_{out} = g(l-1)p_{out} = 96p_{out}$. Este grafo particular é um *benchmark* padrão e, em geral, o primeiro teste aplicado em um novo algoritmo de detecção de comunidades. Em geral, os algoritmos de detecção apresentam bons resultados para baixos valores de z_{out} e começam a falhar a medida que z_{out} se aproxima de 8.

A figura 9 ilustra três configurações diferentes para do modelo BGM. No primeiro caso, as quatro partições são bem definidas uma vez que $z_{out} = 2$, ou seja, o grau médio de cada comunidade é superior ao grau médio entre elas. Já no segundo caso, $z_{out} = 7$, indicando que o número de ligações dentro e fora das partições são próximos, entretanto ainda é possível distinguir visualmente cada uma delas. Diferentemente do último caso, quando $z_{out} = 8$, situação onde o grau das ligações dentro e fora das comunidades são iguais, tornando muito difusa a identificação adequada dos grupos, como era de se esperar dado que estamos no limite da definição expressa em (2.18). Devido ao baixo número de vértices e ao tamanho fixo das comunidades, a precisão deste método é quantificada comparando grupo a grupo para encontrar a maior quantidade de vértices classificados

corretamente.

Outras abordagens foram desenvolvidas a partir deste teste. Uma versão com arestas ponderadas foi proposta por Fan,¹⁴⁷ sendo atribuído diferentes pesos para arestas dentro e fora das comunidades. Já Arenas³³ desenvolveu um modelo hierárquico de dois níveis utilizando 256 vértices ao invés dos originais 128. No primeiro nível, N é dividido em 16 grupos e no segundo, diminui para 4 grupos com mais membros, uma vez que as 4 partições do nível anterior estão contidas neste nível. Em nosso trabalho, utilizamos como teste inicial apenas o modelo BGM e aprofundamos as análises com o próximo algoritmo.

As comunidades geradas por esse modelo seguem uma distribuição de Poisson, logo todos os vértices tem aproximadamente o mesmo grau médio, além de todas as partições serem criadas com o mesmo tamanho. Portanto, redes reais com estruturas modulares não são bem representadas por este método. Para aumentar a similaridade com a realidade, Lancichinetti, Fortunato e Radicchi desenvolveram outro algoritmo,¹⁴⁸ conhecido como *benchmark*. Eles assumiram que a distribuição de grau e o número de vértices nas comunidades seguem uma lei de potências com expoentes γ e β , respectivamente.^{125,149} Além disso, cada vértice compartilha a fração $1 - \mu$ de arestas com vértices do mesmo grupo e μ com vértices de outras partições, sendo $0 \leq \mu \leq 1$ definido como o parâmetro de mistura (*mixing parameter*). O grafo é construído da seguinte forma:

1. Uma sequência de tamanhos de comunidades é definido de acordo com a lei de potência dada por β ;
2. Cada vértice i recebe um grau interno $(1 - \mu)k_i$, com k_i determinado pela lei de potência dada por γ . Desta forma, cada vértice i possui $(1 - \mu)k_i$ conexões disponíveis;
3. Todos dos vértices da mesma comunidade são conectados aleatoriamente entre si até não existir mais conexões livres;
4. Cada vértice i recebe um número adicional de conexões livres igual a μk_i que é conectado aleatoriamente com vértices de comunidades diferentes até todas as arestas serem determinadas.

Claramente, quando μ é baixo, as comunidades são mais bem definidas umas das outras, sendo mais fácil os algoritmos de detecção identifica-las e, à medida que μ aumenta, a qualidade da classificação começa a cair. A figura 10 mostra um exemplo de rede simples gerado pelo modelo LFR, contendo 185 vértices distribuídos em nove comunidades bem definidas com $\beta = 1$, $\gamma = 2$ e $\mu = 0,1$. Vale notar que devido à abrangência deste algoritmo, também é possível reproduzir as redes do modelo BGN.

Como neste caso a distribuição de grau e o tamanho das comunidades seguem leis de potências com diferentes expoentes e o tamanho das redes são maiores, computacionalmente

Figura 10: Exemplo de rede gerada pelo modelo LFR com 185 vértices divididos em nove partições bem definidas dado que as distribuições de grau e de tamanho das comunidades seguem uma lei de escala com $\gamma = 2$ e $\beta = 1$, respectivamente. Além disso, $\mu = 0, 1$.

Fonte: Elaborada pelo autor.

é inviável calcular a precisão do resultado por meio da fração de vértices que foram classificados corretamente (forma utilizada em BGN). Assim, o *benchmark* LFR utiliza a informação mútua normalizada (*NMI*) como medida de precisão.¹⁵⁰ Esta é uma medida de similaridade utilizada em teoria da informação e é definida como

$$NMI = \frac{H(X) + H(Y) - H(X, Y)}{(H(X) + H(Y))/2}, \quad (2.25)$$

onde $H(X)(H(Y))$ é a entropia da variável aleatória $X(Y)$ associada às partições $C_1(C_2)$, enquanto que $H(X, Y)$ é a entropia conjunta. Já as entropias são dadas por

$$H(X) = - \sum_x p(x) \log p(x). \quad (2.26)$$

$$H(X, Y) = - \sum_x \sum_y p(x, y) \log p(x, y). \quad (2.27)$$

A informação mútua normalizada varia no intervalo $[0, 1]$, sendo máxima apenas quando as comunidades C_1 e C_2 forem exatamente iguais.

Em 2009, o modelo LFR foi estendido para geração de redes sobrepostas (possibilidade de um vértice pertencer simultaneamente a mais de uma comunidade), digrafos e redes ponderadas.¹⁵¹ Neste primeiro caso, a principal diferença no algoritmo é a divisão de k_i em k_i^{in} e k_i^{out} , ou seja, a parametrização é análoga ao caso sem direção. No caso

ponderado, são adicionados dois novos parâmetros, τ e μ_w . O parâmetro τ é usado para determinar a força de cada nó i , de acordo com $s_i = k_i^\tau$, uma vez que uma lei de potência entre grau e peso é frequentemente observada em redes reais ponderadas. Já μ_w é utilizado para atribuir a força interna, $s_i^{in} = (1 - \mu_w)s_i$, que é definida como a soma dos pesos entre o vértice i e todos os vizinhos que pertencem a mesma comunidade.

2.1.4 Redes reais

Para aplicar as técnicas desenvolvidas neste trabalho, escolhemos algumas redes construídas a partir de sistemas reais. Estes sistemas apresentam baixo número de vértices e conexões, simplificando a interpretação e visualização dos resultados, além de serem redes bem caracterizadas, uma vez que já foram exploradas por alguns trabalhos. A seguir, descrevemos brevemente as principais características dessas redes.

Clube de karatê Zachary: Rede social de amigos entre os 34 membros de um clube de karatê de uma universidade americana. Este grupo foi observado durante um período de três anos e as 78 arestas indicam a interação dos participantes dentro e fora das atividades do clube.¹⁵² Em dado momento ocorreu um desentendimento financeiro entre o presidente do clube e o principal professor, resultando na divisão do grupo em dois clubes menores. Devido a esta separação, esta rede real é usualmente utilizada como *benchmark* para algoritmos de detecção de comunidades.

Golfinhos nariz de garrafa: Rede social de um grupo de 62 golfinhos nariz de garrafa da região de Doubtful Sound (Nova Zelândia).¹⁵³ Esta rede foi montada durante sete anos de estudo onde as 159 arestas entre os golfinhos representam o resultado estatístico da proximidade de cada animal. Este grupo separou-se naturalmente em duas comunidades depois que os golfinhos deixaram o local de estudo por algum tempo. Devido a esta divisão natural ela também é utilizada como rede de testes para algoritmos de detecção de comunidades e assim como a rede do clube de karatê as arestas não possuem direção nem peso.

Os Miseráveis: Rede de interação entre os principais personagens do romance Os Miseráveis escrito pelo escritor francês Vitor Hugo no século XIX.¹⁵⁴ Neste sistema os 77 vértices representam os personagens e as 254 arestas indicam a participação dos mesmos em um único capítulo. A frequência com que os protagonistas contracenam é indicada pelo peso nas ligações. Contudo, a utilizaremos em dois contextos diferentes, com e sem ponderação.

Liga universitária de futebol americano: Os 115 vértices desta rede representam os times da primeira divisão do futebol americano universitário na temporada de 2000.¹⁸ Cada aresta corresponde a um jogo entre dois deles. Nesta competição os times são organizados em 12 conferências e partidas entre times pertencentes a mesma conferências são mais frequente do que os jogos entre elas. Com isso, uma estrutura de comunidades

surge naturalmente devido à organização do campeonato.

Músicos de jazz: Rede de colaboração de músicos de jazz que gravaram álbuns na mesma banda entre 1912 e 1940.¹⁵⁵ Os 198 vértices correspondem aos músicos e cada uma das 2742 arestas representam qual deles tocaram juntos em uma ou mais bandas. A estrutura modular presente neste grafo reflete duas características, uma segregação racial (os principais grupos contém apenas participantes negros ou brancos) e uma separação geográfica devido aos diferentes locais de gravação.

Verme *Caenorhabditis elegans*: Rede neural direcionada representando a conexão entre neurônios do nematoide *C. Elegans*.^{156,157} Esta rede foi construída baseado em mapeamento por imagem e microscopia eletrônica, onde as arestas são sinapses químicas e elétricas. Este verme tem um corpo longo e tubular, com os neurônios espalhados por toda sua extensão, não apenas em sua cabeça, assim, cada neurônio tende a se ligar com outros próximos a ele. Aqui, os 297 vértices representam os neurônios e as 2359 conexões as sinapses. Apesar desta rede ser direcionada, também utilizamos o caso particular sem direção.

Blogs de política: Rede direcionada de *hyperlinks* entre 1222 *blogs* durante a eleição presidencial americana de 2004.¹⁵⁸ Uma ligação direcionada é feita entre dois destes *sites* quando um *link* (origem), presente na página inicial, referencia outro *link* (destino). Ao final da navegação *Web*, 16714 arestas foram determinadas. Uma característica importante desta rede é a divisão dos *blogs* de acordo com seu alinhamento político, liberal ou conservador. Em nossos estudos utilizamos este grafo com e sem direção.

Guerra dos Tronos: Rede criada a partir do segundo livro, A Tormenta de Espadas, da saga de fantasia escrita por George R. R. Martin,¹⁵⁹ Crônicas de Gelo e Fogo. Os 107 vértices representam os personagens da trama que inclui desde a alta nobreza até guardas e aldeões. A ligação entre estes personagens é feita de acordo com a interação entre eles, que consiste tanto em interagir diretamente ou simplesmente falar um do outro, contudo isso não indica necessariamente que ambos são amigos ou possuem uma boa relação. Com isso, foram definidas 353 arestas ponderadas, no qual maiores valores correspondem a uma forte ligação entre os protagonistas.

Ações na bolsa de valores: Rede ponderada construída a partir do preço de fechamento diário de 433 ações. As conexões entre os vértices (ações) são definidas de acordo com a correlação do retorno histórico compreendido entre Janeiro de 2007 até Janeiro de 2017 (10 anos) e o peso de cada aresta é simplesmente o coeficiente de Pearson entre as respectivas ações. Para encontrar uma rede significativa, escolhe-se um valor de corte para as arestas, ou seja, as ligações que possuem correlação menor que o corte são removidas. Ao final, obtivemos uma rede com a maior componente contendo 121 vértices e 716 arestas. Estes vértices foram categorizados de acordo com o setor de atividade (Financeiro, Indústria, Energia, Materiais e Bens de Consumo). Mais detalhes sobre a

Tabela 2: Resumo de todas as redes utilizadas neste trabalho, indicando o número total de vértices e arestas na maior componente fracamente conectada, sua identificação com relação ao peso nas arestas e se possui direção.

Rede Complexa	N	E	Ponderada	Direcionada
Karatê	34	78	Não	Não
Golfinhos	62	159	Não	Não
Miseráveis	77	254	Sim	Não
Guerra dos Tronos	107	353	Sim	Não
Futebol Americano	115	613	Não	Não
Ações	121	716	Sim	Não
Jazz	198	2742	Não	Não
<i>C. elegans</i>	297	2359	Não	Sim
Blogs de Política	1222	16714	Não	Sim

Fonte: Elaborada pelo autor.

construção desta rede se encontram no apêndice B.

A tabela 2 sumariza as nove redes descritas anteriormente, indicando o número total de vértices e arestas na maior componente fracamente conectada e se ela é ponderada ou um digrafo.

2.2 Passeios aleatórios em redes complexas

O estudo de modelos matemáticos por meio de caminhadas aleatórias é realizado em diversas áreas da ciência. Na física, por exemplo, estes passeios refletem a dinâmica de muitos sistemas naturais como interação de proteínas, cadeia de polímeros e difusão de gases⁵¹. Em redes complexas, os passeios aleatórios são utilizados no estudo estrutural das redes, uma vez que são extraídas informações globais do sistema por meio da exploração local.^{72,160,161} A seguir, introduzimos os conceitos básicos dos dois tipos específicos de passeios aleatórios estudados neste trabalho.

2.2.1 Passeio aleatório

O passeio aleatório sem restrição em redes complexas (RW, do inglês *random walk*) é um processo estocástico de tempo discreto descrito por uma equação mestra.^{71,162} Assim, se um caminhante está no vértice j no instante de tempo $t - 1$ ele escolhe, com probabilidade uniforme, um de seus vizinhos para ocupar no tempo t . Portanto a probabilidade de encontrar um caminhante no vértice i no instante t é

$$\mathcal{P}_i(t) = \sum_j \frac{A_{i,j}}{k_j} \mathcal{P}_j(t-1), \quad (2.28)$$

ou na forma matricial,

$$\mathcal{P}(t) = \mathbf{A}\mathbf{D}^{-1}\mathcal{P}(t-1), \quad (2.29)$$

com $\mathcal{P}$ o vetor com os elementos de $\mathcal{P}_i$ e $\mathbf{D}$ uma matriz diagonal contendo o grau de cada vértice.

No limite de tempos longos, esta dinâmica converge para um estado estacionário, ou seja, as probabilidades não se alteram mais com a evolução temporal do sistema. Nesta condição, temos que

$$\mathcal{P}(t) = \mathcal{P}(t-1) = \mathcal{P}^\infty, \quad (2.30)$$

onde $\mathcal{P}^\infty$ são as probabilidades estacionárias do RW para $t \rightarrow \infty$. Com isso, utilizando a equação 2.29, ficamos com,

$$\mathcal{P}^\infty = \mathbf{A}\mathbf{D}^{-1}\mathcal{P}^\infty, \quad (2.31)$$

que pode ser rearranjada para

$$(\mathbf{I} - \mathbf{A}\mathbf{D}^{-1})\mathcal{P}^\infty = (\mathbf{D} - \mathbf{A})\mathbf{D}^{-1}\mathcal{P}^\infty = \mathcal{L}\mathbf{D}^{-1}\mathcal{P}^\infty = 0, \quad (2.32)$$

onde $\mathcal{L}$ é a matriz Laplaciana¹⁶³ e $\mathbf{D}^{-1}\mathcal{P}^\infty$ seu autovetor com autovalor zero. Assim, temos que $\mathcal{P}^\infty = a\mathbf{D}\mathbf{I}$, com a uma constante e $\mathbf{I} = (1, 1, 1, \dots, 1)$. Então para uma rede conectada, a probabilidade de um caminhante ser encontrado no vértice i no limite de tempos longos é

$$\mathcal{P}_i^\infty = \frac{k_i}{\sum_j k_j}, \quad (2.33)$$

ou seja, vértices de alto grau possuem maior probabilidade de serem visitados, pois há mais formas de se chegar até eles.

Como o estado futuro do caminhante depende apenas do seu estado atual (o caminhante não apresenta memória) esta dinâmica é classificada como um processo estocástico Markoviano.¹⁶⁴ Por esse motivo, há diferentes formas de se determinar computacionalmente o estado estacionário, entre elas, destacamos o método matricial que consiste basicamente em diagonalizar a matriz de transição $\mathbf{T} = \mathbf{A}\mathbf{D}^{-1}$ da equação 2.29 e identificar o autovetor associado ao autovalor unitário (teorema de Perron-Frobenius¹⁶⁵).

2.2.2 Passeio aleatório auto-excludente

Vários modelos alternativos de caminhadas aleatórias em redes complexas vem sendo estudados visando otimizar análises topológicas.^{76–78} Entre estes modelos, destaca-se o passeio aleatório auto-excludente (SAW, do inglês *self-avoiding walk*). Nesta dinâmica, cada vértice da rede não pode ser revisitado durante a mesma caminhada, obrigando o caminhante a sempre procurar um caminho viável através dos vértices que ainda não foram visitados. Caso não exista mais vértices acessíveis, a caminhada termina.

A figura 11 exemplifica esse tipo específico de passeio. Inicialmente em 11(a) temos um caminhante partindo do vértice 3 e atingindo o vértice 4 (neste primeiro passo, além do vértice 4, o caminhante também tinha a opção de chegar em 2). Uma vez em 4, o

Figura 11: Exemplo do passeio auto-excludente em uma rede simples. (a) Ele se inicia no vértice 3 e, no primeiro passo, atinge o nó 4. (b) Uma vez em 4, o caminhante só pode ir para o vértice 2, pois 3 já foi visitado. (c) Partindo de 2, o caminhante atinge obrigatoriamente o nó 1 e consequentemente termina o passeio dado que todos os vértices já foram visitados.

Fonte: Elaborada pelo autor.

caminhante não pode mais voltar para o nó anterior devido a restrição do SAW, portanto ele passa para o vértice 2. Já neste sítio, o caminhante só pode alcançar o vértice 1, pois 3 e 4 já foram visitados e finalmente em 1, o passeio termina dado que não existe mais vértices acessíveis da posição atual.

Diferentemente do RW, nesse caso os passos futuros do caminhantes são influenciados pelos estados anteriores, ou seja, o sistema apresenta memória infinita do caminho percorrido e, portanto, ele é classificado como um processo estocástico não Markoviano. Embora esta classe de problemas apresente grande complexidade matemática, Herrero estudou o SAW para diferentes tipos de redes aleatórias e obteve resultados analíticos para o *attrition length* (número de passos que um caminhante realiza antes de parar).^{88,90,91} Além deste, recentemente, Kim estudou analiticamente o tempo médio de primeira passagem em vários tipos de redes, mostrando ser um recurso exploratório bem eficiente.¹⁶⁶ Aplicações práticas desta dinâmica são encontradas nos trabalhos de Travençolo,^{167,168} que aplicou este modelo aleatório para detecção de bordas em redes complexas e, baseado nas definições de acessibilidades proposta por Rosvall,¹⁶⁹ estudou as características desta definição utilizando SAW aplicado a ruas de cidades. Mais recentemente, Bertozzi¹⁷⁰ estudou a evolução hierárquica de organizações criminosas e aplicou a caminhada auto-excludente como estratégia para capturar os participantes e consequentemente acabar com tais organizações.

3 RESULTADOS

Uma vez apresentados os conceitos básicos e os dados reais utilizados neste trabalho, neste capítulo, discutiremos os resultados provenientes da dinâmica aleatória auto-excludente. Inicialmente, aplicaremos o SAW em algumas redes teóricas para compará-lo com o caminhante sem restrição e para obter informações relevantes tanto da estrutura da rede como da própria dinâmica estocástica. Em seguida, definiremos um método de detecção de comunidades baseado nas análises preliminares e submeteremos este algoritmo a alguns testes em *benchmarks* e redes reais para fim de validação. Ao término do capítulo, generalizamos esta técnica para redes ponderadas, digrafos e sistemas com metadados.

3.1 Caracterização do modelo auto-excludente

Inicialmente, calculamos a probabilidade de encontrar um caminhante no vértice i em um tempo menor ou igual a t para o SAW e para o RW. Sabendo que o processo Markoviano apresenta resultado analítico para t suficientemente grande, dado pela equação (2.33), foi possível verificar a influência da memória infinita na dinâmica do passeio aleatório. Para o SAW, iniciamos 10.000 caminhantes de vértices aleatórios e evoluímos a dinâmica até $t = 5000$ para redes com $N = 5000$. Baseado no trajeto de cada caminhante, calculamos a distribuição de graus P_k de acordo com a expressão (2.6) .

Nos gráficos da figura 12, mostramos a comparação entre a caminhada auto-

Figura 12: Comparação entre as caminhadas RW no estado estacionário (linhas contínuas) e SAW (pontos com linhas tracejada) para redes tipo (a) ER e (b) BA para diferentes graus médios e mesma quantidade de vértices, $N = 5000$. O gráfico interno da figura (b) corresponde a um aumento da região inicial tal que $0 \leq k \leq 230$ e os valores indicados na escala representam o ponto de cruzamento das curvas vermelha, azul e verde respectivamente.

excludente (linha tracejada com pontos) e o RW no estado estacionário, equação (2.33), (linha contínua) para redes tipo ER e BA a medida que aumentamos grau médio da rede ($\langle k \rangle = [8, 10, 14]$). Em ambos os tipos de rede, podemos observar que vértices de alto grau são menos visitados na caminhada auto-excludente, principalmente para rede BA, onde rapidamente ocorre uma saturação na distribuição, resultando em uma grande diferença entre as duas dinâmicas conforme o grau aumenta. Outra diferença para estes modelos pode ser observada quando k é baixo independentemente do $\langle k \rangle$: no caso ER, o P_k se mantém superior no SAW até $k \sim \langle k \rangle$, já no modelo livre de escala a inversão das curvas ocorre para $k \gg \langle k \rangle$. Portanto o efeito da memória infinita aumenta a probabilidade de encontrarmos os caminhantes em vértices de baixo grau, principalmente na topologia livre de escala, onde a presença de *hubs* limita rapidamente o desenvolvimento do passeio.

