UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Detecção de propagadores influentes por aprendizado de máquina

Vitória de Camargo Bugada

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Vitória de Camargo Bugada

Detecção de propagadores influentes por aprendizado de máquina

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestra em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Thomas Kauê Dal'Maso Peron

USP – São Carlos
Abril de 2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

d278 de Camargo Bugada, Vitória
Detecção de propagadores influentes por aprendizado de máquina / Vitória de Camargo Bugada; orientador Thomas Kauê Dal?Maso Peron. -- São Carlos, 2025.
60 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. Redes complexas. 2. Propagadores influentes. 3. Aprendizado de máquina. I. Kauê Dal?Maso Peron, Thomas, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Vitória de Camargo Bugada

Detecting influential spreaders by machine learning

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Master Interagency Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Thomas Kauê Dal'Maso Peron

USP – São Carlos
April 2025

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Defesa de Dissertação de Mestrado da candidata Vitória de Camargo Bugada, realizada em 09/04/2025.

Comissão Julgadora:

Prof. Dr. Thomas Kauê Dal'Maso Peron (USP)

Prof. Dr. Paulino Ribeiro Villas Boas (EMBRAPA)

Prof. Dr. Ruben Interian Kovaliova (UNICAMP)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

AGRADECIMENTOS

Gostaria de expressar minha sincera gratidão aos meus pais, cujo apoio e incentivo foram fundamentais em toda a minha trajetória acadêmica. A sua confiança em meu potencial, a dedicação em me proporcionar as melhores oportunidades e o amor incondicional sempre me deram força para enfrentar os desafios. Sem vocês, este trabalho não seria possível.

Agradeço também aos meus professores, cujos ensinamentos, orientação e exemplo de profissionalismo me motivaram a buscar a excelência. Suas valiosas contribuições durante todo o processo de elaboração deste trabalho foram essenciais para o seu desenvolvimento e aprimoramento.

Por fim, o presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Código de Financiamento 001, ao qual sou imensamente grata pelo suporte e pela confiança em meu potencial de desenvolvimento acadêmico e científico.

A todos, muito obrigada!

RESUMO

BUGADA, V. C. **Detecção de propagadores influentes por aprendizado de máquina**. 2025. 60 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2025.

O estudo de redes complexas é essencial para compreender sistemas interconectados em áreas como biologia, sociologia e tecnologia. Essas redes, formadas por nós e arestas, revelam padrões emergentes importantes. A análise de redes permite otimizar intervenções e melhorar a resiliência de sistemas. A identificação de propagadores influentes em uma rede, como os responsáveis pela disseminação de doenças ou informações, é fundamental e depende da dinâmica do fenômeno. Diferentes abordagens, como a análise de k-shell e modelos de otimização, ajudam a prever esses propagadores. Este estudo busca analisar a relação entre medidas de centralidade e a capacidade de propagação em redes sociais e espaciais, propondo modelos de previsão para identificar os principais influenciadores e otimizar o controle de disseminações.

Palavras-chave: Estatística; Medidas de centralidade; Propagadores.

ABSTRACT

BUGADA, V. C. **Detecting influential spreaders by machine learning**. 2025. 60 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2025.

The study of complex networks is essential to understanding interconnected systems in areas such as biology, sociology and technology. These networks, made up of nodes and edges, reveal important emerging patterns. Network analysis allows you to optimize interventions and improve systems resilience. The identification of influential propagators in a network, such as those responsible for the spread of diseases or information, is fundamental and depends on the dynamics of the phenomenon. Different approaches such as k-shell analysis and optimization models help predict these propagators. This study seeks to analyze the relationship between centrality measures and the propagation capacity in social and spatial networks, proposing prediction models to identify the main influencers and optimize dissemination control.

Keywords: Statistic; Centrality measures; Propagators.

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

O estudo de redes complexas é crucial para entender a estrutura e a dinâmica de sistemas interconectados em diversas áreas, como biologia, sociologia, tecnologia e economia. Essas redes são compostas por múltiplos elementos (nós) conectados por interações (arestas) que, ao serem analisados em conjunto, revelam padrões e propriedades emergentes impossíveis de se perceber ao observar apenas os componentes isoladamente. Por exemplo, na epidemiologia, o estudo de redes ajuda a prever a propagação de doenças [Pastor-Satorras *et al.* 2015], enquanto em redes sociais, auxilia a entender a disseminação de informações e comportamentos. Além disso, em tecnologia, a análise de redes melhora a robustez de sistemas como a internet e redes elétricas. Compreender redes complexas permite otimizar intervenções, aumentar a resiliência de sistemas e melhorar o processo de tomada de decisão em cenários que envolvem múltiplas interações [Barabási 2018].

Redes complexas são relevantes em diversas áreas do conhecimento, de acordo com [Barabási 2018]. O autor apresenta como as redes são formadas e como essas estruturas revelam padrões que afetam diretamente o comportamento do sistema como um todo, explicando que as redes complexas são caracterizadas por uma organização que não é aleatória, mas segue regras específicas, como a presença de nós altamente conectados (hubs). Esses hubs desempenham um papel crucial na robustez e na vulnerabilidade da rede, o que torna o estudo dessas estruturas essencial para prever e mitigar falhas em sistemas interconectados.

Para entender a capacidade de propagação de um sistema e garantir que suas informações sejam difundidas de forma eficiente e controlada, é necessário identificar os mais influentes propagadores da rede subjacente. Esse é um dos desafios ao estudar um sistema e buscar entendê-lo em suas formas: sua estrutura, seu funcionamento e seus respectivos fatores influentes.

Com isso, diversos trabalhos têm surgido apontando que a identificação de propagadores influentes é dependente da dinâmica que está sendo analisada. Em outras palavras, a propagação de diferentes fenômenos, como doenças virais, boatos ou outras doenças, ocorre de maneiras

distintas, mesmo que compartilhem o conceito de propagação. Os propagadores de um fenômeno não necessariamente serão propagadores de outro, pois cada processo envolve fatores específicos. Então, uma pessoa pode espalhar um boato, mas não uma doença, ou vice-versa, devido a diferenças nas formas de transmissão e nas características dos fenômenos, sejam biológicas, sociais ou comunicacionais. Portanto, a dinâmica de propagação varia conforme o contexto.

Recentemente, [Kitsak *et al.* 2010] verificaram que os propagadores mais influentes podem ser previstos por meio da análise de decomposição de k-shell, em que os agentes estão dentro do núcleo da rede, ou seja, não precisa ser necessariamente o nó mais conectado. Dessa forma, encontraram o grau e a acessibilidade desse vértice, relacionando com a propagação da doença. Em contrapartida, [Baños, Borge-Holthoefer e Moreno 2013] identificaram que em modelos padrão de rumores não é possível encontrar os propagadores mais influentes usando as mesmas métricas.

A identificação dos principais propagadores em uma rede desempenha um papel crítico em diversas áreas, desde a epidemiologia até o marketing viral. Vários estudos notáveis contribuíram significativamente para este campo, desenvolvendo métodos e algoritmos que têm o poder de identificar nós influentes. Abaixo, destacamos alguns desses artigos-chave e discutimos suas contribuições.

Um dos trabalhos fundamentais é o estudo de [Kitsak *et al.* 2010], mencionado acima. Este estudo aborda a identificação de propagadores influentes, levando em consideração a dinâmica de propagação em redes reais. Os autores desenvolveram abordagens para identificar os nós que têm o maior impacto na disseminação de informações, ideias ou doenças em uma rede complexa.

Outro artigo relevante é o trabalho de [Kempe, Kleinberg e Tardos 2003]. Esse estudo apresentou uma abordagem algorítmica para identificar os principais propagadores em redes sociais. Eles formularam o problema de maximizar a influência na disseminação de informações como um problema de otimização e propuseram algoritmos eficientes para encontrá-lo, tornando-se um marco no campo da identificação de propagadores influentes.

Além desses artigos principais, há outras contribuições notáveis no campo de redes complexas relacionadas à identificação de propagadores influentes. Por exemplo, o estudo de [Adamic e Huberman 2001] que explorou a estrutura de redes de páginas da web e identificou os principais propagadores de informações na *World Wide Web*.

Outro artigo de destaque é o de [Leskovec *et al.* 2007] que introduziu métodos para detectar surtos em redes complexas, identificando nós influentes que podem atuar como indicadores de eventos significativos.

Em resumo, a identificação de propagadores influentes em redes complexas é um campo de pesquisa crucial em várias disciplinas. Os artigos mencionados, juntamente com outros estudos notáveis, ajudaram a desenvolver métodos e algoritmos que desempenham um papel

vital na compreensão e no controle da propagação de informações e doenças em redes complexas do mundo real.

Neste estudo, realizaremos uma série de experimentos e análises com o objetivo de identificar os principais propagadores em uma rede complexa (ou em um conjunto de redes complexas). Utilizaremos as principais métricas de redes complexas (em nível de nó) para investigar quais dessas medidas estão mais associadas à capacidade de um nó em disseminar informações dentro da rede. Para tal, definiremos a capacidade de propagação de um nó como a fração final de nós recuperados no modelo SIR, considerando que a epidemia se inicia no nó em questão. Mais especificamente, iremos iniciar a propagação 500 vezes no mesmo nó e depois fazemos a média desses 500 valores para avaliar a capacidade de transmissão desse nó. Neste contexto, os termos "informação"e "epidemia"serão tratados como sinônimos, considerando que ambos referem-se ao processo de propagação e disseminação de um fenômeno na rede.

Para uma dada rede, iremos construir um *pipeline* que transforma as informações estruturais em uma tabela, onde a matriz X representa as covariáveis (*features*) extraídas de cada nó da rede, sendo que o elemento X_{ij} indica a *j-ésima* variável preditora (como grau, centralidade ou clustering) associada ao *i-ésimo* nó. O vetor de resposta y corresponderá à fração final de nós recuperados em um processo dinâmico de disseminação (como a recuperação de uma epidemia ou propagação de informação). O objetivo será formular o problema de regressão linear simples, onde buscamos modelar a relação entre as features da rede e o comportamento observado na fração de recuperados. Além disso, será implementado um estudo de *feature importance* utilizando modelo baseado em árvores de decisão, como Random Forest, para identificar as variáveis mais influentes na previsão da fração final de recuperados. Esse estudo permitirá avaliar o impacto relativo das diversas medidas de centralidade e conectividade no desfecho da propagação na rede.

Em resumo, existem diversos trabalhos a respeito de propagadores influentes em redes e como detectá-los, mas não se pode afirmar uma generalização e uma forma única de como determiná-los. Dessa forma, aqui analisaremos a relação de diferentes medidas de centralidade e suas respectivas capacidades de identificar os principais propagadores para redes do tipo espacial (aeroportos) e redes sociais. Criaremos modelos de previsão, onde a variável resposta será o número de infectados ao final da simulação, e enquanto as covariáveis serão medidas de centralidade da rede. Ainda, vale destacar que os trabalhos anteriores mencionados se concentram no efeito de centralidades específicas, enquanto neste trabalho analisamos o poder preditivo de várias medidas conjuntamente.

1.1 Estrutura do trabalho

Na primeira parte desse trabalho, será apresentada uma introdução ao tema, destacando os principais referenciais teóricos que abordam a questão em análise e justificando a relevância

da pesquisa na área. Após isso, serão analisadas quatro redes complexas utilizadas no estudo. Inicialmente, será investigada a correlação entre as medidas aplicadas e a fração final de indivíduos recuperados em cada rede. Em sequência, será realizada uma análise de regressão para prever a fração final de recuperados, utilizando dois modelos de regressão: a regressão linear e o Random Forest.

A partir dessas análises, será possível comparar os resultados obtidos para cada rede, assim como os métodos de modelagem utilizados, permitindo uma avaliação detalhada de cada caso.

CAPÍTULO

2

CONCEITOS BÁSICOS DE REDES COMPLEXAS

Uma rede complexa é um conceito fundamental na teoria dos sistemas complexos e na ciência de redes [Newman 2010]. Em sua essência, redes complexas são representações de sistemas compostos por elementos, denominados nós ou vértices, que estão interligados por meio de conexões, referidas como arestas ou links. Esse domínio de pesquisa envolve uma análise abrangente dessas estruturas a fim de identificar e caracterizar padrões emergentes e propriedades notáveis.

A conectividade é uma propriedade essencial para a caracterização das redes complexas, pois descreve como os componentes de uma rede estão conectados uns aos outros, influenciando a capacidade de comunicação e interação entre os nós, evidenciando uma topologia não trivial. Em algumas redes reais, alguns nós possuem uma quantidade significativamente maior de conexões em comparação com a maioria dos outros, o que é frequentemente descrito pela presença de uma distribuição de grau que segue uma lei de potência. Essa característica é fundamental para a identificação de hubs, nós altamente conectados que desempenham um papel crucial na rede. [Barabási 2018] explica detalhadamente a topologia das redes complexas e a presença de nós altamente conectados, ou hubs, que resultam em uma distribuição de grau que segue a lei de potência. Ele introduz o conceito de redes "livres de escala", onde a maioria dos nós tem poucas conexões, enquanto poucos nós têm muitas, formando a base do estudo de redes complexas.

Um fenômeno notável associado às redes complexas é o chamado "pequeno mundo", onde a distância média entre os nós é surpreendentemente curta, apesar do grande número de nós presentes na rede [Watts e Strogatz 1998]. Isso resulta em uma rápida disseminação de informações ou influência através da rede, caracterizando uma propriedade emergente.

O fenômeno do "pequeno mundo"em redes complexas foi introduzido por [Watts e Strogatz 1998] onde demonstram que redes com muitos nós podem ter uma distância média

surpreendentemente curta entre eles. Esse conceito revela que, mesmo em redes grandes, como redes sociais ou a internet, a maioria dos nós pode ser alcançada por meio de poucos passos. Essa característica facilita a rápida disseminação de informações ou influências dentro da rede, tornando-a altamente eficiente. O fenômeno ocorre devido a uma combinação de alta conectividade local entre nós vizinhos e a existência de algumas conexões de longo alcance, que reduzem drasticamente as distâncias globais.

Além disso, as redes complexas frequentemente exibem alta capacidade de agrupamento (clustering), indicando a formação de comunidades ou grupos de nós altamente interconectados. Esses agrupamentos podem ser cruciais para a funcionalidade e resiliência da rede.

A resiliência é outra característica proeminente de muitas redes complexas, o que significa que essas redes podem resistir a falhas ou ataques direcionados sem uma perda catastrófica de conectividade. A redundância e a robustez na estrutura da rede desempenham um papel importante nesse aspecto. Em suma, a resiliência em redes complexas refere-se à capacidade de uma rede de manter sua conectividade e funcionalidade mesmo diante de falhas ou ataques, especialmente quando essas falhas ocorrem em alguns de seus nós ou conexões. Redes complexas, como as de comunicação, infraestrutura ou biológicas, frequentemente exibem essa característica, permitindo que continuem operando, ainda que algumas partes da rede sejam comprometidas.

