UNIVERSIDADE ESTADUAL DE MARINGÁ
DEPARTAMENTO DE FÍSICA

Dennys Felipe de Oliveira Santicioli Rizzon

Megaregiões no Brasil: Um estudo sobre a formação de estruturas de comunidade em redes pendulares

Maringá, Abril de 2022.

UNIVERSIDADE ESTADUAL DE MARINGÁ
DEPARTAMENTO DE FÍSICA

Dennys Felipe de Oliveira Santicioli Rizzon

Megaregiões no Brasil: Um estudo sobre a formação de estruturas de comunidade em redes pendulares

Trabalho de conclusão de curso apresentado ao Departamento de Física da Universidade Estadual de Maringá como requisito para obtenção do título de Bacharel em Física.

Orientador: Prof. Dr. Haroldo Valentin Ribeiro

Maringá, Abril de 2022.

Agradecimentos

Eu gostaria de agradecer...

À minha família e amigos, em especial à minha irmã, pelo constante apoio e incentivo, mesmo nos momentos mais difíceis.

Ao meu orientador, pela admirável paciência e disposição durante todos esses meses em que trabalhamos juntos.

E também aos demais professores que contribuíram para a minha formação e me proporcionaram a base e a inspiração necessárias para entregar um trabalho de qualidade.

Resumo

Este trabalho apresenta uma investigação sobre alguns aspectos da rede de movimento pendular entre as cidades brasileiras. Nosso objetivo principal foi compreender as características básicas dessa rede complexa e verificar sua estrutura de comunidades. Para isso, inicialmente, realizamos uma revisão dos conceitos mais básicos da teoria de redes complexas e também estudamos um algoritmo para identificação de comunidades em rede denominado Infomap. Além disso, também apresentamos algumas ferramentas computacionais típicas para análise de dados e análise de redes complexas. Entre nossos achados empíricos mais relevantes, destacamos que a rede de movimento pendular brasileira apresenta uma distribuição de grau lei de potência com expoente próximo da unidade, demonstrando a existência de algumas poucas cidades superconectadas (*hubs*) e uma maioria de municípios com poucas conexões. Além de apresentar características de uma rede livre de escala, observamos que a rede de movimento pendular tem características de mundo pequeno que se manifesta em um coeficiente de aglomeração relativamente elevado. Verificamos também que essa rede apresenta uma estrutura de comunidades não-trivial que, embora tenha fortes restrições espaciais, não se limita às divisões geográficas estaduais do Brasil. Contrariamente, argumentamos que essas regiões identificadas naturalmente a partir da estrutura da rede refletem as complexas relações do mercado de trabalho brasileiro. Analisamos ainda as distribuições de tamanho, tanto espacial quanto populacional, das comunidades bem como identificamos uma associação sublinear entre essas duas quantidades. Finalmente, esperamos que nossos resultados contribuam com a criação de políticas públicas que estimulem a migração de pessoas para regiões com menor desenvolvimento econômico e motivem estudos mais aprofundados sobre como se dá a relação econômica entre as cidades.

Palavras-chave: Análise de Dados. Física Estatística. Sistemas Complexos. Redes Complexas. Cidades.

Abstract

This work investigates various aspects of the commuting network between Brazilian cities. Our primary objective was to understand the fundamental characteristics of this complex network and verify its community structure. To achieve this, we first reviewed the fundamental concepts of the theory of complex networks and studied an algorithm for identifying network communities, known as Infomap. Additionally, we presented some standard computational tools for data analysis and complex network analysis. Among our most significant empirical findings, we highlight that the Brazilian commuting network exhibits a power-law distribution with an exponent close to unity, demonstrating the existence of a few highly-connected cities (hubs) and a majority of municipalities with few connections. In addition to displaying scale-free characteristics, we observed that the commuting network has small-world features, reflecting a relatively high clustering coefficient. We also discovered that this network exhibits a non-trivial community structure that, despite having solid spatial restrictions, is not limited to the state geographic divisions of Brazil. Conversely, we argue that the regions naturally identified from the network structure reflect the complex relationships of the Brazilian labor market. We also analyzed the size distributions of the communities, both in terms of their spatial and population sizes, and identified a sublinear association between these two quantities. Finally, we hope that our findings will contribute to the creation of public policies that encourage people to migrate to regions with less economic development and encourage more in-depth studies of the financial relationships between cities.

Keywords: Data Analysis. Statistical Physics. Complex systems. Complex networks. Cities.

Sumário

CAPÍTULO 1

Introdução

O estudo de redes complexas é um dos temas mais atuais da física estatística, e tem sido aplicado em uma ampla gama de problemas do mundo real. Um dos exemplos mais claros de uma rede complexa é a Rede Mundial de Computadores, que conecta um número surpreendente de páginas web por meio de *hiperlinks*. Em 1998, Watts e Strogatz [1] estudaram redes neurais, colaborações de atores e transmissão de energia elétrica, descobrindo características comuns a todos esses fenômenos, como o efeito de mundo pequeno e o alto coeficiente de aglomeração. Já em 1999, Albert, Jeong e Barabási [2] modelaram a estrutura da Rede Mundial de Computadores como um grafo e perceberem que as conexões entre as páginas poderiam ser descritas por uma lei de potência. Essas características foram continuamente observadas em muitas redes complexas, permitindo o desenvolvimento de inúmeros algoritmos, muitos dos quais são usados tanto na pesquisa científica quanto no desenvolvimento de *software*.

Um desses algoritmos é o Infomap, um algoritmo de detecção de comunidades que foi desenvolvido por Rosvall e Bergstrom em 2008 [3]. Esse algoritmo é um método de particionamento de redes complexas, que consiste em dividir uma rede em subredes, ou comunidades, de modo que as conexões entre as comunidades sejam mínimas.

Em 2016, dois pesquisadores da Universidade de Stanford [4] compararam duas abordagens diferentes para compreender a formação de megaregiões no fluxo de viajantes para os Estados Unidos. A primeira abordagem usou um método heurístico, enquanto a segunda usou um algoritmo de particionamento. Os resultados foram encorajadores a favor da metodologia baseada em algoritmos computacionais. É nesse contexto de redes complexas que o presente trabalho está inserido e seu o principal objetivo é compreender alguns aspectos da rede de movimento pendular entre cidades brasileiras e, similarmente ao trabalho de Nel-

son e Rae [4], procurar identificar a formação de megaregiões no Brasil usando algoritmos computacionais de detecção de comunidades em redes complexas.

No que segue neste trabalho, apresentamos uma breve fundamentação teórica sobre conceitos de rede (Capítulo 2), os objetivos desta pesquisa (Capítulo 3), a metodologia que empregamos em nossas análises (Capítulo 4), nossos resultados (Capítulo 5) e a discussão dos mesmos (Capítulo 6). Por fim, encerramos este trabalho com algumas considerações finais (Capítulo 7).

CAPÍTULO 2

Fundamentação teórica

2.1 Redes complexas

De maneira simplificada, redes complexas são estruturas matemáticas que representam relações entre elementos [5]. Essas redes podem ser utilizadas para investigar, modelar e simular uma grande variedade de sistemas, como redes sociais, sistemas biológicos, infraestrutura de transporte e muitos outros [5]. O estudo de redes complexas é um campo interdisciplinar que utiliza teorias e métodos de matemática, física, computação e sociologia.

Redes complexas podem se distinguir por sua topologia, ou seja, a forma como os elementos estão conectados uns aos outros. Algumas propriedades comuns das redes complexas incluem a distribuição de grau, coeficiente de agrupamento, centralidade, entre outras [6].

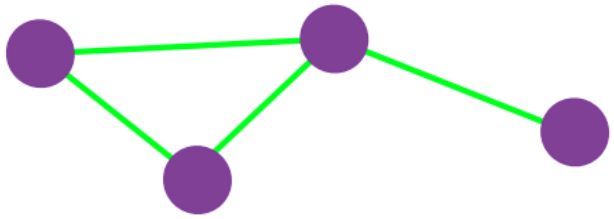
Entender as propriedades e os processos dinâmicos em redes complexas pode ter aplicações em diversos campos, como a previsão de falhas em sistemas, o estudo da evolução de doenças e a identificação de indivíduos críticos em redes sociais. Além disso, o estudo dessas redes também pode fornecer *insights* sobre a natureza da complexidade em sistemas e ajudar na criação de novas estratégias para solução de problemas [5].

2.2 Grafos

Grafos são objetos matemáticos usados para representar redes complexas e, consequentemente, relações entre objetos ou conceitos. Um grafo é composto por um conjunto de vértices ou nós e arestas ou ligações que conectam os nós. A Figura 2.1 mostra um exemplo simples de um grafo formado por 4 vértices e 4 ligações.

Uma maneira comum de representar um grafo é por meio da matriz adjacente, também

Figura 2.1: Ilustração de um grafo simples com 4 nós e 4 arestas.



Adaptado do livro *Network Science* de Barabási, Capítulo 2, Página 5, Figura 2.3 [7].

conhecida como matriz de adjacência. Nessa representação, a i-ésima linha e j-ésima coluna da matriz são iguais a 1 se há uma aresta entre os nós i e j, e 0 caso contrário. Assim, os elementos de uma matriz de adjacência podem ser escritos como

$$A_{ij} = \begin{cases} 1 \text{ se } i \text{ e } j \text{ estão conectados,} \\ 0 \text{ caso contrário.} \end{cases} \tag{2.1}$$

A matriz adjacente é uma maneira eficiente de representar grafos pequenos, mas não é adequada para grafos grandes, pois exige uma quantidade de memória proporcional ao quadrado do número de nós.

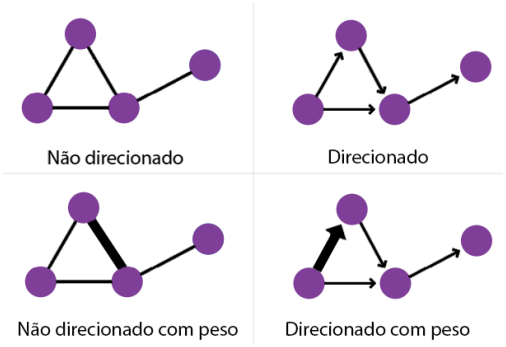
Além da matriz adjacente, um grafo também pode ser representado por uma matriz peso, que armazena o peso (ou custo) das arestas em vez de apenas indicar se existe ou não uma aresta entre dois nós, isto é,

$$P_{ij} = \begin{cases} \text{Peso da conexão se } i \text{ e } j \text{ estão conectados,} \\ 0 \text{ caso contrário.} \end{cases} \tag{2.2}$$

Esta representação é útil para modelar sistemas onde as arestas têm pesos, tais como sistemas de transporte, onde as arestas representam rotas e os pesos representam o tempo ou a distância dessas rotas.

Além disso, as arestas em um grafo podem ser direcionadas ou não-direcionadas. Arestas direcionadas representam relações unidirecionais entre nós, enquanto que arestas não-direcionadas representam relações bidirecionais. Em uma representação por matriz adjacente ou peso, arestas não-direcionadas são representadas por uma matriz simétrica [8], enquanto que uma matriz de adjacência relacionada a arestas direcionadas geralmente não apresenta essa propriedade. A Figura 2.2 ilustra os tipos comuns de redes complexas.

Figura 2.2: Ilustração de grafos direcionados e não direcionados, com e sem peso.

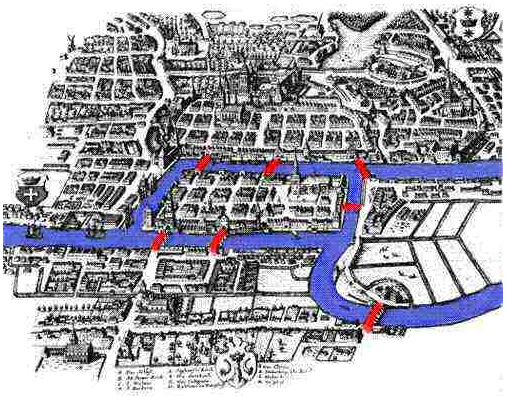


Adaptado do livro *Network Science* de Barabási, Capítulo 2, Página 12, Figura 2.5 [7].