Outra característica do SAW que medimos para valores crescentes do grau médio foi o número de passos desenvolvidos pelos caminhantes. Além dos modelos ER e BA, aqui também evoluímos a dinâmica para redes geradas por meio do *benchmark* LFR com os parâmetros dados por $\beta = 2$, $\gamma = 3$, $\mu = 0,3$ e os tamanhos das comunidades variando entre 1% e 10% do tamanho da rede. Conforme mostrado na figura 13, para os três modelos testados, utilizamos três tamanhos diferentes de redes 5000, 10000 e 20000 representados respectivamente pelas cores vermelha, verde e azul, já para o grau médio, definimos $\langle k \rangle = 10, 20, 30$. A fim de obter uma boa estatística, cada curva foi construída usando 10^5 caminhantes partindo de vértices aleatórios sorteados com probabilidade uniforme.

Como era de se esperar, em todos os casos temos que quanto maiores o grau médio e o tamanho da rede, maior é o número médio de passos dados pelos caminhantes, já que existem mais rotas disponíveis para os *walkers* evoluírem a caminhada. Entretanto nas redes BA, independente do grau, o deslocamento da distribuição acontece de forma muito mais lenta do que nas outras duas. Este efeito está em acordo com o resultado apresentado na figura 12, onde a presença de *hubs* dificulta o desenvolvimento do SAW. Já nos casos ER e LFR as distribuições são muito parecidas, porém no primeiro caso, elas possuem o lado esquerdo mais alongado do que o resultado para o *benchmark*, fazendo com o que o número médio de passos seja levemente menor, ou seja, os caminhantes têm um pouco mais de dificuldade em explorar redes ER do que as redes artificiais com comunidades geradas pelo modelo LFR.

Para complementar o estudo do comprimento dos caminhantes, calculamos, para cada distribuição, os números médio (analogamente a equação 2.7) e máximo de passos. Dada estas medidas e número total de vértices de cada rede (N), encontramos respectivamente a fração média e a fração máxima de vértices alcançados, como mostrado na figura 14. Os modelos ER e LFR, apresentaram um comportamento aproximadamente linear independente do grau médio, contudo quando temos comunidades na rede tanto a fração média como a fração máxima são levemente superiores, evidenciando novamente a maior

Figura 13: Distribuições do número de passos para caminhantes realizando o SAW em três modelos distintos de redes: BA, ER e LFR para três tamanhos diferentes $N = 5000$, 10000 e 20000 e $\langle k \rangle$ definido para 10 , 20 e 30 . Para obter uma boa estatística utilizamos $N = 10^5$ para cada uma das curvas. Ademais, no modelo com comunidades usamos os parâmetros adicionais: $\beta = 2$, $\gamma = 3$, $\mu = 0,3$ e os tamanhos das comunidades variando entre 1% e 10% do tamanho da rede.

Fonte: Elaborada pelo autor.

facilidade de deslocamento dos *walkers* na rede com partições. Por outro lado, no caso BA as amplitudes das duas medidas são uma ordem de grandeza inferiores quando comparada com os outros casos e, além disso, o comportamento das curvas não apresenta sinais de linearidade, caindo consideravelmente à medida que o grau médio aumenta. No mais, também vale observar que nos três modelos existe uma grande diferença de magnitude quando saímos de $\langle k \rangle = 10$ e vamos para $\langle k \rangle = 20$, o que não ocorre de $\langle k \rangle = 20$ para $\langle k \rangle = 30$.

Uma vez observado o impacto da memória e o comprimento dos caminhantes no processo difusivo, o método que utilizamos para estudar com maiores detalhes a topologia das redes através do SAW consiste em iniciar a dinâmica com M_i caminhantes partindo do vértice i . Desta forma, para um tempo suficientemente longo, definimos $m_{i,j}$ como o número de caminhantes que saíram de i e passaram por j antes de parar e $l_{i,j}^w$ a quantidade

Figura 14: Frações média e máxima de vértices alcançados em função do tamanho da rede. Novamente analisamos os resultados aplicados aos modelos ER, BA e LFR para $\langle k \rangle = 10, 20, 30$.

Fonte: Elaborada pelo autor.

de passos necessária para o k -ésimo caminhante w_i^k alcançar j partindo de i . Assim, baseado nestas informações provenientes da dinâmica do passeio aleatório, definimos a probabilidade de transição, $p_{i,j}$, através da fração de caminhantes que saíram de i e passaram por j ,

$$p_{i,j} = \frac{m_{i,j}}{M_i}, \quad (3.1)$$

e seu respectivo número médio de passos

$$\langle l_{i,j} \rangle = \frac{1}{m_{i,j}} \sum_{w_i^k} l_{i,j}^k, \quad (3.2)$$

onde a somatória é feita apenas sobre os w_i^k que efetivamente passam por j . A equação 3.1 mede a facilidade com que os caminhantes atingem o vértice j partindo de i independentemente das distâncias topológicas, então se existem muitas formas de encontrar j ou a conexão entre i e j é trivial, o valor de $p_{i,j}$ tende a ser elevado. Por outro lado, complementarmente a $p_{i,j}$, $\langle l_{i,j} \rangle$ revela as distâncias topológicas da rede por meio do número médio de passos entre o par i, j .

Figura 15: Classificação de um par de vértices de acordo com $p_{i,j}$ e $\langle l_{i,j} \rangle$. Os vértices na cor verde indicam a origem dos caminhantes enquanto os amarelos o destino. (a) TIPO A: Situação na qual a fração de caminhantes e a quantidade média de passos é elevada. (b) TIPO B: Este grafo representa poucos caminhantes atingindo o destino em poucos passos. (c) TIPO C: Nesta rede temos uma baixa fração de caminhantes percorrendo altas distâncias. (d) TIPO D: Finalmente, muitos caminhantes que saíram do vértice verde atingem o amarelo em poucos passos.

Fonte: Elaborada pelo autor.

De forma simplificada, a partir dessas duas grandezas, podemos classificar topologicamente um par de vértices de acordo com quatro situações. Quando a fração de caminhantes que passou por j partindo de i é alta e a quantidade média de passos também é elevada indica que o caminho entre i e j é largo porém muitos viajantes atingem o destino. Em geral, esta situação acontece quando o número de caminhos entre dois vértices é baixo e, além disso, esta distância é longa, mas simplificada, tornando o número de visitas elevado, como mostra a figura 15(a). Na figura 15(b) temos o oposto, o par i, j possui $p_{i,j}$ e $\langle l_{i,j} \rangle$ baixos, mostrando que eles são próximos mas o acesso é baixo. Uma situação em que isto acontece é quando o vértice origem está conectado a um *hub* e o destino é outro vértice ligado ao mesmo *hub*. Já quando poucos caminhantes atingem o destino e o número de passos médio para isso é alto, os vértice i e j estão distantes e não existem muitas rotas conectando-os. Este caso ocorre tipicamente quando i e j pertencem a comunidades distintas, como mostra o grafo da figura 15(c). Já na rede da figura 15(d),

Figura 16: Gráfico de espalhamento construído a partir das medidas definidas nas expressões (3.1) e (3.2) e aplicadas nos modelos BA, ER e LFR, todos com $\langle k \rangle = 8$ e $N = 500$. No último, os parâmetros adicionais foram escolhidos tal que $\beta = 2$, $\gamma = 3$, $\mu = 0, 5$ e o tamanho das comunidades variando entre 10 e 50. Aqui, utilizamos os quatro padrões de classificação entre vértices definidos na figura 15 para distinguir esses três tipos de rede.

Fonte: Elaborada pelo autor.

temos um exemplo de $p_{i,j}$ alto e $\langle l_{i,j} \rangle$ baixo, indicando que o vértice origem apresenta grau baixo (poucas possibilidades de caminho) e está próximo do destino. Para simplificar a notação, identificamos cada uma das quatro possíveis classificações, respectivamente por TIPO A, TIPO B, TIPO C e TIPO D.

Apesar de simplificada, esta interpretação do par de medidas $p_{i,j}$, $\langle l_{i,j} \rangle$ pode ser aplicada para identificar padrões estruturais nos diferentes tipos de redes complexas por meio de um gráfico de espalhamento, como o exposto no gráfico da figura 16. Neste, temos três modelos diferentes de redes (BA, ER e LFR) com $\langle k \rangle = 8$ e $N = 500$, ou seja, cada respectivo conjunto de pontos contém 249.500 valores ($N^2 - N$). Para gerar a rede artificial com comunidades os parâmetros utilizados foram $\beta = 2$, $\gamma = 3$ com os tamanhos das partições no intervalo $[10, 50]$ para $\mu = 0, 5$.

No modelo BA, a nuvem de pontos ficou concentrada no lado esquerdo do gráfico, correspondendo a baixos comprimentos médios, com valores distribuídos por toda extensão de $p_{i,j}$. Com isso, concluímos que devido a efeito *ultra-small world* os caminhantes não conseguem desenvolver a dinâmica, porém alguns poucos vértices são facilmente alcançados por muitos caminhantes, dando a este espalhamento o aspecto de uma gota. Assim, conseguimos identificar nesta rede livre de escala dois padrões, ambos para pequenos valores de $\langle l_{i,j} \rangle$ e para os dois tipos de $p_{i,j}$, representados pelos tipos B e D.

Já para o caso ER, os valores da probabilidade de transição ficaram menos espalhados, com a maior parte dos pontos concentrada entre 0,1 e 0,7. Por outro lado, os caminhantes evoluíram mais o passeio uma vez que o comprimento médio se mostrou superior ao caso BA. Para este tipo de rede identificamos as quatro categorias de classificação, contudo a maior parte dos valores estão mais próximos das classes TIPO A e TIPO C.

Finalmente para o *benchmark*, temos a maior concentração de valores entre os três modelos para ambas as medidas. A presença de comunidades, mesmo que não muito bem definida ($\mu = 0,5$), faz com que os *walkers* explorem mais a rede, aumentando o comprimento médio e também a probabilidade de transição. Além disso, como as partições são relativamente pequenas, temos a predominância de maiores valores para $\langle l_{i,j} \rangle$, pois uma vez que um caminhante escapa de uma comunidade ele tende a dar muitos passos. Desta forma, destaca-se aqui os padrões TIPO C e TIPO D, o primeiro mais comum entre vértices de comunidades distintas e o segundo, no caso de nós que dividem o mesmo grupo.

Portanto, a análise estrutural por meio da combinação das expressões (3.1) e (3.2) possibilita distinguir e até mesmo classificar modelos de redes complexas através das propriedades dinâmicas intrínsecas do SAW.

Considerando essas características da dinâmica do passeio auto-excludente, combinamos as informações provenientes das equações (3.1) e (3.2) em uma única medida e definimos $f_{i,j}$,

$$f_{i,j} = \frac{p_{i,j}}{\langle l_{i,j} \rangle}. \quad (3.3)$$

O inverso do comprimento médio quantifica essencialmente a proximidade entre um par de vértices ("*closeness*"), mas mesmo vértices que podem ser alcançados em poucos passos poderiam ser difíceis de receber a visita de um caminhante pois poucas rotas existem entre este par. Portanto, multiplicamos o inverso da distância média pela probabilidade de transição, para refletir de maneira mais adequada a proximidade entre um par de vértices, com a definição acima atingindo valores mais altos à medida que $p_{i,j}$ cresce e $\langle l_{i,j} \rangle$ diminui.

Na figura 17, temos a distribuição dos valores de $f_{i,j}$ para modelos de redes com a mesma configuração da figura 16, ou seja, redes ER, BA e LFR com $N = 500$, $\langle k \rangle = 8$ e no caso com comunidades, $\beta = 2$, $\gamma = 3$, $\mu = 0,5$ com as partições variando entre [10, 50] vértices cada. O primeiro gráfico (figura 17a) é o resultado para o modelo ER, onde a grande maioria dos valores está no intervalo $0, 0 < f_{i,j} < 0,01$ (esta rede particular possui dois vértices que atingiram $f_{i,j} = 1$). Portanto, com o intuito de obter uma maior resolução da região inicial, mostramos no respectivo gráfico interno os pontos compreendidos entre $[0,0, 0,01]$ e, a partir deste, vemos que nesta região os dados são distribuídos de forma praticamente simétrica com a maior frequência em torno de 0,003. Analogamente, no resultado para rede BA, figura 17b, também ocorreu uma aglomeração de pontos para baixos valores de $f_{i,j}$, sendo necessária a ampliação da região para melhor visualização. Assim, diferentemente do caso anterior, as maiores frequências estão localizadas nas

Figura 17: Distribuição dos valores de $f_{i,j}$ para os modelos (a) ER, (b) BA e (c) LFR. Os parâmetros utilizados aqui foram: $N = 500$, $\langle k \rangle = 8$ e no caso com comunidades, $\beta = 2$, $\gamma = 3$, $\mu = 0,5$ com as partições variando entre $[10, 50]$ vértices cada. Os gráficos internos de (a) e (b) correspondem a ampliação da região inicial das respectivas distribuições.

Fonte: Elaborada pelo autor.

primeiras medidas da escala, entre 0,001 e 0,004; deste ponto em diante a distribuição apresenta uma queda acentuada até atingir uma uniformidade em torno de 0,01 e continuar até 0,02. O último caso, presente na figura 17c, representa a distribuição para o *benchmark* LFR. Aqui, temos os valores de $f_{i,j}$ mais concentrados ($[0,0025, 0,0095]$), diferentemente das redes ER e BA que foi necessário até ampliar o início da distribuição. Além disso, apesar da falta de simetria, as frequências caem suavemente após os valores de pico, encontrados entre 0,003 e 0,004. Outra característica observada nestes gráficos é que mesmo com tipos de estruturas de redes diferentes, a escala de $f_{i,j}$ é similar em todos os casos (quando consideramos a maior parte da distribuição nos casos ER e BA), como podemos ver comparando o eixo das abcissas dos gráficos internos de (a) e (b) com o gráfico de (c).

Note que, $p_{i,j}$ varia entre zero e um e o valor mínimo que $\langle l_{i,j} \rangle$ pode assumir é unitário ($i \neq j$), então $0 \leq f_{i,j} \leq 1$ para $i \neq j$ e, em geral, $f_{i,j} \neq f_{j,i}$ devido a assimetria da dinâmica do SAW. Além disso, quando o número médio de passos entre dois vértices

Figura 18: Norma (equação 3.5) da diferença de $\mathbf{F}$ em função do número médio de caminhantes por vértice para um intervalo de $\Delta M = 100$ aplicado a redes do modelo LFR com tamanhos 100, 500 e 1000. Para melhor visualização, o eixo das ordenadas está em escala logarítmica.

Fonte: Elaborada pelo autor.

começa a crescer $f_{i,j}$ vai a zero independentemente da fração de caminhantes. Por outro lado, quando o caminho médio entre dois vértices é pequeno, $f_{i,j} \sim p_{i,j}$. Quando um vértice j não recebe visitas de caminhantes partindo de i , nesta situação, assumimos que $\langle l_{i,j} \rangle \rightarrow \infty$, ou seja,

$$\lim_{\langle l_{i,j} \rangle \rightarrow \infty} f_{i,j} = 0. \quad (3.4)$$

Outra característica importante da matriz $\mathbf{F}$ é sua convergência, ou seja, qual a quantidade necessária de caminhantes para estabilizar as variações de $f_{i,j}$. Com o intuito de quantificar esta propriedade, utilizamos a norma de Frobenius definida para matrizes,¹⁷¹

$$\|\mathbf{F}\| = \sqrt{\sum_{i,j} |f_{i,j}|^2}, \quad (3.5)$$

e a partir desta, calculamos a norma da diferença de $\mathbf{F}$ em função do número de caminhantes por vértice (M), $\|\mathbf{F}' - \mathbf{F}\|$, com $\mathbf{F}'$ calculado para $M + \Delta M$ onde ΔM é um passo pré definido. O resultado para o modelo LFR com diferentes tamanhos de rede ($N = 100, 500, 1000$) e com $\Delta M = 100$ é mostrado na figura 18 (para uma melhor visualização, o gráfico está com escala logarítmica no eixo y). Nesta, constatamos que após uma década e meia o valor da norma da diferença começa a diminuir lentamente, independente do tamanho da rede. Assim, temos a opção de usar esta medida como um parâmetro de precisão de $\mathbf{F}$ e, ao invés de fixarmos o número de caminhantes por vértice, podemos escolher uma tolerância de convergência. Também é possível concluir que com um número relativamente baixo de caminhantes, obtemos uma baixa variação de $f_{i,j}$, por exemplo:

para uma tolerância de 0,001, precisamos de 3920, 4590 e 4860 caminhantes por vértice para redes com $N = 100, 500, 1000$. Isto também mostra a baixa dependência de M com o tamanho da rede, conforme o tamanho da rede cresce a norma da diferença fica praticamente constante para uma dada precisão.

3.2 Detecção de comunidades através do passeio auto-excludente

Redes que possuem subgrafos com alta densidade de ligações entre seus membros, em comparação com o restante do grafo, apresentam padrões na caminhada aleatória auto-excludente que refletem em valores relativamente altos de $f_{i,j}$ entre elementos do subgrafo e baixos valores com o restante da rede. Este comportamento está relacionado diretamente com a definição de comunidades (equação 2.18). De fato, um conjunto de vértices na mesma comunidade tende a apresentar altos valores de $p_{i,j}$ e baixos valores de $\langle l_{i,j} \rangle$ (muitos caminhantes alcançam os membros de uma mesma comunidade em poucos passos) devido a grande concentração de arestas dentro da partição.

Essas características são exemplificadas nas matrizes da figura 19. Nela temos em (a) um grafo com cinco comunidades bem definidas com dez vértices cada, sequencialmente numerados para favorecer a visualização dos conjuntos. Os resultados diretos do SAW aplicado a esta rede, $p_{i,j}$ e $\langle l_{i,j} \rangle$, são mostrados respectivamente em (b) e (c). Na probabilidade de transição temos a diagonal nula (o caminhante nunca retorna para o vértice origem) e o restante das medidas concentradas em um estreito intervalo entre aproximadamente 0,6 e 1,0, ou seja, apenas com $p_{i,j}$ não é possível identificar claramente a presença dos subgrafos. Já para o comprimento médio dos *walkers* temos uma amplitude maior, sendo possível verificar que os caminhantes realizam trajetórias menores dentro das comunidades (tons de bordô) e maiores entre elas (tons de amarelo). Finalmente em (d), mostramos a matriz $\mathbf{F}$, resultado da equação 3.3, onde podemos verificar facilmente que as regiões com os maiores valores de $f_{i,j}$ correspondem aos vértices de uma mesma partição. Mesmo com a magnitude desta reduzida ($p_{i,j} \ll \langle l_{i,j} \rangle$), a razão entre estas duas medidas ressalta os padrões topológicos e facilita a futura identificação sistemática destas estruturas. Como vemos na diagonal das matrizes de (b) e (c), os valores de $p_{i,i}$ e $\langle l_{i,i} \rangle$ são nulos, deixando a diagonal de $\mathbf{F}$ indefinida. Para resolver esta divergência, utilizamos para $i = j$ a seguinte expressão,

$$f_{i,i} = \max_{\substack{1 \leq j \leq N \\ j \neq i}} f_{i,j}, \quad (3.6)$$

ou seja, atribuímos à $f_{i,i}$ o maior valor $f_{i,j}$ tal que j é o vértice mais facilmente alcançado por i . Assim, temos todos os elementos de $\mathbf{F}$ variando no intervalo $0 \leq f_{i,j} \leq \max_i f_{i,i}$.

Figura 19: (a) Grafo com uma estrutura de comunidades bem definida com dez vértices em cada uma das cinco partições (estes estão numerados sequencialmente para favorecer a visualização dos padrões em $f_{i,j}$). (b) Imagem representando a matriz das probabilidades de transições, como todos os valores são altos não é possível distinguir claramente os cinco grupos. (c) Matriz com os valores de $\langle l_{i,j} \rangle$, aqui fica evidente que os caminhantes realizam trajetórias menores dentro das comunidades e maiores entre elas. (d) Figura indicando a matriz $\mathbf{F}$, as sub-matrizes com valores mais elevados indicam os vértices de uma mesma partição. Os elementos da diagonal são determinados através da expressão 3.6.

Fonte: Elaborada pelo autor.

3.2.1 Descrição do Método

A semelhança entre as linhas de $\mathbf{F}$ refletem a perspectiva dos caminhantes que partiram do mesmo grupo, assim, matematicamente, as comunidades podem ser determinadas identificando os vértices que apresentam linhas similares. Portanto, utilizando esta grandeza como parâmetro para uma medida de similaridade, desenvolvemos um método de clusterização hierárquica aglomerativa (apêndice A) capaz de encontrar a distribuição de comunidade mais adequada por meio da modularidade, (equação 2.19).