Segundo [Cohen e Havlin 2010], essa robustez decorre principalmente da distribuição de grau assimétrica observada em muitas redes complexas, como as redes "livres de escala". Nesses sistemas, a maioria dos nós tem poucas conexões, enquanto poucos nós, chamados de hubs, possuem um grande número de conexões. Isso implica que, mesmo que alguns nós sejam removidos aleatoriamente, a rede pode permanecer globalmente conectada, já que os hubs tendem a manter a estrutura da rede intacta. No entanto, ataques direcionados a esses hubs podem causar um colapso mais significativo da conectividade, ressaltando a importância estratégica de nós altamente conectados na estrutura e resiliência da rede.

Em termos de dinâmica, as redes complexas não são estáticas, mas sim evoluem ao longo do tempo, com conexões sendo adicionadas ou removidas. Essa dinâmica é observada em uma variedade de contextos, como redes sociais, redes de transporte e redes de colaboração científica, e é essencial para entender a adaptação e evolução dessas redes.

No estudo avançado de redes complexas, também é comum investigar estruturas hierárquicas e análises multiescala, visando a compreensão profunda das interações e padrões em diferentes níveis de organização da rede. As redes complexas são aplicadas em várias disciplinas, incluindo biologia, sociologia, física, ciência da computação e muitas outras. Elas desempenham um papel crucial na modelagem e na compreensão de sistemas complexos, fornecendo insights valiosos sobre a estrutura e o comportamento de sistemas interconectados do mundo real.

2.1 Referenciais relevantes sobre Redes

A pesquisa em redes complexas tem experimentado um notável destaque nas últimas décadas, impulsionada por trabalhos inovadores que estabeleceram as bases teóricas e metodológicas fundamentais para este campo em constante evolução. Vamos destacar alguns desses artigos-chave que desempenharam um papel significativo no avanço das redes complexas.

Um dos artigos mais influentes na área é o trabalho de [Watts e Strogatz 1998]. Este artigo revolucionou a compreensão das redes ao introduzir o conceito de redes de mundo pequeno. Essas redes combinam a eficiência das redes regulares com a conectividade surpreendente das redes aleatórias. Os autores demonstraram que muitos sistemas do mundo real exibem essa propriedade de mundo pequeno, o que tem implicações significativas em áreas que vão desde a disseminação de informações até a propagação de doenças.

Outro marco importante na teoria das redes complexas é o artigo de [Barabási e Albert 1999]. Este trabalho descreveu a descoberta das redes livres de escala e sua prevalência em sistemas do mundo real. Redes livres de escala são aquelas em que alguns nós (hubs) têm um número significativamente maior de conexões do que outros, seguindo uma distribuição de poder. Essa descoberta transformou a maneira como vemos a topologia das redes, influenciando áreas como redes sociais, biologia, e muitas outras.

Além desses dois artigos principais, há outros trabalhos notáveis que também contribuíram substancialmente para o campo das redes complexas. Por exemplo, o estudo de [Newman e Girvan 2004], trouxe métodos cruciais para identificar grupos distintos de nós em uma rede, com aplicações em várias disciplinas, desde análise de redes sociais até biologia de sistemas.

Outro artigo relevante é o de [Milo *et al.* 2004], que investigou as semelhanças e diferenças entre redes biológicas e artificiais, ajudando a elucidar os princípios subjacentes à organização de redes complexas em sistemas naturais e criados pelo homem.

O estudo de [Arruda *et al.* 2014], investiga a relação entre diferentes medidas de centralidade e a capacidade de disseminação de informação ou epidemias em redes complexas. A partir da análise de nove métricas distintas, os autores demonstram que a eficácia de uma centralidade em identificar nós influentes depende tanto da dinâmica de propagação (como modelos de epidemias ou rumores) quanto da topologia da rede (espacial ou não espacial). Em redes não espaciais, medidas como grau e k-core mostraram melhor desempenho para processos epidêmicos, enquanto centralidades como closeness, acessibilidade e grau médio da vizinhança destacaram-se em cenários de propagação de rumores. Já em redes espaciais, a acessibilidade foi consistentemente a métrica mais eficaz, independentemente da dinâmica. Esses resultados evidenciam a importância de considerar tanto a estrutura da rede quanto o tipo de processo dinâmico ao escolher métricas para identificar disseminadores influentes.

Por fim, o trabalho de [Rodrigues *et al.* 2025] propõe uma metodologia inovadora que utiliza algoritmos de aprendizado de máquina para prever resultados de processos dinâmicos

com base apenas em características estruturais de redes complexas. Os autores demonstram que é possível estimar, com alta precisão, observáveis dinâmicos como o tamanho de surtos epidêmicos iniciados por um único nó. Além disso, o estudo identifica quais métricas topológicas da rede são mais relevantes para essas previsões, oferecendo um ranking de importância das métricas com maior acurácia do que abordagens anteriores. Essa abordagem é geral e pode ser aplicada a qualquer processo dinâmico em redes complexas, representando um avanço significativo na aplicação de técnicas de aprendizado de máquina para compreender padrões dinâmicos emergentes em sistemas interconectados.

Em resumo, os avanços na pesquisa em redes complexas foram moldados por uma série de artigos fundamentais. Esses trabalhos pioneiros continuam a inspirar e orientar pesquisadores em todo o mundo, enquanto outros artigos notáveis, como os mencionados acima, enriquecem ainda mais nossa compreensão desse fascinante campo interdisciplinar.

2.2 Medidas de Centralidade

Neste estudo, utilizamos as seguintes medidas de centralidade como variáveis preditoras. Mas antes, é importante destacar que a matriz de adjacência é uma representação fundamental de uma rede que descreve as conexões entre os nós. Para uma rede com N nós, a matriz de adjacência A é uma matriz $N \times N$, onde cada elemento A_{ij} indica a presença de uma conexão entre o nó i e o nó j: se existe uma ligação direta entre os dois nós, então $A_{ij} = 1$, e se não existe uma ligação direta, então $A_{ij} = 0$.

Grau: A medida de centralidade Grau (ou *degree*) em uma rede representa o número de conexões que um nó específico possui. Em outras palavras, o grau de um nó indica quantas arestas estão conectadas a ele. É uma medida fundamental na análise de redes, pois ajuda a identificar os nós mais conectados e, portanto, potencialmente mais influentes ou importantes na rede. Quanto maior o grau de um nó, mais central ele é na rede em termos de conectividade. O grau de um nó é simplesmente o número de arestas conectadas a ele, denotado como k_i para o nó i na rede, sendo calculado como:

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

[Newman 2010]

Centralidade de proximidade: A medida de centralidade por proximidade (ou *Closeness Centrality*) em uma rede quantifica o quão próximo um nó está de todos os outros nós na rede. Ela é calculada como o inverso da soma das distâncias mais curtas entre o nó em questão e todos os outros nós na rede. Quanto menor a soma das distâncias, maior a centralidade de proximidade do nó. Em resumo, a centralidade de proximidade identifica nós que estão geograficamente próximos ou acessíveis a outros nós em uma rede, sendo útil para identificar

a eficiência de comunicação ou influência de um nó na rede. Para um nó i, a centralidade de proximidade é calculada como:

$$C(i) = \frac{1}{\sum_{j=1}^{n}(dist(i,j))} \tag{2.2}$$

onde $dist(i,j)$ representa a distância geodésica entre o nó i e todos os outros nós j na rede [Newman 2010].

Centralidade de intermediação: A medida de centralidade por intermediação (ou *Betweenness Centrality*) em uma rede quantifica o quanto um nó atua como um intermediário na comunicação entre outros nós. É calculada contando o número de caminhos mais curtos que passam por um nó específico em relação ao total de caminhos mais curtos possíveis na rede. Nós com alta Betweenness Centrality têm um papel importante na manutenção da conectividade e na facilitação da comunicação entre diferentes partes da rede [Newman 2010]. Em resumo, essa medida destaca a importância de um nó como um intermediário crítico na transmissão de informações na rede. A centralidade de intermediação de um nó i é calculada como:

$$C_B(i) = \sum_{s \neq t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}} \tag{2.3}$$

onde σ_{st} é o número total de caminhos mais curtos entre os nós s e t, $\sigma_{st}(i)$ é o número de caminhos mais curtos entre s e t que passam pelo nó i, e a soma é feita para todos os pares distintos de nós s e t, excluindo o nó i.

Centralidade do autovetor: A medida de centralidade do autovetor (ou *Eigenvector Centrality*) em uma rede atribui uma pontuação a cada nó com base na importância dos seus vizinhos. Nós com uma alta Eigenvector Centrality são aqueles que estão conectados a outros nós igualmente ou mais importantes na rede. Isso significa que a centralidade de autovetor destaca nós que têm conexões com outros nós influentes, em vez de apenas contar o número de conexões diretas. Em resumo, a Eigenvector Centrality identifica nós que estão ligados a outros nós de alta importância na rede. A centralidade de vetores próprios é calculada como um vetor próprio associado ao maior autovalor da matriz de adjacência da rede [Newman 2010].

Núcleo k: A medida de centralidade núcleo k (ou *k-core*) em uma rede é uma técnica usada para identificar grupos ou subgrafos de nós que são altamente interconectados. Um k-core é um subconjunto da rede onde cada nó tem pelo menos k conexões com outros nós dentro desse subconjunto. Isso ajuda a revelar a estrutura de subgrupos densamente conectados na rede e é útil para identificar regiões-chave de interação em redes complexas. Em resumo, a medida de centralidade k-core destaca grupos de nós altamente coesos dentro de uma rede. O k-core é uma estrutura iterativa na qual os nós com grau menor que k são removidos da rede até que todos os nós tenham grau maior ou igual a k [Seidman 1983].

Classificação de página: A medida por classificação da página (ou *PageRank*) é uma medida de centralidade usada principalmente em motores de busca e redes da web, que avalia a importância relativa dos nós em uma rede com base na ideia de que os nós que são apontados por outros nós importantes também são importantes. Ele atribui uma pontuação a cada nó, calculada iterativamente, levando em consideração as ligações de entrada e a importância dos nós que apontam para ele. O PageRank é usado para classificar páginas da web em motores de busca, onde páginas com maior PageRank geralmente são consideradas mais importantes e aparecem mais alto nos resultados de pesquisa. Em resumo, o PageRank avalia a relevância e influência dos nós em uma rede com base em suas conexões e é amplamente utilizado na classificação de páginas da web e em análises de rede. O PageRank é um algoritmo iterativo em que a pontuação de PageRank de um nó i é calculada como:

$$PR(i) = \frac{1-d}{N} + d \sum_{j=1}^{n} \frac{PR(j)}{L(j)} \tag{2.4}$$

onde N é o número total de nós na rede, d é um fator de amortecimento (tipicamente entre 0,85 e 0,95), $PR(j)$ é o PageRank do nó j e $L(j)$ é o número de links de saída do nó j [Meyer 2000].

Grau Médio de Vizinhança A medida de centralidade grau médio de vizinhança (ou *Average Neighbor Degree*) em uma rede avalia a média dos graus dos vizinhos de um nó específico na rede. Em outras palavras, calcula-se o grau médio dos nós que estão diretamente conectados ao nó em questão. Essa medida ajuda a entender o nível de interconexão dos vizinhos imediatos de um nó na rede, o que pode indicar sua posição e influência relativas. Para um nó i, o grau médio dos vizinhos é calculado como a média dos graus dos nós vizinhos de i, denotado como:

$$k_{avg}(i) = \sum_{j=1}^{n} \frac{k_j}{k_i} \tag{2.5}$$

onde k_j é o grau dos vizinhos de i e k_i é o grau de i [Barabási e Albert 1999].

2.3 Modelo SIR

O modelo SIR, que significa Susceptíveis-Infectados-Recuperados, é um modelo de propagação fundamental na epidemiologia que descreve a disseminação de doenças infecciosas em uma população. Ele divide a população em três compartimentos principais: Susceptíveis (S): Este grupo inclui indivíduos que ainda não foram expostos à doença e, portanto, são suscetíveis a serem infectados caso entrem em contato com pessoas infectadas; Infectados (I): Este compartimento engloba indivíduos que estão atualmente infectados com a doença e, consequentemente, têm o potencial de transmiti-la a outras pessoas; Recuperados (R): Neste

grupo estão aqueles que já se recuperaram da doença e, geralmente, adquiriram imunidade, o que os torna não suscetíveis a uma reinfecção. Em algumas variantes do modelo, também podem ser incluídas pessoas que faleceram devido à doença.

O modelo SIR é regido por um conjunto de equações diferenciais que descrevem as taxas de mudança de cada compartimento ao longo do tempo. Essas equações fundamentais são as seguintes:

- Taxa de mudança de Susceptíveis $\frac{dS}{dt} = -\beta SI$;
- Taxa de mudança de Infectados $\frac{dI}{dt} = \beta S(I - \mu)I$;
- Taxa de mudança de Recuperados $\frac{dR}{dt} = \mu I$.

Onde: β representa a taxa de transmissão da doença, refletindo a probabilidade de uma pessoa suscetível ser infectada por uma pessoa infectada em um determinado intervalo de tempo. μ denota a taxa de recuperação, descrevendo a probabilidade de uma pessoa infectada se recuperar ou falecer em um determinado período de tempo. Embora o modelo SIR seja uma simplificação da realidade epidemiológica complexa, ele serve como uma ferramenta fundamental para a compreensão teórica da disseminação de doenças infecciosas em populações. Vale ressaltar que em aplicações do mundo real, podem ser necessárias variações e extensões deste modelo para considerar aspectos adicionais, como vacinação, variação na transmissibilidade, intervenções de saúde pública e outras complexidades epidemiológicas [Pastor-Satorras *et al.* 2015].

Quando o modelo é implementado em redes complexas, ele oferece uma abordagem mais realista para capturar a dinâmica de infecções, levando em consideração as interações não homogêneas entre os indivíduos da população, onde é representado:

- Nós (vértices): Representam os indivíduos de uma população.
- Arestas (ligações): Representam as interações ou conexões sociais entre indivíduos, por onde a infecção pode se propagar.

As redes podem apresentar diferentes topologias, cada uma com características distintas que influenciam a dinâmica de propagação de informações ou doenças. Em redes aleatórias, como as descritas por Erdős-Rényi, as conexões entre os nós são formadas de maneira aleatória, seguindo uma probabilidade fixa. Essa configuração pode dificultar a propagação rápida de informações. [Newman 2010].

Por outro lado, nas redes de mundo pequeno, introduzidas por [Watts e Strogatz 1998], há uma alta densidade de conexões locais, complementadas por algumas ligações de longa distância. Essa combinação cria o fenômeno conhecido como "pequeno mundo", onde a maioria dos nós

pode ser alcançada em poucos passos. Como resultado, a infecção ou a informação podem se espalhar rapidamente, uma vez que as conexões de longa distância reduzem as distâncias médias na rede, mesmo em um ambiente com muitas conexões locais.