2.3 O problema das pontes de Konigsberg

O problema das pontes de Konigsberg é considerado a origem da teoria dos grafos. A cidade de Konigsberg, na Prússia (hoje Kaliningrado, na Rússia), era dividida em duas ilhas e duas margens do rio Pregel, onde sete pontes se conectam a duas margens e duas ilhas. A Figura 2.3 representa essas pontes em um mapa da cidade. O problema de Konigsberg consiste em determinar se era possível atravessar cada uma das pontes uma única vez e retornar ao ponto de partida [9].

Figura 2.3: Ilustração das pontes de Konigsberg.

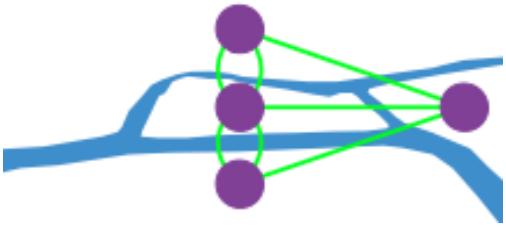


Disponível em: <https://mathshistory.st-andrews.ac.uk/Extras/Konigsberg>

Leonhard Euler (1707-1783) resolveu o problema aplicando seu conceito de teoria dos grafos. Para isso, ele representou a cidade de Konigsberg e suas sete pontes como um grafo, onde as ilhas e as margens eram os nós e as pontes eram as arestas. A Figura 2.4 mostra um diagrama com essa representação de Euler.

Euler então analisou as conexões entre os nós e fez uma importante observação: um caminho que cruza cada ponte uma única vez é possível apenas se todos os nós tiverem um grau par. O grau de um nó é o número de arestas que incidem nele e, nesse caso, todos os

Figura 2.4: Representação do problema das pontes de Konigsberg como um grafo.



Adaptado do livro *Network Science* de Barabási, Capítulo 2, Página 3, Figura 2.1 [7].

nós devem ter um grau par para que um caminho seja possível.

Entretanto, na cidade de Konigsberg, haviam dois nós com grau ímpar, o que significa que uma solução para o problema não era possível. Euler mostrou que esse era o caso usando raciocínio matemático e seguindo uma sequência lógica de passos. Ele demonstrou que uma solução não era possível e, ao fazer isso, forneceu as bases para o estudo da teoria dos grafos [7].

Até hoje, o trabalho de Euler sobre as pontes de Konigsberg continua sendo um dos exemplos mais clássicos do poder do raciocínio matemático e da aplicação da matemática a problemas do mundo real, além de ter levado ao estudo de problemas mais gerais, como a existência de caminhos eulerianos em grafos. Esses conceitos são aplicados em diversos campos, incluindo ciência da computação, pesquisa operacional e análise de redes [9].

2.4 Propriedades de redes complexas

2.4.1 Grau

O grau de uma rede complexa é uma medida do número de conexões que um nó (ou vértice) tem dentro de uma rede. É uma medida simples sobre a estrutura local de redes e pode ser usada para entender a influência e a centralidade de um nó específico. Matematicamente, o grau k_i de um nó i em uma rede pode ser representado usando a matriz de adjacência (A_{ij}) como [8]:

$$k_i = \sum_{j}^{N} A_{ij} \,, \tag{2.3}$$

com N sendo o número de vértices da rede.

Também é bastante comum a investigação da distribuição de probabilidade associada ao grau de todos os vértices de uma rede, a chamada distribuição de grau [5]. Uma rede com uma distribuição de grau que segue uma distribuição de potência, como redes de citações de artigos científicos [10], é normalmente denominada uma rede livre de escala. Por outro lado, uma rede com uma distribuição de grau que segue uma distribuição normal ou qualquer

distribuição com desvio padrão finito é usualmente considerada uma rede de escala fixa. Discutiremos mais sobre os tipos de redes nas próximas seções.

2.4.2 Coeficiente de agrupamento

O coeficiente de agrupamento fornece informações sobre a tendência de formação de triângulos em uma rede, ou seja, se os nós da rede tendem a estar conectados aos seus vizinhos próximos. Uma das formas de definir o coeficiente de agrupamento de um nó i é:

$$C_i = \frac{2E_i}{k_i(k_i - 1)}\,, \tag{2.4}$$

no qual E_i é o número de ligações entre os k_i vizinhos de i [5].

O coeficiente de agrupamento médio de uma rede é dado pelo valor médio de C_i, ou seja,

$$C = \frac{1}{N}\sum_{i=1}^{N} C_i \tag{2.5}$$

com N sendo o número de nós na rede. Esse coeficiente pode ser utilizado para comparar diferentes redes, permitindo uma compreensão mais profunda sobre suas estruturas. Por exemplo, redes com coeficiente de agrupamento médio elevado tendem a ter uma estrutura mais coesa e conexa, enquanto redes com coeficiente de agrupamento médio baixo tendem a ter uma estrutura mais dispersa ou esparsa [8].

2.4.3 Medidas de centralidade

De modo geral, uma medida de centralidade descreve a posição de um nó (ou vértice) na estrutura da rede. Em outras palavras, a centralidade mede a "importância" de um nó em relação à estrutura da rede. Existem vários tipos de centralidade, cada um com sua própria abordagem para quantificar o conceito relativo de "importância" de um nó na rede [8]. A seguir, apresentamos as principais medidas de centralidade [8].

Centralidade de grau

A centralidade de grau é o método mais simples para medir a centralidade em uma rede complexa. Ela é definida como o número de ligações (ou arestas) que um nó tem na rede, ou seja, ela é o grau do nó. Em outras palavras, quanto maior o número de ligações de um nó, maior é sua centralidade de grau. Assim, a centralidade de grau pode ser representada por

$$C_{grau}(i) = k_i\,, \tag{2.6}$$

na qual k_i é o grau do nó i.

Centralidade de Katz

A centralidade de Katz é uma medida de centralidade que se baseia na ideia de que um nó é importante se recebe ligações de outros nós importantes e também se faz ligações a outros nós importantes. Matematicamente, a centralidade de Katz pode ser representada por:

$$C_i = \alpha \sum_{j=1}^{n} A_{ij} C_j + \beta \,, \tag{2.7}$$

na qual A_{ij} é a matriz de adjacência da rede, α é um fator de escala e β é uma constante. Observamos que essa medida é definida de forma recursiva e conduz a uma centralidade mais sofisticada do que a centralidade de grau, pois leva em consideração não apenas o número de ligações de um nó, mas também a importância dos nós aos quais ele está ligado e dos nós aos quais ele faz ligações.

Centralidade de PageRank

A centralidade de PageRank é uma medida de centralidade baseada na teoria de um sistema de buscas na Internet. Similarmente ao caso anterior, essa medida se baseia na ideia de que um nó é importante se recebe ligações de outros nós importantes. Em outras palavras, quanto maior o número de ligações de outros nós importantes, maior é a centralidade de PageRank de um nó. Matematicamente, a centralidade de PageRank pode ser representada por:

$$C_{pagerank}(i) = \frac{1-\alpha}{N} + \alpha \sum_{j} \frac{C_{pagerank}(j)}{k_{out}(j)} \,, \tag{2.8}$$

na qual α é o chamado fator de amortecimento, N é o número total de nós na rede, $k_{out}(j)$ é o grau de saída do nó j e a somatória é realizada sobre todos os nós j que estão ligados ao nó i.

2.5 Tipos de redes complexas

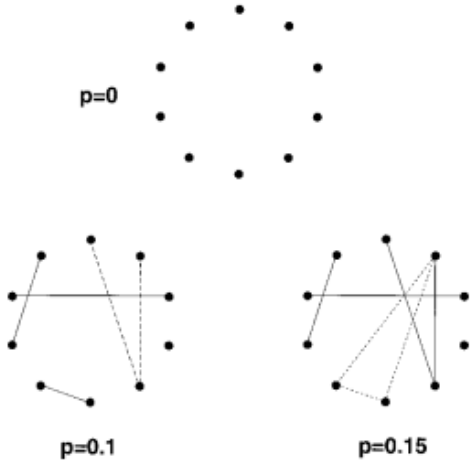
Nesta seção, apresentamos os principais tipos e modelos de redes complexas.

2.5.1 Redes aleatórias

O grafo de Erdös e Rényi é um modelo de geração de redes aleatórias baseado apenas na probabilidade de uma ligação existir entre dois nós. Dessa forma, partindo de um grupo de N nós desconexos, dois nós são escolhidos aleatoriamente e conectados de acordo com a probabilidade p, com cada par de nós sendo considerado apenas uma vez. Esse modelo resulta em uma rede com $pN(N-1)/2$ ligações [5] em média, sendo $N(N-1)/2$ o número

possíveis pares entre os N nós da rede. Consequentemente, a estrutura de uma rede aleatória gerada pelo modelo de Erdös-Rényi é altamente homogênea, pois todas as conexões têm a mesma probabilidade de ocorrer. A Figura 2.5 mostra alguns exemplos simples de redes geradas a partir do modelo de Erdös-Rényi com três valores para o parâmetro p.

Figura 2.5: Ilustração de uma rede aleatória conforme p aumenta.



Retirado de *Statistical Mechanics of Complex Networks* de Albert e Barabási [5].

O grau médio de um nó em uma rede aleatória de Erdös e Rényi é dado por:

$$\langle k \rangle = (N-1)p \tag{2.9}$$

sendo N o número de nós na rede e p a probabilidade de uma ligação existir entre dois nós [8].

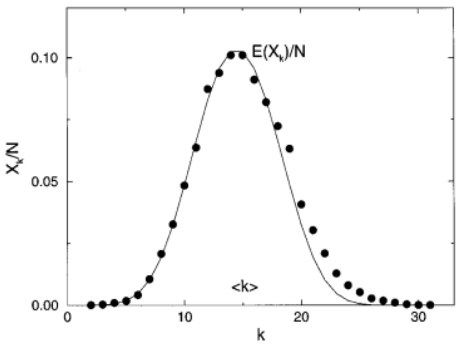
Se um nó específico no grafo aleatório tem conexão independente com probabilidade p a cada um dos demais nós, ou seja, $(N-1)$ nós, a probabilidade de estar conectado a k dos outros nós e não a nenhum outro é $p^k(1-p)^{N-1-k}$. Há $(N-1)$ formas de selecionar esses k outros nós, então a probabilidade total de ter conexão exata com k nós é [8]:

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}\,. \tag{2.10}$$

Essa é, portanto, a distribuição de grau de uma rede de Erdös-Rényi. Trata-se de uma distribuição binomial para o número de conexões de um nó. Assim, conforme o número de nós N aumenta e a média de conectividade é mantida constante, a distribuição do número de conexões tende a seguir a distribuição de Poisson, conforme ilustra a Figura 2.6.

Se considerarmos cada nó em uma rede aleatória juntamente com seus vizinhos, a probabilidade de que dois desses vizinhos também estejam conectados entre si é a mesma de dois nós escolhidos aleatoriamente na rede. Como resultado, o coeficiente de agrupamento

Figura 2.6: Distribuição de grau de uma simulação numérica de uma rede aleatória.



Retirado de *Statistical Mechanics of Complex Networks* de Albert e Barabási [5].

de um grafo aleatório é igual a:

$$C = \frac{\langle k \rangle}{(N-1)}\,. \tag{2.11}$$

O modelo de Erdös e Rényi também apresenta uma transição de fase, na qual, para valores de p maiores que um certo limiar p_c, o grafo tende a se tornar uma rede totalmente conectada, enquanto que, para valores de p menores que p_c, o grafo tende a ser composto por vários componentes isolados [5]. O valor do limiar p_c é, aproximadamente, $p_c \approx 1/N$.

Como sugere Newman [11], as redes do mundo real não são aleatórias e, consequentemente, o modelo é insuficiente para fornecer estruturas de rede realistas. Ainda assim, conhecer as características de uma rede aleatória e compará-las com aquelas de redes reais é um processo muito utilizado para caracterizar e compreender redes empíricas.