Essa matriz pode ser interpretada como um conjunto de N amostras (vértices origem dos caminhantes) por N dimensões (vértices destino dos caminhantes que partiram de i). Entretanto, como temos um espaço com muitas dimensões, os dados estão esparsamente distribuídos, prejudicando o desempenho do nosso classificador. Assim, visando melhorar a separação das comunidades e consequentemente a eficiência do método, utilizamos a análise de componentes principais (PCA) para remover possíveis redundâncias e reduzir a

dimensionalidade do sistema, evitando a “maldição da dimensionalidade”.¹⁷²

O PCA é um método paramétrico de extração de informações relevantes em dados redundantes.¹⁷³ Seu principal objetivo é re-expressar o conjunto original em uma nova base na qual as observações não são correlacionadas e a maior parte da informação se concentra em poucas dimensões. Para isso, este método propõe a seguinte transformação linear

$$\tilde{\mathbf{F}} = \mathbf{P} \cdot \mathbf{F}, \quad (3.7)$$

onde $\mathbf{P}$ é a nova base composta pelas componentes principais mais relevantes de $\mathbf{F}$ e $\tilde{\mathbf{F}}$ são os dados originais projetados em $\mathbf{P}$.

Uma vez realizado o pré-processamento por meio do PCA e definido a modularidade como parâmetro de qualidade, podemos descrever nosso algoritmo de clusterização hierárquica aglomerativa da seguinte forma,

1. Execute um número suficiente de SAWs começando de cada vértice;
2. Calcule $\mathbf{F}$ através das equações (3.3) e (3.6);
3. Aplique o método do PCA em $\mathbf{F}$;
4. Utilize duas componentes principais de $\mathbf{F}$ para determinar $\tilde{\mathbf{F}}$;
5. Determine a clusterização hierárquica aglomerativa a partir de $\tilde{\mathbf{F}}$;
6. Calcule a modularidade de cada nível do dendrograma;
7. Salve a configuração de maior modularidade;
8. Aumente o número de componentes principais de $\mathbf{F}$;
9. Repita do passo 5 até todas as componentes serem usadas.

O resultado do processo é um dendrograma que possui a distribuição de comunidades que maximiza a modularidade para um determinado número de componentes principais do PCA.

Como descrito no apêndice A, na clusterização hierárquica, precisamos definir um critério para determinar a proximidade de dois vértices (métrica) e a forma com que os dados serão agrupados (método). Para realizar esta escolha combinamos três diferentes técnicas de agrupamento (*single*, *average* e *complete*) com cinco medidas de proximidade (Euclideana, Manhattan, Canberra, Chebyhev e Bray-Curtis) e executamos o algoritmo descrito acima para um conjunto de sete redes reais e uma rede artificial LFR; os resultados são mostrados na tabela 3. Para o *benchmark* foram realizadas 100 execuções em diferentes redes ($N = 500$, $\beta = 2$, $\gamma = 3$, $\mu = 0,3$ e o tamanho das comunidades no intervalo $[10, 50]$)

Tabela 3: Comparação de diferentes métodos e métricas a fim de escolher o par mais adequado para a clusterização hierárquica aglomerativa por meio do SAW. Apesar da modularidade ser estatisticamente igual para as redes do modelo LFR (aproximadamente $0,6393 \pm 0.0005$), a métrica de Bray-Curtis e o método do *average linkage* (destacados em negrito) obtiveram os melhores resultados em todas as sete redes reais e, conseqüentemente, foram adotados como medidas padrão do nosso algoritmo de detecção de comunidades.

Método	Métrica	Redes Complexas							
		LFR	Karatê	Golfinhos	Miseráveis	F. Americano	Jazz	C. Elegans	B. Política
<i>Single</i>	Euclidean	0,6394	0,3786	0,4649	0,5348	0,6044	0,4060	0,2784	0,4158
	Manhattan	0,6393	0,3635	0,4519	0,5362	0,6031	0,4039	0,2715	0,4136
	Canberra	0,6393	0,3990	0,4824	0,5376	0,6006	0,3261	0,3301	0,4135
	Chebyshev	0,6393	0,3763	0,4500	0,5348	0,6005	0,3666	0,3002	0,4203
	Bray-Curtis	0,6394	0,4197	0,5101	0,5404	0,5918	0,4223	0,3307	0,4249
<i>Average</i>	Euclidean	0,6393	0,4188	0,4962	0,5361	0,6043	0,4267	0,3626	0,4233
	Manhattan	0,6393	0,4197	0,5086	0,5416	0,6043	0,4333	0,3476	0,4244
	Canberra	0,6393	0,3990	0,4944	0,5408	0,6035	0,4348	0,3576	0,4249
	Chebyshev	0,6393	0,4119	0,4962	0,5342	0,6007	0,4364	0,3475	0,4240
	Bray-Curtis	0,6394	0,4197	0,5277	0,5467	0,6044	0,4428	0,3746	0,4268
<i>Complete</i>	Euclidean	0,6398	0,4188	0,4656	0,5467	0,6044	0,4314	0,3433	0,4231
	Manhattan	0,6398	0,4188	0,5237	0,5467	0,6044	0,4137	0,3653	0,4234
	Canberra	0,6397	0,4197	0,4877	0,5422	0,6035	0,4179	0,3402	0,4197
	Chebyshev	0,6397	0,4119	0,4912	0,5303	0,5975	0,4392	0,3570	0,4243
	Bray-Curtis	0,6398	0,3939	0,5222	0,5498	0,6035	0,4422	0,3626	0,4252

Fonte: Elaborada pelo autor.

para todas as combinações de medidas, contudo os resultados foram estatisticamente iguais (aproximadamente $0,6393 \pm 0.0005$). Entretanto, para as redes reais a diferença de modularidade foi mais significativa com a métrica de Bray-Curtis¹⁷⁴ e o método do *average linkage*¹⁷⁵ (destacados em negrito) que obtiveram sempre os melhores resultados. Por conseqüência, definimos este par de medidas para a etapa 5 do nosso algoritmo de detecção de comunidades.

A complexidade deste algoritmo não é tão simples de ser determinada devido aos vários processos utilizados. Além disso, o número de passos na dinâmica do SAW não é determinístico, mas no pior caso estimamos o cálculo de $\mathbf{F}$ em $\mathcal{O}(MN^2)$, onde M é o número de caminhantes por vértice ($M_i = M$, ou seja, o número de caminhantes que partem de cada nó é o mesmo) realizando o máximo de aproximadamente N passos cada, como apresentado nas figuras 13 e 14. Já o método do PCA é $\mathcal{O}(N^3)$ por causa do cálculo da matriz de covariância e sua decomposição em autovalores. Depois disso, a clusterização hierárquica e a modularidade são calculadas, cada uma, em $\mathcal{O}(N^2 \log N)$, para redes reais, N vezes (uma para cada componente principal). Portanto, a complexidade dessa técnica é $\mathcal{O}(MN^2 + N^3 \log N)$ mas de acordo com o gráfico da figura 18, M pode ser reescrito como αN com α uma constante positiva tal que $\alpha \ll N$, então finalmente encontramos $\mathcal{O}(N^3 \log N)$. Para atingir uma melhor performance, esse método pode ser parametrizado fixando o número de componentes principais e reduzindo a quantidade total de caminhantes na rede, porém a redução do custo computacional implica na queda da precisão do algoritmo.

Uma vantagem deste método com relação a outros algoritmos tradicionais de detecção de comunidades é a fácil generalização para redes ponderadas e com direção uma vez que toda informação topológica é mapeada por meio do SAW e expressa em $\tilde{\mathbf{F}}$, sendo trabalho do classificador apenas dividir os vértices em grupos de acordo com esta informação estrutural e a respectiva modularidade.

Assim, esse processo de detecção para redes ponderadas consiste basicamente em duas alterações. A primeira, se dá no cálculo da modularidade, onde agora utilizamos a definição que considera o peso das arestas, equação (2.21) (ao invés da expressão 2.20). Já a segunda e principal mudança está no sorteio do vértice j que será visitado por um caminhante. Nas redes simples, esta escolha é uniforme entre todos os vizinhos do nó i que o caminhante se encontra, ou seja, a probabilidade é

$$\mathcal{P}_j(t+1) = \frac{1}{k_i}, \quad (3.8)$$

contudo, quando a rede é ponderada consideramos que o caminhante tem uma chance maior de visitar o nó mais fortemente ligado ao vértice i , então a equação anterior se torna

$$\mathcal{P}_j(t+1) = \frac{w_{i,j}}{\sum_{j'} w_{i,j'}}, \quad (3.9)$$

com j' representando os vértices ligados a i que ainda não foram visitados. No limite de $w_{i,j'} = 1$ para qualquer j' , recuperamos a expressão (3.8). Devido a esta alteração, a complexidade numérica do cálculo da matriz $\mathbf{F}$ aumenta consideravelmente deixando o algoritmo como um todo mais lento. Além disso, o peso das arestas deve ser sempre um número real positivo para a equação 3.9 valer.

No caso direcionado, também precisamos fazer duas alterações similares à rede ponderada; trocamos a equação do cálculo da modularidade, que aqui obedece a expressão (2.22) e os caminhantes devem respeitar a direção das ligações, deixando a probabilidade de transição da seguinte forma,

$$\mathcal{P}_j(t+1) = \mathcal{P}_j^{out}(t+1) = \frac{1}{k_i^{out}}, \quad (3.10)$$

ou seja, os caminhantes podem apenas percorrer as ligações de saída do vértice atual. Devido a esta restrição, uma rede direcionada pode conter vértices que nunca serão atingidos durante uma caminhada ($k_i^{in} = 0$) ou vértices que não permitem nem o início da caminhada ($k_i^{out} = 0$), para evitar estes dois contratemplos e realizar a classificação adequada do grafo em comunidades, utilizamos em todas as redes apenas a maior componente fortemente conectada, garantindo que exista pelo menos um caminho entre um par de vértices, como ilustrado na figura 5(b).

A seguir mostraremos os resultados da aplicação dessa técnica para estes três tipos de redes. Em todos os casos, utilizamos $M_i = M = 10.000$, totalizando MN caminhantes

Figura 20: Fração de vértices classificados corretamente para os diferentes algoritmos de detecção de comunidades submetidos ao *benchmark* de Girvan-Newman. Cada ponto do gráfico é uma média de 100 grafos e as barras de erro correspondem ao desvio padrão. Os métodos Wt e SAW apresentaram os melhores resultados, perdendo precisão em torno de $z_{out} = 7,5$. Por outro lado o IMap, único algoritmo que não minimiza a modularidade, começou a perder qualidade na classificação para $z_{out} = 6$.

Fonte: Elaborada pelo autor.

por rede para garantir a estabilidade de $p_{i,j}$ e $\langle l_{i,j} \rangle$. Ademais, em todos os casos testados temos $N \ll M$.

3.2.2 Redes simples

Através de grafos gerados pelo *benchmark* de Girvan-Newman (128 vértices divididos em quatro comunidades com o mesmo número de elementos tal que $\langle k \rangle = 16$), comparamos a eficiência dos diferentes métodos de detecção de comunidade com o nosso processo de partição baseado em SAW. Nesta comparação, variamos o grau médio das ligações entre as partições, z_{out} , e observamos como a precisão (fração de vértices que foram classificados corretamente) diminui à medida que o número de ligações dentro de cada comunidade se torna igual ao número de conexões entre comunidades, ou seja, quando $z_{out} = z_{in} = 8$. O gráfico da figura 20 compara o resultado do novo método utilizando caminhada auto-excludente (SAW) com os algoritmos de Girvan-Newman (GN), *Fast Greedy* (FG), *Walktrap* (Wt) e *Infomap* (IMap). O FG, apesar de ser o mais rápido, foi o primeiro a perder precisão, em torno de $z_{out} \sim 4$, seguido pelo IMap que começa a cair em $z_{out} = 6$, e a partir deste ponto, a qualidade da classificação se torna a pior entre todos os algoritmos até $z_{out} = 8$ (situação na qual o grau interno (z_{in}) é igual ao grau externo (z_{out}) e, conseqüentemente, as comunidades não são bem definidas). O GN, por sua vez, começou a falhar de forma significativa para $z_{out} > 6$. Já o Wt e o SAW apresentaram uma melhor precisão na classificação do que os demais, uma vez que a fração de vértices classificados corretamente começou a decair apenas nas proximidades de $z_{out} = 7,5$.

Figura 21: Comparação de cinco algoritmos de detecção de comunidades para redes geradas por meio do *benchmark* LFR para diferentes valores de β e γ com $N = 500$ e $\langle k \rangle = 16$. O tamanho das comunidades variam entre 10 e 50 para cada uma das 100 realizações por μ com seus respectivos desvios padrões. O método do SAW, assim como no BGN, também produziu bons resultados para este conjunto de testes.

Fonte: Elaborada pelo autor.

Apesar do *benchmark* de Girvan-Newman ser muito utilizado para testes de algoritmos de detecção de comunidades, ele é limitado quanto à variação de parâmetros e não representa bem as redes reais com estruturas de comunidades. Com isso, para verificar a eficiência de um novo modelo, apenas a utilização deste conjunto de teste não é suficiente. Portanto, também verificamos o desempenho do nosso método com o *benchmark* LFR.

A figura 21 mostra o resultado da comparação entre os cinco diferentes métodos aplicados a estas redes artificiais com $N = 500$, $\langle k \rangle = 16$ e os tamanhos das comunidades variando entre 10 e 50 participantes. Para os expoentes que controlam a distribuição de tamanhos das comunidades e a distribuição de graus, utilizamos respectivamente, $1 \leq \beta \leq 2$ e $2 \leq \gamma \leq 3$, valores tipicamente encontrados em sistemas reais.^{176,177} Assim como no BGN, o algoritmo FG produziu os piores resultados, independente da combinação de parâmetros β e γ . Enquanto que na figura 20 esta técnica sempre mostrou precisão compatível com os demais para comunidades bem definidas (baixos valores de z_{out}), neste caso, a NMI sempre começa em valores consideravelmente baixos (0,8) quando comparado com os outros métodos, mostrando que o FG perde muita qualidade de classificação quando sujeito a redes mais heterogêneas. O IMap se mostrou o método mais sensível a variações do parâmetro β . Para $\beta = 1$, a NMI começa a decair para $\mu = 0,5$ e com uma grande barra

Figura 22: (a) Dendrograma resultante do nosso método de detecção de comunidade para rede do clube de karatê Zachary. As cores azul e vermelho correspondem à divisão da rede após o rompimento do clube, com exceção do vértice 2. (b) Estrutura de comunidade que apresentou o maior valor de modularidade ($Q = 0,4197$). As áreas coloridas são referência à estrutura de mesma cor mostrada no dendrograma.

Fonte: Elaborada pelo autor.

de erro vai a zero quando $\mu > 0,6$. Para $\beta = 2$ a queda de precisão ocorre bruscamente para $\mu = 0,6$ indo a zero logo em $\mu = 0,7$. O método GN mostrou bons resultados para este *benchmark*: em todos os casos ele foi o que menos perdeu precisão, começando em $\mu = 0,5$. O Wt e o SAW tiveram novamente comportamento similares com uma boa precisão em todas as situações (especialmente para $\gamma = 3$), tendo uma perda significativa na NMI a partir de $\mu = 0,6$.

Baseado nestas duas tradicionais referências para testes de algoritmos para detecção de comunidade, observamos que a classificação de comunidades por meio do SAW produziu bons resultados em redes artificiais, similar ou superando as outras quatro técnicas, dependendo das características da rede. Desta forma, por seguinte aplicamos nosso algoritmo em algumas redes reais.

O primeiro caso é a rede social do clube de karatê Zachary. O dendrograma da estrutura hierárquica encontrado pelo nosso método para esta rede é mostrado na figura 22(a). Identificamos facilmente dois grandes grupos (regiões em vermelho e azul) que, com exceção do vértice 2 (usualmente classificado de forma incorreta), representam o clube depois do desentendimento entre os dois principais membros. Apesar desta classificação se aproximar do resultado real, ela não representa a melhor divisão encontrada pelo nosso método. Já na figura 22(b) temos o particionamento do grafo em quatro grupos, que resultou na distribuição de comunidades que maximiza a modularidade, $Q = 0,4197$. Esta configuração está contida na divisão real, indicada na imagem por meio das áreas coloridas e ela é o valor mais alto já encontrado na literatura por meio de várias outras técnicas heurísticas.^{178–180} Para atingir esta configuração, nosso método utilizou três componentes principais do PCA.

Figura 23: Distribuição de comunidades para a rede dos Miseráveis. As cores representam a melhor divisão encontrada pelo nosso algoritmo, que alcançou $Q = 0,5467$ com 7 partições. As áreas limitadas pelas linhas trastejadas é a distribuição resultante do método GN e indicam 11 grupos com $Q = 0,5380$

Fonte: Elaborada pelo autor.

A próxima aplicação foi feita na rede de interação entre os personagens da trama *Les Misérables* e o resultado está presente na figura 23. As cores correspondem à estrutura de comunidade encontrada pelo método do SAW ($Q = 0,5467$) e as áreas pontilhadas são as subdivisões identificadas pelo algoritmo GN ($Q = 0,5380$). Apesar destas detecções apresentarem resultados similares para a modularidade, as partições são completamente diferentes. Enquanto nossa abordagem resultou em 7 comunidades com pelo menos seis membros, o GN projetou 11 divisões onde duas delas possuem apenas um vértice. Por meio deste exemplo, podemos perceber que estruturas de comunidades com valores muito próximos de modularidade podem ser bem diferentes.

A divisão em 7 grupos revela uma variedade de interações social ao longo da trama. A comunidade principal concentra os três personagens principais, o protagonista, Jean Valjean, o oficial de polícia Javert e a jovem Cosette, além de alguns personagens menores. Fantine, mãe de Cosette, é o centro do grupo composto pelas suas amigas e seu marido Tholomyes. Outro importante *subplot* é a partição que reflete os Amigos do ABC, associação de revolucionários liderada por Enjoras e o estudante de direito Marius. A família de Marius ficou separada em uma partição específica (cor rosa) assim como a família adotiva de Cosette, que compartilha a comunidade marrom com a quadrilha de bandidos *Patron Minette* e o marginal Brujon. O bispo Myriel, por sua vez, é o representante principal da comunidade laranja e a menor divisão (em violeta) têm um juiz e alguns bandidos. Para atingir esta distribuição de comunidades foram utilizadas sete componentes principais.

Observando os dois resultados presentes na figura 23, encontramos três diferenças principais. Primeira, comunidades com poucos membros foram unidas a uma maior, isto ocorreu nas comunidades vermelha (*Child1*, *Child2*, *MmeBurgon*, *Jondrette* e *MotherPlutarch*), marrom (*Boulatruelle*) e verde (*Fauchelevant*, *MotherInnocent* e *Gribier*). Segunda, no caso GN, Valjean, Javert e Cosette pertencem a grupos distintos diferentemente da nossa distribuição onde eles compartilham a mesma partição. Isto acontece pois estes vértices são *hubs* e caminhantes que partem de um deles rapidamente atingem os outros, portanto a perspectiva da rede é muito semelhante para estes três vértices, logo eles foram classificados na mesma comunidade. Finalmente, no terceiro caso, temos alguns personagens individuais que foram classificados de forma distinta. *Woman2* e *Toussaint* possuem conexões apenas com Javert, Valjean e Cosette, deste modo eles também estão no grupo verde. *Georges Pontmercy* (pai de Marius), por sua vez, apresenta interações com Marius, *MmePontmercy* e *Thernardier*, todos em grupos distintos, contudo nosso resultado o colocou juntamente com sua família enquanto o algoritmo GN o pôs com a família de Cosette. A última exceção é a irmã *Simplice*, ela possui quatro arestas com personagens em duas comunidades diferentes (azul: *Fantine* e *Perpetue*; verde: *Valjean* e *Javert*), o SAW a classificou com *Fantine* e o GN a colocou junto com *Valjean*. Nesta situação, as duas classificações fazem sentido, já que ela é próxima tanto de *Fantine* como de *Valjean*.

O último exemplo da aplicação do nosso método para redes simples é mostrado na figura 24, na rede da Liga Universitária de Futebol Americano. As cores representam a estrutura de comunidades determinada pelo método do SAW e as formas geométricas indicam a conferência de cada time. Conseguimos classificar perfeitamente todos os times filiados em cinco conferências (*Atlantic Coast*, *Big Ten*, *Pacific Ten*, *Southeastern* e *Big Twelve*). Os times da *Sun Belt's* foram divididos em duas comunidades, três membros ficaram na mesma comunidade da maioria das universidades da *Conference USA* e a outra parte (quatro times) se juntou com a *Mountain West*. Isto aconteceu devido à

Figura 24: Rede da liga universitária de futebol americano para a temporada do ano 2000. Os vértices são coloridos de acordo com a detecção de comunidades do SAW e e as formas geométricas representam as respectivas conferências. O mapa dos Estados Unidos mostra a influência da localização geográfica na nossa classificação (universidades geograficamente próximas disputam mais partidas entre si). Esta distribuição de comunidades atingiu $Q = 0,6044$ com 10 divisões.

Fonte: Elaborada pelo autor.

localização geográfica dos participantes, como mostrado no mapa dos Estados Unidos, onde times com a mesma cor estão geograficamente próximos apesar de pertencerem a conferências diferentes. Da mesma forma, classificamos corretamente oito membros da comunidade amarela, porém duas universidades desta mesma conferência foram colocadas em outra partição e, além disso, a *TCU Horned Frogs* (bola amarela) ficou neste grupo dominado pelas universidades texanas, mesmo ela pertencendo a *Conference USA*. Outro fato interessante aconteceu com os times independentes: como eles não pertencem a nenhuma conferência, o algoritmo tende a agrupá-los com a associação mais próxima. De fato, em 2002, o *UCF Knights* se juntou ao *Mid-American* (estrela) e em 2003 o *Utah State Aggies* se associou ao *Sun Belt*. Esta distribuição com 10 comunidades atingiu $Q = 0,6044$, por meio de onze componentes principais.