Já nas redes sem escala, como as desenvolvidas por [Barabási e Albert 1999], a distribuição de grau segue uma lei de potência, onde a maioria dos indivíduos possui poucas conexões, enquanto alguns nós, chamados de "hubs", têm um número desproporcionalmente alto de conexões. A presença desses hubs é crucial, pois eles atuam como pontos estratégicos que podem acelerar a propagação, conectando uma vasta quantidade de indivíduos e facilitando a disseminação de informações ou doenças por meio da rede. Assim, cada uma dessas topologias apresenta dinâmicas de propagação distintas, influenciadas pela forma como os nós estão interligados.

O modelo SIR em redes complexas segue a mesma lógica básica do modelo SIR tradicional, mas é adaptado para a estrutura de redes. Em vez de lidar com uma população homogênea, as interações entre os indivíduos são determinadas pela topologia da rede, ou seja, pela conectividade entre os nós.

A dinâmica do modelo SIR pode ser descrita por um conjunto de equações diferenciais que consideram as mudanças nas frações da população nos estados de S (Susceptível), I (Infectado) e R (Recuperado) ao longo do tempo.

- Variação de Susceptíveis (S): $\frac{dS_i}{dt} = \beta \sum_{j=1}^{n} A_{ij} S_i I_j$
- Variação de Infectados (I): $\frac{dI_i}{dt} = \beta \sum_{j=1}^{n} A_{ij} S_i I_j - \mu\ I_i$
- Variação de Recuperados (R): $\frac{dR_i}{dt} = -\ \mu\ I_i$

As equações acima descrevem como a fração de suscetíveis, infectados e recuperados varia ao longo do tempo para cada nodo i da rede. O termo $\sum_{j=1}^{n} A_{ij} S_i I_j$ representa a soma de todas as interações entre o nodo i e seus vizinhos j que estão infectados, ponderado pelas ligações da rede (definidas pela matriz de adjacência A). Já o fator β controla a taxa de transmissão da infecção ao longo das conexões da rede, e a recuperação é dada pela taxa μ, que remove os indivíduos infectados para o compartimento de recuperados.

As equações diferem do modelo SIR clássico, que assume uma mistura homogênea, porque aqui o contato entre os indivíduos depende da topologia da rede (interações não homogêneas). Essas equações podem ser integradas numericamente para estudar a propagação da infecção em diferentes tipos de redes (aleatórias, sem escala, de mundo pequeno) e avaliar o impacto da topologia na dinâmica da doença.

CAPÍTULO

3

RELAÇÃO ENTRE MEDIDAS E FRAÇÃO FINAL DE RECUPERADOS

Neste capítulo, concentraremos nossa análise na correlação entre a fração final de indivíduos recuperados no modelo SIR e uma série de medidas de centralidade específicas. Essas medidas incluem Degree, Closeness Centrality, Betweenness Centrality, Eigenvector Centrality, k-core, PageRank e Average Neighbor Degree. O objetivo principal é avaliar qual dessas medidas de centralidade está mais fortemente correlacionada com a capacidade de um nó se tornar um propagador chave da doença ou influência em uma rede.

Para atingir esse objetivo, examinaremos uma variedade de redes complexas que representam sistemas do mundo real. Faremos uso de redes de aeroportos, capturando as conexões entre diferentes locais, bem como redes sociais, que retratam as interações entre indivíduos. Essa diversidade de contextos permitirá uma análise comparativa robusta e enriquecedora.

Para termos relevância estatística nos resultados, a simulação do modelo SIR foi realizada 500 vezes para cada nó, e a fração final de recuperados desse nó é a média desses 500 valores obtidos. Ainda, é importante destacar que em um primeiro momento consideramos um valor fixo de $\beta = 0.3$ e depois consideramos um intervalo de β entre 0.05 e 0.95 para podermos comparar de maneira mais efetiva o que esse valor fixo e essa variação podem causar no nossos resultados. Por fim, o valor de μ é fixo em todos os casos, consideramos $\mu = 1$, ou seja, estamos considerando que na última iteração não há mais infectados, uma fez que o μ significa taxa de recuperação. Ao considerá-lo igual a 1, estamos considerando que todos serão recuperados.

Os principais métodos de análise consistirão em construir gráficos de dispersão para cada medida de centralidade em relação à fração final de recuperados. Além disso, uma tabela detalhada será compilada, descrevendo as características de cada rede, a medida de centralidade correspondente e a correlação associada. Também será analisado um *heatmap* com o valor da correlação variando os valores de β do modelo SIR. As conclusões extraídas dessas análises não

apenas lançarão luz sobre quais medidas de centralidade estão mais alinhadas com a identificação de propagadores-chave, mas também contribuirão para a compreensão mais profunda das dinâmicas de propagação em diferentes tipos de redes complexas.

Embora as conclusões finais possam ser abrangentes em natureza, enfatizaremos as medidas de centralidade que apresentam as correlações mais fortes e consistentes com a fração final de recuperados. Esses resultados podem ter implicações importantes em várias áreas, desde a prevenção de doenças até a disseminação eficaz de informações em redes sociais.

Em seguida, apresentamos uma análise das correlações entre as medidas de centralidade e a fração final de recuperados.

3.1 Análise de correlação

3.1.1 Netscience

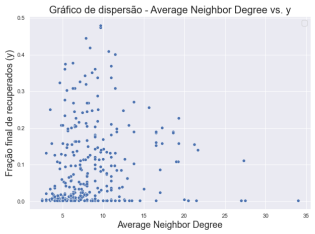
Em um primeiro momento será utilizada a maior componente conectada (LCC) da rede netscience (http://konect.cc/networks/dimacs10-netscience/), que consiste em uma rede de coautoria entre cientistas que trabalham com redes complexas. Essa rede é uma projeção do grafo bipartido de autores e suas publicações.

A rede Netscience é uma rede de coautoria de cientistas que trabalham no campo da teoria e experimentação de redes. Cada nó representa um cientista, e uma aresta não direcionada entre dois nós indica que esses cientistas co-autoraram pelo menos um artigo juntos. Esta rede foi compilada por [Newman 2001] e é frequentemente utilizada em estudos que analisam a estrutura e dinâmica da colaboração científica.

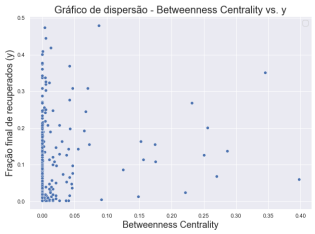
A identificação de nós influentes na rede de coautoria Netscience permite compreender com maior profundidade os mecanismos de disseminação de informação científica, colaborações e impacto intelectual. Na prática, isso pode auxiliar na alocação mais estratégica de recursos para fomento à pesquisa, no fortalecimento de redes colaborativas produtivas e na promoção de difusão acelerada de ideias inovadoras. Além disso, compreender quais autores exercem papel central na propagação de conhecimento possibilita intervenções direcionadas, como convites para liderança de projetos interdisciplinares, políticas de incentivo à cooperação entre diferentes grupos, ou mesmo a antecipação de tendências científicas emergentes. Dessa forma, a análise de propagadores influentes vai além da teoria e pode ajudar de forma prática no planejamento e na organização da produção científica, tanto em instituições quanto em políticas públicas.

Medida	**Valor**
Número de nós	379
Número de arestas	914
Grau médio	4.82
Densidade	0.0128
Coeficiente de aglomeração médio	0.7412
Comprimento médio de caminho	6.041
Diâmetro	17

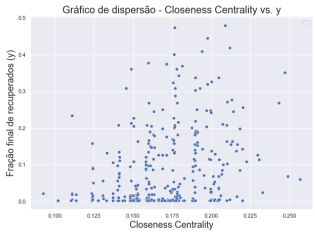
Tabela 1 – Medidas globais da rede *Netscience*



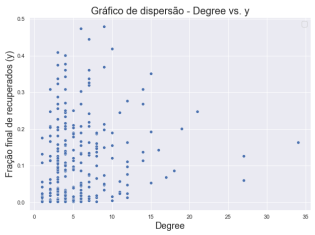
(a) Average Neighbor Degree.



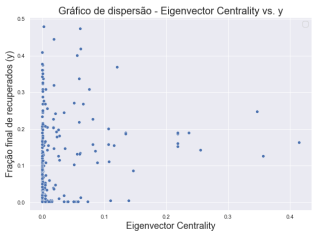
(b) Betweenness Centrality.



(c) Closeness Centrality.



(d) Degree.



(e) Eigenvector Centrality.

(f) k-core.

(g) PageRank.

Figura 1 – Gráficos de dispersão entre medidas e a fração de recuperados, considerando $\mu = 1$ e $\beta = 0.3$

A figura 1 e a tabela 2 foram criadas utilizando um valor de β fixo em 0.3, e dentre as correlações observadas na tabela 1, destacam-se as mais fortes correlações positivas entre Degree e a fração final de recuperados (0.6248), bem como Eigenvector Centrality (0.6166). Essas correlações indicam que nós mais conectados e influentes na rede tanto em termos de número de conexões quanto da qualidade dessas conexões, desempenham um papel fundamental na propagação e recuperação do comportamento nessa rede de difusão de informações. Esses nós podem ser essenciais para a eficácia de intervenções ou para o controle da propagação em

Tabela 2 – Correlação das medidas de centralidade com o número de infectados para a rede Netscience

Medida de Centralidade	Valor
Degree	0.6248
Closeness Centrality	0.5765
Betweenness Centrality	0.3691
Eigenvector Centrality	0.6166
k-core	0.6019
PageRank	0.5061
Average Neighbor Degree	0.4108

sistemas complexos.

Além disso, k-core também apresenta uma correlação positiva (0.6019). Isso sugere que nós que estão em subgrupos altamente conectados estão associados a uma maior fração de nós recuperados, ou seja, esses nós em núcleos densos tendem a ser mais eficazes na propagação do comportamento que está sendo modelado.

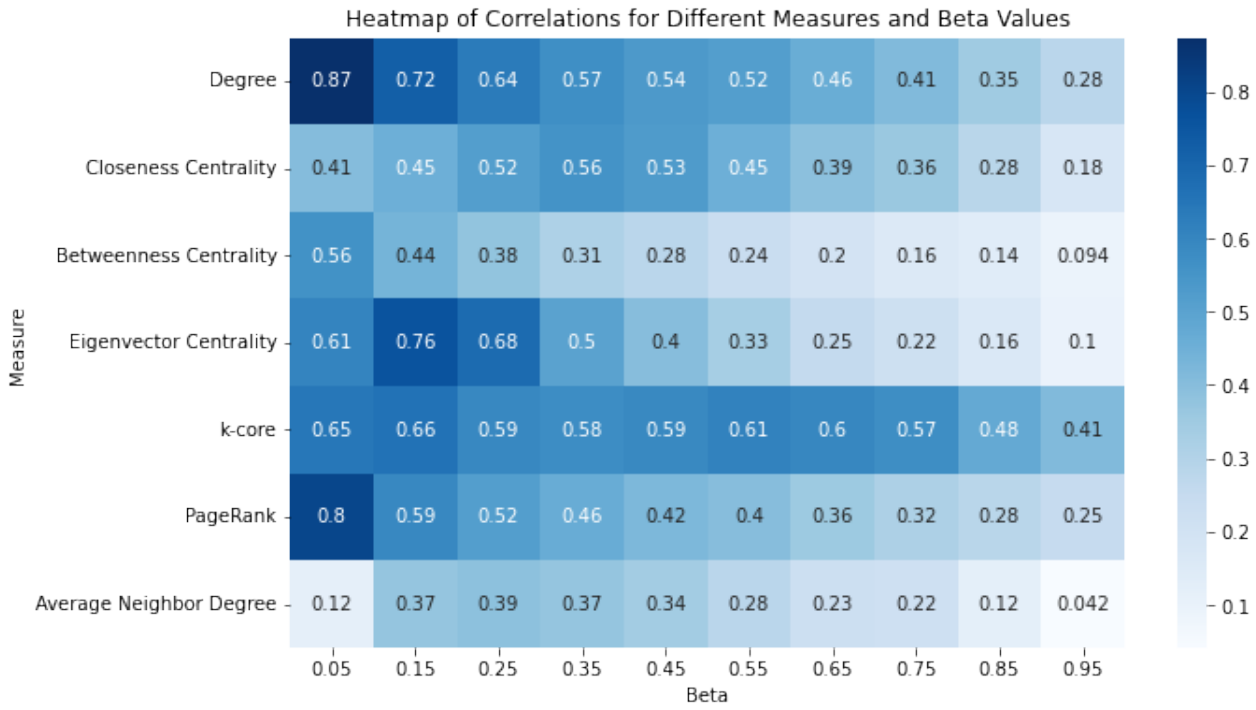
Ainda, temos Closeness Centrality (0.5765) e PageRank (0.5061) como correlações positivas e valores medianos, indicando que nós mais acessíveis (Closeness Centrality) e mais influentes (PageRank) são mais eficazes em promover a recuperação em uma propagação de informações. Esses nós têm um papel fundamental em acelerar a disseminação da recuperação. Portanto, nós com alta Closeness Centrality e PageRank podem ser mais estratégicos para a recuperação global na rede.

Em resumo, a figura 1 e a tabela 2 foram geradas com um valor fixo de $\beta = 0.3$, destacando correlações positivas entre a fração final de recuperados e várias medidas de centralidade. As correlações mais fortes foram entre Degree (0.6248) e Eigenvector Centrality (0.6166), indicando que nós mais conectados e influentes desempenham um papel crucial na propagação e recuperação de comportamentos na rede. A k-core também apresentou uma correlação positiva (0.6019), sugerindo que nós em subgrupos densamente conectados são mais eficazes na propagação. Já as correlações medianas de Closeness Centrality (0.5765) e PageRank (0.5061) indicam que nós mais acessíveis e influentes são estratégicos para acelerar a recuperação e difusão de comportamentos na rede.

Na figura 2, o valor de β não foi afixado com o intuito de comparar com os resultados anteriores. Essa figura mostra maior correlação com a medida de centralidade Degree, mesmo variando β entre 0.05 e 0.95. Porém, uma observação interessante é que nas medidas Closeness Centrality e Average Neighbor Degree a correlação apresenta um comportamento não monotômico, isto é, não é nem diretamente e nem inversamente proporcional com o valor de β como nas outras medidas.

O comportamento do aumento seguido de decréscimo na correlação entre a Closeness

Figura 2 – Correlação para diferentes valores de β na rede *Netscience*



Centrality e a fração final de recuperados, conforme o parâmetro β aumenta, pode indicar uma relação não linear entre a influência da Closeness Centrality e a propagação da epidemia.