2.5.2 Redes de mundo pequeno

O modelo de Watts-Strogatz é a forma mais utilizada de se gerar uma rede de mundo pequeno e foi proposta por Duncan J. Watts e Steven H. Strogatz em 1998 [1]. O modelo é construído a partir de uma rede regular, ou seja, uma rede na qual cada nó tem o mesmo número de vizinhos e todos os nós estão conectados em uma sequência.

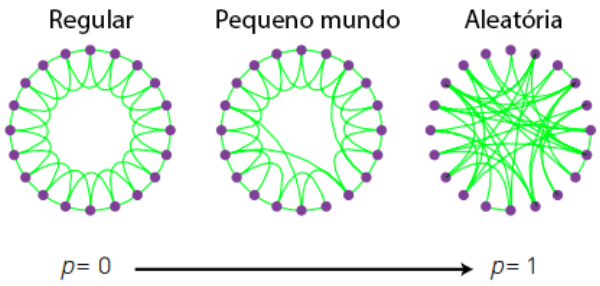
A ideia principal do modelo de Watts-Strogatz é reconectar algumas ligações de forma aleatória, de modo a gerar caminhos curtos (atalhos) e conexões aleatórias na rede. A reconexão é controlada por um parâmetro p, que representa a probabilidade de uma ligação ser reconectada.

Para construir uma rede pelo modelo de Watts-Strogatz, podemos seguir os seguintes passos:

- construir uma rede regular ($p = 0$), na qual os N nós e k vizinhos são conectados, formando $2k$ conexões em cada nó;
- em seguida, reconectar aleatoriamente cada aresta com probabilidade fixa p.

A Figura 2.7 mostra exemplos simples de redes geradas pelo modelo de Watts-Strogatz com diferentes valores para o parâmetro p.

Figura 2.7: Representação das redes do modelo de Watts-Strogatz organizadas em ordem crescente de p.



Adaptado de *Network Science* de Barabási, Capítulo 3, Página 28, Figura 3.14 [7].

O grau médio de um nó em uma rede de Watts-Strogatz é dado por [1]:

$$\langle k \rangle = 2k\,. \tag{2.12}$$

Como a rede regular possui $2k$ conexões, Barrat e Weigt [12] mostraram que o número de conexões entre esses nós vizinhos é $N_0 = 3k(k-1)/2$, quando $p = 0$. Dessa forma, o coeficiente de agrupamento nessa mesma condição fica

$$C_0 = \frac{3(k-1)}{2(2k-1)}\,. \tag{2.13}$$

No entanto, quando $p > 0$, dois vizinhos que estavam conectados quando $p = 0$ continuam conectados e são vizinhos do mesmo nó com probabilidade $(1-p)^3$ já que aquelas arestas devem permanecer conectadas. Logo, o coeficiente de agrupamento de uma rede de Watts-Strogatz, para qualquer valor de p é dado por

$$C = \frac{3(k-1)}{2(2k-1)}(1-p)^3\,. \tag{2.14}$$

Por sua vez, a distribuição de grau de uma rede de Watts-Strogatz é dada por [11]

$$p_j = \sum_{N=0}^{min(j-k,k)} \binom{k}{N} (1-p)^N p^{k-N} \frac{(pk)^{j-k-N}}{(j-k-N)!} e^{-pk}\,, \tag{2.15}$$

para $j \geq k$ e $p_j = 0$ para $j < k$.

O modelo de Watts-Strogatz permite controlar a quantidade de reconexões na rede e, assim, investigar como essa quantidade afeta as propriedades da rede. Por exemplo, ao

aumentar o valor de p, a quantidade de reconexões na rede aumenta, resultando em uma maior quantidade de caminhos curtos e conexões aleatórias na rede. Além disso, ao aumentar o valor de p, o diâmetro médio da rede tende a diminuir, o que também é uma característica de redes de mundo pequeno.

Esse modelo pode ser aplicado a vários tipos de redes, incluindo redes sociais, biológicas, econômicas, entre outras. Isso torna o modelo uma ferramenta útil para a compreensão de como as propriedades de mundo pequeno podem afetar a dinâmica em diferentes tipos de redes [5].

2.5.3 Redes livres de escala

Redes livres de escala são um tipo de rede com distribuição de grau seguindo uma função lei de potência. Essas estruturas são encontradas em muitos sistemas naturais, sociais e tecnológicos, incluindo a internet, a ciência e a economia [13]. O modelo de Barabási-Albert é um dos modelos mais conhecidos para a descrição de redes livres de escala e é amplamente utilizado para entender as propriedades estatísticas dessas redes [14].

O modelo de Barabási-Albert é baseado na hipótese de que redes livres de escala são formadas a partir de uma adição contínua de nós, com novos nós tendendo a se ligar a nós já existentes com mais ligações [15], um processo conhecido como ligação preferencial ou princípio de Mateus.

A distribuição de grau desse modelo é descrita por

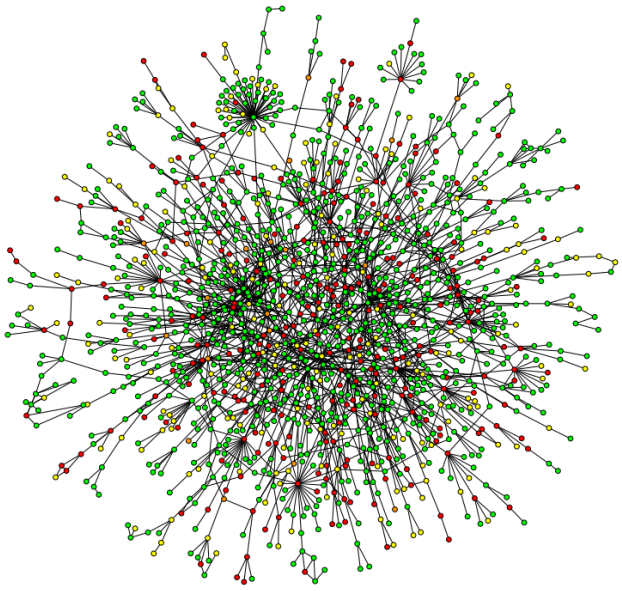
$$P(k) \propto k^{-\gamma}\,, \tag{2.16}$$

com $\gamma \approx 3$.

Em sua proposta original, esse modelo foi usado para explicar a estrutura da internet, isto é, como páginas da web tendem a ter muitos links a partir de páginas populares, mas poucos links a partir de páginas menos populares [16]. Além disso, ele também foi usado para entender a estrutura de redes biológicas, como as redes de interação proteína-proteína e ajudou a identificar proteínas-chave que desempenham um papel crítico em várias doenças [17]. A Figura 2.8 mostra um exemplo de rede de interação proteína-proteína, na qual podemos observar a existência de nós que recebem um grande número de conexões (localizados na parte mais central da visualização) e outros com um pequeno número de conexões (localizados na periferia da rede).

Por fim, é necessário salientar que os tipos de rede não são mutuamente exclusivos, ou seja, é possível que uma rede real apresente características de diferentes modelos de rede. Por exemplo, uma rede livre de escala pode exibir propriedades de mundo pequeno, como um alto coeficiente de agrupamento e um caminho médio curto.

Figura 2.8: Exemplo de uma rede livre de escala relacionada a interação proteína-proteína.



Retirado de *Lethality and centrality in protein networks* de Jeong, Hawoong e outros [18].

2.6 Lei de Zipf em redes complexas

A Lei de Zipf é uma lei matemática que foi originalmente usada para descrever a distribuição de frequência de palavras em um texto [19]. Ela foi proposta por George Kingsley Zipf em 1949 e é baseada na ideia de que a frequência de uma palavra é inversamente proporcional ao seu *ranking* em uma lista das palavras mais usadas [19].

A Lei de Zipf é frequentemente encontrada em diversas áreas, incluindo linguística, economia e ciência da informação [20]. A lei de Zipf também foi estudada em redes complexas, por exemplo, para descrever a distribuição de tamanho de componentes conectados em uma rede [21]. Nesse caso, a Lei de Zipf pode ser descrita matematicamente como

$$s_i \propto i^{-\alpha}\,, \tag{2.17}$$

na qual s_i é o tamanho da i-ésima componente conectada de uma rede complexa e α é o chamado expoente de Zipf [21].

Estudos têm mostrado que o valor de α geralmente é aproximadamente igual a 1 em redes complexas [21]. Isso significa que a maior componente conectada tem aproximadamente o mesmo tamanho que a soma de todas as demais componentes conectadas.

A Lei de Zipf em redes complexas também pode ser usada para descrever a distribuição de tamanho de comunidades em uma rede. De modo geral, uma comunidade é definida como um grupo de nós que estão fortemente conectados entre si, mas pouco conectados com nós de outras comunidades [22]. Existem vários métodos computacionais para identificar

comunidades em redes complexas. A seguir, apresentamos um desses métodos conhecido por Infomap [3].

2.7 Detecção de comunidades em redes com o Infomap

O algoritmo de *clustering* (agrupamento) Infomap é um método usado para identificar comunidades em redes complexas. Esse algoritmo foi desenvolvido por Martin Rosvall e Carl T. Bergstrom em 2008 [3]. O Infomap é baseado em teoria da informação e usa uma abordagem de codificação de fluxo de informação para identificar comunidades na rede.

A ideia básica do algoritmo é mapear o fluxo de informação da rede e identificar os grupos de nós que têm maior fluxo de informação interno do que externo. Esses grupos são considerados comunidades na rede. O algoritmo funciona da seguinte maneira:

1. Inicialmente, cada nó é considerado uma comunidade independente e, a partir desses nós, o algoritmo define um conjunto de caminhadas aleatórias na rede, com cada caminho representado por uma sequência de nós. A Figura 2.9 ilustra essas caminhadas aleatórias ou *random walks* sobre uma rede.

2. Em seguida, o algoritmo usa essas informações para calcular a probabilidade de transição de um nó para outro. Para fazer isso, contabiliza-se o número de vezes que cada nó é visitado em cada *random walk* e usa-se essa informação para construir uma matriz de probabilidade de transição.

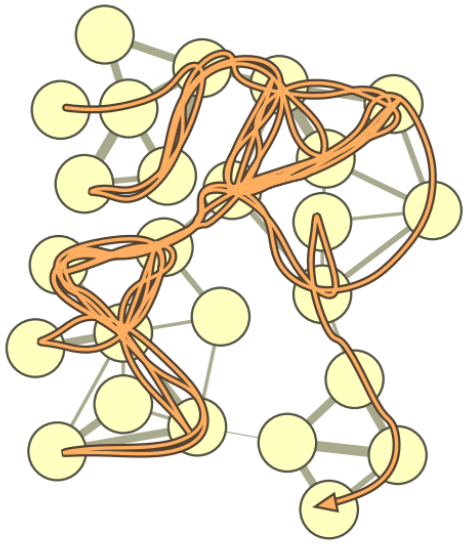
3. Usando essa matriz de probabilidade de transição, o algoritmo constrói uma representação de fluxo de informação na rede, na qual cada nó é representado por uma série de fluxos entrando e saindo dele.

4. O algoritmo então usa uma técnica de compressão de informação para nós em comunidades, a fim de minimizar a quantidade de informação necessária para representar o fluxo de informação na rede.

5. O processo de compressão de informação se repete várias vezes até que não sejam mais encontradas oportunidades de compressão.

O algoritmo Infomap também é capaz de detectar comunidades em múltiplos níveis, o que significa que é possível encontrar comunidades de diferentes tamanhos e hierarquias na rede. Para isso, o algoritmo começa detectando as comunidades no nível básico, como descrito anteriormente, e depois aplica a mesma técnica de compressão de informação aos grupos de nós inicialmente encontrados. Isso permite encontrar comunidades de níveis mais altos, que englobam comunidades de níveis mais baixos, ou seja, comunidades de comunidades.

Figura 2.9: Representação do *random walk* feito pelo algoritmo Infomap em uma rede.



Retirado de *Maps of random walks on complex networks reveal community structure* de Rosvall, Martin e Bergstrom [18].

O processo é repetido várias vezes, a fim de identificar comunidades em vários níveis de hierarquia. Ao final do processo, o algoritmo produz uma árvore hierárquica que representa a estrutura de comunidades na rede. Cada nível da árvore representa uma camada de comunidades de diferentes tamanhos [23].