A tabela 4 mostra a performance da detecção de comunidades através do SAW em comparação com os outros métodos populares aplicados em algumas redes reais. Além

Tabela 4: Comparação da modularidade e do número de partições para os diferentes algoritmos de detecção de comunidades aplicados em algumas redes reais.

Rede	N	FG	GN	Wt	IMap	SAW
Karatê	34	0,3807 (3)	0,4013 (5)	0,3532 (5)	0,4020 (3)	0,4197 (4)
Golfinhos	62	0,4955 (4)	0,5194 (5)	0,4888 (4)	0,5247 (6)	0,5277 (5)
Miseráveis	77	0,5006 (5)	0,5380 (11)	0,5214 (8)	0,5461 (9)	0,5467 (7)
F. Americano	115	0,5497 (6)	0,5996 (10)	0,6029 (10)	0,6005 (12)	0,6044 (10)
Jazz	198	0,4389 (4)	0,4051 (39)	0,4384 (11)	0,2800 (7)	0,4428 (4)
<i>C. elegans</i>	297	0,3723 (5)	0,3018 (33)	0,3532 (22)	0,3858 (8)	0,3746 (5)
B. Política	1222	0,4256 (10)	0,4179 (205)	0,4253 (11)	0,4218 (40)	0,4268 (4)

Fonte: Elaborada pelo autor.

do Clube de karatê Zachary, Os Miseráveis e a Liga Universitária de Futebol Americano (F. Americano), também usamos a rede social dos Golfinhos nariz de garrafa, a rede dos músicos de jazz, a rede neural do verme *Caenorhabditis elegans* (*C. elegans*) e a rede de Blogs de Política (B. Política). Nesta última, assim como na rede *C. elegans*, ignoramos a direção dos 1222 vértices, ou seja, estamos considerando apenas o alinhamento político (liberal ou conservador) e ignorando quem cita quem.

Com exceção da *C. elegans*, em todos os outros casos nosso método apresentou o maior valor de modularidade, apesar de em algumas situações o Q de outros algoritmos terem sido muito próximos. Contudo, esta pequena diferença é suficiente para produzir resultados consideravelmente diferentes, por exemplo, na rede do jazz onde a modularidade para o FG e Wt difere do SAW apenas por $\Delta Q \sim 0,004$, onze grupos foram encontrados no Wt enquanto que o FG e o SAW dividiram o grafo em quatro partes. A mesma situação também acontece na rede dos Miseráveis com os métodos GN e SAW, como discutido na figura 23. Por outro lado, FG e SAW obtiveram modularidades muito similares ($Q = 0,3723$ e $Q = 0,3746$ respectivamente) na rede *C. elegans*, mas o IMap atingiu a melhor modularidade com oito partições. Outro fato interessante acontece para a rede B. Política, na qual temos a maior variação do número de comunidades entre os algoritmos testados. O SAW alcançou a maior modularidade ($Q = 0,4268$) com o número de partições mais próximo do ideal (já que neste sistema temos dois grupos bem definidos), diferentemente do GN, que com $Q = 0,4179$ identificou mais de duzentos grupos. Também é relevante notar como o algoritmo Wt, que atingiu bons resultados para redes artificiais, não obteve o mesmo sucesso para as redes reais.

Os resultados dessa tabela sugerem que o método aqui proposto é capaz de determinar uma boa distribuição de comunidade, sempre coerente com o sistema real estudado. Porém, vale ressaltar que o caminhante restrito pode ter um alto custo computacional. Então, esta nova abordagem dinâmica é indicada para determinar a melhor estrutura de comunidades para redes não muito grandes.

Figura 25: Resultados comparativos da qualidade da detecção de comunidades por meio dos métodos Wt, IMap, GN e SAW em redes artificiais geradas pelo modelo LFR com peso a medida que variamos os parâmetros β e γ . Cada ponto corresponde a média de 50 redes com $N = 500$, $\langle k \rangle = 16$, $\tau = 1, 5$, $0, 3 \leq \mu = \mu_w \leq 0, 8$ e o tamanho das comunidades variando entre 10 e 50. As barras de erros foram determinadas através do desvio padrão.

Fonte: Elaborada pelo autor.

3.2.3 Redes com peso

Assim como na subseção anterior, submetemos o algoritmo ponderado proposto a testes com redes artificiais e reais além de compararmos os resultados com os mesmos métodos de detecção de comunidades aplicado neste tipo específico de rede. Porém como mostrado na tabela 1, não foi possível utilizar a técnica do FG pois ela não é definida para redes ponderadas.

Desta forma, na figura 25 temos os resultados da aplicação do novo método de detecção baseado em SAW contraposto com os outros três algoritmos (Wt, IMap e GN em suas versões ponderadas) para redes com peso produzidas pelo *benchmark* LFR. Os parâmetros utilizados foram análogos ao caso simples, $N = 500$, $\langle k \rangle = 16$, $1 \leq \beta \leq 2$, $2 \leq \gamma \leq 3$ e o tamanho das comunidades variando entre 10 e 50. Esta versão possui dois parâmetros adicionais, τ : expoente que controla a força de cada nó i tal que $s_i = k_i^\tau$ e μ_w : utilizado para atribuir a força interna s_i^{in} do nó i (soma dos pesos de todos os vizinhos que pertencem a mesma comunidade). Aqui, fixamos $\tau = 1, 5$ para controlar a distribuição de pesos e, para simplificar, adotamos $\mu_w = \mu$ no intervalo $[0, 3; 0, 8]$, ou seja, à medida que

as comunidades vão ficando mais difusas a força interna de cada vértice diminui deixando o peso das arestas entre comunidades maior do que o peso intra comunidade (no caso de $\mu_w = 0,5$ a força s_i é igualmente dividida dentro e fora das partições). Para todas as combinações de parâmetros o algoritmo GN foi o que apresentou a menor perda de amplitude, tendo uma queda significativa quando $\mu = 0,6$ e se mantendo praticamente constante até $0,8$ onde o conceito de comunidade já se perdeu. Novamente os métodos Wt e SAW tiveram desempenhos similares, com uma queda contínua que se inicia a partir de $\mu = 0,5$. Entretanto apesar da proximidade entre estas duas curvas, o Wt sempre se mantém acima do SAW, indicando uma classificação ligeiramente superior. Já o IMap continua se mostrando inferior aos demais nos testes com redes artificiais, com uma precisão compatível com os outros até $\mu = 0,5$. Porém, a partir deste ponto, a qualidade da classificação cai drasticamente até $\mu = 0,7$. De forma geral, o comportamento dos quatro gráficos foram parecidos, indicando uma baixa sensibilidade com a variação dos parâmetros β e γ , além disso nosso método não se mostrou muito atrativo quando comparado ao GN e ao Wt, porém ele partiu da generalização do algoritmo sem peso, cabendo algumas otimizações como a troca da métrica e do método da hierarquização. Contudo, como veremos a seguir, quando ele é aplicado a redes reais sua utilização se torna justificável.

O primeiro exemplo de rede real ponderada que utilizamos foi a rede dos Miseráveis. No caso sem peso discutido na figura 23, não nos preocupamos com a quantidade de interações entre os personagens mas apenas se um contracenava com o outro. Já aqui, o peso da conexão entre dois atores representa a frequência com que eles contracenam, ou seja, acrescentamos mais esta informação que influencia diretamente o passeio do SAW e, conseqüentemente, pode mudar significativamente o resultado obtido anteriormente. A figura 26 mostra a distribuição de comunidades para esta rede, com cada cor indicando uma partição distinta. Diferentemente do caso sem peso, no qual obtivemos sete comunidades, aqui encontramos seis grupos, onde quatro deles se mantiveram inalterados (partições do bispo Myriel, da Fantine, da família adotiva de Cosette e do juiz com alguns marginais) e dois apresentaram mudanças significativas, que correspondem na imagem aos vértices representados por elipses. A primeira diferença ocorre com o estudante Marius: no caso anterior ele havia sido classificado junto com o grupo revolucionário Amigos do ABC, ao qual de fato fazia parte, entretanto suas fortes conexões com Jean Valjean e principalmente Cosette fizeram com que ele fosse classificado na comunidade mais central (verde) juntamente com estes dois protagonistas mesmo possuindo mais ligações com a comunidade vermelha. A outra distinção aconteceu na comunidade representada pela família de Marius que, assim como o estudante, também pertence à comunidade verde. Esta união acontece justamente pela ponderação das arestas: como Gillenormand e MlleGillenormand são fortemente conectados entre si e com Marius, a grande maioria dos caminhantes da antiga comunidade rosa (figura 23) atingem em poucos passos o *hub* Marius, que por sua vez, leva estes caminhantes até Valjean e Cosette, deixando a perspectiva da caminhada muito

Figura 26: Detecção de comunidades via SAW na rede do romance Os Miseráveis em sua versão ponderada, onde a força das arestas indicam a frequência com que dois atores contracenam. Diferentemente do caso sem peso, quando obtivemos sete grupos, aqui encontramos seis comunidades. As principais diferenças se deram nos vértices indicados pelas elipses.

Fonte: Elaborada pelo autor.

semelhante entre todos estes vértices, o que resulta na formação de uma grande partição. Apesar das diferenças, esta estrutura de comunidades também está coerente com a drama desenvolvida por Vitor Hugo.

A próxima rede real ponderada que analisamos foi a rede do romance Guerra dos Tronos, apresentada na figura 27. Através das 7 comunidades identificadas, podemos inferir a influência política e geográfica das quatro principais famílias (Stark, Lannister, Baratheon e Targaryen) de *Westeros* na luta pelo poder e conquista do trono de ferro. A maior comunidade (amarelo), representa a região da capital do reino, onde se passa grande parte da trama, dominada pela casa Lannister, tendo como destaque os irmãos desta família, Tyrion, Jaime e Cersei, por possuírem grande influencia nesta partição (muitas ligações com alto peso), já que são respectivamente conselheiro real, chefe da guarda e mãe do rei. Além destes, outros dois importantes personagens presentes neste grupo com fortes ligações são o atual rei, o jovem Joffrey Lannister, e sua esposa Sansa Stark. À direita, temos em azul a menor comunidade encontrada, que caracteriza a região de Pedra do Dragão pela presença do aspirante a rei Stannis Baratheon. Próximo a esta, observamos a comunidade em vermelho com a presença da exilada Daenerys Targaryen juntamente com seu exército e seu conselheiro Jorah, com quem tem maior afinidade. Assim como a comunidade azul, este conjunto se mostra bem definido, pois este *plot* acontece em um conjunto de terras livres afastadas do continente principal. No centro do grafo nosso algoritmo identificou duas comunidades, referentes aos acontecimentos desenvolvidos nas Terras Fluviais. Em laranja, identificamos a região dos castelos das Gêmeas e Correrrio, onde se encontra o rei do Norte (Robb Stark) juntamente com seus aliados e sua mãe Catelyn. Próximo dali, nas mediações do castelo de Harrenhal, está Arya Stark, principal membro da comunidade bordô, que possui ligações fortes com seu amigo Gendry e seu raptor Sandor Clegane. Mais ao norte, na partição marrom, estão os personagens que habitam a principal cidade da região, Winterfell, que está sob supervisão do paraplégio Bran Stark, que possui forte ligação com seu ajudante Hodor. Finalmente, na comunidade verde está presente a patrulha da noite, que controla a grande muralha de gelo, comandada por Jon Snow (bastardo da família Stark) e os membros dos povos livres juntamente com seu líder Mance Rayder, que é muito próximo de John Snow como mostra o peso da aresta entre eles. Para facilitar a localização geográfica destas sete comunidades, incluímos na figura 27 o mapa de onde se passa a trama com cada partição representada por um círculo de mesma cor. Um caso de destaque nesta rede foi a classificação de Eddard Stark: apesar dele possuir mais ligações com os membros da comunidade amarela (4), o antigo conselheiro do rei Robert foi classificado junto com sua filha, Arya, devido ao peso da aresta entre eles.

O último exemplo da detecção de comunidades por meio do SAW em redes ponderadas é mostrado na figura 28. Nesta, temos a rede de ações (apêndice B) onde cada vértice representa uma empresa de capital aberto e o peso das arestas indica a correlação

Figura 27: Resultado da detecção de comunidades obtida pelo método do SAW na rede ponderada da Guerra dos Tronos. Foram identificadas sete comunidades que representam a influência geopolítica das quatro principais famílias de Westeros (Stark, Lannister, Baratheon e Targaryen) pela disputa do poder. A esquerda temos o mapa onde se passa o romance com os círculos coloridos indicando a região geográfica de cada comunidade.

Fonte: Elaborada pelo autor.

Figura 28: Divisão da rede de ações em comunidades por meio do algoritmo do SAW ponderado, as cores indicam as partições e as formas geométricas o setor econômico de cada empresa, com o peso das arestas representando a correlação entre duas companhias.

Fonte: Elaborada pelo autor.

entre duas delas. As seis comunidades que identificamos são destacadas pelas cores e as formas geométricas representam grandes setores do mercado americano. De forma geral, esta classificação se dá de acordo com as características econômicas de cada companhia. Em nosso caso temos cinco categorias,

- Bens de Consumo: bens e serviços que são considerados não essenciais para o consumidor, incluindo por exemplo bens duráveis, entretenimento e lazer, vestuário, e automóveis.
- Mercado Financeiro: consiste em bancos, companhias de seguros, emissoras de cartão de crédito, financiadoras entre outras centradas em movimentações monetárias.
- Energia: composto por empresas que fornecem, perfuram, extraem e refinam as commodities primárias necessárias para manter um país, como petróleo e gás.
- Materiais: empresas que abastecem os outros setores com as matérias-primas, desde metais preciosos, papel e produtos químicos até recipientes de transporte e minério industrial.
- Indústria: engloba das ferrovias e das companhias aéreas às armas militares e aos conglomerados industriais.

Assim, vemos que as comunidades que foram encontradas apresentam forte relação com estes setores econômicos. A partição em azul se mostrou a mais bem definida do grafo com apenas três conexões fora do grupo e todos os integrantes pertencentes ao setor de energia, mostrando que esta categoria econômica é provavelmente menos sujeita às flutuações naturais de mercado dos outros setores, especialmente o de Bens de Consumo, com quem não compartilha nenhuma ligação. Esta, por sua vez, é uma categoria muito diversificada e, desta forma, foi representada pelas duas menores comunidades (laranja e bordô) que nem possuem arestas em comum. O respectivo conjunto com cinco membros representa empresas de entretenimento como canais de televisão e parques temáticos, já o menor possui companhias de turismo sendo dois hotéis e duas agências de viagens marítimas. Por outro lado, na divisão em vermelho observa-se dois setores distintos, sendo o da Indústria o predominante e o de Materiais o minoritário. Apesar de nosso algoritmo não ser capaz de resolver estas categorias econômicas, neste caso, mesmo com arestas de baixo peso, temos muitas conexões entre os vértices destes dois setores, mostrando que um exerce influência sobre o outro, tornando-se plausível a classificação na mesma comunidade. Finalmente, as partições em amarelo e verde também representam o mesmo setor econômico, o Financeiro. As ligações intra e entre estes grupos são relativamente fortes e densas, indicando que ambos naturalmente deveriam compor apenas uma comunidade. De fato, em um primeiro nível de análise de setores, o SAW falhou em dividir o maior setor em duas partições. Porém, apesar deste contratempo, conseguimos fracionar de forma

Tabela 5: Comparação da otimização da modularidade em três redes reais ponderadas submetidas aos métodos GN, Wt, IMap e SAW. Em todos os casos testados nosso algoritmo foi o que apresentou o maior valor de modularidade.

Rede	N	GN	Wt	IMap	SAW
Miseráveis	77	0,5016 (5)	0,5402 (9)	0,5571 (11)	0,5662 (6)
Guerra dos Tronos	103	0,4309 (4)	0,5989 (8)	0,5999 (7)	0,5999 (7)
Ações	121	0,6311 (7)	0,6265 (9)	0,6279 (8)	0,6355 (6)

Fonte: Elaborada pelo autor.

coerente uma rede construída a partir de correlações temporais. Isto pode ser aplicado, por exemplo, na construção de uma carteira de ativos baseada em renda variável, uma vez que o investidor procura compor seu investimento buscando ativos sem correlação, o que, segundo a teoria moderna dos portfólios, minimiza o risco de mercado.^{181,182}

A análise comparativa da modularidade entre os diferentes métodos de detecção de comunidades aqui testados aplicados nessas três redes reais estão sumarizados na tabela 5. Nas redes de Os Miseráveis e de A Guerra dos Tronos, obtivemos resultados semelhantes; com o algoritmo GN atingindo a menor modularidade, respectivamente, para 5 e 4 comunidades, seguido pelo Wt, que apresentou um aumento de Q em 0,04 para uma distribuição de 9 comunidades no primeiro caso e 0,168 com 8 partições no segundo. Já o IMap e o SAW alcançaram as maiores modularidades, com uma pequena diferença na rede de Os Miseráveis ($\Delta Q \sim 0,009$) a favor de nosso algoritmo, que por sua vez atingiu um número mais coerente de comunidades, 6, à medida que o IMap identificou 11. Já em A Guerra dos Tronos, o resultado foi idêntico, $Q = 0,5999$ com 7 grupos. Na rede de ações, diferentemente das anteriores, o Wt e o IMap foram os métodos com menor modularidade, respectivamente 0,6265 e 0,6279, na sequência, o GN ficou pouco atrás do SAW, que novamente foi a melhor abordagem. Assim, na contramão do resultado para redes artificiais discutido na figura 25, aqui temos o IMap com uma pequena melhora e, principalmente, o SAW sendo superior aos demais em todos os casos testados, se mostrando uma boa alternativa na otimização da modularidade em redes reais ponderadas.

3.2.4 Redes com direção

A próxima variação do algoritmo que propusemos para detecção de comunidades foi em redes direcionadas. Dentre as técnicas padrões de particionamento que utilizamos para comparação e validação do nosso método, as únicas que se aplicam a esta categoria de redes são o GN e o IMap. Portanto, a seguir mostramos os resultados para algumas variações de redes artificiais e duas redes reais submetidas a estes métodos.

A figura 29 mostra a evolução da informação mútua normalizada para quatro conjuntos de redes geradas pelo modelo LFR direcionado submetido à classificação em comunidades dos três métodos descritos anteriormente. Assim como nos casos ponderado e

Figura 29: Resultados da aplicação dos métodos GN, IMap e SAW em redes artificiais geradas pelo modelo LFR direcionado para diferentes valores de β e γ a medida que variamos μ em redes com $N = 500$ e $\langle k \rangle = 16$. Cada ponto corresponde a média de 100 redes, o respectivo desvio padrão foi tão baixo que não pôde ser observado nos gráficos.

Fonte: Elaborada pelo autor.

simples, os parâmetros selecionados foram: $N = 500$, $\langle k \rangle = 16$, $1 \leq \beta \leq 2$, $2 \leq \gamma \leq 3$, e o tamanho das comunidades variando entre o mínimo de 10 e o máximo de 50. Novamente, o comportamento das curvas é semelhante à medida que variamos os parâmetros β e γ , porém a diferença entre os três algoritmos é muito significativa. Diferentemente do resultado discutido na figura 25, o GN apresentou a pior precisão, com uma queda muito rápida na NMI logo para $\mu = 0,4$ que atinge o mínimo em seguida para $0,5$ e se mantém constante em $NMI \sim 0,2$ até o final do intervalo definido. O IMap, por outro lado, mostrou bons resultados até as proximidades de $\mu = 0,6$, onde a partir deste ponto a precisão vai a zero abruptamente. Já em nosso método, assim como no IMap, uma boa classificação se mantém até $\mu \sim 0,6$, mas daí em diante cai suavemente até $NMI \sim 0,2$. Portanto, neste modelo de rede artificial o SAW se mostrou superior aos demais em todo o intervalo estudado.

Agora, para exemplificar esta variação de nosso algoritmo em uma rede real, utilizamos a rede direcionada dos Blogs de política, onde a direção indica quem referencia quem na *Web*. Quando estudamos sua forma mais simplificada e ignoramos a direção, dividimos os 1222 vértices presentes na maior componente fracamente conectada em 4 comunidades, como mostrado na tabela 4. Porém, devido à restrição dos caminhantes, devemos reduzir o sistema para a maior componente fortemente conectada, assim o

Figura 30: Rede direcionada dos Blogs de política, onde a direção representa quem citou quem na *Web*. A divisão em duas comunidades está de acordo com o alinhamento político das páginas referenciadas, em vermelho temos a predominância de *sites* com ideologia de esquerda, enquanto que em verde, vemos os vértices de tendência mais conservadora, ficando evidente que blogs de mesmo alinhamento político costumam citar outros que concordam com seus ideais.

Fonte: Elaborada pelo autor.

Tabela 6: Comparação entre as modularidades encontradas por três diferentes métodos (GN, IMap e SAW) nas redes direcionadas dos Blogs de política e do verme *C. elegans*. Mais uma vez nosso algoritmo se mostrou uma boa alternativa a estas técnicas mais tradicionais.