Esse padrão sugere que em valores menores de β (transmissibilidade baixa), a Closeness Centrality pode ter um papel relevante em controlar a propagação da infecção, pois representa o quão rapidamente um nó pode alcançar os outros. Contudo, à medida que β cresce e a transmissibilidade aumenta, a importância dessa medida diminui, pois a infecção pode se espalhar mais livremente pela rede, alcançando nós mesmo com menor Closeness Centrality. A partir de um certo ponto, outros fatores podem começar a ter mais peso no resultado da propagação, como a interconectividade geral da rede e outras medidas de centralidade (como Degree e Eigenvector Centrality).

Este comportamento sugere que a eficiência de medidas como Closeness Centrality depende do nível de transmissibilidade da infecção. Em níveis moderados de β, ela contribui mais fortemente para a dinâmica da propagação, mas em transmissibilidades altas, a rede inteira pode se tornar rapidamente saturada, minimizando a relevância dessa medida.

Ainda, em valores baixos de β, a Average Neighbor Degree têm uma influência ligeiramente maior, mas ainda assim pequena, indicando que a conectividade média dos vizinhos de um nó pode ter um impacto leve na capacidade do nó de participar na propagação inicial da infecção. No entanto, conforme β aumenta, a infecção pode se espalhar de maneira mais homogênea pela rede, independentemente da média dos graus dos vizinhos. Isso faz com que a relevância dessa medida diminua, refletindo que, em transmissibilidades mais altas, o papel da conectividade local média dos vizinhos perde importância, pois a infecção não depende mais da estrutura local

imediata dos nós, mas sim de um fator mais geral e abrangente da rede. Resumidamente, para um valor de β baixo, a epidemia não sobrevive, portanto um nó com a doença só contamina os seus primeiros vizinhos: essa é a razão pela qual o grau apresenta uma correlação alta para a transmissibilidade baixa.

3.1.2 Air traffic control

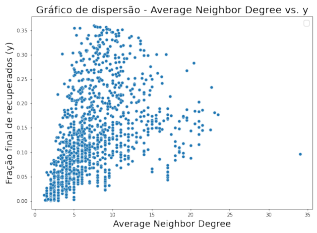
A rede *Air traffic control* representa o sistema de controle de tráfego aéreo dos EUA (http://konect.cc/networks/maayan-faa/). Os nós são instalações de controle de tráfego aéreo (como torres, centros e controles de aproximação), e as arestas representam links de comunicação ou transferências de controle entre eles. Essa rede reflete a estrutura operacional da gestão do tráfego aéreo nos EUA e é útil para estudar a resiliência e eficiência dos sistemas de controle de tráfego aéreo.

É importante destacar aqui que, para a modelagem, foi retirada a direção e a ponderação, uma vez que desejamos comparar diferentes redes sob uma base comum e simplificada e queremos focar apenas na estrutura topológica da rede.

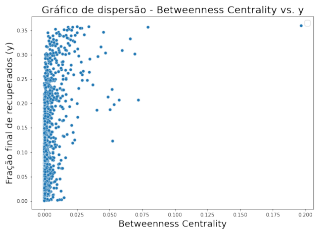
Estudar os propagadores influentes na rede Air Traffic Control pode ter aplicações diretas na melhoria da segurança, eficiência e resiliência do tráfego aéreo. Ao identificar os nós mais críticos para a disseminação de informações ou potenciais falhas (como aeroportos ou centros de controle com alta conectividade), é possível otimizar rotas de comunicação e planejar estratégias mais eficazes de resposta a eventos adversos. Isso também contribui para a alocação inteligente de recursos em infraestrutura e treinamento, garantindo que os pontos mais estratégicos da rede estejam preparados para lidar com situações de alta demanda ou emergência. Em um setor onde segundos podem salvar vidas, compreender a estrutura da rede e seus elementos-chave é essencial para decisões mais precisas e seguras.

Medida	**Valor**
Número de nós	1226
Número de arestas	2408
Grau médio	3.93
Densidade	0.0032
Coeficiente de aglomeração médio	0.0675
Comprimento médio de caminho	5.92
Diâmetro	17

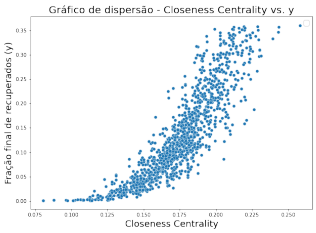
Tabela 3 – Medidas globais da rede Air traffic control



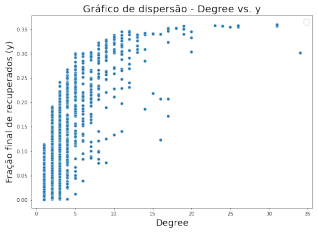
(a) Average Neighbor Degree.



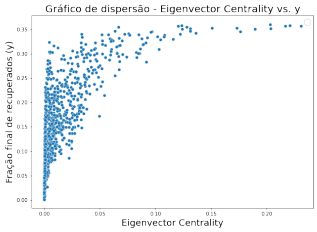
(b) Betweenness Centrality.



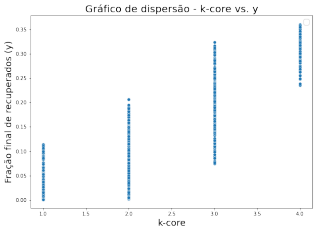
(c) Closeness Centrality.



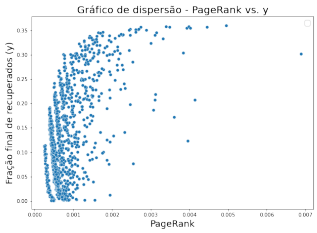
(d) Degree.



(e) Eigenvector Centrality.



(f) k-core.



(g) PageRank.

Figura 3 – Gráficos de dispersão entre medidas e a fração de recuperados, considerando $\mu = 1$ e $\beta = 0.3$

Tabela 4 – Medidas de Centralidade rede Air traffic control

Medida de Centralidade	Valor
Degree	0.7210
Closeness Centrality	0.8666
Betweenness Centrality	0.4126
Eigenvector Centrality	0.6899
k-core	0.8492
PageRank	0.5605
Average Neighbor Degree	0.4393

Para a rede Air traffic control, o grau apresenta uma correlação positiva e robusta de 0.72 quando β é fixo em 0.3. Isso sugere que indivíduos com um número maior de conexões diretas na rede tendem a estar associados a uma fração final mais elevada de recuperados. Essa medida reflete a influência direta da conectividade individual na disseminação e recuperação.

A Closeness Centrality (Centralidade de Proximidade) exibe uma correlação ainda mais forte, alcançando 0.87. Isso indica que a proximidade física na rede desempenha um papel crucial

na propagação da doença e na subsequente recuperação. Indivíduos geograficamente próximos a muitos outros têm uma forte associação com uma fração mais alta de recuperados.

A Betweenness Centrality (Centralidade de Intermediação), com uma correlação de 0.41, mostra uma associação mais modesta. Isso sugere que, embora os intermediários desempenhem um papel na propagação, sua contribuição para a fração final de recuperados pode não ser tão proeminente quanto outras medidas.

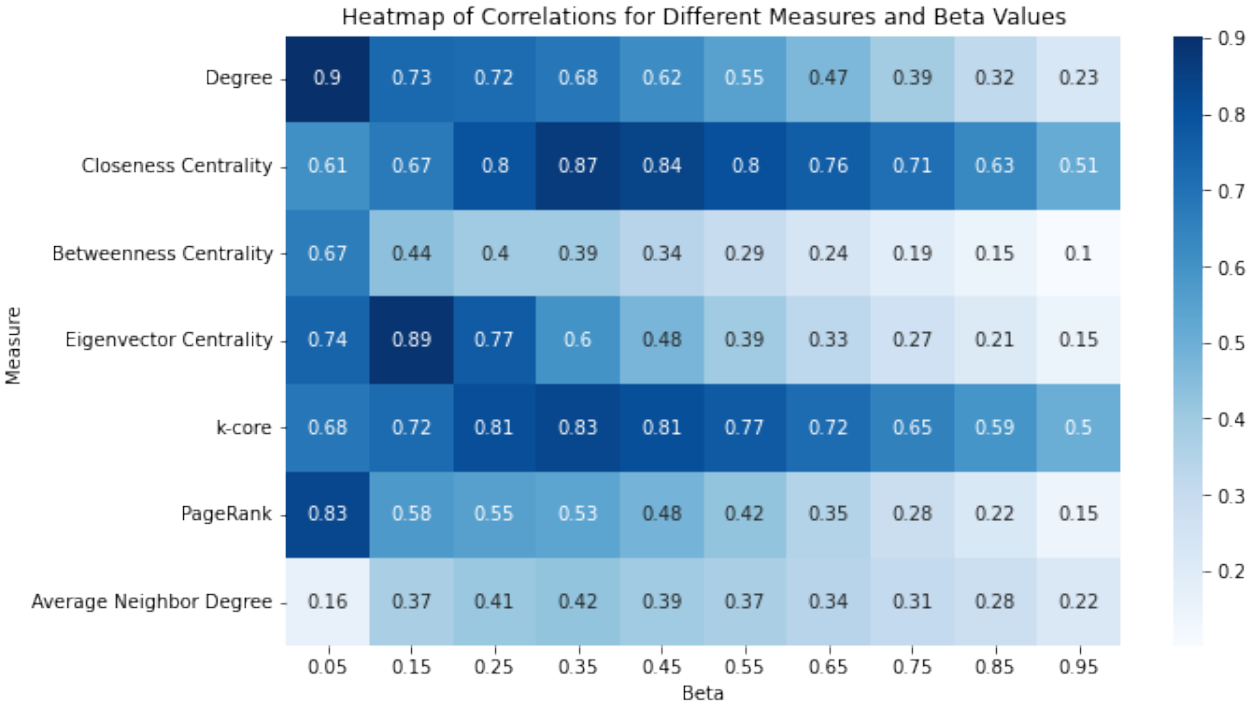
A Eigenvector Centrality (Centralidade de Autovetor), com uma correlação de 0.69, destaca a importância da posição estratégica na rede. Indivíduos com alta centralidade de autovetor estão associados a uma fração maior de recuperados.

A medida de k-core revela uma correlação significativa de 0.85. Isso destaca que a presença de núcleos mais densos e conectados está fortemente associada a uma maior fração final de recuperados, indicando a importância dos grupos altamente interligados na disseminação e recuperação.

A PageRank apresenta uma correlação de 0.5605, indicando que a propagação de recuperação está positivamente associada à importância atribuída pela métrica PageRank. Isso sugere que indivíduos identificados como recorrentes em um passeio aleatório na rede têm uma contribuição significativa para a fração final de recuperados.

Finalmente, a Average Neighbor Degree (Grau Médio dos Vizinhos) exibe uma correlação de 0.4393. No entanto, valores altos para essa medida de centralidade não indicam necessariamente que um alcance alto da epidemia.

Figura 4 – Correlação para diferentes valores de β na rede Air traffic control



Na figura 4, para diferentes valores de β, as medidas de centralidade de mais impacto foram Degree e Closeness Centrality. Ainda, é possível observar que neste caso mais medidas de centralidade tiveram o comportamento de aumentar a correlação até determinado valor de β e depois diminuiu. São elas: Closeness Centrality, Eigenvector Centrality, k-core e Average Neighbor Degree.

A Eigenvector Centrality mede a importância de um nó não apenas com base em suas conexões diretas, mas também na importância dos nós aos quais ele está conectado. No contexto da propagação de uma infecção, ela é relevante em níveis de transmissibilidade moderados a altos, pois destaca os nós que estão conectados a partes altamente interconectadas da rede. Esses nós podem atuar como hubs de propagação e facilitar a disseminação para grandes porções da rede.

O aumento inicial na correlação indica que, conforme o β cresce, a Eigenvector Centrality é cada vez mais importante para determinar como a infecção se espalha pela rede. Isso faz sentido, pois em regimes de transmissibilidade moderada, a infecção começa a se propagar mais amplamente, mas ainda depende dos nós mais influentes para atingir novas partes da rede.

Porém, o decréscimo em níveis muito altos de β reflete que, quando a transmissibilidade é extremamente alta, quase todos os nós na rede se tornam capazes de propagar a infecção com eficiência. Nesse caso, a importância dos nós centrais em termos de eigenvetores diminui, pois a propagação se torna generalizada e menos dependente de hubs específicos. A infecção se espalha tão rapidamente e de forma tão abrangente que a centralidade estrutural dos nós perde relevância, e a propagação acontece independentemente das características individuais destacadas pela Eigenvector Centrality.

Já o k-core é mais influente em níveis de transmissibilidade intermediários, onde a propagação depende da estrutura central da rede e de nós mais densamente conectados. No entanto, em níveis de transmissibilidade muito altos, essa importância se dilui, pois a infecção se espalha de forma quase universal, minimizando a relevância da estrutura de conectividade central. Isso resulta no aumento seguido de decréscimo na correlação entre o k-core e a fração final de recuperados conforme o β cresce.

3.1.3 OpenFlights

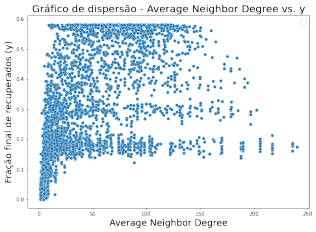
A rede OpenFlights representa o mapa global de rotas aéreas. Os nós são aeroportos ao redor do mundo, e as arestas direcionadas representam voos diretos entre eles. Dados são provenientes do site OpenFlights (http://konect.cc/networks/opsahl-openflights/), que coleta e fornece dados abertos sobre aeroportos, companhias aéreas e rotas.

A análise de propagadores influentes dessa rede pode trazer benefícios práticos importantes para a gestão do transporte aéreo e da logística global. Identificar os aeroportos com maior potencial de disseminação de fluxo (de passageiros, cargas ou até eventos disruptivos) permite

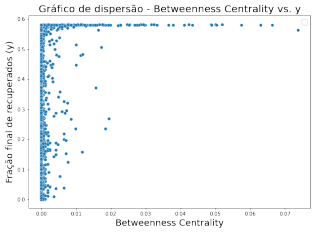
otimizar rotas comerciais, fortalecer pontos estratégicos da malha aérea e reduzir vulnerabilidades operacionais. Além disso, essa análise pode apoiar políticas de prevenção a pandemias, priorizando *hubs* críticos para controle sanitário, bem como orientar investimentos em infraestrutura nos aeroportos com maior impacto sistêmico. Em um cenário de crescente interdependência global, entender como a estrutura da rede influencia a propagação é essencial para decisões mais eficientes e resilientes no setor aéreo.