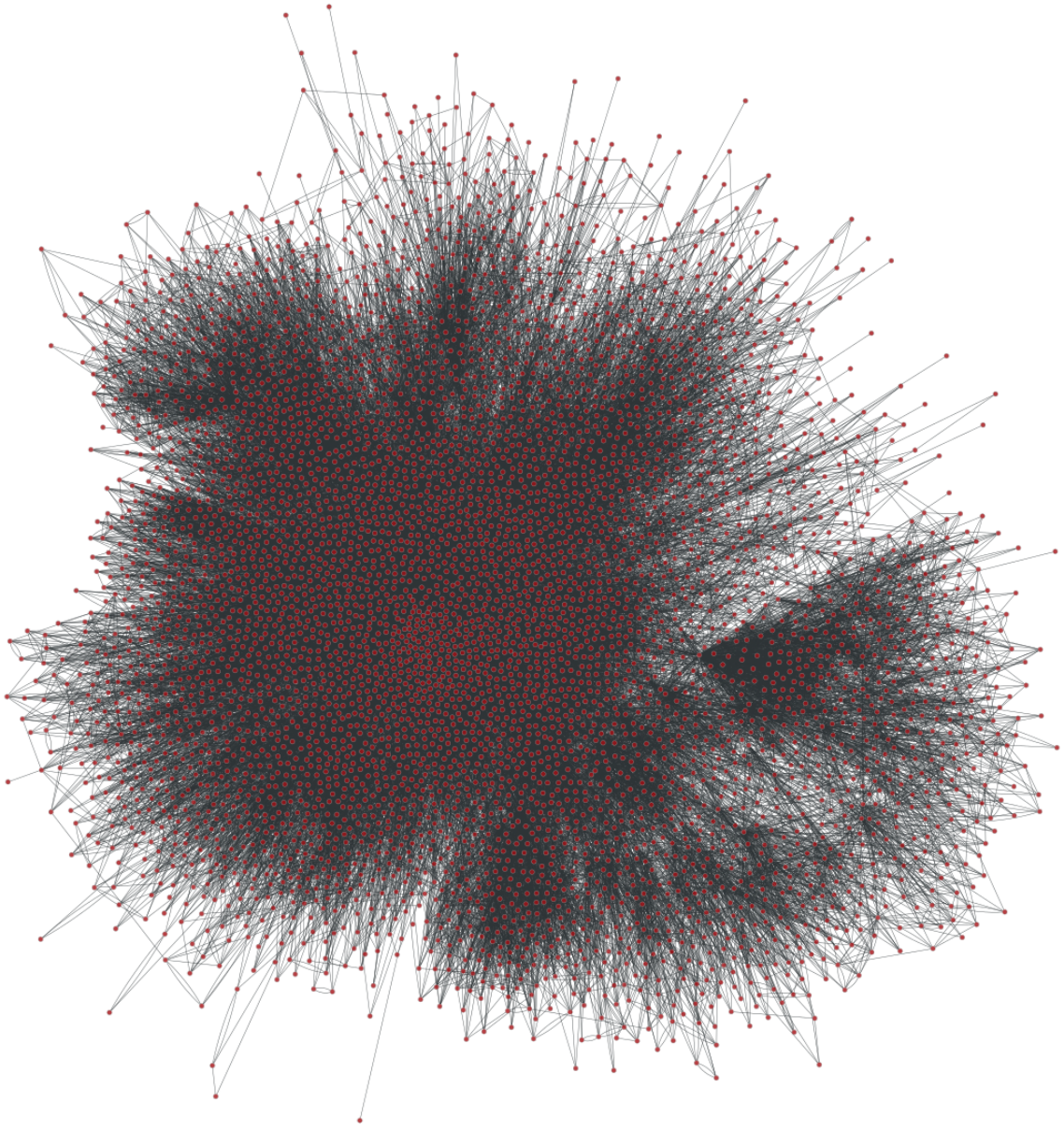
O algoritmo de detecção de comunidades do Infomap é considerado um dos métodos mais eficientes para identificação de comunidades em redes complexas, pois leva em consideração o fluxo de informação na rede ao invés de apenas considerar as conexões entre os nós. Além disso, o Infomap é uma ferramenta escalável e pode ser aplicado a redes de diferentes tamanhos e complexidades [3].

2.8 Rede de movimento pendular

Conforme já mencionamos, nesse estudo utilizamos uma rede de movimento pendular entre cidades brasileiras para identificar a formação de comunidades. Essas redes são exemplos de redes complexas que têm sido estudadas por pesquisadores em diversas áreas, como geografia, economia e sociologia. Nessas redes, por exemplo, os nós são as cidades brasileiras e as conexões entre elas indicam a quantidade de pessoas que se deslocam de uma cidade para outra para trabalhar. A Figura 2.10 ilustra essa rede.

A identificação de comunidades em redes de movimentos pendulares é útil para compreender a dinâmica dessas redes e prever sua evolução. As comunidades identificadas podem fornecer informações sobre as áreas de origem e destino dos trabalhadores, bem como sobre a distribuição de oportunidades de emprego.

Figura 2.10: Rede complexa dos movimentos pendulares entre as cidades do Brasil.



CAPÍTULO 3

Objetivos

3.1 Objetivos gerais

O objetivo geral desse trabalho de conclusão de curso é compreender as características de uma rede complexa relacionada ao movimento pendular de trabalhadores entre as cidades brasileiras. Para alcançar esse objetivo, nosso estudo irá identificar e analisar a formação de comunidades ou megaregiões a partir da análise empírica de redes e do uso de algoritmos de particionamento de redes complexas.

3.2 Objetivos específicos

Os objetivos específicos consistem em:

1. Construção da rede complexa de movimento pendular de trabalhadores entre as cidades brasileiras por meio de dados relevantes coletados de fontes confiáveis;
2. Análise das propriedades dessa rede a fim de caracterizar suas propriedades;
3. Identificação de comunidades ou mega-regiões presentes nessa rede a fim de entender como as cidades brasileiras estão conectadas entre si;
4. Análise das propriedades dessas comunidades da rede.

CAPÍTULO 4

Metodologia

4.1 Dados

Os dados utilizados nesse trabalho foram obtidos a partir do Censo de 2010 do Instituto Brasileiro de Geografia e Estatística (IBGE). Eles foram submetidos a um processo de tratamento adequado e separados em duas tabelas distintas. A primeira tabela consiste em informações gerais sobre as cidades brasileiras, incluindo suas características demográficas e geográficas. Já a segunda tabela concentra as características específicas para a construção da rede de movimento pendular.

A Tabela 4.1 ilustra as informações gerais sobre as cidades brasileiras. A tabela completa contém 5563 linhas de informações, cada uma representando um dos municípios brasileiros. As informações são divididas em nove colunas, que incluem:

- $CODE$: Código do município;
- $STATE$: Sigla do estado ao qual ela pertence;
- $CITY$: Nome do município;
- LON: Longitude;
- LAT: Latitude;
- X: Coordenada X;
- Y: Coordenada Y;
- POP: População;

- GDP: Produto Interno Bruto (PIB), valor de bens e serviços produzidos no município (não utilizado em nosso trabalho).

Essas informações foram úteis para categorizar as partições obtidas por meio dos algoritmos utilizados nesse trabalho. Além disso, esses dados foram fundamentais para a criação de gráficos, fornecendo, por exemplo, a localização das cidades.

Tabela 4.1: Ilustração das informações gerais sobre as cidades brasileiras usadas em nosso estudo.

CODE	STATE	CITY	LON	LAT	X	Y	POP	GDP
1100015	RO	ALTA FLORESTA D OESTE	-61.999824	-11.935540	608.909154	-1319.622430	24392	335644
1100023	RO	ARIQUEMES	-63.033269	-9.908463	496.352821	-1095.292033	90353	1293436
1100031	RO	CABIXI	-60.544314	-13.499763	765.828075	-1493.734191	6313	99399
1100049	RO	CACOAL	-61.442944	-11.433865	669.864550	-1264.410418	78574	1168442
1100056	RO	CEREJEIRAS	-60.818426	-13.195033	736.438268	-1459.731707	17029	272423
1100064	RO	COLORADO DO OESTE	-60.555067	-13.130564	765.065167	-1452.859684	18591	226177
1100072	RO	CORUMBIARA	-60.948701	-12.997520	722.490861	-1437.757543	8783	195054
1100080	RO	COSTA MARQUES	-64.231654	-12.436014	366.133198	-1375.077840	13678	135416
1100098	RO	ESPIGAO D OESTE	-61.020173	-11.528555	715.929229	-1275.168812	28729	366718
1100106	RO	GUAJARA MIRIM	-65.323952	-10.773884	245.865787	-1191.941671	41656	650142

Representação das 10 primeiras linhas da tabela.

Já a Tabela 4.2 ilustra as informações relacionadas ao fluxo de pessoas a trabalho entre as cidades brasileiras. A tabela completa possui 55243 linhas. As informações estão organizadas em três colunas distintas, que representam:

- $SOURCE$: Código do município onde as pessoas residem;
- $TARGET$: Código do município onde as pessoa trabalham;
- $WEIGHT$: Quantidade de pessoas que se deslocam diariamente a trabalho entre os dois municípios.

Essa tabela foi fundamental para a criação dos grafos, pois fornece as ligações entre os municípios.

Tabela 4.2: Ilustração da informações sobre os deslocamentos pendulares entre as cidades brasileiras.

source	target	weight
1100015	1100049	1.0
1100015	1100379	1.0
1100015	1101484	1.0
1100015	5107578	1.0
1100023	1100114	2.0
1100023	1100130	1.0
1100023	1100205	3.0
1100023	1100262	3.0
1100023	1100338	1.0
1100023	1100403	9.0

Representação das 10 primeiras linhas da tabela.

4.2 Linguagem de programação e bibliotecas de funções

Por se tratar de um trabalho que utiliza algoritmos computacionais, é importante especificar a linguagem de programação utilizada, bem como as bibliotecas e versões utilizadas.

Nesse trabalho, foi escolhido utilizar a linguagem Python (versão 3.9.7), principalmente pela sua praticidade e facilidade de compreensão. Embora essa linguagem possua limitações de velocidade em comparação com outras linguagens, esse problema é mitigado pela utilização de bibliotecas como a Pandas, que oferece ferramentas para manipulação eficiente e rápida de grandes quantidades de dados em forma tabular. Estas características tornam o Python uma das linguagens mais utilizadas no estudo de sistemas complexos.

Além disso, diversas bibliotecas foram utilizadas com vários objetivos ao longo desta pesquisa. Elas foram empregadas na criação e visualização de redes, bem como na elaboração de gráficos. Para facilitar a compreensão e acompanhamento do trabalho, listamos na Tabela 4.3 as bibliotecas que utilizamos em nosso trabalho.

4.3 Principais funções construídas ao longo desse trabalho

Em trabalhos que envolvem programação é comum a necessidade de repetir um algoritmo diversas vezes ao longo de uma análise, para isso é útil a criação de funções. Nesse trabalho, foram construídas diversas funções. A seguir, listamos apenas uma breve descrição de algumas funções que produzimos bem como o objetivo principal de cada uma delas:

Tabela 4.3: Tabela apresentando o nome e versão dos pacotes utilizados nesse trabalho.

Biblioteca	**Versão**	**Descrição**
Hull	1.0	Cálculo de envolventes convexas (disponível em `https://github.com/jsmolka/hull`).
SciPy	1.5.4	Fornece funções avançadas para álgebra linear, integração, otimização, etc.
NumPy	1.20.3	Computação científica com suporte a *arrays* multidimensionais.
Pandas	1.2.2	Visualização, manipulação e tratamento de dados em forma tabular.
Matplotlib	3.3.3	Ferramentas para produção de gráficos e outros tipos de visualizações.
NetworkX	2.5	Manipulação e análise de redes complexas.
Graph-Tool	2.34	Ferramentas de alto desempenho para visualização e análise de redes complexas.
Infomap	1.6.0	Análise de redes e detecção de comunidades.
Statsmodels	0.13.5	Modelagem estatística, modelagem de séries temporais, etc.
Seaborn	0.11.0	Visualização de dados estatísticos, interface de alto nível para a biblioteca Matplotlib.
Scikit-Learn	0.24.1	Aprendizado de máquina.
Shapely	1.6.4	Processamento de geometria espacial.