Rede	N	GN	IMap	SAW
B. Política	792	0,0144 (697)	0,0217 (28)	0,4338 (2)
<i>C. elegans</i>	239	0,0488 (61)	0,2389 (16)	0,3725 (6)

Fonte: Elaborada pelo autor.

número de vértices que compõem esta rede cai para 792. O resultado da detecção de comunidades por meio do SAW é exibido na figura 30, onde as duas partições identificadas são representadas pelas cores vermelha e verde. A comunidade vermelha contém 342 vértices dos quais 332 são liberais (esquerda) e apenas 10 são conservadores (direita), por outro lado, no grupo verde selecionamos 451 blogs dos quais 432 são de direita e 19 de esquerda. Portanto, fica evidente a forte divisão entre o posicionamento políticos dos *sites*, ou seja, em geral páginas de mesmo alinhamento tendem a citar outras que concordam com seus ideais.

A tabela 6 compara os valores da modularidade para os três diferentes métodos nas redes dos Blogs de política e do verme *C. elegans* (ambas as redes compostas pela

componente fortemente conectada). Neste dois sistemas estudados, assim como nas redes do modelo LFR, figura 29, nosso algoritmo foi muito superior aos demais, atingindo $Q = 0,4338$ na primeira rede e $Q = 0,3725$ na segunda, enquanto que o IMap conseguiu apenas $Q = 0,0217$ e $Q = 0,2389$, mostrando que foi incapaz de identificar este tipo de estrutura na rede dos blogs. O GN, por sua vez, fracassou na identificação de grupos nestas redes direcionadas, pois nos dois casos o valor da modularidade foi muito baixo e o número de comunidades completamente descasado com as características individuais de cada sistema, na rede de política chegou a 697 e no verme a 61. Assim, o método de detecção de comunidades por meio do SAW demonstrou mais uma vez ser uma boa alternativa aos métodos tradicionais na busca por estruturas modulares.

3.2.5 Redes com metadados

Em geral, redes são frequentemente tratadas como um objeto puramente topológico baseado apenas nos padrões de interações. Porém, as fontes originais de onde são construídas estas redes contém mais informações do que apenas as conexões. Estas informações adicionais descrevem propriedades dos vértices como, por exemplo, idade dos indivíduos, gênero, opção sexual ou etnia em redes sociais, tipo de dieta alimentar ou massa corpórea das espécies em redes de cadeias alimentares, capacidade de dados ou localização física dos vértices em redes de Internet. Estes atributos que enriquecem o sistema são conhecidos como metadados e podem ser incluídos no grafo para melhorar as análises e o entendimento da estrutura da rede.^{183–187} Aqui, utilizaremos os metadados considerando que eles possuem alguma correlação com a estrutura de comunidades para assim melhorar a classificação dos vértices em grupos.

Na prática, essas características adicionais podem ser representadas como a ponderação das arestas e indicar a força de relacionamento entre os indivíduos da rede. Assim, inclui-se na estrutura informações exclusivas dos vértices que não são consideradas pelos algoritmos de detecção de comunidades. Com isso, a partir desta consideração e utilizando a vantagem do *random walk* em explorar os vértices da rede de maneira aleatória, porém não necessariamente uniforme, utilizamos nosso método baseado no SAW para identificar comunidades considerando os metadados.

O peso das arestas é distribuído da seguinte forma: seja ψ um metadado, se os vértices i e j possuem o mesmo ψ e $a_{i,j} \neq 0$, atribuímos a aresta $a_{i,j}$ um peso inteiro positivo maior que zero ζ , ou seja,

$$w_{i,j} = \begin{cases} a_{i,j} \times \zeta, & \text{se } \psi_i = \psi_j, \\ a_{i,j}, & \text{caso contrário.} \end{cases} \quad (3.11)$$

A escolha do valor de ζ fica a critério do usuário, dado que quanto maior o valor desta variável, maior é a ênfase empregada aos atributos durante a classificação e quando este

assume o valor unitário, obtemos a rede sem peso e sem influência dos metadados. Desta forma, o mapeamento da estrutura com metadados fica a cargo do SAW ponderado e a escolha da melhor distribuição de comunidades fica com a respectiva modularidade (equação 2.21) assim como aplicado nas redes da subseção 3.2.3. A única alteração significativa com relação os casos apresentados anteriormente é que agora fixamos o número de comunidade de acordo com a quantidade de ψ , já que assumimos que existe correlação entre as partições e os atributos utilizados. Esta alteração, apesar de simples, melhora o desempenho do algoritmo pois a modularidade é calculada apenas uma vez para cada componente do PCA e não mais para todos os níveis hierárquicos de cada componente principal, alterando a complexidade desta etapa de $\mathcal{O}(N^3 \log N)$ para $\mathcal{O}(N^3)$.

Para validar esta implementação em condições controladas, utilizamos o algoritmo LFR sem peso e sem direção como modelo base das redes artificiais, porém esta implementação não inclui metadados na construção das redes. Deste forma, elaboramos um algoritmo estocástico a fim de distribuir atributos correlacionados entre os vértices das redes geradas por esse *benchmark*. Inicialmente, fixamos a quantidade de metadados (N_ψ) igual ao número de comunidades (N_c), tal que se todos os membros de cada comunidade possuírem o mesmo atributo, a correlação entre comunidade e metadado é máxima. Assim, para um parâmetro controlado $p \in \mathfrak{R} \mid 0 \leq p \leq 1$ e uma rede artificial com $\mathcal{C} = (c_1, c_2, c_3, \dots, c_{N_c})$ comunidades, cada vértice i de uma dada comunidade $c \in \mathcal{C}$ tem uma probabilidade menor que p de receber um metadado c , ou $1 - p$ de receber um rótulo aleatório referente as outras partições, c' . Com isso, a partir de uma rede gerada pelo modelo LFR, esta lógica de atribuição pode ser descrita pelo seguinte algoritmo,

1. Define-se $N_\psi = N_c$ e p
2. Para cada comunidade $c \in \mathcal{C}$
3. Para cada vértice $i \in c$
4. Sorteia-se um número aleatório, $r \in \mathfrak{R} \mid 0 \leq r \leq 1$
5. Se $r < p$, $\psi_i \leftarrow c$, senão $\psi_i \leftarrow c'$

Desta forma conseguimos controlar de forma direta o nível de correlação entre metadado e comunidades por meio do parâmetro p , como exemplificado nos grafos da figura 31. Nos três casos, temos a mesma rede com $N = 40$ dividida em $N_c = 5$ partições com 8 vértices cada (representadas pelas cores) e $N_\psi = N_c$ metadados, diferenciados pelas formas geométricas. Em (a) fixamos $p = 0,4$, resultando em numa variedade de metadados em todas as comunidades. Já em (b), para $p = 0,6$, podemos identificar a dominância do principal atributo da respectiva comunidade, porém ainda observa-se uma certa heterogeneidade dos metadados dentro dos grupos. Finalmente em (c), onde temos $p = 0,8$, encontram-se

Figura 31: Exemplo da distribuição aleatória de metadados em uma rede com $N = 40$ dividida em 5 comunidades de mesmo tamanho para (a) $p = 0, 4$ (b) $p = 0, 6$ (c) $p = 0, 8$. As cores indicam as partições e as formas geométricas os respectivos atributos.

Fonte: Elaborada pelo autor.

duas comunidades (azul e vermelha) com todos os membros de cada compartilhando do mesmo metadado (respectivamente quadrado e triângulo para baixo), duas delas (amarela e verde) com dois vértices diferenciados cada e a comunidade laranja composta por três vértices com diferentes atributos.

Outra característica do parâmetro fixado p é sua relação linear com o coeficiente de correlação de Pearson, que pode ser escrito como,

$$\rho = \frac{\langle XY \rangle - \langle X \rangle \langle Y \rangle}{\sqrt{(\langle X^2 \rangle - \langle X \rangle^2) (\langle Y^2 \rangle - \langle Y \rangle^2)}}, \quad (3.12)$$

onde X e Y representam as distribuições de comunidade e de metadados. Esta relação linear é mostrada no gráfico da figura 32. Neste, vemos que à medida que aumentamos o valor de p , ρ cresce na mesma proporção, atingindo o máximo em 1, situação na qual todos os membros de cada comunidade c possuem o mesmo atributo $\psi = c$. Portanto, controlando p também definimos qual o valor da correlação de Pearson entre atributos e partições.

Uma vez estudado o comportamento do parâmetro p , precisamos encontrar um valor adequado para a ponderação das arestas entre vértices com mesmo metadado (ζ). Para determinar esta grandeza no modelo LFR, buscamos um inteiro que aumente consideravelmente a NMI sem perdas acentuadas na modularidade, ou seja, procuramos um equilíbrio entre a topologia original da rede e a influência da informação adicional introduzida no sistema. Para isso, fixamos uma rede artificial com $N = 500$, $\beta = 1$, $\gamma = 3$ e $\mu = 0, 7$. A escolha elevada deste último (parâmetro de mistura) se dá justamente para deixar a separação em grupos muito difusa e, com isso, determinar um ζ que de fato tenha uma contribuição relevante na estrutura da rede.

Os resultados da variação de ζ em função de Q e da informação mútua normalizada para valores crescentes de p são mostrados respectivamente nos gráficos 33(a) e 33(b). No

Figura 32: Gráfico mostrando a relação linear entre o parâmetro p e o coeficiente de correlação de Pearson. Para cada valor de p foram realizadas 100 amostragens. A rede base para a produção da distribuição de ψ foi gerada pelo modelo LFR com $N = 500$, $\beta = 1$, $\gamma = 3$ e $\mu = 0,5$.

Fonte: Elaborada pelo autor.

Figura 33: (a) Modularidade em função da ponderação artificial entre vértices com mesmo metadado (ζ) à medida que aumentamos a probabilidade p . (b) Gráfico representando a NMI para a classificação de comunidades à medida que variamos (ζ) para $p \in \{0,4; 0,6; 0,8\}$. O valor ideal que encontramos para esta grandeza foi $\zeta = 3$, pois mantém a modularidade elevada e aumenta de forma significativa a informação mútua normalizada. Apesar do baixo número amostral (5 redes por ponto), refletido nas elevadas barras de erros, as curvas não se sobrepõem, deixando o resultado satisfatório.

Fonte: Elaborada pelo autor.

Figura 34: Comparação da influência dos metadados em redes LFR a medida que aumentamos o parâmetro de mistura (μ) para rede sem metadados (SAW) e com correlação entre os atributos adicionais e as comunidades ($p \in \{0, 4; 0, 6; 0, 8\}$). Cada ponto corresponde a uma média de 100 redes com $N = 500$, $\beta = 1$ e $\gamma = 3$, onde as barras de erros indicam o respectivo desvio padrão.

Fonte: Elaborada pelo autor.

primeiro caso, para as três curvas podemos observar um pico entre $\zeta = 2$ e $\zeta = 3$ e, a partir deste ponto, inicia-se uma leve queda. Já para o segundo gráfico, temos um comportamento diferenciado, um rápido crescimento da NMI de $\zeta = 1$ para $\zeta = 2$, sendo mais acentuado conforme p aumenta e daí em diante as três curvas se mantêm praticamente constantes, mostrando que para $\zeta \geq 3$ a relação entre comunidade e metadados não apresenta melhora significativa. Assim, concluímos que para o modelo LFR, $\zeta = 3$ mantém a modularidade relativamente elevada e também melhora de forma considerável a detecção de comunidades, refletida pelo aumento do valor da NMI. Em ambos os gráficos é notável o grande desvio padrão proveniente do resultado de 5 redes artificiais por ζ , porém mesmo com este baixo número amostral, uma curva não sobrepõe a outra, deixando, portanto, o resultado satisfatório.

Com o peso artificial da aresta que conecta vértices com mesmo metadado fixado em 3, construímos o gráfico da figura 34 variando o parâmetro de mistura no intervalo compreendido entre $0, 3 \leq \mu \leq 0, 8$ e analisamos a informação mútua normalizada para os casos sem metadado (SAW) e para $p \in \{0, 4; 0, 6; 0, 8\}$ em redes com $N = 500$, $\beta = 1$ e $\gamma = 3$, de forma similar ao estudado na figura 21. Nesse, podemos observar que para $\mu \leq 0, 5$ a NMI é muito alta independente dos atributos adicionados no sistema. Porém, deste ponto em diante (onde estamos além da definição tradicional de comunidades), o efeito dos metadados na estrutura do sistema começa a se destacar. Enquanto que no caso sem metadados (preto) a curva cai rapidamente até atingir $NMI \sim 0, 1$, à medida que

Tabela 7: Comparação do resultado da detecção de comunidades por meio do SAW entre quatro redes reais com e sem a presença de metadados. Como era de se esperar, observamos uma queda na modularidade devido ao número fixo de comunidades (número de metadados distintos) e um aumento significativo da NMI, revelando uma boa identificação com sistema real. Aqui, assim como no modelo LFR, utilizamos $\zeta = 3$.

Redes	SAW			
	Q		NMI	
	Sem Metadado	Com Metadado	Sem Metadado	Com Metadado
Karatê	0,4197 (4)	0,3599 (2)	0,5856	1,0000
Golfinhos	0,5277 (5)	0,4791 (3)	0,2243	0,2122
F. Americano	0,6044 (10)	0,5983 (12)	0,8923	0,9361
B. Política	0,4268 (4)	0,4233 (2)	0,6682	0,7534

Fonte: Elaborada pelo autor.

aumentamos o valor de p , ou seja, introduzimos mais correlação entre atributos e partições, o valor da NMI se mantém mais elevado, mostrando que a informação externa adicional resultou em um aumento considerável no número de vértices classificados corretamente quando temos grupos muito difusos.

Para exemplificar nosso método de detecção de comunidades com o auxílio de metadado em redes reais, submetemos quatro redes sem peso, sem direção e com atributos bem conhecidos a esta técnica. Os resultados comparativos da modularidade e da NMI com e sem metadados foram concentrados na tabela 7. Na rede do clube de Karatê associamos a cada vértice a sua respectiva academia (duas), para a rede dos Golfinhos atribuímos aos participantes seu gênero (macho, fêmea ou desconhecido), já na liga Universitária de Futebol Americano temos as conferencias de cada time como atributo e finalmente na rede dos Blogs de Política consideramos como metadado o direcionamento político (direita ou esquerda). Assim como no *benchmark* LFR, utilizamos $\zeta = 3$ e, exceto pela rede dos Golfinhos, ocorreu uma queda na modularidade devido ao número fixo de comunidades (a quantidade de comunidades é igual ao número de metadados distintos) e um aumento significativo da informação mútua normalizada em todos os casos testados. A rede dos Golfinhos, por sua vez, assim como as demais apresentou uma queda na modularidade, porém a presença dos três tipos de metadados diminuiu o valor da NMI em 0,0121, isto se justifica pois aqui temos quatro vértices de sexo desconhecido em pontos isolados do sistema, o que não caracteriza uma comunidade, então o número fixo de partições em três foi inadequado para buscar uma correlação entre os dois gêneros e os grupos estruturais.

Apesar destas poucas redes reais serem simples e apresentarem alta correlação entre metadados e comunidades, nosso objetivo aqui foi mostrar a extensão de nosso algoritmo de classificação de comunidades em possíveis situações onde a densidade de arestas é elevada, mas existem na rede características adicionais dos vértices que podem ser utilizadas para melhorar uma divisão em grupos. Devido à flexibilidade do SAW em

explorar redes com diferentes características, com poucas alterações no algoritmo, será possível generaliza-lo para redes ponderadas, digrafos e até mesmo sistemas que apresentam diversos metadados por nós, tornando a análise estrutural de redes complexas por meio da caminhada auto-excludente muito versátil e de grande auxílio para futuras pesquisas nessa área.

4 CONCLUSÕES E PERSPECTIVAS

A grande eficácia das redes complexas em representar os mais diversos efeitos e fenômenos presentes na natureza originou muitos trabalhos interdisciplinares dentro das últimas décadas. Um fenômeno recorrente em sistemas do mundo real é a formação de grupos onde seus membros compartilham alguma ou várias características em comum. Devido à importância destes conjuntos, o estudo de comunidades se tornou um dos principais temas explorados em redes complexas. Porém, a falta de uma definição formal e a dificuldade computacional para identificar tais partições ainda não gerou uma solução satisfatória para o problema, deixando uma série de importantes questões ainda em aberto. Entretanto, os resultados encontrados por métodos heurísticos de detecção de comunidades tem mostrado alta compatibilidade e coerência com as características predominantes de partições em sistemas variados (sociais, biológicos, econômicos e literários).

Trabalhos recentes mostram que a caminhada aleatória auto-excludente, dinâmica que impede o caminhante de visitar vértices que já foram alcançados, é muito eficiente na busca e exploração de redes, sugerindo que ela poderia ser um bom mecanismo para encontrar estruturas modulares. Assim, neste trabalho estudamos propriedades destas estruturas modulares baseados nesta caminhada aleatória restritiva. Mais precisamente, buscamos explorar as vantagens da dinâmica do SAW a fim de aprimorar as análises qualitativa e quantitativa de sistemas reais que apresentam ou possam apresentar propriedades modulares. Para isso, definimos uma medida que concentra algumas informações topológicas do sistema por meio da dinâmica aleatória auto-excludente e, a partir desta, elaboramos um método de detecção de comunidade que otimiza a modularidade e classifica os vértices utilizando a técnica da clusterização hierárquica aglomerativa. Devido a flexibilidade deste tipo de passeio, generalizamos nosso algoritmo para identificar partições em redes ponderadas, digrafos e sistemas que apresentam metadados como informação adicional dos vértices.

Nossas primeiras análises tiveram como objetivo caracterizar a caminhada auto-excludente em redes sintéticas com o intuito de compará-la com o passeio sem restrição e para obter informações relevantes tanto da estrutura como da própria dinâmica estocástica quando submetida a modelos controlados. Começamos comparando a probabilidade de encontrar um caminhante em um vértice qualquer da rede após um grande número de passos para os passeios restritivo e tradicional, no qual conhecemos no resultado analítico nestas condições, em redes tipo ER e BA com diferentes valores de grau médio. No caso Markoviano, sabemos que o processo dispersivo para tempos longos não depende da topologia da rede, apenas do grau de cada vértice. Contudo, quando introduzimos memória infinita na dinâmica, o resultado se alterou consideravelmente. Tanto para rede

ER como para rede BA, a probabilidade de encontrar um caminhante em vértices de grau elevado é inferior ao passeio tradicional, principalmente na topologia livre de escala, onde a presença de *hubs* limitou drasticamente o desenvolvimento da caminhada, uma vez que muitos vértices deixam de ser explorados quando um *hub* é visitado. Por outro lado, a lembrança do trajeto percorrido aumenta a probabilidade de encontrarmos os caminhantes em vértices de baixo grau, sendo que, no caso ER o $P(k)$ do SAW se mantém superior até $k \sim \langle k \rangle$ e no caso livre de escala a inversão da probabilidade acontece para $k \gg \langle k \rangle$.

Outra medida do SAW que quantificamos foi o número de passos desenvolvidos pelos caminhantes em função do grau médio. Para este caso, além dos modelos ER e BA também utilizamos redes geradas pelo *benchmark* LFR. Como era de se esperar, verificamos que quanto maior o $\langle k \rangle$ e a quantidade de vértices da rede, maior é o número médio de passos desenvolvido pelos caminhantes já que existem mais rotas disponíveis para a evolução da caminhada. Porém, o deslocamento das distribuições de passos nas redes BA foi mais lento, ressaltando a dificuldade da evolução da dinâmica devido a presença de *hubs*. Já para os modelos ER e LFR os resultados foram parecidos, com o segundo caso sendo um pouco mais explorado pelos caminhantes, principalmente para valores menores de grau médio. Complementarmente a esta análise, a partir das distribuições de comprimento dos *walkers*, calculamos os números médio e máximo de passos em função do tamanho da rede. Nos casos ER e LFR estas duas medidas apresentaram um comportamento aproximadamente linear para qualquer valor de $\langle k \rangle$, mas para as redes produzidas pelo *benchmark*, mais vértices foram alcançados no passeio, mostrando que a presença de comunidades favoreceu a evolução da dinâmica. Por outro lado, no caso BA a fração de visitas foi uma ordem de grandeza inferior aos demais, além dela cair consideravelmente à medida que aumentamos o tamanho da rede. Nos três modelos testados ocorreu uma grande diferença de magnitude quando aumentamos o grau médio de 10 para 20, o que não ocorre quando saímos de 20 e vamos para 30.