Medida	**Valor**
Número de nós	3425
Número de arestas	19256
Grau médio	11.24
Densidade	0.0033
Coeficiente de aglomeração médio	0.4871
Comprimento médio de caminho	4.10
Diâmetro	13

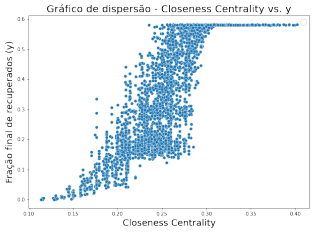
Tabela 5 – Medidas globais da rede OpenFlights



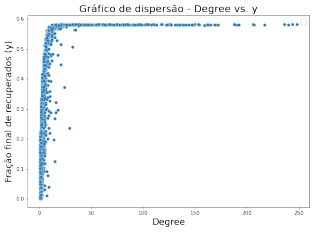
(a) Average Neighbor Degree.



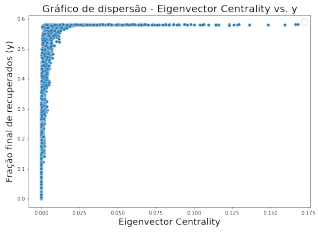
(b) Betweenness Centrality.



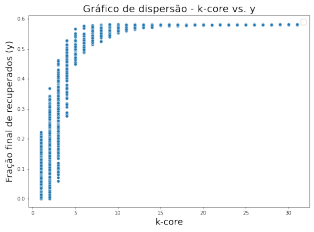
(c) Closeness Centrality.



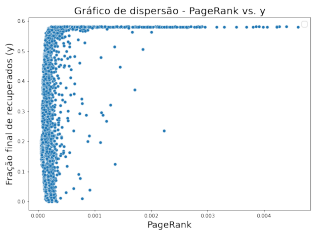
(d) Degree.



(e) Eigenvector Centrality.



(f) k-core.



(g) PageRank.

Figura 5 – Gráficos de dispersão entre medidas e a fração de recuperados, considerando $\mu = 1$ e $\beta = 0.3$

Tabela 6 – Medidas de Centralidade rede Openflights

Medida de Centralidade	Valor
Degree	0.4980
Closeness Centrality	0.8244
Betweenness Centrality	0.2250
Eigenvector Centrality	0.4739
k-core	0.7190
PageRank	0.4370
Average Neighbor Degree	0.3726

Dentre as correlações observadas na tabela 6, destacam-se as mais fortes correlações positivas entre Closeness Centrality e a fração final de recuperados (0.82), bem como k-core (0.72). Essas correlações positivas sugerem que, à medida que essas centralidades aumentam, a fração final de recuperados tende a aumentar também. Isso indica que indivíduos que estão em posições mais centrais na rede em termos de proximidade têm maior influência na propagação de informações ou ações que levam à recuperação. Ou ainda, grupos de nós que são altamente conectados entre si também desempenham um papel influente na rede.

Eigenvector Centrality (0.47), PageRank (0.44) e Degree (0.50) exibem correlações positivas moderadas, o que implica que a importância dessas medidas na rede também está associada a uma maior fração final de recuperados.

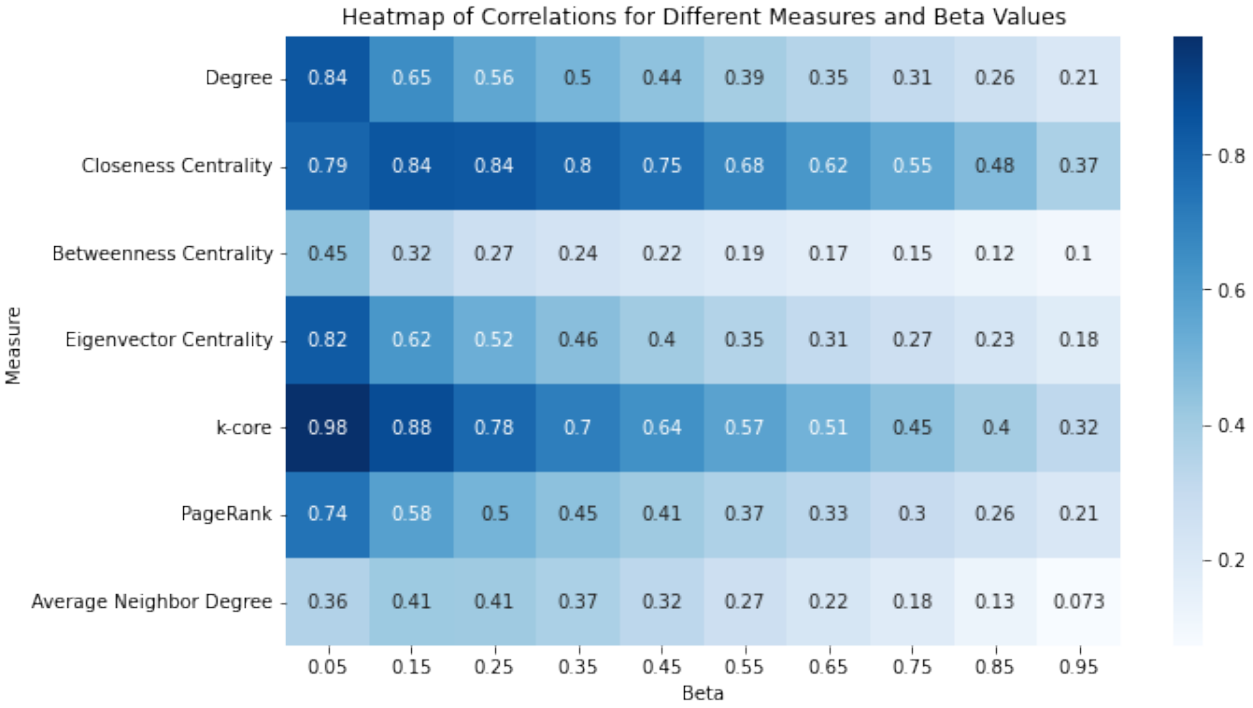
A Betweenness Centrality (0.23), possui um valor positivo, porém ela é a correlação mais fraca entre todas as medidas analisadas, indicando que um nó que atua como intermediário de outros nós não necessariamente demonstra sua influência de propagação, não estando altamente correlacionado com a fração final de recuperados.

Por fim, embora a Average Neighbor Degree apresente uma correlação positiva (0.37), é um valor baixo, sugerindo que a média dos graus dos vizinhos dos nós pode ter uma influência menos pronunciada na fração final de recuperados.

De acordo com a figura 6, as medidas de centralidade que apresentaram maior correlação foram Closeness Centrality e k-core, para diferentes valores de β, mas novamente é possível observar a distribuição atípica em Closeness Centrality e Average Neighbor Degree.

Novamente, o padrão de correlação crescente indica que em regimes de transmissibilidade moderada, a Closeness Centrality é altamente relevante porque representa a capacidade de um nó alcançar outros rapidamente, o que facilita a propagação da infecção. Isso faz com que a centralidade de Closeness tenha uma alta correlação com a fração de recuperados durante esses níveis intermediários de β. Já a Average Neighbor Degree, possui seu impacto limitado e a queda em relevância com valores mais altos de β indicam que ela não é um bom preditor da propagação da infecção em regimes de alta transmissibilidade, onde a propagação se torna mais

Figura 6 – Correlação para diferentes valores de β



global e menos dependente de detalhes locais da rede.

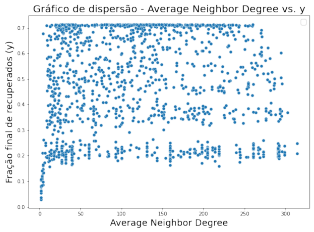
3.1.4 US airports

A rede de Aeroportos dos EUA foca no sistema de transporte aéreo dentro dos Estados Unidos. Os nós representam aeroportos dos EUA, e as arestas representam voos diretos entre eles. É uma rede direcionada e ponderada, opcionalmente por volume de passageiros ou frequência de voo. Há aproximadamente 1.574 aeroportos nos EUA e esses dados são coletados da *Federal Aviation Administration* (FAA) e do *Bureau of Transportation Statistics* (http://konect.cc/networks/opsahl-usairport/).

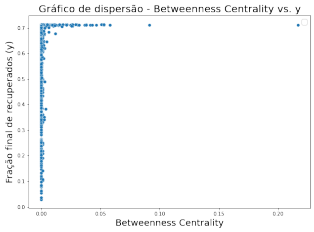
Analisar os propagadores influentes dessa rede pode oferecer insights valiosos para melhorar a eficiência, a segurança e a resiliência do transporte aéreo nacional. Identificar os aeroportos mais centrais na propagação de fluxos — sejam eles de passageiros, mercadorias ou informações — permite otimizar a malha aérea, reduzir atrasos e direcionar investimentos de forma mais estratégica. Além disso, essa análise é fundamental para a elaboração de planos de contingência em situações de emergência, como eventos climáticos extremos, falhas operacionais ou surtos epidêmicos. Ao compreender quais aeroportos funcionam como nós críticos na rede, gestores públicos e operadores logísticos podem tomar decisões mais rápidas e eficazes, com impactos diretos na mobilidade e na economia do país.

Medida	**Valor**
Número de nós	1572
Número de arestas	17214
Grau médio	21.90
Densidade	0.0139
Coeficiente de aglomeração médio	0.5048
Comprimento médio de caminho	3.1151
Diâmetro	8

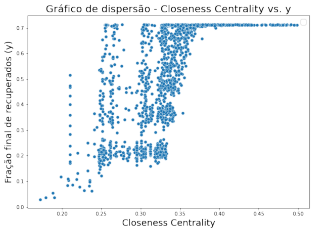
Tabela 7 – Medidas globais da rede US airports



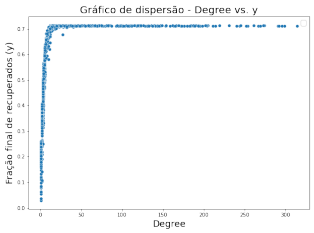
(a) Average Neighbor Degree.



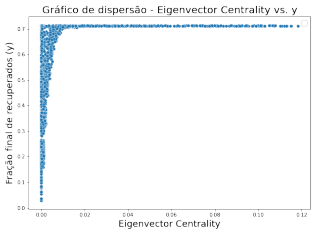
(b) Betweenness Centrality.



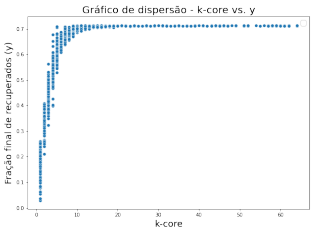
(c) Closeness Centrality.



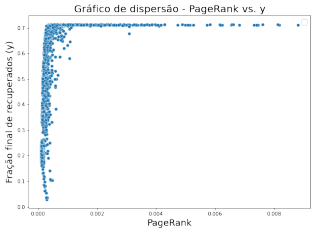
(d) Degree.



(e) Eigenvector Centrality.



(f) k-core.



(g) PageRank.

Figura 7 – Gráficos de dispersão entre medidas e a fração de recuperados, considerando $\mu = 1$ e $\beta = 0.3$

Tabela 8 – Medidas de Centralidade rede US airports

Medida de Centralidade	Valor
Degree	0.4551
Closeness Centrality	0.6345
Betweenness Centrality	0.1681
Eigenvector Centrality	0.4436
k-core	0.5960
PageRank	0.4538
Average Neighbor Degree	-0.0472

Sobre as correlações observadas na tabela 8, destacam-se as mais fortes correlações positivas entre Closeness Centrality e a fração final de recuperados (0.63), bem como k-core (0.60). Essas correlações positivas sugerem que, à medida que essas medidas de centralidade aumentam, a fração final de recuperados tende a aumentar também. Isso indica que indivíduos que estão em posições mais centrais na rede em termos de proximidade, têm maior influência na propagação de informações ou ações que levam à recuperação. Ou ainda, grupos de nós que são altamente conectados entre si também desempenham um papel influente na rede.

Eigenvector Centrality (0.4436), PageRank (0.4538) e Degree (0.4551) exibem correlações positivas moderadas, o que implica que a importância dessas medidas na rede também está associada a uma maior fração final de recuperados.

A Betweenness Centrality (0.1681), possui um valor positivo, porém ela é a correlação mais fraca entre todas as medidas analisadas, indicando que um nó que atua como intermediário de outros nós não necessariamente demonstra sua influência de propagação, não estando altamente correlacionado com a fração final de recuperados.

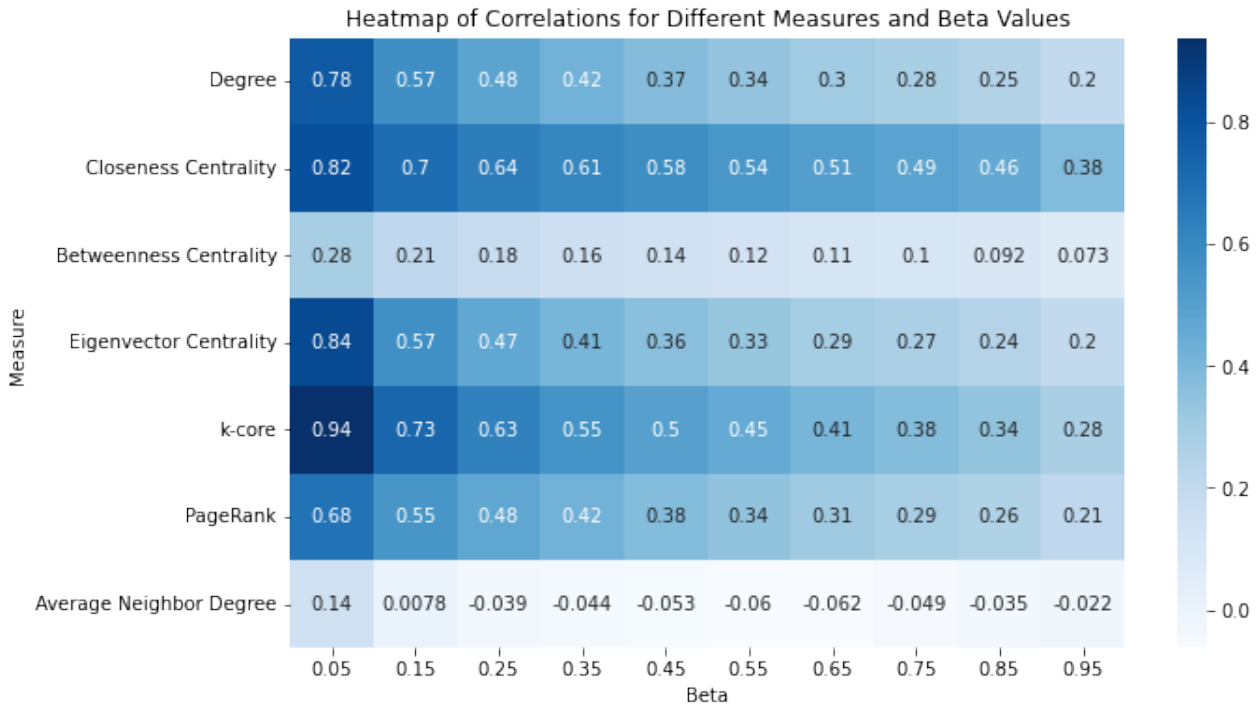
Por fim, a Average Neighbor Degree apresenta uma correlação negativa (-0.04727), porém indica um valor muito próximo de zero. Sendo assim, pode-se considerar que não há correlação entre essa medida e a fração final de recuperados.