As versões das bibliotecas se referem às utilizadas na data da realização da pesquisa.

1. Criar um modelo de regressão linear para prever a relação entre duas variáveis, com opção de ser robusto ou comum, e retornar resultados incluindo coeficientes, incertezas e correlação entre as variáveis.

2. Criar uma planilha que contém informações sobre um determinado objeto do tipo Infomap e transformar as informações contidas nesse objeto em uma planilha de fácil manipulação.

3. Usar a biblioteca Infomap para encontrar as comunidades em uma rede. Retornar os resultados da análise.

4. Aplicar o algoritmo Infomap em um grafo e retornar as informações sobre os nós desse grafo em formato de planilha.

5. Agrupar os dados com base em uma coluna específica, calcular a área de uma envoltória côncava, salvar informações adicionais como coordenadas, códigos e população, e aplicar uma análise de regressão linear. O resultado final é uma planilha com informações adicionais, incluindo o resultado da análise de regressão linear.

4.4 Alguns procedimentos realizados para análise dos dados

4.4.1 Carregamento e análise das tabelas

Com o objetivo de verificar possíveis incongruências ou valores ausentes, as tabelas de dados foram analisadas usando a biblioteca Pandas do Python. Tanto a tabela de informações (Tabela 4.1) gerais quanto a tabela sobre os deslocamentos pendulares (Tabela 4.2) foram examinadas. Alguns poucos casos de incongruências (menos de 0.1% dos dados) ou

valores faltantes foram observados e essas linhas foram excluídas de nossas análises. Esse procedimento garante o funcionamento correto dos algoritmos de particionamento de outros que utilizamos.

4.4.2 Definição das regiões geográficas brasileiras

Foi criado um dicionário em Python para separar os viajantes a trabalho em tabelas distintas de acordo com suas regiões geográficas no Brasil. Esse passo foi fundamental para a construção das redes para cada região, permitindo uma análise mais focada em cada uma das regiões geográficas brasileiras bem como a observação de possíveis diferenças.

4.4.3 Construção e visualização das redes pendulares

Utilizando as bibliotecas NetworkX e Graph-Tool, as redes dos movimentos pendulares foram criadas inicialmente ao considerar os dados de todos os municípios brasileiros e, posteriormente, considerando as cidades em cada uma das cinco regiões geográficas do Brasil. Esse procedimento permitiu a visualização das redes pendulares, facilitando a compreensão de sua estrutura e topologia, bem como a utilização das mesmas como dados de entrada dos algoritmos de particionamento.

4.4.4 Obtenção das propriedades gerais das redes

Utilizando as bibliotecas NetworkX e Graph-Tool, foram obtidas informações gerais sobre as redes, como número de nós, número de arestas, grau médio e coeficiente de agrupamento. Além disso, também construímos um gráfico de distribuição de grau da rede e ajustamos essa distribuição empírica a um modelo lei de potência.

4.4.5 Identificação da estrutura de comunidades

Com o algoritmo de detecção de comunidades Infomap de múltiplos níveis, foi possível identificar as diferentes estruturas hierárquicas de comunidades da rede de movimento pendular entre cidades brasileiras. Esse procedimento foi o passo mais relevante para nosso estudo, pois possibilitou a identificação e a consequente análise das comunidades presentes na rede. Na seção de discussões, apresentaremos mais detalhes sobre essas análises.

4.4.6 Gráficos para análise das comunidades

A partir da estrutura de comunidades identificada, foram criados diversos gráficos. Dentre os quais destacamos:

1. Gráfico (Figura 5.4) para representar a distribuição dessas comunidades em seus vários níveis. Esse gráfico está ainda geo-localizado pelas coordenadas geográficas das cidades;

2. Gráfico (Figura 5.5) em escala semi-logarítmica (no eixo y) para representar o tamanho característico das comunidades em relação à raiz quadrada da área dessas comunidades;

3. Gráfico (Figura 5.6) em escala logarítmica do tamanho característico das comunidades em relação a população total nessas comunidades;

4. Gráfico (Figura 5.7) de dispersão para representar a relação entre raiz quadrada da área das comunidades e a população das mesmas.

Além disso, cada um desses gráficos foi confeccionado para cada uma das cinco regiões geográficas do Brasil (Figuras 5.8, 5.9, 5.10 e 5.11).

CAPÍTULO 5

Resultados

Por uma questão de organização e clareza, nossos resultados foram divididos em três seções. A primeira seção é referente ao estudo das propriedades gerais da rede de movimento pendular do Brasil como um todo. A segunda e terceira seções referem-se às análises sobre a estrutura de comunidades da rede de movimento pendular do Brasil como um todo, bem como das redes construídas para as cinco regiões geográficas do país. Além disso, nesse capítulo não haverá uma discussão aprofundada sobre nossos resultados, a qual será apresentada no capítulo seguinte.

5.1 Propriedades básicas da rede de movimento pendular brasileira

Buscando caracterizar a rede de movimento pendular brasileira, foram calculadas informações gerais sobre a rede. A Tabela 5.1 mostra os valores para o número de nós, o número de arestas, o grau médio e o coeficiente de agrupamento. Também foram identificados os nós mais influentes da rede, ou seja, aqueles que recebem (Tabela 5.2) ou exportam (Tabela 5.3) mais trabalhadores.

Tabela 5.1: Tabela contendo as propriedades gerais da rede de movimento pendular brasileira.

Propriedade	**Valor**
Número de nós	5551
Número de arestas	55244
Grau médio	19.9
Coeficiente de agrupamento	0.32

Tabela 5.2: Dez cidades que mais recebem trabalhadores de outros municípios diariamente.

Trabalhadores Recebidos	Cidade	Estado	População	GDP
68656	SAO PAULO	SP	11253503	443600102
39784	RIO DE JANEIRO	RJ	6320446	190249043
28389	BELO HORIZONTE	MG	2375151	51661760
22332	PORTO ALEGRE	RS	1409351	43038100
20522	CURITIBA	PR	1751907	53106497
19444	RECIFE	PE	1537704	30032003
15698	BRASILIA	DF	2570160	149906319
11312	GOIANIA	GO	1302001	24445744
11174	VITORIA	ES	327801	24969295
10020	CAMPINAS	SP	1080113	36688629

Tabela 5.3: Dez cidades que mais exportam trabalhadores para outros municípios diariamente.

Trabalhadores Exportados	Cidade	Estado	População	GDP
8492	SAO JOAO DE MERITI	RJ	458673	4826212
8301	CARAPICUIBA	SP	369584	3429411
7606	SAO GONCALO	RJ	999728	10340756
7538	MAUA	SP	417064	7352093
7419	RIBEIRAO DAS NEVES	MG	296317	1926219
6923	SAO PAULO	SP	11253503	443600102
6673	DIADEMA	SP	386089	11254523
6449	SAO VICENTE	SP	332445	3277443
6243	CARIACICA	ES	348738	4904147
6191	OLINDA	PE	377779	3108010

Além disso, também estimamos a distribuição de grau da rede (Figura 5.1) bem como a distribuição acumulada (CCDF) do grau (Figura 5.2). Para isso, consideramos apenas as ligações entre as cidades, ignorando os pesos e as direções das arestas da rede. No caso da CCDF, procuramos ajustar esses dados a uma distribuição lei de potência e encontramos um bom acordo com uma função lei de potência com expoente unitário.

Figura 5.1: Gráfico da distribuição empírica de grau da rede de movimento pendular brasileira.

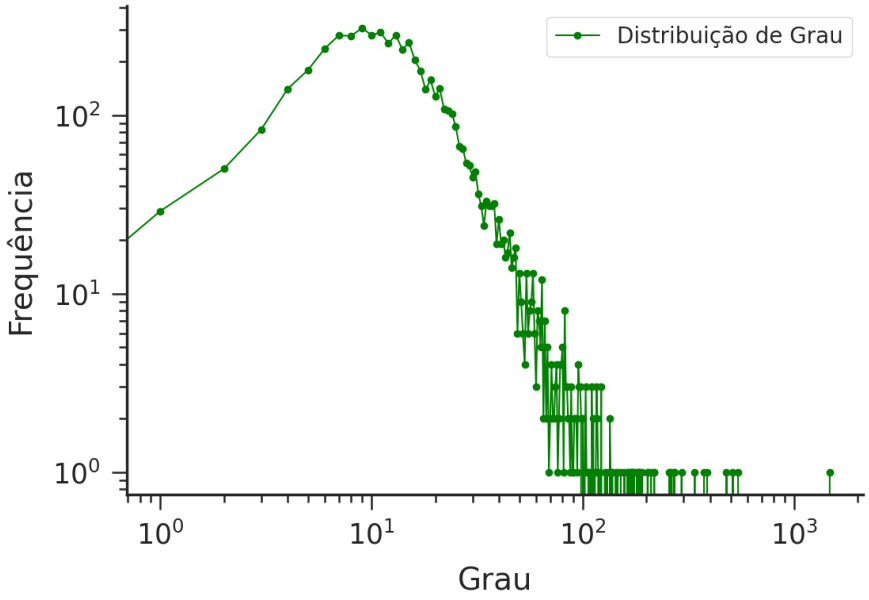
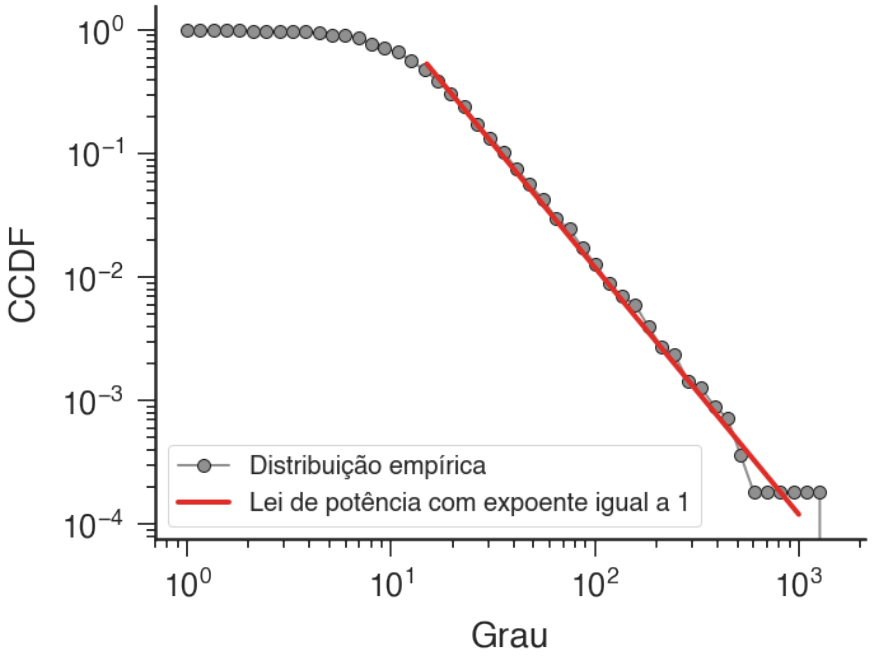


Figura 5.2: Gráfico da distribuição empírica acumulada (CCDF) de grau da rede de movimento pendular brasileira em comparação com uma distribuição lei de potência com expoente unitário.



5.2 Estrutura de comunidades da rede de movimento pendular do Brasil

Inicialmente foram detectadas as comunidades na rede ao considerar todos os municípios brasileiros. A Figura 5.4 mostra o nível mais baixo da hierarquia (*Level* 0) da estrutura de comunidades obtidas pelo algoritmo do Infomap, isto é, os maiores grupos. Por outro lado, a Figura 5.4 mostra a comunidades cada um dos níveis da hierarquia encontrada pelo Infomap. Além disso, a Tabela 5.4 mostra o número de comunidades detectadas em cada um dos níveis da hierarquia.

De modo geral, nesse estudo vamos nos concentrar principalmente nos nível mais alto da hierarquia (anterior ao nível das cidades), pois acreditamos que nesses níveis é possível encontrar informações mais detalhadas para o objetivo de nosso estudo.