Após observarmos o efeito da memória infinita e o comprimento dos caminhantes no processo difusivo, estudamos a caminhada auto-excludente com maiores detalhes definindo duas grandezas: $p_{i,j}$, fração de caminhantes que saíram do vértice i e passaram pelo vértice j antes do final do percurso, e $\langle l_{i,j} \rangle$, o respectivo número médio de passos. A primeira medida quantifica a facilidade com que os caminhantes atingem j partindo de i , independente das distâncias topológicas. Complementarmente, a segunda definição revela as distâncias topológicas do sistema através do número médio de passos. Assim, de forma simplificada, classificamos pares de vértices em quatro situações (TIPO A, TIPO B, TIPO C e TIPO D) de acordo com os valores destas grandezas e, posteriormente, aplicamos esta definição para identificar padrões estruturais em diferentes modelos de redes complexas por meio de um gráfico de espalhamento. Para o caso BA, o efeito *ultra small world* não permitiu o desenvolvimento do passeio, deixando qualquer par i, j com baixos valores de comprimento médio, porém poucos vértices são facilmente alcançados por muitos caminhantes (*hubs*)

enquanto a maioria dificilmente é visitado. Com isso, $p_{i,j}$ apresentou valores em toda sua extensão, nos levando a concluir que neste modelo de rede livre de escala dois padrões se destacaram, os TIPO B e TIPO D. Já no caso ER, a probabilidade de transição ficou menos espalhada, entretanto, os caminhantes evoluíram mais a caminhada pois os valores de $\langle l_{i,j} \rangle$ foram superiores quando comparados ao modelo anterior. Deste modo, a maior parte dos vértices ficaram próximas às classes TIPO A e TIPO C. Comparamos, por fim, o *benchmark* LFR, onde a presença de comunidades fez com que os caminhantes explorassem mais a rede, aumentando o comprimento médio e também a probabilidade de transição. Aqui destacaram-se os padrões TIPO C e TIPO D, com o primeiro mais comum entre vértices de comunidades distintas e o segundo, no caso de membros de uma mesma partição. Portanto, mostramos que a combinação de $p_{i,j}$ e $\langle l_{i,j} \rangle$ permite distinguir e até mesmo classificar modelos de redes complexas por meio das propriedades dinâmica intrínsecas do SAW.

De forma a concentrar as informações estruturais mapeadas pelo SAW em uma única medida, definimos $f_{i,j}$ como a razão entre $p_{i,j}$ e $\langle l_{i,j} \rangle$. Multiplicamos o inverso do comprimento médio pela probabilidade de transição para refletir de forma mais adequada a proximidade de dois vértices, com $f_{i,j}$ atingindo valores mais altos conforme $p_{i,j}$ cresce e $\langle l_{i,j} \rangle$ diminui. Por meio da norma de Frobenius verificamos a estabilidade desta grandeza, quantidade mínima necessária de caminhantes para estabilizar variações estatísticas, e mostramos a possibilidade de utilizar esta norma como um parâmetro de tolerância para convergência de $f_{i,j}$ em alternativa ao número de caminhantes por vértices. Além disso, concluímos que a dependência entre o número de caminhantes necessário para a convergência e o tamanho da rede é baixa, ou seja, à medida que N aumenta poucos caminhantes precisam ser incluídos na dinâmica para obter uma boa estabilidade em $f_{i,j}$.

Devido à presença de altos valores em $f_{i,j}$ entre membros de subgrafos com alta densidade de conexões, usamos esta grandeza como parâmetro de uma medida de similaridade para desenvolver um novo método que utiliza a clusterização hierárquica aglomerativa para a detecção de comunidades. Interpretamos $f_{i,j}$ como um conjunto de N amostras por N medidas representado pela matriz $\mathbf{F}$. Feito isso, realizamos o processo do PCA para remover a redundância dos dados e reduzir a dimensionalidade do sistema para, posteriormente, utilizar a clusterização hierárquica aglomerativa com a dissimilaridade de *Bray-Curtis* como métrica de proximidade entre as amostras e o *average linkage* como o método de união dos grupos. Ao término da última etapa, calculamos a modularidade em cada nível do dendrograma para identificar a divisão que apresentou melhor qualidade. O processo de hierarquização é repetido até todas as componentes principais serem usadas. Assim, o resultado final deste processo gera o dendrograma que possui distribuição de comunidades que maximiza a modularidade para um determinado número de componentes principais do PCA. A complexidade deste algoritmo não é simples de ser estimada, uma vez que o número exato de passos do SAW não é bem determinado, apesar disto, estimamos

a complexidade em $\mathcal{O}(N^3 \log N)$. Para atingir uma performance superior pode-se fixar o número de componentes principais e reduzir a quantidade de caminhantes na rede, porém isto implica na queda de precisão da classificação dos grupos.

Uma vantagem deste método com relação aos demais é a fácil generalização para redes ponderadas e digrafos, já que toda informação topológica é mapeada pelo SAW, sendo função do classificador apenas identificar os grupos dada a informação estrutural e a respectiva modularidade. Para as redes ponderadas foi necessário realizar duas alterações: a primeira se dá no cálculo da modularidade, onde agora utilizamos sua definição que considera o peso das arestas, e a segunda, está no sorteio aleatório da escolha do próximo vértice que será visitado que, neste caso, não é mais uniforme, mas ponderada pelos pesos das arestas dos vizinhos acessíveis. Dada estas alterações, a complexidade numérica aumentou, deixando o método mais lento, além disso, a ponderação deve ser um número real positivo para que o novo sorteio possa ser utilizado. Nos digrafos também foi necessário realizar duas modificações: substituímos a equação de cálculo da modularidade para a forma mais adequada e, durante o passeio, os *walkers* devem respeitar a direção das ligações, ou seja, eles podem apenas percorrer as arestas de saída do vértice atual. Devido a esta última restrição, apenas digrafos com a maior componente fortemente conectada podem ser submetidos a este processo de detecção, pois assim evitamos situações onde vértices nunca serão atingidos por um caminhante ($k_i^{in} = 0$) e condições iniciais que impedem o início da dinâmica ($k_i^{out} = 0$).

Para realizar a validação do nosso algoritmo em redes simples, o submetemos aos *benchmarks* BGN e LFR juntamente com outros quatro tradicionais métodos de detecção de comunidades, Girvan–Newman, *Fast Greedy*, *Walktrap* e *Infomap*. Para a rede artificial BGN, onde os parâmetros são fixos com 128 vértices divididos em quatro comunidades com o mesmo número de elementos tal que $\langle k \rangle = 16$, os métodos baseados em caminhadas aleatórias (SAW e Wt) foram os que apresentaram melhores resultados, seguido respectivamente pelos algoritmos de GN, FG e, finalmente, Imap. Já para o *benchmark* LFR, geramos quatro configurações de redes com estruturas de comunidades variando de acordo com os parâmetros $1 \leq \beta \leq 2$ e $2 \leq \gamma \leq 3$ para $N = 500$ e $\langle k \rangle = 16$. Neste caso, o algoritmo FG produziu os piores resultados independentemente da combinação de β e γ , o IMap se mostrou o método mais sensível a variações do parâmetro β , o GN, por sua vez, foi o que menos perdeu precisão para todos os casos testados e, novamente, o Wt e o SAW tiveram comportamento semelhante, com uma boa precisão independente da combinação de parâmetros. Com isso, concluímos que a classificação de comunidades por meio do SAW produziu bons resultados em ambos os tipos de redes artificiais, sendo até superior as outras técnicas dependendo da combinação de parâmetros.

A primeira rede do mundo real que analisamos foi a do clube de karatê Zachary, onde identificamos dois grandes grupos que, com exceção de um vértice, representa o clube

depois do desentendimento entre os dois principais membros. Entretanto a divisão que maximizou a modularidade ($Q = 0,4197$) possui quatro grupos e está contida na divisão real. A seguir, aplicamos nosso algoritmo na rede de interação entre os personagens da trama *Les Misérables* e comparamos o resultado com a distribuição encontrada pelo método GN. Apesar de ambas as técnicas terem encontrado valores próximos de modularidade ($Q_{SAW} = 0,5467$ e $Q_{GN} = 0,5380$) o resultado foi bem distinto, indicando que variações da modularidade pode levar a resultados e interpretações diferentes. Nossa abordagem identificou sete grupos que estão relacionados com as variedades de núcleos e interações sociais presentes na trama, com destaques para a comunidade com os três personagens principais (Valjean, Javert e Cosette), a partição com Fantine e suas amigas e outra com os estudantes revolucionários liderados por Enjoras e Marius. O último exemplo da aplicação de nosso método em redes simples foi na rede da Liga Universitária de Futebol Americano, onde identificamos dez comunidades contendo as doze conferências da liga. Dentre estas doze, cinco foram perfeitamente classificadas e as demais ficaram distribuídas nos outros cinco grupos devido a influência da localização geográfica dos times. Também notamos que os times independentes foram classificados naturalmente com a associação mais próxima e, na temporada seguinte, dois times acabaram sendo incorporados à respectiva liga. No mais, comparamos a modularidade encontrada por todos os métodos em mais quatro redes (Golfinhos, Jazz, *C. elegans* e Blog de Política) e, com exceção do *C. elegans*, a técnica que desenvolvemos apresentou a maior modularidade com distribuições de comunidades coerente com o sistema real. Porém, devido a necessidade de utilizarmos muitos caminhantes para estabilizar $f_{i,j}$, o SAW possui alto custo computacional sendo indicado para redes não muito grandes.

Uma vez que testamos e validamos a classificação via SAW, aplicamos o algoritmo ponderado proposto a testes com redes artificiais e reais. Porém não foi possível utilizar a técnica FG, pois ela não é definida para este tipo de rede. Por meio do *benchmark* LFR geramos redes com peso para a mesma combinação de parâmetros do caso anterior e fixamos $\tau = 1,5$ para controlar a distribuição de pesos e, para simplificar, adotamos $\mu_w = \mu$, ou seja, a medida que as comunidades vão ficando mais difusas, a força interna de cada vértice diminui deixando o peso das arestas entre comunidades maior do que o peso intra comunidades. Aqui, observamos uma baixa sensibilidade dos resultados com a variação dos parâmetros β e γ , com o GN apresentando os melhores resultados, o Wt e o SAW novamente com desempenho similar e o IMap também continua com a menor precisão na classificação. Nosso método não se mostrou muito atrativo quando comparado ao GN e ao Wt (mais rápido), contudo quando aplicado a redes reais sua utilização se tornou justificável.

Para este caso, o primeiro sistema real ponderado que utilizamos foi a rede do romance *Les Misérables*, onde o peso indica a frequência com que os personagens contracenam. Nesta foram identificadas seis comunidades, das quais quatro permaneceram inalteradas

com relação ao caso sem peso, mas em duas observamos mudanças: Marius foi classificado na comunidade principal junto com Valjean, Javert e Cosette e, conseqüentemente, a partição que representa a família deste estudante também se juntou no grupo principal devido às fortes interações com Marius. A rede seguinte foi a do romance Guerra dos Tronos, onde encontramos sete comunidades referentes às relações políticas e geográficas entre os personagens das principais famílias. A seguir, construímos uma rede de ações com o peso indicando a correlação entre duas empresas de capital aberto. Aqui, identificamos seis comunidades que estão relacionadas com cinco grandes setores do mercado americano. Uma destas partições possui dois setores distintos (Indústria e Materiais) que exercem muita influência um sobre o outro, além disso, as empresas relativas ao Mercado Financeiro foram divididas em duas partições. Apesar destes detalhes, fomos capazes de fracionar de forma coerente uma rede construída a partir de correlações temporais. Por fim, aplicamos as outras técnicas de particionamento nessas três redes do mundo reais e, mais uma vez, o SAW foi superior aos demais nas redes de ações ($Q = 0,6355$) e dos Miseráveis ($Q = 0,5662$) e obteve o mesmo resultado que o IMap para a rede da Guerra dos Tronos ($Q = 0,5999$), se mostrando uma boa alternativa na classificação de vértices em comunidades para redes reais ponderadas.

A próxima variação do algoritmo que testamos foi a versão para digrafos. Os únicos métodos tradicionais que possuem uma implementação para este tipo de rede são o GN e o IMap, então, primeiramente, comparamos o desempenho destas três técnicas por meio do *benchmark* LFR para redes direcionadas utilizando a mesma configuração definida para redes simples. Assim como no caso ponderado, os resultados foram semelhantes conforme variamos os parâmetros β e γ , no entanto, o algoritmo GN apresentou uma perda na precisão muito acentuada, não sendo capaz de resolver as comunidades pré definidas, o IMap e o SAW já apresentaram bons resultados, porém nosso algoritmo se manteve com a NMI elevada para valores maiores de μ . Para exemplificar a aplicação desta variação do nosso método em redes reais, utilizamos a rede direcionada dos Blogs de política, onde a direção indica quem referencia quem na *Web*. Conseguimos distinguir com clareza a presença de duas comunidades que reflete o posicionamento político dos *sites*, já os outros dois algoritmos divergiram muito do caso real, com o GN encontrando 697 grupos e o IMap 28. Além desta rede social, também usamos a rede do verme *C. elegans*, e mais uma vez o SAW relevou ser o mais adequado para a detecção de comunidades no caso direcionado, atingindo 6 comunidades ($Q = 0,3725$) enquanto que o IMap revelou 16 ($Q = 0,2389$) e o GN 61 ($Q = 0,0488$).

Na última parte desta tese incluímos informações adicionais relativas aos vértices (metadados) para aprimorar o entendimento da estrutura da rede, uma vez que os metadados apresentam correlações com a distribuição de comunidades. Na prática, utilizamos esta informação adicional como uma ponderação das arestas, ou seja, incluímos na estrutura topológica da rede informações exclusivas dos vértice que, em geral, não

são consideradas pelos algoritmos de detecção de comunidades. O valor da ponderação (ζ) é parametrizável e quanto mais elevado, maior é a ênfase empregada aos atributos durante a classificação. Assim, basta aplicar a detecção de comunidades utilizando o SAW em sua versão ponderada para identificar as partições. Contudo, como temos uma correlação entre metadado e comunidades, o número de grupos é fixado de acordo com a quantidade de atributos. Esta simples alteração no algoritmo melhora seu desempenho pois agora a modularidade é calculada apenas uma vez por componente do PCA, levando a complexidade de $\mathcal{O}(N^3 \log N)$ para $\mathcal{O}(N^3)$.

Para validar estas alterações em ambiente controlado, utilizamos o método LFR sem peso e sem direção, porém este *benchmark* não inclui metadados durante a construção das redes. Com isso, construímos um algoritmo para distribuir atributos aos vértices de forma a correlaciona-los com a estrutura de comunidades. Este algoritmo consiste basicamente em atribuir rótulos aleatoriamente de acordo com o método de Monte Carlo, onde cada vértice tem uma probabilidade fixada p de possuir um metadado correlacionado com seu respectivo grupo. Observamos que este parâmetro pré-definido têm uma relação linear com o coeficiente de correlação de Pearson. Assim, usamos esta abordagem para encontrar o valor de ζ , tal que não ocorra uma perda acentuada da modularidade e aumente consideravelmente o valor da NMI. Diferentemente dos testes realizados anteriormente, aqui definimos $\mu = 0,7$ para deixar a separação entre as partições mais difusa e, com isso, determinar um ζ que realmente seja impactante na estrutura da rede. O valor mais adequado desta grandeza entre as redes geradas foi de $\zeta = 3$, ou seja, ele manteve a modularidade relativamente elevada e melhorou a detecção de comunidades dado que a NMI aumentou. Uma vez definida esta variável, comparamos a detecção de comunidades por meio do SAW em redes LFR para valores crescentes de p e verificamos que para $\mu \leq 0,5$ os valores de NMI são elevados em todos os casos, porém para $\mu > 0,5$ as maiores precisões ocorrem para os maiores valores de p , mostrando que a informação extra resultou em um aumento considerável no número de vértices classificados corretamente quando temos grupos mais difusos.

Finalmente, para exemplificar a aplicação desta variação de nosso algoritmo em redes reais, utilizamos quatro redes simples com atributos bem conhecidos. Na rede do clube de Karatê, associamos a cada vértice a sua respectiva academia, para a rede de Futebol Americano Universitário, consideramos as doze associações como os metadados, já na rede dos Blogs de Política atribuímos o direcionamento político aos participantes e na rede dos Golfinhos consideramos o gênero de cada animal à característica externa. Assim como nas redes LFR, fixamos $\zeta = 3$, e, com exceção desta última rede, observamos uma queda representativa da modularidade e um aumento significativo da NMI. A rede dos Golfinhos também apresentou queda em Q , mas a presença de três animais com gênero desconhecido em pontos isolados do sistema prejudicou o classificados, pois três não é o número adequado de comunidades para este caso. Apesar do baixo número e da

simplicidade das redes do mundo real que utilizamos, nosso objetivo foi mostrar a extensão de nosso algoritmo em situações onde a identificação das partições é complicada devido à alta densidade de arestas, mas o sistema apresenta informações extras que podem ser incluídas na estrutura para melhorar a classificação dos vértices.

Em razão da grande flexibilidade do SAW e aos resultados aqui encontrados, a análise estrutural de redes complexas por meio desta dinâmica ainda pode ser muito explorada, como por exemplo caracteriza-la em modelos além dos tradicionais BA e ER, já que observamos uma forte dependência topológica na evolução do passeio. Além disso, trabalhos futuros podem analisar suas propriedades em sistemas artificiais ponderados e direcionados para também compara-la com a dinâmica Markoviana do caminhante sem restrição. Outra possibilidade de estudo estrutural se dá a partir de $\mathbf{F}$ para além da área de comunidades uma vez que esta matriz concentra muita informação proveniente dos padrões de ligações do sistema.

Para a análise de comunidades, um interessante tópico para um estudo mais aprofundado é buscar compreender por que nosso algoritmo produz resultados consistentemente melhores em redes reais quando comparado às outras técnicas e, a partir disso, desenvolver ou propor melhorias na construção de modelos artificiais de redes com comunidades. Apesar deste bom desempenho, a possibilidade de melhorar a precisão da detecção por meio de outras medidas de similaridade ou técnicas de classificação continua em aberto, principalmente na variação para o caso de redes ponderadas, onde não obtivemos resultados muito satisfatórios no *benchmark*. Como esse método possui um alto custo computacional, também recomendamos o estudo de formas para otimizar sua performance e tornar possível a exploração de redes maiores. No mais, ainda é viável adaptar o processo de classificação para buscar partições sobrepostas, comunidades de vértices individuais e expandir a utilização de metadados para redes ponderadas, digrafos e redes com várias informações adicionais presentes nos vértices.

REFERÊNCIAS

- 1 EULER, L. Solutio problematis ad geometriam situs pertinentis. **Commentarii Academiae Scientiarum Petropolitanae**, v. 8, p. 128–140, 1741.
- 2 HAMILTON, W. Letter to John T. Graves on the Icosian (17th october 1856). In: HALBERSTAM, H.; INGRAM, R. (Ed.). **Mathematical papers**. Cambridge: Cambridge University Press, 1967. v. 3, p. 612–625.
- 3 APPEL K.; HAKEN, W. Every planar map is four colorable. part I: discharging. **Journal of Mathematics**, v. 21, p. 429–490, 1977.
- 4 SOLOMONOFF, R. A. Y.; RAPOPORT, A. Connectivity of random nets. **Bulletin of Mathematical Biology**, v. 13, n. 2, p. 107–117, 1951.
- 5 BOLLOBAS, B. **Random graphs**. Cambridge: Cambridge University Press, 2001. ISBN 9780511814068.
- 6 ERDŐS, P.; RÉNYI, A. On the evolution of random graphs. **Hungarian Academy of Sciences**, v. 5, n. 1-2, p. 17–61, 1960.
- 7 ERDŐS P.; RÉNYI, A. On random graphs i. **Publicationes Mathematicae Debrecen**, v. 6, p. 290–297, 1959.
- 8 ERDOS P.; RÉNYI, A. On the strength of connectedness of a random graph. **Acta Mathematica Scientia Hungary**, n. 1-2, p. 261–267.
- 9 HARARY, F. **Graph theory**. New York: Addison Wesley, 1995.
- 10 BOLLOBAS, B. **Modern graph theory**. New York: Springer, 1998.
- 11 FALOUTSOS, M.; FALOUTSOS, P.; FALOUTSOS, C. On power-law relationships of the internet topology. **Computer Communication Review**, v. 29, n. 4, p. 251–262, 1999.
- 12 ALBERT, R.; JEONG, H.; BARABASI, A. L. Internet: diameter of the World-Wide Web. **Nature**, v. 401, n. 6749, p. 130–131, 1999.
- 13 BARABÁSI, A.; ALBERT, R. Emergence of scaling in random networks. **Science**, v. 286, n. 5439, p. 509–512, 1999.
- 14 BARABASI, A. L.; ALBERT, R.; JEONG, H. Mean-field theory for scale-free random networks: Statistical mechanics and its applications. **Physica A**, v. 272, n. 1–2, p. 173–187, 1999.
- 15 BARABASI, A. L.; ALBERT, R. Topology of evolving networks: local events and universality. **Physical Review Letters**, v. 85, n. 24, p. 5234–5237, 2000.
- 16 WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world' networks. **Nature**, n. 393, p. 440–442, 1998.
- 17 MILGRAM, S. The small-world problem. **Psychology Today**, v. 1, n. 1, p. 110, 1967.