Na figura 8, a correlação que expressa maior força é a Closeness Centrality, seguida da k-core, ou seja, os resultados alcançados ao variar o valor de β são similares aos anteriores, quando o seu valor estava fixado em 0.3. Ainda, nesta rede não houve o aumento seguido de decréscimo em algumas medidas de centralidade como nas redes anteriores.

A ausência de um padrão de aumento seguido de decréscimo nas correlações para todas as medidas de centralidade nesta imagem pode indicar que a propagação da infecção no modelo SIR, em relação às diferentes centralidades, não passa por uma transição clara onde essas centralidades perdem sua relevância à medida que a transmissibilidade β aumenta.

Isso pode significar um impacto consistente das centralidades, isto é, as correlações permanecem estáveis ou em declínio gradual, sugerindo que as centralidades continuam a ter

Figura 8 – Correlação para diferentes valores de β



influência sobre a propagação da infecção, mas essa influência se torna cada vez menor à medida que β aumenta. Em transmissibilidades mais altas, a importância dessas medidas parece não passar por um pico e posterior queda acentuada, mas sim por uma redução mais linear e constante. Também pode significar a ausência de uma fase de transição marcante, onde uma medida de centralidade começa a perder relevância repentinamente em níveis altos de β pode significar que a rede tem uma estrutura que permite a propagação da infecção de forma mais uniforme, sem depender exclusivamente dos nós mais conectados ou influentes. Por fim, esse comportamento também pode indicar que a estrutura da rede estudada não é tão hierárquica ou que os nós mais centrais não exercem um papel desproporcional na propagação em níveis médios a altos de β. Redes com menos hubs dominantes ou uma conectividade mais distribuída tendem a mostrar um comportamento onde a importância das centralidades se dilui de forma mais uniforme. Com isso, a sua estrutura pode ser diferente das anteriores, justificando esse comportamento.

3.2 Regressão e previsão da fração final de recuperados

Em vez de nos concentrarmos apenas na análise de correlações, abordaremos o problema por meio de uma análise de regressão. Isso nos permitirá quantificar o impacto de cada medida de centralidade na previsão da fração final de recuperados. Cada nó em nossa rede será tratado como uma observação individual, com suas características representadas pelas medidas de centralidade. A fração final de recuperados será a variável de resposta que desejamos prever.

3.2.1 Métodos de Regressão

A classificação e a regressão são métodos fundamentais para analisar relações entre variáveis e prever resultados. Neste contexto, utilizaremos esses métodos para analisar a relação entre medidas de centralidade e a fração de recuperados em redes complexas.

3.2.1.1 Regressão linear

A regressão linear é um dos métodos mais básicos e amplamente utilizados na análise estatística. O objetivo é modelar a relação linear entre uma variável de resposta (dependente) e uma ou mais variáveis independentes (covariáveis). No contexto deste estudo, a variável de resposta é a fração final de recuperados, e as covariáveis são as medidas de centralidade de cada nó na rede.

Para analisar várias medidas de centralidade simultaneamente, podemos estender o modelo para uma regressão linear múltipla:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$$

onde:

- Y é a variável de resposta (fração final de recuperados);
- β_0 é o intercepto;
- $\beta_1, \beta_2, ..., \beta_p$ são os coeficientes das respectivas covariáveis x;
- $X_1, X_2, ..., X_p$ são as covariáveis respectivas a cada medida de centralidade;
- ε é o erro aleatório, em que a hipótese do erro é dada por $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

O objetivo é estimar os coeficientes $\beta_i (i = 1, ..., p)$ que melhor se ajustam aos dados. Ao interpretar os coeficientes estimados, podemos determinar a direção e a magnitude da influência de cada medida de centralidade na previsão da fração de recuperados.

Para garantir comparabilidade entre variáveis com escalas distintas, foi aplicada a normalização padrão (z-score), na qual cada variável teve sua média subtraída e foi dividida pelo desvio padrão. Essa transformação torna as variáveis adimensionais, evita que atributos com maiores magnitudes dominem a regressão e facilita a interpretação dos coeficientes em termos de importância relativa. Ainda, é importante destacar que aqui consideramos apenas o valor fixo de $\beta = 0.3$ e não fizemos considerando sua variação.

O foco principal desta análise será a interpretação dos coeficientes resultantes da regressão. Cada coeficiente representará a mudança média na fração final de recuperados associada a uma mudança unitária na medida de centralidade correspondente, mantendo as outras medidas

constantes. Ao analisar a magnitude e o sinal dos coeficientes, poderemos determinar quais medidas de centralidade têm um impacto mais significativo na previsão da fração de recuperados.

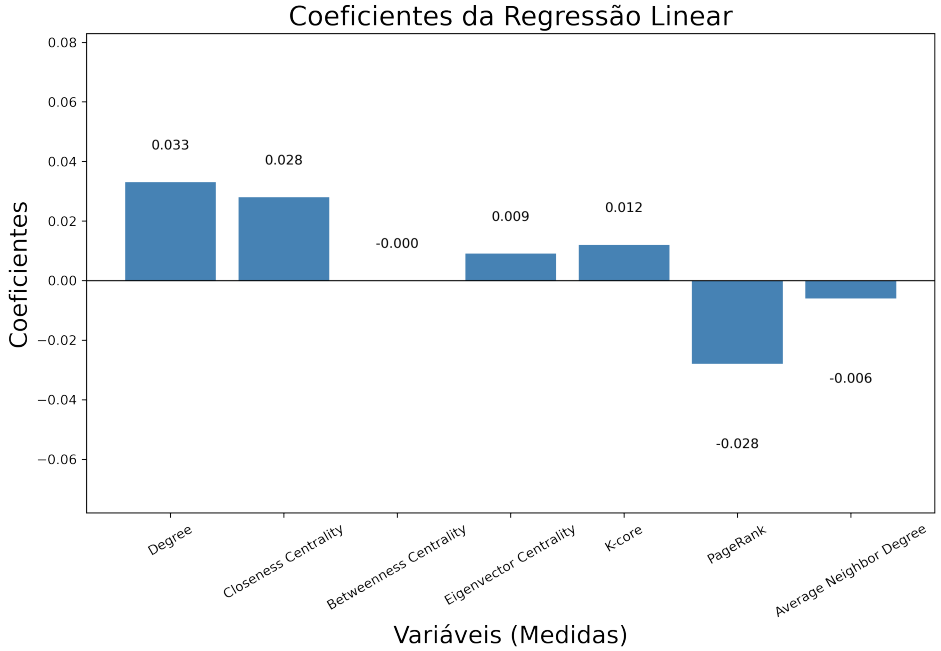
É importante ressaltar que este capítulo visa identificar e quantificar as medidas de centralidade mais influentes, sob a premissa de que uma mudança em uma medida específica influenciará diretamente a propagação. Embora a modelagem de redes complexas seja inerentemente complexa e sujeita a limitações, esta abordagem nos fornecerá insights valiosos sobre quais características dos nós têm maior impacto na dinâmica de propagação.

À medida que avançamos neste capítulo, exploraremos os resultados da análise de regressão para cada medida de centralidade e discutiremos as implicações desses resultados em diferentes contextos de rede. A identificação das medidas de centralidade mais influentes não apenas aprofundará nossa compreensão teórica das redes complexas, mas também poderá ter aplicações práticas em estratégias de intervenção e controle.

Para garantir a comparabilidade das medidas de centralidade, normalizaremos os dados antes da análise. Isso nos permitirá avaliar o impacto relativo de cada medida de centralidade, independentemente de suas escalas originais. A normalização também ajuda a evitar viés devido a diferenças nas magnitudes das medidas.

3.2.1.2 Netscience

Figura 9 – Coeficientes da regressão linear



Na figura 9, ao considerarmos os valores dos coeficientes de regressão relacionados a cada medida de centralidade (Degree, Closeness Centrality, Betweenness Centrality, Eigenvector Centrality, k-core, PageRank e Average Neighbor Degree) em relação à fração final de recupera-

Variável	**Coef. (β_i)**	**Erro Padrão (SE_{β_i})**	**t-Estat.(β_i/SE_{β_i})**	**Significativa**
Degree	0.033	0.0073	4.52	Sim
Closeness Centrality	0.028	0.0015	17.75	Sim
Betweenness Centrality	-0.000	0.0020	0.00	Não
Eigenvector Centrality	0.009	0.0017	5.19	Sim
k-core	0.012	0.0023	5.16	Sim
PageRank	-0.026	0.0061	-4.22	Sim
Average Neighbor Degree	-0.006	0.0017	-3.36	Sim

Tabela 9 – Coeficientes estimados, erros padrão, estatísticas t e significância com $\alpha = 0.05$.

dos, podemos obter uma visão aprofundada da influência dessas métricas no comportamento do sistema em análise.

Observamos que as medidas de Degree e Closeness Centrality têm coeficientes positivos significativos, com valores de 0.0329 e 0.0281, respectivamente. Isso sugere que um aumento nessas medidas está associado a um aumento na fração final de recuperados. Essas descobertas coincidem com as correlações de Pearson anteriormente discutidas, fortalecendo a ideia de que indivíduos que ocupam posições mais centrais ou próximas na rede desempenham um papel fundamental na promoção da recuperação.

Por outro lado, a medida de Betweenness Centrality, apesar de apresentar uma correlação positiva anteriormente, possui um coeficiente de regressão próximo a zero (-0.0004), indicando que seu impacto direto na fração final de recuperados pode ser limitado em comparação com outras métricas.

Eigenvector Centrality e k-core também têm coeficientes positivos, com valores de 0.0093 e 0.0116, respectivamente. Embora esses valores sejam menores em magnitude em comparação com Degree e Closeness Centrality, eles ainda indicam uma influência positiva na fração final de recuperados.

Por outro lado, PageRank apresenta um coeficiente de regressão negativo considerável de -0.0282, o que sugere que, à medida que o PageRank aumenta, a fração final de recuperados tende a diminuir. Isso pode ser interpretado como uma indicação de que uma maior importância na rede, conforme medida pelo PageRank, pode não ser benéfica para o aumento da recuperação.

Finalmente, a Average Neighbor Degree tem um coeficiente de regressão negativo de -0.0058, indicando uma influência ligeiramente negativa na fração final de recuperados.

Em resumo, os coeficientes de regressão nos fornecem insights sobre como cada medida de centralidade contribui para a previsão da fração final de recuperados. Degree e Closeness Centrality surgem como os principais impulsionadores da recuperação, enquanto PageRank parece ter um impacto adverso. Entre todas as métricas, essas informações podem ajudar na tomada de decisões sobre a importância relativa das medidas de centralidade na análise de redes

complexas.

A tabela 9 apresenta alguns elementos para análise de coeficientes em um modelo estatístico. São eles:

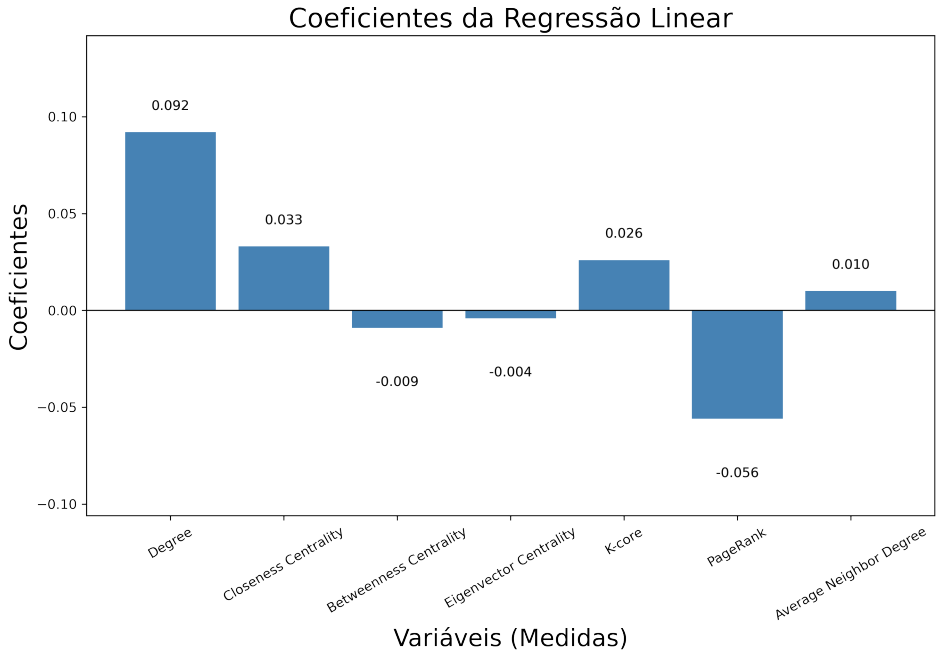
- Coeficiente: mostra os valores estimados dos coeficientes do modelo;
- Erro Padrão: indica a variabilidade estimada dos coeficientes;
- t-Estatística: mostra a razão entre o coeficiente e seu erro padrão. Essa métrica é usada para avaliar a significância estatística do coeficiente;
- Significância: indica se o coeficiente é estatisticamente significativo, considerando um nível de significância de $\alpha = 0.05$.

É importante salientar que os valores de t-estatística fora de uma faixa próxima de zero ($|t| > 1.96$) são considerados significativos para $\alpha = 0.05$, de acordo com a distribuição normal padrão. Ainda, o coeficiente -0.000 tem uma t-estatística de 0.00 indicando que ele não é significativo.

Agora, analisando a figura 9 e a tabela 9 conjuntamente, temos que a variável Betweenness Centrality não é significativa, a PageRank e a Average Neighbor Degree têm impacto negativo, enquanto as outras variáveis possuem impacto positivo no modelo.

3.2.1.3 *Air traffic control*

Figura 10 – Coeficientes da regressão linear



Variável	**Coef. (β_i)**	**Erro Padrão (SE_{β_i})**	**t-Estat.(β_i/SE_{β_i})**	**Significativa**
Degree	0.092	0.0062	14.61	Sim
Closeness Centrality	0.033	0.0015	20.89	Sim
Betweenness Centrality	-0.009	0.0013	-6.71	Sim
Eigenvector Centrality	-0.004	0.0015	-2.54	Sim
k-core	0.026	0.0015	17.16	Sim
PageRank	-0.056	0.0050	-11.05	Sim
Average Neighbor Degree	0.010	0.0011	8.47	Sim

Tabela 10 – Coeficientes estimados, erros padrão, estatísticas t e significância com $\alpha = 0.05$.