Figura 5.3: Comunidades detectadas na rede de movimento pendular brasileira no nível mais baixo (*Level* 0) da hierarquia obtida pelo algoritimo Infomap.

Figura 5.4: Comunidades detectadas na rede de movimento pendular brasileira em cada um dos níveis hierárquicos obtidos pelo algoritimo Infomap.

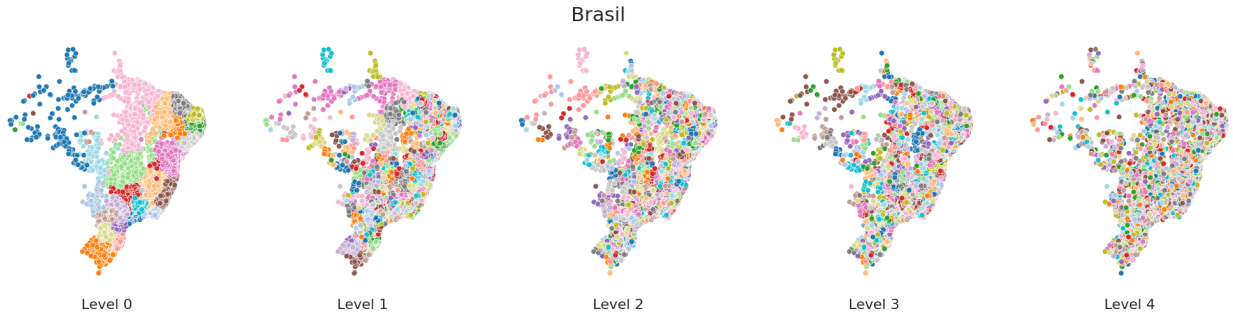


Tabela 5.4: Quantidade de comunidades detectadas na rede de movimento pendular brasileira em cada um dos níveis hierárquicos obtidos pelo algoritimo Infomap.

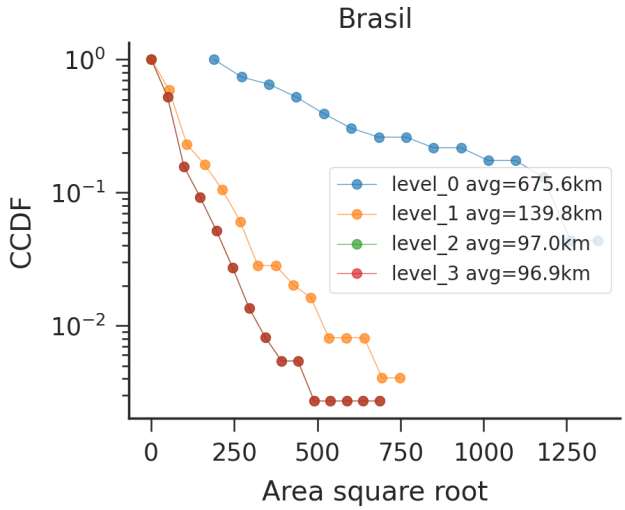
Nível	**Comunidades detectadas**
0	24
1	422
2	677
3	680
4	5551 (nível das cidades)

5.2.1 Resultados para a rede envolvendo todos os municípios brasileiros

Nessa seção serão apresentados os gráficos com os resultados de análises das comunidades detectadas no Brasil como um todo. Utilizou-se da função de distribuição cumulativa complementar (CCDF), a qual indica a probabilidade de que um evento (nesse caso, o tamanho da comunidade) exceda um determinado limite. No contexto de nosso estudo, a CCDF foi usada para determinar a probabilidade de que uma comunidade tenha um determinado tamanho, seja medido pela raiz quadrada da área do envelope que envolve todas as cidades de uma comunidade ou pela soma das populações das cidades em cada um das comunidades.

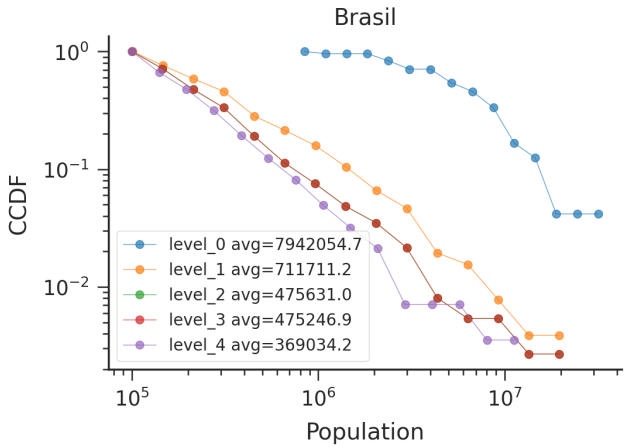
O gráfico da Figura 5.5 mostra a CCDF da raiz quadrada da área das comunidades detectadas.

Figura 5.5: CCDF do tamanho característico medido pela raiz quadrada da área das comunidades.



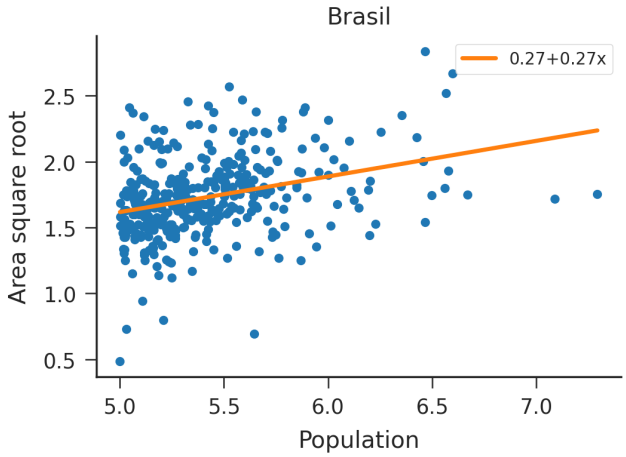
O gráfico da Figura 5.6 mostra a CCDF da população em cada uma das comunidades detectadas.

Figura 5.6: CCDF do tamanho característico medido pela pela população total das comunidades.



Por fim, o gráfico da Figura 5.7 mostra a relação entre a raiz quadrada da área e a população das comunidades, no qual as quantidades estão representadas após o cálculo do logaritmo na base 10. Além disso, a linha contínua mostra um ajuste linear aos dados, de modo que, é possível aproximar a raiz quadrada da área por uma potência da população das comunidades.

Figura 5.7: Relação entre a raiz quadrada da área e a população das comunidades no nível mais alto da hierarquia (anterior ao nível das cidades) do Infomap.



5.3 Estrutura de comunidades da rede de movimento pendular de cada região geográfica do Brasil

Nessa seção, serão apresentados os resultados referentes às cinco regiões geográficas do Brasil, bem como os gráficos criados para analisar as comunidades detectadas. O objetivo de refazer esses gráficos para cada uma das regiões é tentar entender se existem diferenças características entre cada região geográfica.

5.3.1 Resultados para a rede envolvendo cada uma das regiões brasileiras

A Figura 5.8 mostra a estrutura de comunidades detectada pelo Infomap para cada uma das cinco regiões geográficas do Brasil. Além disso, a Tabela 5.5 mostra o número de comunidades detectadas em cada um dos níveis da hierarquia para cada uma das regiões brasileiras. Fica evidente que a divisão estadual não é um fator definitivo para a estrutura das comunidades, pois é possível observar que grande parte das comunidades se estendem por mais de um estado. Essa observação será discutida de maneira mais detalhada no capítulo seguinte.

Figura 5.8: Comunidades detectadas nas redes de movimento pendular em cada uma das cinco regiões geográficas brasileiras e em cada um dos níveis hierárquicos obtidos pelo algoritimo Infomap.