- 18 GIRVAN, M. et al. Community structure in social and biological networks. **Proceedings of the National Academy of Sciences of the United States of America**, v. 99, n. 12, p. 7821–7826, 2002.
- 19 RIVES, A. W.; GALITSKI, T. Modular organization of cellular networks. **Proceedings of the National Academy of Sciences**, v. 100, n. 3, p. 1128–1133, 2003.
- 20 SPIRIN, V.; MIRNY, L. A. Protein complexes and functional modules in molecular networks. **Proceedings of the National Academy of Sciences**, v. 100, n. 21, p. 12123–12128, 2003.
- 21 CHEN, J.; YUAN, B. Detecting functional modules in the yeast protein-protein interaction network. **Bioinformatics**, v. 22, n. 18, p. 2283–90, 2006.
- 22 FLAKE, G. W. et al. Self-organization and identification of web communities. **Computer**, v. 35, n. 3, p. 66–71, 2002.
- 23 DOURISBOURE, Y.; GERACI, F.; PELLEGRINI, M. Extraction and classification of dense communities in the web. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 16, WWW '07, Banff, Canada. **Proceedings...** New York: 2007. p. 461–470.
- 24 TRAUD, A. L. et al. Comparing community structure to characteristics in online collegiate social networks. **SIAM Review**, v. 53, n. 3, p. 17, 2008.
- 25 BEGUERISSE-DÍAZ, M. et al. Interest communities and flow roles in directed networks: the twitter network of the uk riots. **Journal of The Royal Society Interface**, v. 11, n. 101, 2014.
- 26 WEISS, R. S.; JACOBSON, E. A method for the analysis of the structure of complex organizations. **American Sociological Review**, v. 20, n. 6, p. 661–668, 1955.
- 27 RICE, S. A. The identification of blocs in small political bodies. **American Political Science Review**, v. 21, n. 3, p. 619–627, 1927.
- 28 FRENCH, D. **Homan's theory of the human group**. New York: Brandeis University, 1964.
- 29 NEWMAN, M. E. J. Modularity and community structure in networks. **Proceedings of the National Academy of Sciences**, v. 103, n. 23, p. 8577–8582, 2006.
- 30 NEWMAN, M.; GIRVAN, M. Finding and evaluating community structure in networks. **Physical Review E**, v. 69, n. 2, p. 026113, 2004.
- 31 REICHARDT, J.; BORNHOLDT, S. Statistical mechanics of community detection. **Physical Review E**, v. 74, n. 1, p. 016110, 2006.
- 32 STAUFFER, D.; AHARONY, A. **Introduction to percolation theory**. New York: Oxford University Press, 1971.
- 33 ARENAS, A.; DÍAZ-GUILERA, A.; PÉREZ-VICENTE, C. J. Synchronization reveals topological scales in complex networks. **Physical Review Letters**, v. 96, n. 11, p. 114102, 2006.

-
- 34 REICHARDT, J.; BORNHOLDT, S. Detecting fuzzy community structures in complex networks with a Potts model. **Physical Review Letters**, v. 93, n. 21, p. 218701, 2004.
 - 35 BOCCALETTI, S. et al. Detecting complex network modularity by dynamical clustering. **Physical Review E**, v. 75, n. 4, p. 1–4, 2007.
 - 36 ZHOU, H. Distance, dissimilarity index, and network community structure. **Physical Review E**, v. 67, n. 6, p. 061901, 2003.
 - 37 NEWMAN, M.; BARABASI, A.-L.; WATTS, D. J. **The structure and dynamics of networks**. New Jersey: Princeton University Press, 2006.
 - 38 TRAVIESO, G.; COSTA, L. da F. Spread of opinions and proportional voting. **Physical Review E**, v. 74, n. 3, p. 036112, 2006.
 - 39 TADIĆ, B.; RODGERS, G. J.; THURNER, S. Transport on complex networks: flow, jamming and optimization. **International Journal of Bifurcation and Chaos**, v. 17, n. 7, p. 2363–2385, 2007.
 - 40 ALBERT, R.; JEONG, H.; BARABASI, A. Error and attack tolerance of complex networks. **Nature**, v. 406, n. 6794, p. 378–382, 2000.
 - 41 RODRIGUES, F. A.; COSTA, L. D. F. Surviving opinions in sznajd models on complex networks. **International Journal of Modern Physics C**, v. 16, n. 11, p. 1785–1792, 2005.
 - 42 HONG, H.; CHOI, M. Y.; KIM, B. J. Synchronization on small-world networks. **Physical Review E**, v. 65, n. 2, p. 026139, 2002.
 - 43 DOROGOVTSSEV, S. N.; GOLTSEV, A. V.; MENDES, J. F. F. Critical phenomena in complex networks. **Review Modern Physics**, v. 80, n. 4, p. 1275–1335, 2008.
 - 44 STANLEY, H. E. **Introduction to phase transitions and critical phenomena**. Oxford: Oxford University Press, 1987.
 - 45 MARRO, J.; DICKMAN, R. **Nonequilibrium phase transitions in lattice models**. Cambridge: Cambridge University Press, 1999.
 - 46 COHEN, R. et al. Breakdown of the internet under intentional attack. **Physical Review Letters**, v. 86, n. 16, p. 3682–3685, 2001.
 - 47 ECHENIQUE, P.; GÓMEZ-GARDEÑES, J.; MORENO, Y. Dynamics of jamming transitions in complex networks. **Europhysics Letters**, v. 71, n. 2, p. 325, 2005.
 - 48 CASTELLANO, C.; FORTUNATO, S.; LORETO, V. Statistical physics of social dynamics. **Review Modern Physics**, v. 81, n. 2, p. 591–646, 2009.
 - 49 COHEN, R. et al. Resilience of the internet to random breakdowns. **Physical Review Letters**, v. 85, n. 21, p. 4626–4628, 2000.
 - 50 WEISS, G. H. **Aspects and applications of the random walk**. New York: North-Holland Publishing Co., 1994. ISBN 0444816062.

- 51 SPITZER, F. **Principles of random walk**. New York: Springer, 2001. ISBN 0387951547.
- 52 BACHELIER, L. Théorie de la spéculation. **Annales Scientifiques de L'Ecole Normale Supérieure**, v. 17, n. 2, p. 21–88, 1900.
- 53 PEARSON, K. The problem of the random walk. **Nature**, v. 72, n. 1865, p. 294, 1905.
- 54 EINSTEIN, A. **Investigations on the theory of the brownian movement**. New York: Dover Publications, 1905.
- 55 GENNES, P. de. **Scaling concepts in polymer physics**. New York: Cornell University Press, 1979. ISBN 080141203X.
- 56 DOI, M.; EDWARDS, S. **The theory of polymer dynamics**. Oxford: Clarendon Press, 1988. ISBN 9780198520337.
- 57 EVANS, M. R.; MAJUMDAR, S. N. Diffusion with stochastic resetting. **Physical Review Letters**, v. 106, n. 16, p. 160601, 2011.
- 58 AVRAHAM, D. ben; HAVLIN, S. **Diffusion and reactions in fractals and disordered systems**. Cambridge: Cambridge University Press, 2000. ISBN 0521617200.
- 59 LAWLER, G. F.; LIMIC, V. **Random walk: a modern introduction**. Cambridge: Cambridge University Press, 2010. ISBN 0521519187.
- 60 DUSSUTOUR VINCENT FOURCASSIÉ, D. H. A.; DENEUBOURG, J.-L. Optimal traffic organization in ants under crowded conditions. **Nature**, v. 428, p. 70–73, 2004.
- 61 GONZALEZ, M. C.; HIDALGO, C. A.; BARABASI, A.-L. Understanding individual human mobility patterns. **Nature**, v. 453, n. 7196, p. 779–782, 2008.
- 62 MISSIURO, P. V. et al. Information flow analysis of interactome networks. **PLOS Computational Biology**, v. 5, n. 4, p. 1–15, 2009.
- 63 GFELLER, D. et al. Complex network analysis of free-energy landscapes. **Proceedings of the National Academy of Sciences**, v. 104, n. 6, p. 1817–1822, 2007.
- 64 REDNER, S. **A guide to first-passage processes**. Cambridge: Cambridge University Press, 2001.
- 65 SOOD, V.; REDNER, S.; AVRAHAM, D. ben. First-passage properties of the erdős–rényi random graph. **Journal of Physics A**, v. 38, n. 1, p. 109, 2005.
- 66 TISHBY, I.; BIHAM, O.; KATZAV, E. The distribution of first hitting times of randomwalks on erdős–rényi networks. **Journal of Physics A**, v. 50, n. 11, p. 115001, 2017.
- 67 BACCO, C. D.; MAJUMDAR, S. N.; SOLLICH, P. The average number of distinct sites visited by a random walker on random graphs. **Journal of Physics A**, v. 48, n. 20, p. 205004, 2015.
- 68 KAHN, J. D. et al. On the cover time of random walks on graphs. **Journal of Theoretical Probability**, v. 2, n. 1, p. 121–128, 1989.

-
- 69 LAHTINEN, J.; KERTÉSZ, J.; KASKI, K. Scaling of random spreading in small world networks. **Physical Review E**, v. 64, n. 5, p. 057105, 2001.
- 70 ALMAAS, E.; KULKARNI, R.; STROUD, D. Scaling properties of random walks on small-world networks. **Physical Review E**, v. 68, n. 5, p. 056105, 2003.
- 71 NOH, J. D. Random walks on complex networks. **Physical Review Letters**, v. 92, n. 11, p. 118701, 2004.
- 72 MASUDA, N.; KONNO, N. Return times of random walk on generalized random graphs. **Physical Review E**, v. 69, n. 6, p. 1–7, 2004.
- 73 TEJEDOR, V.; BÉNICHOU, O.; VOITURIEZ, R. Global mean first-passage times of random walks on complex networks. **Physical Review E**, v. 80, n. 6, p. 065104, 2009.
- 74 LAMBIOTTE, R.; SALNIKOV, V.; ROSVALL, M. Effect of memory on the dynamics of random walks on networks. **Journal of Complex Networks**, n. 3, 2014.
- 75 WENG, T. et al. Time-series analysis of networks: exploring the structure with random walks. **Physical Review E**, v. 90, n. 2, p. 022804, 2014.
- 76 YANG, S.-J. Exploring complex networks by walking on them. **Physical Review E**, v. 71, n. 1, p. 016107, 2005.
- 77 COSTA, L. D. F.; TRAVIESO, G. Exploring complex networks through random walks. **Physical Review E**, v. 75, n. 1, p. 016102, 2007.
- 78 López Millán, V. M. et al. A model of self-avoiding random walks for searching complex networks. **Networks**, v. 60, n. 2, p. 71–85, 2011.
- 79 GUO, Q. et al. Levy random walks on multiplex networks. **Scientific Reports**, v. 6, 2016.
- 80 KIM, K.; KYOUNG, J.; LEE, D.-S. Self-attracting walk on heterogeneous networks. **Physical Review E**, v. 93, n. 5, p. 052310, 2016.
- 81 SYKES, M. F. et al. The asymptotic behaviour of selfavoiding walks and returns on a lattice. **Journal of Physics A**, v. 5, n. 5, p. 653, 1972.
- 82 GUTTMANN, A. J. Bounds on connective constants for self-avoiding walks. **Journal of Physics A**, v. 16, n. 10, p. 2233, 1983.
- 83 SLADE, G. The diffusion of self-avoiding random walk in high dimensions. **Communications in Mathematical Physics**, v. 110, n. 4, p. 661–683, 1987.
- 84 CLISBY, N.; LIANG, R.; SLADE, G. Self-avoiding walk enumeration via the lace expansion. **Journal of Physics A**, v. 40, n. 36, p. 10973, 2007.
- 85 CLISBY, N. Calculation of the connective constant for self-avoiding walks via the pivot algorithm. **Journal of Physics A**, v. 46, n. 24, p. 245001, 2013.
- 86 BINDER, K. **Monte Carlo and molecular dynamics simulations in polymer science**. New York: Oxford University Press, 1995. ISBN 0195094387.

- 87 MADRAS, N.; SLADE, G. **The self-avoiding walk**. Boston: Birkhauser Boston, 1996. (Probability and Its Applications).
- 88 HERRERO, C. P.; SABOYÁ, M. Self-avoiding walks and connective constants in small-world networks. **Physical Review E**, v. 68, n. 2, p. 026106, 2003.
- 89 TISHBY, I.; BIHAM, O.; KATZAV, E. The distribution of path lengths of self avoiding walks on erdős–rényi networks. **Journal of Physics A**, v. 49, n. 28, p. 285002, 2016.
- 90 HERRERO, C. P. Kinetic growth walks on complex networks. **Journal of Physics A**, v. 38, n. 20, p. 4349–4364, 2005.
- 91 HERRERO, C. Self-avoiding walks on scale-free networks. **Physical Review E**, v. 71, n. 1, p. 016103, 2005.
- 92 ESTRADA, E. **The structure of complex networks: theory and applications**. New York: Oxford University Press, Inc., 2011.
- 93 NEWMAN, M. **Networks: an introduction**. New York: Oxford University Press, Inc., 2010.
- 94 FORTUNATO, S. Community detection in graphs. **Physics Reports**, v. 486, n. 3-5, p. 75–174, 2010.
- 95 MALLIAROS, F. D.; VAZIRGIANNIS, M. Clustering and community detection in directed networks: a survey. **Physics Reports**, v. 533, n. 4, p. 95–142, 2013.
- 96 FORTUNATO, S.; HRIC, D. Community detection in networks: a user guide. **Physics Reports**, v. 659, p. 1–44, 2016.
- 97 AHUJA, R. K.; MAGNANTI, T. L.; ORLIN, J. B. **Network flows: theory, algorithms, and applications**. Upper Saddle River: Prentice-Hall, 1993.
- 98 BELLMAN, R.; HOLLAND, J.; KALABA, R. On an application of dynamic programming to the synthesis of logical systems. **Journal of the ACM**, v. 6, n. 4, p. 486–493, 1959.
- 99 DIJKSTRA, E. W. A note on two problems in connexion with graphs. **Numerische Mathematik**, v. 1, n. 1, p. 269–271, 1959.
- 100 JOHNSON, D. B. Efficient algorithms for shortest paths in sparse networks. **Journal of the ACM**, v. 24, n. 1, p. 1–13, 1977.
- 101 FLOYD, R. W. Algorithm 97: shortest path. **Communications of the ACM**, v. 5, n. 6, p. 345, 1962.
- 102 BEAUCHAMP, M. A. An improved index of centrality. **Behavioral Science**, v. 10, n. 2, p. 161–163, 1965.
- 103 BONACICH, P. Power and centrality: a family of measures. **American Journal of Sociology**, v. 92, n. 5, p. 1170–1182, 1987.
- 104 KATZ, L. A new status index derived from sociometric analysis. **Psychometrika**, v. 18, n. 1, p. 39–43, 1953.

- 105 BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual Web search engine. **Computer Networks and ISDN Systems**, v. 30, n. 1–7, p. 107–117, 1998.
- 106 FREEMAN, L. C. A set of measures of centrality based on betweenness. **Sociometry**, v. 40, n. 1, p. 35–41, 1977.
- 107 BOCCALETTI, S. et al. Complex networks: structure and dynamics. **Physics Reports**, v. 424, n. 4–5, p. 175–308, 2006.
- 108 JEONG, H. et al. Lethality and centrality in protein networks. **Nature**, v. 411, n. 6833, p. 41–42, 2001.
- 109 BARABÁSI, A.-L.; OLTVAI, Z. N. Network biology: understanding the cell's functional organization. **Nature Reviews Genetics**, v. 5, n. 2, p. 101–113, 2004.
- 110 JEONG, H. et al. The large-scale organization of metabolic networks. **Nature**, v. 407, n. 6804, p. 651–654, 2000.
- 111 BARABÁSI, A. L. et al. Evolution of the social network of scientific collaborations. **Physica A**, v. 311, n. 3–4, p. 590 – 614, 2002.
- 112 BOLLOBAS, B.; RIORDAN, O. The diameter of a scale-free random graph. **Combinatorica**, v. 24, n. 1, p. 5–34, 2004.
- 113 Cohen, R.; Havlin, S. Scale-free networks are ultrasmall. **Physical Review Letters**, v. 90, n. 5, p. 058701, 2003.
- 114 STRUCTURE of growing networks with preferential linking. **Physical Review Letters**, v. 85, n. 21, p. 4633–4636, 2000.
- 115 KRAPIVSKY, P. L.; REDNER, S.; LEYVRAZ, F. Connectivity of growing random networks. **Physical Review Letters**, v. 85, n. 21, p. 4629–4632, 2000.
- 116 LILJEROS, F. et al. The web of human sexual contacts. **Nature**, v. 411, n. 6840, p. 907–908, 2001.
- 117 MAYER, A.; PULLER, S. L. The old boy (and girl) network: social network formation on university campuses. **Journal of Power Electronics**, v. 47, n. 3, p. 169–184, 2007.
- 118 NEWMAN, M. E. J.; LEICHT, E. A. Mixture models and exploratory analysis in networks. **Proceedings of the National Academy of Sciences**, v. 104, n. 23, p. 9564–9569, 2007.
- 119 TRAUD, A. L. et al. Comparing community structure to characteristics in online collegiate social networks. **SIAM Review**, v. 53, n. 3, p. 526–543, 2011.
- 120 SARVARI, H. et al. **Constructing and analyzing criminal networks**. In: IEEE SYMPOSIUM ON SECURITY AND PRIVACY WORKSHOPS, 2014. Disponível em: <http://www.ieee-security.org/TC/SPW2014/papers/5103a084.PDF>. Acesso em: 14 mar. 2018.
- 121 ZHANG, A. **Protein interaction networks: computational analysis**. Cambridge: Cambridge University Press, 2009. ISBN 9780521888950.

- 122 HUTTLIN, E. L. et al. Architecture of the human interactome defines protein communities and disease networks. **Nature**, v. 545, n. 7655, p. 505–509, 2017.
- 123 JONSSON, P. F. et al. Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. **BMC Bioinformatics**, v. 7, p. 2, 2006.
- 124 WILKINSON, D. M.; HUBERMAN, B. A. A method for finding communities of related genes. **Proceedings of the National Academy of Sciences**, v. 10, n. 1073, p. 1–8, 2004.
- 125 DANON, L. et al. Comparing community structure identification. **Journal of Statistical Mechanics**, v. 2005, p. 9008, 2005.
- 126 GÓMEZ, S.; JENSEN, P.; ARENAS, A. Analysis of community structure in networks of correlated data. **Physical Review E**, v. 80, n. 1, p. 016114, 2009.
- 127 LI, Y.; JIA, C.; YU, J. A parameter-free community detection method based on centrality and dispersion of nodes in complex networks. **Physica A**, v. 438, p. 321–334, 2015.
- 128 FORTUNATO, S.; BARTHELEMY, M. Resolution limit in community detection. **Proceedings of the National Academy of Sciences**, v. 104, n. 1, p. 36–41, 2007.
- 129 KUMPULA, J. M. et al. Limited resolution in complex network community detection with potts model approach. **The European Physical Journal B**, v. 56, n. 1, p. 41–45, 2007.
- 130 GOOD, B. H.; De Montjoye, Y. A.; CLAUSET, A. Performance of modularity maximization in practical contexts. **Physical Review E**, v. 81, n. 4, p. 1–19, 2010.
- 131 LANCICHINETTI, A.; FORTUNATO, S. Limits of modularity maximization in community detection. **Physical Review E**, v. 84, n. 6, p. 066122, 2011.
- 132 LUXBURG, U. A tutorial on spectral clustering. **Statistics and Computing**, v. 17, n. 4, p. 395–416, 2007.
- 133 HASTINGS, M. B. Community detection as an inference problem. **Physical Review E**, v. 74, n. 3, p. 035102, 2006.
- 134 PEIXOTO, T. P. Hierarchical block structures and high-resolution model selection in large networks. **Physical Review X**, v. 4, n. 1, p. 011047, 2014.
- 135 NEWMAN, M. E. J. Fast algorithm for detecting community structure in networks. **Physical Review E**, v. 69, n. 6, 2004.
- 136 CLAUSET, A.; NEWMAN, M. E. J.; MOORE, C. Finding community structure in very large networks. **Physical Review E**, v. 70, n. 6 2, p. 1–6, 2004.
- 137 PONS, P.; LATAPY, M. Computing communities in large networks using random walks. **Journal of Graph Algorithms and Applications**, v. 10, p. 191–218, 2005.
- 138 WARD, J. H. Hierarchical grouping to optimize an objective function. **Journal of the American Statistical Association**, v. 58, n. 301, p. 236–244, 1963.