Na figura 10, observa-se que o coeficiente para a medida de Degree é positivo (0.092), indicando que um aumento no grau de um nó na rede está associado a um aumento na fração final de recuperados. Este resultado é coerente com a correlação positiva anterior, reforçando a ideia de que a conectividade direta na rede desempenha um papel significativo na propagação e recuperação.

A Closeness Centrality também possui um coeficiente positivo (0.033), corroborando a correlação forte observada. Isso reafirma que a proximidade física na rede está associada a uma fração maior de recuperados, indicando a importância das interações próximas na dinâmica de propagação.

Para a Betweenness Centrality, o coeficiente é negativo (-0.009), sugerindo que, na regressão, a medida possui uma relação inversa com a fração final de recuperados. Este resultado contrasta com a correlação positiva anterior, indicando que o papel dos intermediários na propagação e recuperação pode não ser tão forte quando considerado em conjunto com outras medidas.

A Eigenvector Centrality mostra um coeficiente negativo (-0.004), em desacordo com a correlação positiva anterior. Isso pode indicar que, na presença de outras variáveis, a relação entre a posição estratégica na rede e a fração final de recuperados é menos pronunciada.

O coeficiente para a variável k-core é positivo (0.026), reforçando a ideia de que pertencer a núcleos mais densos na rede está associado a uma maior fração final de recuperados, alinhando-se com a correlação anterior.

O coeficiente negativo para a PageRank (-0.056) contradiz a correlação positiva. Isso sugere que, ao considerar outras variáveis na regressão, a importância atribuída pela métrica PageRank pode não ser tão determinante para a fração final de recuperados.

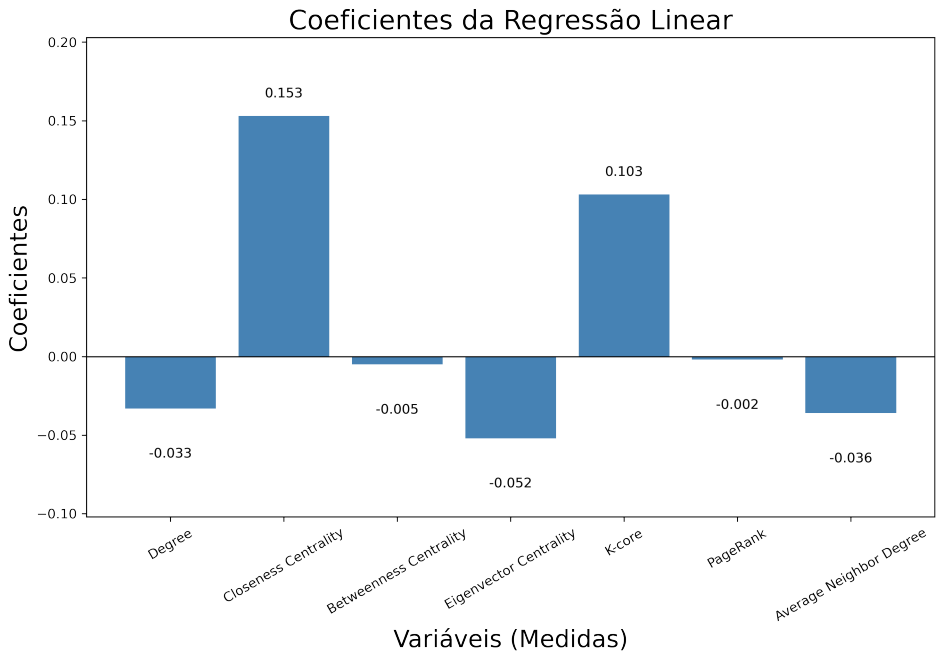
Finalmente, o coeficiente para a Average Neighbor Degree é positivo (0.01), indicando que indivíduos cujos vizinhos têm um alto grau médio de conexões estão associados a uma fração maior de recuperados. Este resultado está alinhado com a correlação positiva anterior.

Em relação à tabela 10, temos que todas as métricas são significativas, porém algumas

positivas e outras negativas. Mais especificamente, Degree, Closeness Centrality, k-core e Average Neighbor Degree têm impacto positivo no modelo, enquanto Betweenness Centrality, Eigenvector Centrality e PageRank têm impacto negativo.

3.2.1.4 OpenFlights

Figura 11 – Coeficientes da regressão linear



Variável	**Coef. (β_i)**	**Erro Padrão (SE_{β_i})**	**t-Estat.(β_i/SE_{β_i})**	**Significativa**
Degree	-0.033	0.0100	-3.27	Sim
Closeness Centrality	0.153	0.0031	49.27	Sim
Betweenness Centrality	-0.005	0.0029	-1.69	Não
Eigenvector Centrality	-0.052	0.0056	-9.26	Sim
k-core	0.103	0.0035	28.66	Sim
PageRank	-0.002	0.0074	-0.27	Não
Average Neighbor Degree	-0.036	0.0023	-15.58	Sim

Tabela 11 – Coeficientes estimados, erros padrão, estatísticas t e significância com $\alpha = 0.05$.

Na figura 11 podemos observar que o coeficiente para a medida Closeness Centrality (0.153) possui o valor mais alto, reafirmando a alta correlação mostrada anteriormente, com a fração final de recuperados. Isso reafirma que a proximidade física na rede está associada a uma fração maior de recuperados.

Ainda, o coeficiente para k-core (0.103) também é um valor considerável, mais uma vez reafirmando o que vimos quanto a sua correlação com a fração final de recuperados, na tabela de correlação de Pearson.

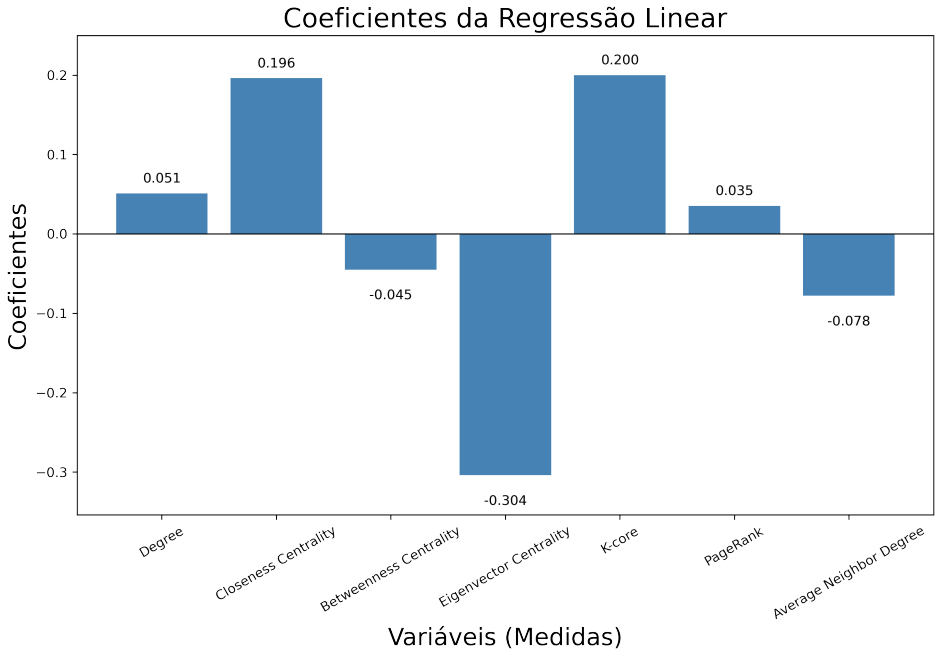
Já os coeficientes para Degree (0.033), Eigenvector (0.052) e Average Neighbor Degree (0.036) apresentam valores positivos mas menos impactantes, portanto não possuem influência tão forte com a fração final de recuperados quanto as medidas anteriores.

Por fim, temos os coeficientes mais fracos que são Betweenness Centrality (0.005) e PageRank (0.002). Nesse último caso, se diferenciou da tabela de correlação de Pearson, na qual a medida PageRank possuía correlação significativa. Isso sugere que, ao considerar outras variáveis na regressão, a importância atribuída pela métrica PageRank pode não ser tão relevante para a fração final de recuperados.

A respeito da tabela 11, podemos observar que as variáveis Betweenness Centrality e PageRank não são significativas no modelo. Ainda, apenas Closeness Centrality e k-core apresentam impacto positivo no modelo, o restante das métricas apresentam impacto negativo.

3.2.1.5 US airports

Figura 12 – Coeficientes da regressão linear



Na figura 12, o coeficiente para a métrica Eigenvector Centrality possui o valor mais forte entre todas as outras métricas (0.304). Isso é o oposto do que obtemos com a correlação de Pearson, indicando que na regressão outras variáveis são mais significativas.

Os coeficientes da Closeness Centrality (0.196) e da k-core (0.200) são significativas na regressão e possuem valores bem próximos. Relacionando com a correlação de Pearson, ambas se destacam no momento anterior, assim como na regressão.

Já os coeficientes de Degree (0.051), Betweenness Centrality (0.045), PageRank (0.035) e Average Neighbor Degree (0.078) são pouco significativas na regressão. Ainda, é importante

Variável	**Coef.** (β_i)	**Erro Padrão** (SE_{β_i})	**t-Estat.**(β_i/SE_{β_i})	**Significativa**
Degree	0.051	0.0306	1.66	Não
Closeness Centrality	0.196	0.0066	29.26	Sim
Betweenness Centrality	-0.045	0.0048	-9.36	Sim
Eigenvector Centrality	-0.304	0.0223	-13.59	Sim
k-core	0.200	0.0105	18.91	Sim
PageRank	0.035	0.0184	1.90	Não
Average Neighbor Degree	-0.078	0.0047	-16.27	Sim

Tabela 12 – Coeficientes estimados, erros padrão, estatísticas t e significância com $\alpha = 0.05$.

destacar essa última métrica em que, aqui, obtemos um valor próximo de zero e positivo, enquanto na correlação de Pearson, obtivemos um valor negativo.

Observando a tabela 12, Degree e PageRank não são significativas, enquanto Closeness Centrality e k-core são significativas e positivas. O restante, são significativas porém negativas no modelo.

3.2.2 Random Forest

O *Random Forest* é um algoritmo de aprendizado de máquina baseado em árvores de decisão [Breiman 2001]. Ele opera construindo um conjunto diversificado de árvores de decisão, cada uma treinada em uma amostra aleatória dos dados de entrada. As árvores individuais são treinadas para fazer previsões, e a previsão final do *Random Forest* é a média (ou a moda, no caso de classificação) das previsões das árvores individuais.

A importância do *Random Forest* reside na sua capacidade de lidar com complexidade e não-linearidades nos dados. Além disso, ele pode lidar com conjuntos de dados grandes e de alta dimensionalidade. Para avaliar a importância das características (medidas de centralidade) em uma previsão, o *Random Forest* utiliza o conceito de "*feature importance*"(importância das características).

As importâncias das covariáveis no *Random Forest* são calculadas com base na redução da impureza (ou ganho de informação) que cada feature proporciona ao dividir os dados durante a construção das árvores. Quanto mais uma feature contribui para a redução da impureza nos nós da árvore, maior é sua importância. O cálculo específico varia, mas em geral, pode-se medir o ganho de informação usando critérios como a impureza de Gini ou entropia [Breiman 2001]. A importância da covariável é então derivada da média ponderada desses ganhos em todas as árvores da floresta. Covariáveis que são mais frequentemente usadas para decisões e resultam em divisões mais significativas terão importâncias mais altas. Essa métrica nos ajuda a entender quais medidas de centralidade contribuem mais para a previsão da fração de recuperados, permitindo a identificação dos nós mais influentes na propagação.

Ambos os métodos, regressão linear e *Random Forest*, nos ajudarão a compreender como as medidas de centralidade impactam a dinâmica de propagação em redes complexas. A combinação de técnicas estatísticas tradicionais com abordagens de aprendizado de máquina mais avançadas nos permitirá obter uma descrição mais completa das relações entre essas variáveis-chave.

A análise da relação entre medidas de centralidade e a fração de recuperados em redes complexas requer métodos sofisticados para capturar a complexidade subjacente. Neste capítulo, expandiremos nossa investigação, concentrando-nos na utilização do algoritmo *Random Forest* para avaliar a importância das características (medidas de centralidade) na previsão da fração de recuperados.

O conceito fundamental por trás da avaliação da importância das características pelo algoritmo Random Forest é a compreensão de como cada medida de centralidade contribui para a variabilidade na previsão da fração de recuperados. A importância da característica é medida pela magnitude de sua influência na melhoria da precisão da previsão quando essa característica é considerada no modelo.

O algoritmo *Random Forest* opera construindo um conjunto diversificado de árvores de decisão, cada uma treinada com uma amostra de dados e uma seleção aleatória de características. A importância de cada característica é calculada observando como a precisão das previsões diminui quando essa característica é aleatoriamente embaralhada nos dados de teste. Quanto maior a diminuição na precisão, mais importante é a característica para o modelo.

Ao analisar as importâncias das covariáveis (*feature importances*), poderemos identificar quais medidas de centralidade têm o maior impacto na previsão da fração de recuperados. Isso nos permitirá classificar as medidas de acordo com sua relevância, destacando quais são os nós mais influentes na propagação de doenças ou informações na rede. A interpretação das *feature importances* também nos dá compreensão sobre as interações complexas entre as medidas de centralidade e como elas afetam a dinâmica de propagação.

Ao combinarmos a análise de regressão linear com a avaliação das *feature importances* pelo *Random Forest*, obteremos uma compreensão abrangente das medidas de centralidade que desempenham um papel crucial na previsão da fração final de recuperados. Através desses métodos, seremos capazes de compreender as características dos nós mais influentes em diferentes tipos de redes complexas, contribuindo para o avanço do entendimento das dinâmicas de propagação e tomada de decisões informadas em diversas áreas.

Após as análises iniciais do *Random Forest*, será aplicado o SHAP Values. Os valores SHAP (*SHapley Additive exPlanations*) de uma observação são calculados usando um conceito da teoria dos jogos chamado Valores de Shapley. Esse método distribui "pagamentos"(no contexto de modelos de *machine learning*, as previsões) entre os "jogadores"(as *features*) de forma que cada um receba uma parcela justa, baseada na sua contribuição para o resultado final [Shapley

1953]. Aqui está uma visão geral de como um valor SHAP é calculado para uma observação específica:

- Conjunto de Coalizões: Cada "coalizão"é um subconjunto possível de covariáveis que poderiam ser incluídas no modelo. Por exemplo, para três covariáveis, as coalizões possíveis incluem: apenas a covariável 1, apenas a covariável 2, apenas a covariável 3, as covariáveis 1 e 2, as covariáveis 1 e 3, as covariáveis 2 e 3, todas as três covariáveis, ou nenhuma covariáveis.