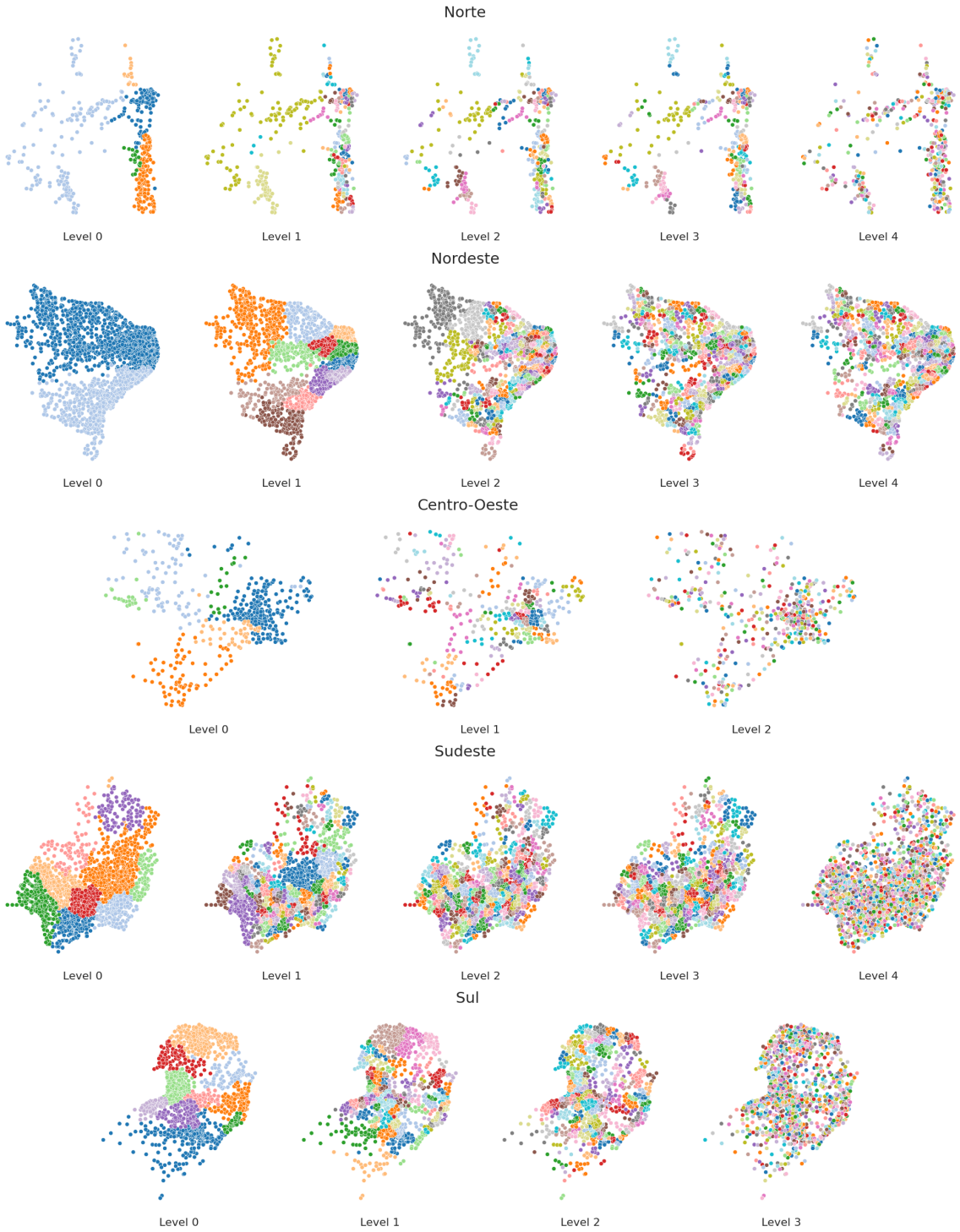


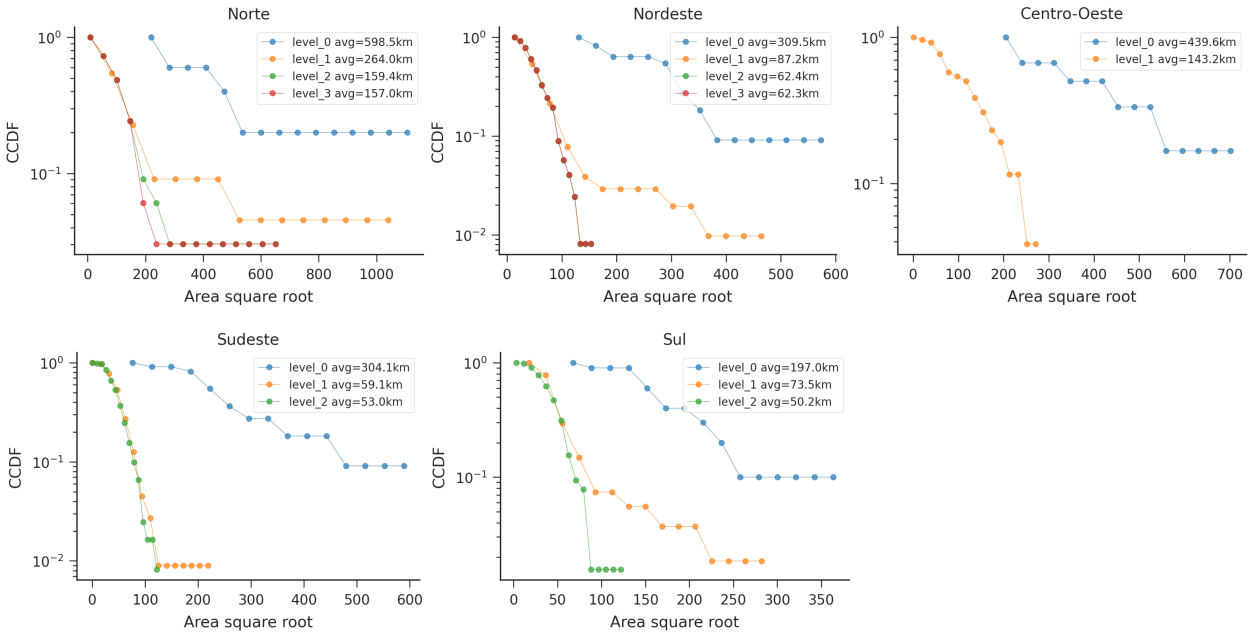
Tabela 5.5: Quantidade de comunidades detectadas na rede de movimento pendular brasileira de cada região geográfica do Brasil e para cada um dos níveis hierárquicos obtidos pelo algoritimo Infomap.

Nível	Norte	Nordeste	Centro-Oeste	Sudeste	Sul
0	5	2	6	9	10
1	47	12	71	126	95
2	64	166	460	202	154
3	65	222	N/D	207	1188
4	437	225	N/D	1666	N/D

N/D: Não detectado, indica que não foram detectadas comunidades em um determinado nível.

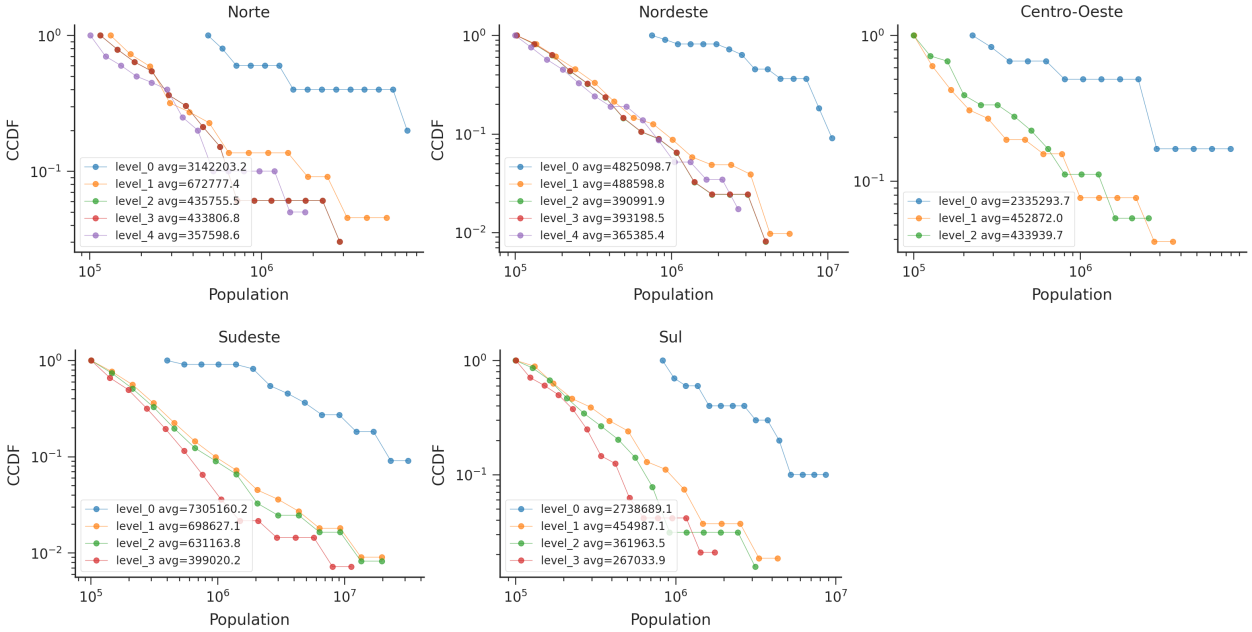
O gráfico da Figura 5.9 mostra a CCDF da raiz quadrada da área das comunidades em cada um dos níveis obtidos pelo Infomap e em cada uma das cinco regiões brasileiras.

Figura 5.9: CCDF da raiz quadrada da área das comunidades em cada região geográfica do Brasil.



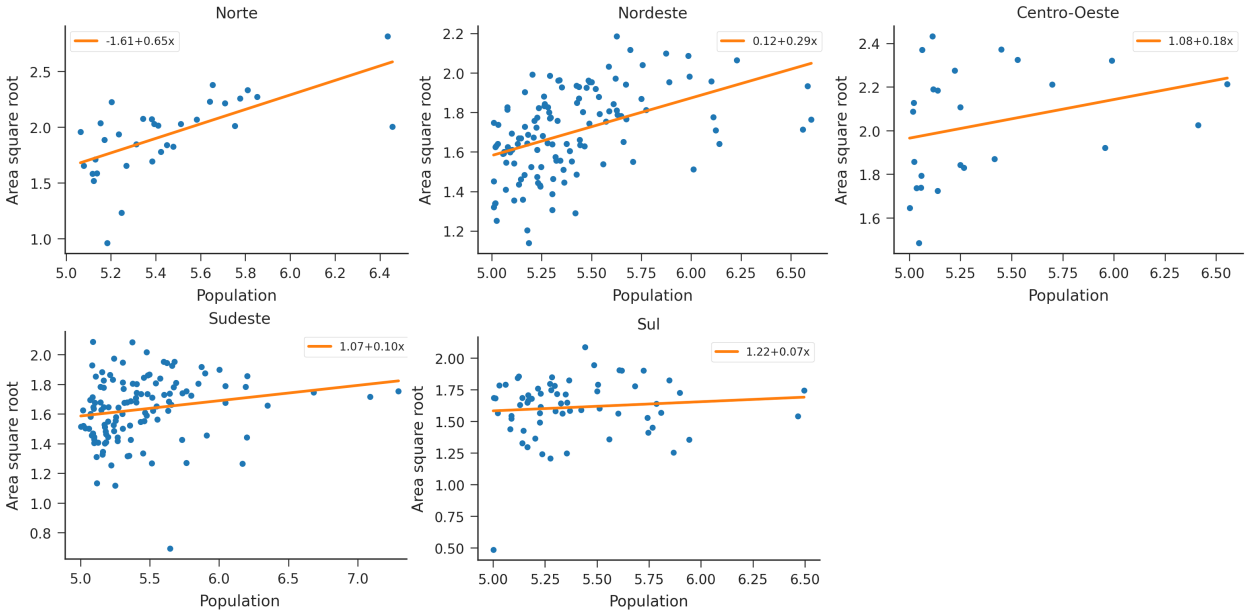
O gráfico da Figura 5.10 mostra a CCDF da população total das comunidades em cada um dos níveis obtidos pelo Infomap e em cada uma das cinco regiões brasileiras.

Figura 5.10: CCDF da população total das comunidades em cada região geográfica do Brasil.



Por fim, os gráficos da Figura 5.11 mostram as relações entre a raiz quadrada da área e a população das comunidades de cada região brasileira. Essas quantidades estão representadas após o cálculo do logaritmo na base 10. Além disso, as linhas contínuas mostram ajuste lineares aos dados de cada região, de modo que, é possível aproximar raiz a quadrada da área por uma potência da população das comunidades.

Figura 5.11: Relação entre a raiz quadrada da área e população das comunidades no nível mais alto da hierarquia (anterior ao nível das cidades) do Infomap. Cada gráfico mostra a associação entre as duas quantidades em cada região do Brasil.



CAPÍTULO 6

Discussão

Na tentativa de manter nossa discussão clara, esse capítulo será dividido em duas seções. A primeira seção será dedicada à discussão das características gerais da rede. Já a segunda seção será dedicada à discussão das características geográficas, o que inclui as comunidades detectadas e os municípios de maior influência.

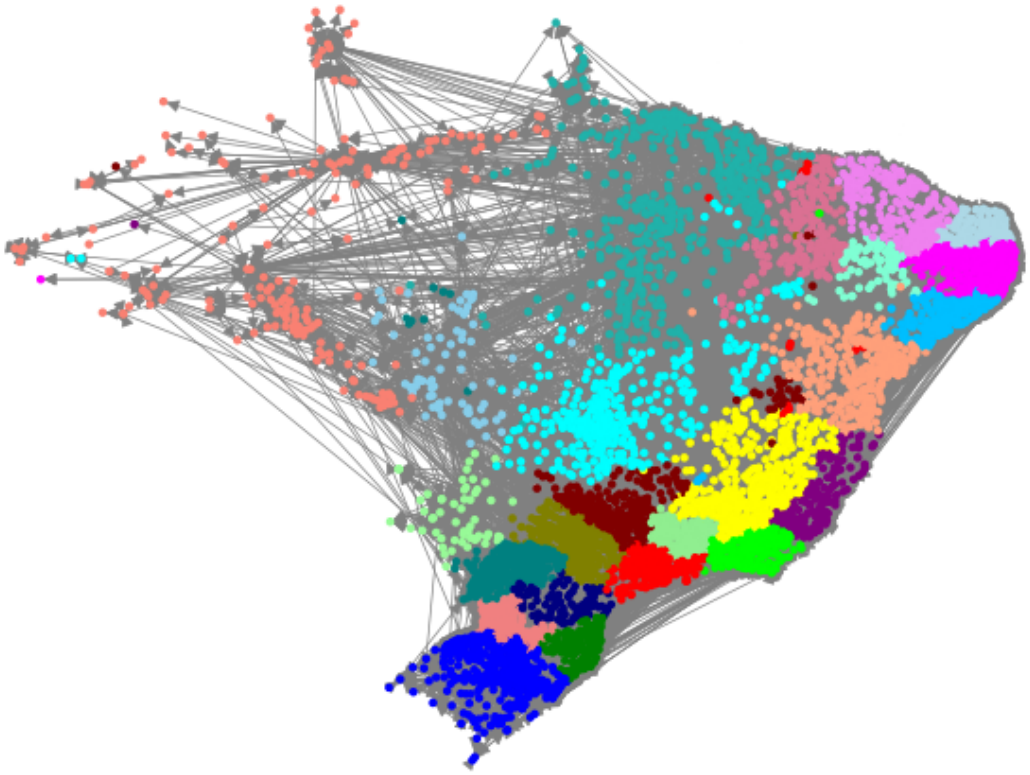
6.1 Características gerais da rede de movimento pendular

A primeira observação a ser feita sobre a rede de movimento pendular brasileira é que, por ser uma rede de viajantes a trabalho, ela é caracterizada por arestas direcionadas, já que as informações fornecem a direção do fluxo de trabalhadores. Além disso, as arestas da rede são ponderadas, uma vez que o peso de cada aresta representa a quantidade de trabalhadores que viajam entre as cidades de origem e destino.

Ao observar a tabela de propriedades gerais (Tabela 5.1), podemos notar que o número de nós é de 5551, esse número representa o número de municípios considerados no estudo. Enquanto o número de arestas é de 55244, indicando a rede de movimento pendular brasileira é relativamente densa. A Figura 6.1 mostra uma visualização geoespacial dessa rede, na qual as cores correspondem às comunidades no nível mais baixo da hierarquia Infomap.