- 139 ROSVALL, M.; BERGSTROM, C. T. Maps of random walks on complex networks reveal community structure. **Proceedings of the National Academy of Sciences**, v. 105, n. 4, p. 1118–1123, 2008.
- 140 HUFFMAN, D. A. A method for the construction of minimum-redundancy codes. **Proceedings of the Institute of Radio Engineers**, v. 40, n. 9, p. 1098–1101, 1952.
- 141 ROSVALL, M.; AXELSSON, D.; BERGSTROM, C. T. The map equation. **The European Physical Journal Special Topics**, v. 178, n. 1, p. 13–23.
- 142 SHANNON, C. E. A mathematical theory of communication. **Bell System Technical Journal**, v. 27, n. 3, p. 379–423, 1948.
- 143 ROSVALL, M.; BERGSTROM, C. T. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. **PLOS ONE**, v. 6, n. 4, p. 1–10, 2011.
- 144 ESQUIVEL, A. V.; ROSVALL, M. Compression of flow can reveal overlapping-module organization in networks. **Physical Review E**, v. 1, n. 2, p. 021025, 2011.
- 145 PERSSON, C. et al. **Maps of sparse Markov chains efficiently reveal community structure in network flows with memory**, 2016. Disponível em: <https://arxiv.org/abs/1606.08328>. Acesso em: 14 mar. 2018.
- 146 CONDON, A.; KARP, R. M. Algorithms for graph partitioning on the planted partition model. **Random Structures and Algorithms**, v. 18, n. 2, p. 116–140, 2001.
- 147 FAN, Y. et al. Accuracy and precision of methods for community identification in weighted networks. **Physica A**, v. 377, n. 1, p. 363 – 372, 2007.
- 148 LANCICHINETTI, A.; FORTUNATO, S.; RADICCHI, F. Benchmark graphs for testing community detection algorithms. **Physical Review E**, v. 78, n. 4, p. 1–5, 2008.
- 149 DANON, L.; DÍAZ-Guilera, A.; ARENAS, A. The effect of size heterogeneity on community identification in complex networks. **Journal of Statistical Mechanics**, v. 2006, n. 11, p. 11010, 2006.
- 150 LANCICHINETTI, A.; FORTUNATO, S.; KERTÉSZ, J. Detecting the overlapping and hierarchical community structure in complex networks. **New Journal of Physics**, v. 11, p. 1–18, 2009.
- 151 LANCICHINETTI, A.; FORTUNATO, S. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. **Physical Review E**, v. 80, n. 1, p. 1–8, 2009.
- 152 ZACHARY, W. An information flow model for conflict and information fission in small groups. **Journal of Anthropological Research**, v. 33, n. 4, p. 452–473, 1977.
- 153 LUSSEAU, D. The emergent properties of a dolphin social network. **Proceedings Biological Sciences**, v. 270 Suppl. 2, p. S186–S188, 2003.
- 154 KNUTH, D. E. **The Stanford graphbase: a platform for combinatorial computing**. Stanford: Addison-Wesley Reading, 1993. v. 37. ISBN 0321606329 9780321606327.

- 155 GLEISER, P.; DANON, L. Community structure in jazz. **Advances in Complex Systems**, v. 6, n. 04, p. 12, 2003.
- 156 WHITE, J. et al. The structure of the nervous system of the nematode *C. elegans*. **Philosophical Transactions Royal Society London**, v. 314, p. 1–340, 1986.
- 157 SZIGETI, B. OpenWorm: an open-science approach to modeling *Caenorhabditis elegans*. **Frontiers in Computational Neuroscience**, v. 8, p. 1–7, 2014.
- 158 ADAMIC, L. A.; GLANCE, N. The political blogosphere and the 2004 u.s. election: divided they blog. In: INTERNATIONAL WORKSHOP ON LINK DISCOVERY, 3, 2005, Chicago. **Proceedings...** New York: ACM, 2005. p. 36–43.
- 159 BEVERIDGE, A.; SHAN, J. Network of thrones. **Mathhorizon**, n. 4, p. 18–22, 2016.
- 160 PERKINS, T. J. et al. A scaling law for random walks on networks. **Nature Communications**, v. 5, p. 5121, 2014.
- 161 ARRUDA, G. F. de et al. The role of centrality for the identification of influential spreaders in complex networks. **Physical Review E**, v. 032812, n. 3, p. 17, 2014.
- 162 LOVÁSZ, L. Random walks on graphs: a survey. **Combinatorics**, v. 2, n. 2, p. 353–398, 1993.
- 163 CHUNG, F. R. K. **Spectral graph theory**. New York: American Mathematical Society, 1997. ISBN 0821803158.
- 164 NORRIS, J. R. **Markov chains**. Cambridge, UK: Cambridge University Press, 1997.
- 165 MEYER, C. D. **Matrix analysis and applied linear algebra (solution)**. Filadelfia: Society for Industrial and Applied Mathematics, 2004. ISBN 0898714540 9780898714548.
- 166 KIM, Y.; PARK, S.; YOOK, S.-H. Network exploration using true self-avoiding walks. **Physical Review E**, v. 94, n. 4, p. 042309, 2016.
- 167 TRAVENÇOLO, B. a. N.; VIANA, M. P.; COSTA, L. D. F. Border detection in complex networks. **New Journal of Physics**, v. 11, n. 6, p. 063019, 2009.
- 168 TRAVENÇOLO, B. A. N.; COSTA, L. d. F. Accessibility in complex networks. **Physics Letters**, v. 373, n. 1, p. 89–95, 2008.
- 169 ROSVALL, M. et al. Networks and cities: An information perspective. **Physical Review Letters**, v. 94, n. 2, p. 1–4, 2005.
- 170 MARSHAK, C. Z. et al. Growth and containment of a hierarchical criminal network. **Physical Review E**, v. 93, n. 2, p. 022308, 2016.
- 171 GOLUB, G.; Van Loan, C. **Matrix computations**. Baltimore: Johns Hopkins University Press, 1989. ISBN 0801854148.
- 172 KEOGH, E.; MUEEN, A. **Curse of Dimensionality**. Boston: Springer Verlag, 2010, p. 257–258.

- 173 JOLLIFFE, I. **Principal component analysis**. New York: Springer Verlag, 2002.
- 174 KINDT, R.; COE, R. Analysis of differences in species composition. In: _____. **Tree Diversity Analysis**. Nairobi: World Agroforestry Centre, 2005. p. 123–138.
- 175 SOKAL, R. R.; MICHENER, C. D. A statistical method for evaluating systematic relationships. **University of Kansas Science Bulletin**, v. 38, p. 1409–1438, 1958.
- 176 PALLA, G. et al. Uncovering the overlapping community structure of complex networks in nature and society. **Nature**, v. 435, n. 7043, p. 814–818, 2005.
- 177 RADICCHI, F. et al. Defining and identifying communities in networks. **Proceedings of the National Academy of Sciences**, v. 101, n. 9, p. 2658, 2004.
- 178 DONETTI, L.; MUNOZ, M. A. Detecting network communities: a new systematic and efficient algorithm. **Journal of Statistical Mechanics**, v. 2004, n. 10, p. 10012, 2004.
- 179 SHEN, H.-W.; CHENG, X.-Q.; GUO, J.-F. Quantifying and identifying the overlapping community structure in networks. **Journal of Statistical Mechanics**, v. 2009, n. 07, p. P07042, 2009.
- 180 ZHANG, S. Hierarchical modular structure identification with its applications in gene coexpression networks. **The Scientific World Journal**, v. 2012, p. 523706, 2012.
- 181 MARKOWITZ, H. Portfolio selection. **Journal of Finance**, v. 7, n. 1, p. 77–91, 1952.
- 182 BLACK, F.; LITTERMAN, R. Global portfolio optimization. **Financial Analysts Journal**, v. 48, n. 5, p. 28–43, 1992.
- 183 STEINHAEUSER, K.; CHAWLA, N. V. Identifying and evaluating community structure in complex networks. **Pattern Recognition Letters**, v. 31, n. 5, p. 413–421, 2010.
- 184 HRIC, D.; DARST, R. K.; FORTUNATO, S. Community detection in networks: Structural communities versus ground truth. **Physical Review E**, v. 90, n. 6, p. 062805, 2014.
- 185 NEWMAN, M. E. J.; PEIXOTO, T. P. Generalized communities in networks. **Physical Review Letters**, v. 115, n. 8, p. 088701, 2015.
- 186 NEWMAN, M. E. J.; CLAUSET, A. Structure and inference in annotated networks. **Nature Communications**, v. 7, p. 11863, 2015.
- 187 HRIC, D.; PEIXOTO, T. P.; FORTUNATO, S. Network structure, metadata, and the prediction of missing nodes and annotations. **Physical Review X**, v. 6, n. 3, p. 031038, 2016.
- 188 JAIN, a. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM Computing Surveys**, v. 31, n. 3, p. 264–323, 1999.
- 189 YAHOO finance - business finance, stock market, quotes, news. . Disponível em: <<https://finance.yahoo.com/>>. Acesso em: 14 mar. 2018.

- 190 MANTEGNA, R. N. Hierarchical structure in financial markets. **The European Physical Journal B**, v. 11, n. 1, p. 193–197, 1999.
- 191 TUMMINELLO, M. et al. Correlation based networks of equity returns sampled at different time horizons. **The European Physical Journal B**, v. 55, n. 2, p. 209–217, 2007.
- 192 PERON, T. K. D.; COSTA, L. da F.; RODRIGUES, F. A. The structure and resilience of financial market networks. **Chaos**, v. 22, n. 1, 2012.
- 193 MACMAHON, M.; GARLASCHELLI, D. Community detection for correlation matrices. **Physical Review X**, v. 5, n. 2, p. 1–34, 2015.
- 194 HULL, J. **Options, futures, and other derivatives**. New Jersey: Pearson Prentice Hall, 2006. ISBN 0133456315.
- 195 UTSUGI, A.; INO, K.; OSHIKAWA, M. Random matrix theory analysis of cross correlations in financial markets. **Physical Review E**, v. 70, n. 2, p. 026110, 2004.
- 196 KIM, D.-H.; JEONG, H. Systematic analysis of group identification in stock markets. **Physical Review E**, v. 72, n. 4, p. 046133, 2005.
- 197 PAN, R. K.; SINHA, S. Collective behavior of stock price movements in an emerging market. **Physical Review E**, v. 76, n. 4, p. 046116, 2007.
- 198 S&P companies with information. . Disponível em: <<http://data.okfn.org/data/core/s-and-p-500-companies#readme>>. Acesso em: 14 mar. 2018.

Apêndices

APÊNDICE A – CLUSTERIZAÇÃO HIERÁRQUICA

Clusterização é a classificação de um conjunto de dados em grupos distintos de acordo com as características dos elementos estudados. Este conjunto de dados é composto por amostras e suas características são denominadas atributos, por exemplo, em uma classificação entre homens e mulheres, os indivíduos catalogados são as amostras e os atributos podem ser peso, altura, tipo de cabelo e cor da pele. O processo de clusterização normalmente envolve os passos descritos a seguir.¹⁸⁸

Inicialmente são identificados os atributos mais relevantes do conjunto de dados inicial. Nesta etapa, variáveis redundantes são eliminadas e o resultado é organizado na forma de uma matriz $l \times n$, onde l é o número de amostras e n a quantidade de atributos. Geometricamente, esta matriz representa l pontos em um espaço n -dimensional, sendo a tarefa da clusterização organizar conjuntos nesse espaço. Com a matriz de dados pronta, precisamos adotar um critério para determinar a proximidade de dois conjuntos de medidas (quantificar a semelhança entre os atributos das diferentes amostras). Esta proximidade pode ser representada por uma distância verdadeira (métrica), $d_{x,y}$, que deve satisfazer os seguintes axiomas para qualquer par x, y, z ,

- $d_{x,y} \geq 0$ (positiva definida),
- $d_{x,y} = d_{y,x}$ (simétrica),
- $d_{x,y} \leq d_{x,z} + d_{z,y}$ (satisfazer a desigualdade triangular).

Entretanto, muitas medidas que não satisfazem a terceira condição são consideradas métrica, mas como não são distâncias verdadeiras, elas são chamadas de dissimilaridades. Quanto menor a distância ou a dissimilaridade, mais próximas duas amostras estão, indicando que elas apresentam características parecidas.

Sejam $A = (a_1, a_2, \dots, a_n)$ e $B = (b_1, b_2, \dots, b_n)$ dois conjuntos de dados, pode-se definir algumas métricas populares como as distância Euclideana,

$$d_{AB}^E = \sqrt{\sum_{k=1}^n (a_k - b_k)^2}, \quad (\text{A.1})$$

Manhattan

$$d_{AB}^M = \sum_{k=1}^n |a_k - b_k|, \quad (\text{A.2})$$

Chebyshev

$$d_{AB}^{Ch} = \max_{k=1, \dots, n} (|a_k - b_k|), \quad (\text{A.3})$$

Figura 35: Exemplo de três métodos de união de grupos usado na hierarquização aglomerativa. (a) *Single linkage*: união de dois membros de acordo com a menor distância. (b) *Complete linkage*: Combina dois dados pela maior distância. (c) *Average linkage*: A junção de duas partições é feita conforme o menor valor médio.

Fonte: Elaborada pelo autor.

Canberra

$$d_{AB}^{Cn} = \sum_{k=1}^n \frac{|a_k - b_k|}{|a_k| + |b_k|}, \quad (\text{A.4})$$

e a dissimilaridade de Bray-Curtis

$$d_{AB}^{BC} = \frac{\sum_{k=1}^n |a_k - b_k|}{\sum_{k=1}^n a_k + b_k}. \quad (\text{A.5})$$

Ao final da comparação entre as n amostras, tem-se uma matriz de proximidade $\mathbf{D}$ de dimensão $n \times n$.

A partir de $\mathbf{D}$, define-se como os dados serão agrupados. Este processo pode ser realizado através de algoritmos de clusterização hierárquica ou por meio de algoritmos de particionamento. A hierarquização pode ser dividida em duas categorias:

1. Aglomerativa, quando cada grupo é iterativamente aglutinado até restar apenas um que contém todos os dados;
2. Divisiva, quando cada grupo é iterativamente dividido até o sistema apresentar n grupos com apenas um dado cada.

Assim como os dados são analisados de acordo com um critério de similaridade, na hierarquização aglomerativa os *clusters* também são integrados utilizando algum critério. No método *single linkage*, dois *clusters* se unem quando a proximidade $d_{i,j}$ é mínima entre o membro i de um grupo com o membro j do outro. Caso contrário, se dois *clusters* se unem quando $d_{i,j}$ é máximo entre os diferentes grupos, o método é chamado de *complete linkage*. Já quando o critério é o menor valor médio entre os membros, tem-se o *average linkage*. A figura 35 ilustra esses três métodos de união dos grupos.

Finalmente, a qualidade das divisões é analisada utilizando-se critérios estatísticos e comparando o resultado a outros algoritmos. A análise do resultado pode levar a alterações no pré-processamento e nas escolhas da métrica e do método de hierarquização. Em geral,

Figura 36: Exemplo do resultado da hierarquização aglomerativa (figura à esquerda) por meio de um dendrograma (imagem à direita). As cores indicam os níveis de agrupamento a medida que se sobe na árvore.

Fonte: Elaborada pelo autor.

o resultado final da clusterização hierárquica é representado através de um dendrograma, uma árvore onde as folhas representam os grupos formados por apenas um elemento e a altura indica o grau de proximidade entre as partições como mostra a figura 36. A medida que se sobe na árvore *clusters* cada vez maiores são formados até todos os dados não apresentarem mais divisões.

Com isso, pode-se resumir o processo de clusterização hierárquica aglomerativa através do seguinte algoritmo:

1. Inicie cada dado em um *cluster*;
2. Calcule a proximidade $\mathbf{D}$ de cada partição;
3. Encontre os dois grupos mais semelhantes e os junte;
4. Atualize $\mathbf{D}$;
5. Repita os passos 3 e 4 até todos os dados pertencerem a uma única divisão.

APÊNDICE B – REDE DE AÇÕES

Nossa base de dados de mercado corresponde a 2518 preços diários de fechamento de ações (Janeiro de 2007 a Janeiro de 2017) negociadas na bolsa de valores de *New York*, adquiridos através do portal *Yahoo!*¹⁸⁹ As ações analisadas foram as pertencentes ao índice *Standard & Poor's 500* (S&P 500), pois este representa os 500 ativos mais importantes da bolsa devido principalmente ao seu tamanho de mercado e sua liquidez. Além disso, este índice é importante para o mercado porque é sensível a mudanças de cenários econômicos em que a economia fica exposta.

As redes de mercado financeiro são geralmente construídas por meio das matrizes de correlação.^{190–193} Esta metodologia considera o coeficiente de correlação entre séries temporais de duas ações i e j , que é dado pelo coeficiente de correlação de Pearson,

$$\rho_{i,j} = \frac{\langle Y_i Y_j \rangle - \langle Y_i \rangle \langle Y_j \rangle}{\sqrt{(\langle Y_i^2 \rangle - \langle Y_i \rangle^2)(\langle Y_j^2 \rangle - \langle Y_j \rangle^2)}}, \quad (\text{B.1})$$

onde Y_i é o retorno do ativo i , definido por

$$Y_i = \frac{P_i(t)}{P_i(t-1)} - 1, \quad (\text{B.2})$$

com $P_i(t)$ o preço de fechamento do ativo i no dia t . Mas como, na prática, os preços das ações entre os instantes t e $t-1$ são muito próximos, temos $Y_i(t) \ll 1$, permitindo reescrever B.2 da seguinte forma,

$$Y_i = \ln \left(\frac{P_i(t)}{P_i(t-1)} \right). \quad (\text{B.3})$$

Além de ser uma boa aproximação, a utilização do log-retorno tem algumas vantagens matemática,¹⁹⁴ como gerar uma distribuição normal de retornos e simplificar cálculos financeiros, uma vez que estes são aditivos no tempo.

Assim, se dois ativos tem comportamento similar ao longo do tempo, ρ tende a um, por outro lado, se a série temporal apresentar comportamento contrário, ρ tende a menos um. Estes valores entre N ações formam uma matriz de correlação simétrica $N \times N$, que é utilizada diretamente para construir a matriz de pesos $\mathbf{W}$. Contudo, para determinar o peso das arestas entre dois ativos, estamos interessados apenas no valor da correlação (se um ativo influencia outro) e não em sua direção (positiva ou negativa), portanto utilizamos o módulo de $\rho_{i,j}$ como o peso de cada ligação, $w_{i,j} = |\rho_{i,j}|$. Desta forma, a ponderação varia no intervalo $[0, 1]$, com as arestas mais fortes conectando ações com forte influência entre si.

A imagem 37 exemplifica o processo de formação deste tipo de rede até a atribuição de pesos nas arestas. No gráfico, temos em (a) a comparação entre o preço diário de três

Figura 37: Ilustração do método de construção de uma rede de mercado financeiro a partir de três ativos. Por meio da série temporal do preço, determina-se o retorno diário e posteriormente as conexões entre as ações, ponderadas pela correlação destes retornos.

Fonte: Elaborada pelo autor.

ações distintas que apresentam visivelmente grande correlação, visto que quando uma começa um movimento (queda ou uma elevação) de alta magnitude no preço, as outras duas seguem a mesma tendência. Já em (b) é mostrado o resultado da equação B.3 aplicada aos mesmos ativos de (a), e finalmente em (c) o grafo formado por estes três ativos sendo a ponderação dada pela correção dos retornos mostrados em (b).

Ao final desse processo é gerada uma rede completa ponderada de acordo com a correlação entre as séries temporais de cada ação. Para obter um grafo com significância estatística escolhe-se um valor de corte (ν) e se faz uma seleção das ligações tal que as arestas com $w_{i,j} < \nu$ são removidas. A vantagem deste método consiste na filtragem de ruídos, uma vez que as baixas correlações, mais sujeitas a flutuações aleatórias, são descartadas. Por outro lado, a principal dificuldade aqui é determinar da maneira mais adequada o corte ν . O critério de escolha deste *threshold* foi baseado na distribuição estatísticas das correlações, que é mostrada na figura 38. A partir desta, utilizamos a média (μ) e o desvio padrão (σ) para definir o corte da seguinte forma,

$$\nu = \mu + \eta\sigma, \quad (\text{B.4})$$

Figura 38: Distribuição estatística do coeficiente de Pearson (eq. B.1) para todos os pares de ações do índice S&P500 em um período de 10 anos. A média e o desvio padrão deste sistema são respectivamente 0.37 e 0.12.

Fonte: Elaborada pelo autor.

onde η é uma grandeza que corresponde ao nível de significância estatística. Assim, para essa distribuição temos $\mu = 0.37$, $\sigma = 0.12$ e escolhemos $\eta = 2.5$, resultando em um corte de $\nu = 0.67$, ou seja, somente correlações maiores que este valor foram mantidas no grafo (aproximadamente 1%). O resultado final foi uma rede complexa com 433 vértices e 855 arestas, com sua maior componente possuindo 121 ações e 716 conexões ponderadas.

Uma vez que o grafo está bem definido, atribuímos a cada vértice o setor de atuação da respectiva empresa na indústria. Isso foi feito pois estamos interessados em estudar a presença de comunidades neste tipo de rede e trabalhos anteriores^{195–197} já identificaram certa correlação de preços entre ativos que atuam no mesmo setor. Estes dados foram obtidos *online*¹⁹⁸ e foram divididos em cinco grandes segmentos: Bens de consumo, Financeiro, Indústria, Energia e Materiais.

Anexos

ANEXO A – INFORMES COMPUTACIONAIS

A elaboração das simulações realizadas neste trabalho foram desenvolvidos principalmente na linguagem de programação Python na versão 2.7 e, quando necessitamos de maior desempenho computacional, foi utilizada a linguagem C/C++. Todos os *scripts* e programas foram executados no sistema operacional Linux 64 bits e, quando necessário, recorreremos ao *cluster* disponibilizado pela Universidade de São Paulo pela Superintendência de Tecnologia da Informação (STI). A seguinte tabela sumariza os principais *sotwares* e pacotes de terceiros usados neste trabalho.

Tabela 8: *Sotwares* e pacotes de terceiros utilizados para geração de *scripts* de análises de dados e simulações, construção de gráficos e ilustrações de redes complexas.

Pacotes	Versão	<i>Website</i>
Numpy	1.9.2	www.numpy.org
SciPy	0.16.1	www.scipy.org
Matplotlib	1.4.3	www.matplotlib.org
Pandas	0.17.1	pandas.pydata.org
Scikit-learn	0.19.1	scikit-learn.org
Igraph	0.7.1	www.igraph.org
NetworkX	1.9.1	www.networkx.github.io
Gephi	0.9.1	www.gephi.org
Inkscape	0.91	www.inkscape.org

Fonte: Elaborada pelo autor.