Conforme já discutimos, o coeficiente de agrupamento mede a probabilidade de dois vizinhos de um nó estarem conectados entre si. A rede de movimento pendular brasileira tem um coeficiente de agrupamento igual a 0.32, um valor relativamente alto. Esse resultado também está de acordo com o fato de termos encontrado diversas comunidades nessa rede.

Figura 6.1: Representação gráfica da rede de trabalhadores.



Por outro lado, o grau médio da rede é 19.9, novamente um valor relativamente alto. Entretanto, quando se observa a distribuição de grau dessa rede (Figura 5.2), fica claro que essa distribuição se parece com uma distribuição livre de escala e que pode ser razoavelmente aproximada por uma distribuição lei de potência com expoente da ordem de 1. O que significa que existem alguns nós com um grande número de conexões e muitos outros com poucas conexões. Essa é uma característica típica de redes livres de escala. Como vimos, os tipos de redes não são mutuamente exclusivos, isto é, uma rede livre de escala também pode ser considerada uma rede de mundo pequeno, o que acreditamos ser o caso da rede de movimento pendular brasileira, tendo em vista seu alto coeficiente de aglomeração.

6.2 Características geográficas

6.2.1 Municípios mais influentes

O termo "influente", nesse estudo, será usado para se referir à importância dos municípios em atrair e enviar trabalhadores. A Tabela 5.2 apresenta os municípios com maior capacidade de atrair trabalhadores e, portanto, mais influentes. Não surpreendente, verificamos que os maiores polos econômicos do país, como São Paulo, Rio de Janeiro e Belo Horizonte, são os municípios que recebem o maior número de trabalhadores de outras cidades. Já a Tabela 5.3 apresenta os municípios com maior capacidade de enviar trabalhadores. Nesse caso, verificamos que os municípios que mais enviam trabalhadores para outras cidades costumam ser aqueles nas proximidades de cidades mais influentes. Esse é o caso, por exemplo, de São João de Meriti e Carapicuíba, ambas cidades das regiões metropolitanas de São Paulo e Rio de Janeiro, respectivamente.

6.2.2 Análise das comunidades detectadas

Ao observar a Figura 5.3, podemos notar que as comunidades detectadas pelo algoritmo do Infomap não se limitam às divisões geográficas estaduais. Contrariamente, as divisões obtidas a partir da rede emergem naturalmente devido a outras divisões possivelmente relacionadas aos diferentes polos econômicos regionais e nacionais do Brasil. É interessante destacar que existem algumas áreas geograficamente distantes, mas que fazem parte da mesma comunidade.

A Figura 5.4 mostra como as comunidades se tornam mais localizadas com o aumento do nível hierárquico das partições obtidas pelo Infomap. Essa informação poderia ser útil para uma análise mais precisa da relação entre essas cidades, permitindo assim a compreensão dos indicadores que são responsáveis por essa relação. Essa característica fica ainda mais evidente ao observar a Figura 5.8, que apresenta a divisão por região do Brasil.

Os gráficos das Figuras 5.5 e 5.9 mostram a distribuição do tamanho característico das comunidades ao considerar todo o Brasil e cada região geográfica, conforme medido pela raiz quadrada da área das comunidades em cada nível do Infomap. Observamos que essas figuras estão em escala semi-logarítmica (o logaritmo está aplicado ao longo do eixo y). Portanto, o comportamento linear dos dados indica um decaimento exponencial. Esse padrão pode ser observado tanto ao considerar os dados do Brasil como um todo, quanto ao considerar cada região geográfica do Brasil. Como as comunidades representam grupos de cidades que trocam muitos trabalhadores entre si, a raiz quadrada da área dessas comunidades indica uma espécie de distância máxima para os deslocamentos dos trabalhadores em cada região. Sendo assim, o valor médio da raiz quadrada da área das comunidades (o qual também pode ser pensado como o único parâmetro da distribuição exponencial que aproxima a distribuição

empírica dessa quantidade) pode ser considerado uma medida da distância típica de interação entre as cidades. Na Tabela 6.1 mostramos esses valores típicos ao considerar os níveis mais altos da hierarquia Infomap para o Brasil como um todo e para as cidades em cada região geográfica brasileira.

Tabela 6.1: Distância típica de interação entre as cidades nos níveis mais altos da hierarquia Infomap para cada região do Brasil e para o país como um todo.

Região	**Brasil**	**Norte**	**Nordeste**	**Centro-Oeste**	**Sudeste**	**Sul**
Distância média de interação (km)	96.9	157.0	62.3	143.2	53.0	50.2

Notamos que a distância média de interação entre as cidades está em conformidade com o tamanho de cada região, exceto pela região Centro-Oeste. Essa medida também pode ser pensada como um indicador de integração entre as regiões. Nesse caso, valores menores mostram que os trabalhadores precisam realizar deslocamentos menores para encontrar oportunidades de emprego. Levando essa ideia adiante, podemos estabelecer que as regiões Sul, Sudeste e Nordeste apresentam um mercado de trabalho mais desenvolvido e integrado. Por outro lado, as regiões Norte e Centro-Oeste seriam menos integradas nesse quesito.

Por sua vez, os gráficos das Figuras 5.6 e 5.10 retratam a distribuição de população dessas comunidades, tanto para o Brasil como um todo, quanto para suas regiões geográficas. Notamos que esses gráficos estão em escala log-log e, portanto, a forma aproximadamente linear dessas distribuições é um indicativo de comportamento lei de potência, tal qual aquele que discutimos no contexto da lei de Zipf. Vale notar ainda que a lei de Zipf também foi verificada em inúmeros trabalhos relacionados à população de cidades ao redor do mundo. Nesse contexto, uma distribuição lei de potência das comunidades mostra que o agrupamento de cidades que compartilham de um mesmo mercado de trabalho ainda mantém essa característica das distribuições de populações urbanas, promovendo ainda mais universalidade para a descrição da lei de Zipf no contexto de cidades.

Por fim, os gráficos das Figuras 5.7 e 5.11 mostram que a associação entre a raiz da área das comunidades e a população desse grupos de cidades, tanto para o Brasil como um todo quanto para as suas regiões geográficas, é não linear. Novamente, como essas quantidades foram representadas em escala logarítmica, as relações aproximadamente lineares que observamos indicam que a raiz da área e população das comunidades estão relacionadas por uma função lei de potência ou uma relação alométrica. Nesses gráficos também incluímos um ajuste linear, de modo que o coeficiente angular desses ajustes representa o expoente da relação lei de potência entre a raiz da área e a população. Ao considerar o Brasil, temos um expoente aproximadamente igual a 0.27. Esse valor inferior a unidade representa uma associação sublinear, ou seja, um aumento de 1% da população é acompanhado por um aumento de apenas 0.27% na raiz da área das comunidades. Esse tipo de relação é muitas

vezes denominado de relação de economia de escala e mostra que comunidades mais populosas apresentam um tamanho espacial típico per capita menor do que comunidades com população pequena.

Ao analisar os resultados por região Brasileira, encontramos o mesmo padrão de relação lei de potência. Porém, cada região apresenta um coeficiente angular específico e, consequentemente, um expoente lei de potência para a relação alométrica. Observamos que o expoente alométrico para a região Norte (0.65) é significativamente maior que os valores obtidos para as demais regiões. Isso indica que no Norte o tamanho espacial típico das comunidades cresce muito mais rapidamente com a população do que nas demais regiões. No espectro oposto, temos as regiões Sul e Sudeste como os menores valores para esse expoente (0.07 e 0.10, respectivamente). Portanto, os tamanhos típicos das comunidades dessas duas são os menos afetados pelas respectivas populações das comunidades. Finalmente, as regiões Centro-Oeste e Nordeste apresentam valores intermediários para o expoente alométrico (0.18 e 0.29, respectivamente). Além disso, essas duas regiões apresentam relações alométricas parecidas com aquela observada ao considerar todos os municípios do Brasil em conjunto.

CAPÍTULO 7

Considerações finais

Apresentamos nesse trabalho uma análise da rede de movimento pendular entre as cidades brasileiras. Entre nossos principais resultados, destacamos que essa rede pode ser considerada como uma rede livre de escala e com características comuns a redes de mundo pequeno. Também pode-se destacar o fato dessa rede apresentar uma estrutura de comunidades intrincadas. Entre outras observações, a estrutura de comunidades que encontramos permitiu identificar a distância média de interação entre as cidades brasileiras como um todo e também as diferenças regionais com relação a essa quantidade.

Acreditamos que esses resultados podem fornecer uma base para a criação de políticas públicas que estimulem a migração de pessoas para regiões com menor desenvolvimento econômico, bem como para estudos mais aprofundados sobre como se dá a relação econômica entre as cidades e como essa relação pode ser aprimorada.

Importante ressaltar que este estudo foi realizado com base em dados do Censo do IBGE publicado em 2010. Portanto, existem dados mais atualizados e a iminente divulgação no novo Censo brasileiro pode proporcionar uma oportunidade ímpar para investigar a evolução temporal da rede de movimento pendular e também de suas estruturas de comunidades. Por fim, acreditamos que o desenvolvimento do presente trabalho proporcionou uma boa introdução aos conceitos de redes complexas e às principais ferramentas para a investigação empírica de redes. Conhecimento esse que pode tornar-se útil para futuras investigações no tema de redes pendulares ou, de modo geral, para o estudo de outros sistemas complexos.

Referências Bibliográficas

[1] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[2] R. Albert, H. Jeong, and A.-L. Barabási, "Diameter of the world-wide web," *nature*, vol. 401, no. 6749, pp. 130–131, 1999.

[3] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the national academy of sciences*, vol. 105, no. 4, pp. 1118–1123, 2008.

[4] G. D. Nelson and A. Rae, "An economic geography of the united states: From commutes to megaregions," *PloS one*, vol. 11, no. 11, p. e0166083, 2016.

[5] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, no. 1, p. 47, 2002.

[6] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," *Physics reports*, vol. 424, no. 4-5, pp. 175–308, 2006.

[7] A.-L. Barabási and M. Pósfai, *Network Science*. Cambridge University Press, 4th printing ed., 2017.

[8] M. Newman, *Networks: An Introduction*. Oxford University Press, 2010.

[9] D. B. West, *Introduction to Graph Theory*. Upper Saddle River, NJ: Prentice Hall, 2nd ed., 2000.

[10] D. J. D. S. Price, "Networks of scientific papers: The pattern of bibliographic references indicates the nature of the scientific research front.," *Science*, vol. 149, no. 3683, pp. 510–515, 1965.

[11] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[12] A. Barrat and M. Weigt, "On the properties of small-world network models," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 13, pp. 547–560, 2000.

[13] R. Albert, H. Jeong, and A.-L. Barabási, "Error and attack tolerance of complex networks," *nature*, vol. 406, no. 6794, pp. 378–382, 2000.

[14] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[15] A.-L. Barabási, R. Albert, and H. Jeong, "Scale-free characteristics of random networks: the topology of the world-wide web," *Physica A: statistical mechanics and its applications*, vol. 281, no. 1-4, pp. 69–77, 2000.

[16] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," *ACM SIGCOMM computer communication review*, vol. 29, no. 4, pp. 251–262, 1999.

[17] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, "The large-scale organization of metabolic networks," *Nature*, vol. 407, no. 6804, pp. 651–654, 2000.

[18] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.

[19] G. K. Zipf, *Human Behavior and the Principle of Least Effort*. Addison-Wesley, 1949.

[20] W. Li, "Random texts exhibit zipf's-law-like word frequency distribution," *IEEE Transactions on information theory*, vol. 38, no. 6, pp. 1842–1845, 1992.

[21] K.-I. Goh, B. Kahng, and D. Kim, "Universal behavior of load distribution in scale-free networks," *Physical review letters*, vol. 87, no. 27, p. 278701, 2001.

[22] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[23] M. Rosvall and C. T. Bergstrom, "Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems," *PloS one*, vol. 6, no. 4, p. e18209, 2011.