- Contribuição Marginal: Para cada coalizão, o modelo é avaliado duas vezes: uma vez com a *feature* de interesse incluída e uma vez sem ela. A diferença nas previsões dessas duas avaliações é a contribuição marginal da *feature* para aquela coalizão específica.

- Média Ponderada das Contribuições: A contribuição marginal de uma *feature* é calculada para cada possível coalizão de outras *features*. O valor SHAP de uma *feature* é a média ponderada dessas contribuições marginais, onde a ponderação é determinada pelo número de maneiras que a coalizão poderia ser formada. Isso assegura que todas as possíveis ordens de entrada das *features* sejam consideradas.

- Permutações de covariáveis: Para calcular essas médias ponderadas, todas as permutações possíveis das covariáveis são consideradas. Isso garante que o efeito de cada covariável seja avaliado de forma justa, independentemente da ordem em que as covariáveis entram no modelo.

O gráfico gerado pelo SHAP Value fornece uma visão geral de como as diferentes *features* (ou variáveis) contribuem para o resultado do modelo.

- Importância das *Features*: As *features* são listadas no eixo Y, ordenadas da mais importante para a menos importante, de cima para baixo.

- Valores SHAP no Eixo X: O eixo X mostra os valores SHAP, que indicam quanto cada *feature* impacta a previsão do modelo.

- Distribuição dos Pontos (Cor e Densidade): Cada ponto no gráfico representa um único exemplo (ou linha) do conjunto de dados. A cor dos pontos indica o valor da *feature* para aquele exemplo: azul representa valores baixos da *feature*, enquanto rosa representa valores altos. As cores mais escuras (tipicamente azuis) indicam valores mais baixos da *feature* para a observação específica. Isso significa que, para esses pontos, a *feature* tem um valor menor. Já as cores mais claras (tipicamente rosas ou vermelhas), indicam valores mais altos da *feature*. Para esses pontos, a *feature* tem um valor maior.

- Espalhamento dos Pontos: O espalhamento horizontal dos pontos mostra a variabilidade no impacto das *features*. Covariáveis com uma maior dispersão horizontal têm um impacto mais variável nas previsões.

Os valores de SHAP no eixo X representam o impacto quantitativo que cada covariável tem sobre a previsão do modelo. Eles são medidos em termos de mudança na saída do modelo causada por ter a covariável em questão presente em uma determinada previsão.

Analisando o significado dos valores do SHAP, o primeiro passo é uma visão crítica do impacto positivo ou negativo.

- Valores Positivos de SHAP: Indicam que a presença da *feature* no modelo contribui para aumentar a previsão do modelo em relação ao valor base (ou expectativa). Por exemplo, se estivermos prevendo o preço de uma casa, uma *feature* com um valor SHAP positivo aumentaria o preço previsto.
- Valores Negativos de SHAP: Indicam que a presença da *feature* contribui para diminuir a previsão do modelo em relação ao valor base. Usando o mesmo exemplo, isso significaria uma redução no preço previsto da casa.

Ainda, sobre a magnitude dos valores de SHAP:

- Magnitude Alta: Significa que a *feature* tem um impacto substancial na alteração da previsão do modelo. Quanto maior o valor absoluto do SHAP (positivo ou negativo), maior é o impacto da *feature*.
- Magnitude Baixa: Indica um impacto menor da *feature* na previsão do modelo.

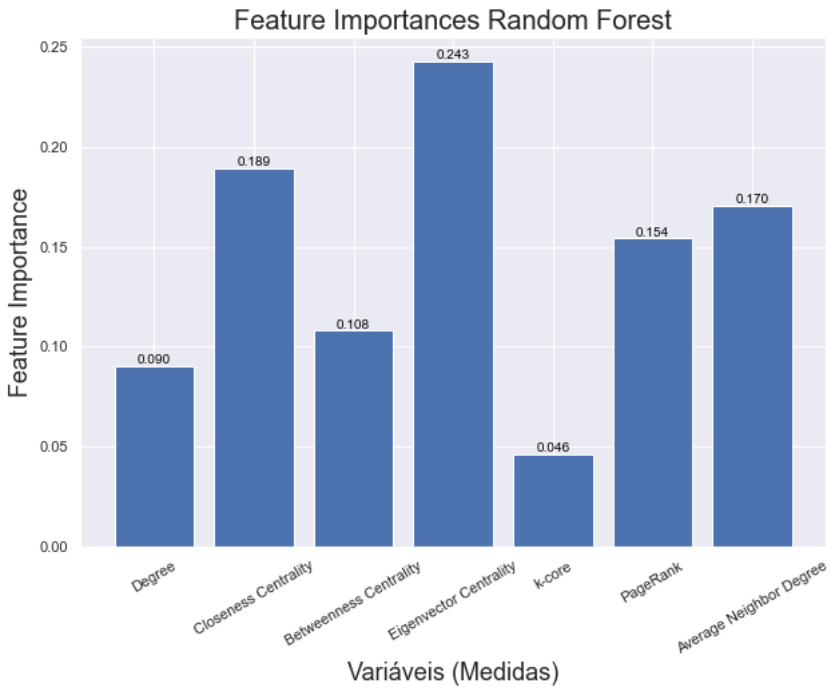
Por fim, há o valor base (Expected Value), o qual se refere ao valor médio de previsão do modelo quando nenhuma das features está sendo considerada. Cada valor SHAP é uma contribuição à diferença entre a previsão real de uma instância e esse valor médio.

3.2.2.1 Netscience

Ao avaliar os valores das feature importances em um modelo de Random Forest em relação às medidas de centralidade (Degree, Closeness Centrality, Betweenness Centrality, Eigenvector Centrality, k-core, PageRank e Average Neighbor Degree) e sua influência na fração final de recuperados, podemos discernir quais métricas têm maior peso na capacidade do modelo de fazer previsões.

As *feature importances* indicam o grau de contribuição de cada variável para as previsões do modelo. Nesse contexto, destacam-se os seguintes resultados:

Figura 13 – Features importance



A métrica de Eigenvector Centrality se destaca com uma *feature importance* significativa, atingindo um valor de 0.24. Isso sugere que, de acordo com o modelo de *Random Forest*, a Eigenvector Centrality desempenha um papel preponderante na previsão da fração final de recuperados.

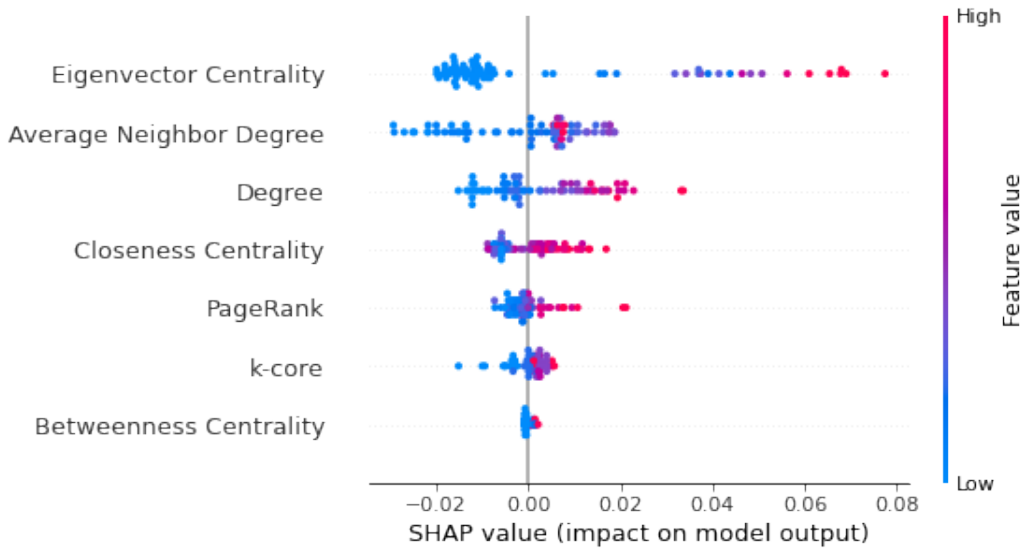
Closeness Centrality e Average Neighbor Degree também apresentam *feature importances* consideráveis, com valores de 0.19 e 0.17, respectivamente. Isso indica que essas métricas têm uma influência significativa nas previsões do modelo em relação à fração final de recuperados.

Degree, PageRank e Betweenness Centrality, embora tenham feature importances menores em comparação com as métricas acima, ainda contribuem para a capacidade do modelo de fazer previsões. Seus valores são, respectivamente, 0.09, 0.15 e 0.11.

Em resumo, a análise das *feature importances* destaca que a Eigenvector Centrality, Closeness Centrality e Average Neighbor Degree são as métricas que mais influenciam as previsões do modelo em relação à fração final de recuperados, de acordo com o modelo de *Random Forest* utilizado. Por outro lado, Degree, PageRank e Betweenness Centrality também têm um papel a desempenhar, embora com menos peso em comparação com as métricas mencionadas anteriormente.

A figura 14 é um gráfico gerado com a dispersão dos valores SHAP. Por meio dele, é possível observar que a característica Eigenvector Centrality tem um impacto predominantemente negativo no modelo, especialmente quando seus valores são baixos (azul). À medida que os valores aumentam (em direção ao vermelho), o impacto se torna mais próximo de zero, mas ainda existem alguns casos onde o impacto pode ser positivo.

Figura 14 – SHAP Value



A Closeness Centrality e a Degree têm uma distribuição de impacto bastante balanceada entre positivo e negativo, sugerindo que o impacto pode variar bastante dependendo do valor da característica.

Já a PageRank e a Betweenness Centrality têm impactos menores e menos variáveis no modelo, sugerindo que, em média, estas características são menos importantes ou têm menos influência na previsão do modelo.

Enquanto isso, a k-core, embora tenha menos variabilidade de impacto, ainda pode contribuir significativamente para as previsões, especialmente quando os valores são mais altos.

No geral, o gráfico sugere que Eigenvector Centrality e Degree são características influentes no modelo, enquanto Betweenness Centrality tem um impacto menor.

3.2.2.2 Air traffic control

Notavelmente, a proximidade física (Closeness Centrality) e a pertença a núcleos mais densos (k-core) são características mais determinantes, refletindo-se em altas importâncias. Isso contrasta com a contribuição mais modesta do grau individual (Degree) e da centralidade de intermediação (Betweenness Centrality). Embora a correlação e os coeficientes da regressão linear forneçam perspectivas valiosas, as importâncias das características destacam como o modelo Random Forest avalia a relevância de cada medida na previsão da dinâmica de propagação e recuperação em redes complexas durante surtos epidêmicos.

Este novo summary plot (figura 16) dos valores SHAP mostra a importância e o impacto das características no modelo de machine learning, com algumas diferenças em relação ao gráfico anterior. Vamos interpretar os principais pontos.

- k-core: esta característica apresenta um impacto positivo e significativo no modelo, espe-

Figura 15 – Features importance

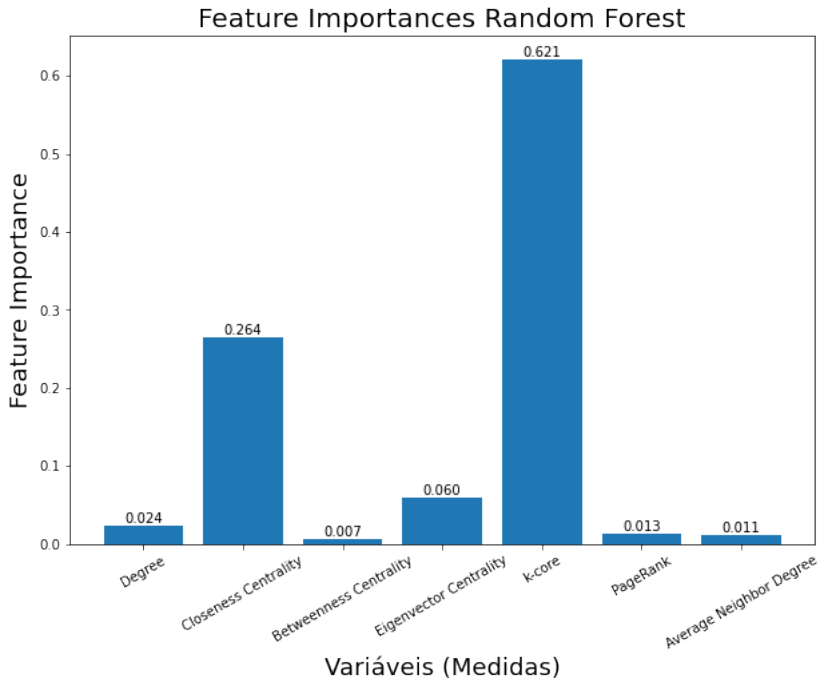
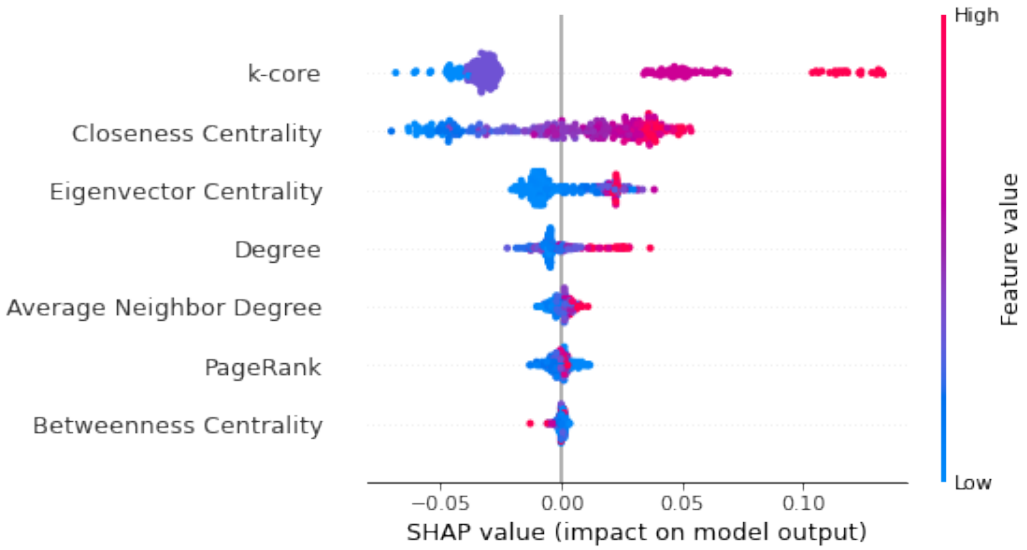


Figura 16 – SHAP Value



cialmente para valores altos (vermelho). Valores baixos (azul) tendem a ter um impacto negativo, mas menos pronunciado. Aparece no topo da lista, indicando ser uma das características mais importantes para o modelo.

- Closeness Centrality: Tem um impacto positivo quando os valores são altos (vermelho) e um impacto negativo quando os valores são baixos (azul). A dispersão dos pontos sugere uma variação considerável do impacto. Também é uma característica importante, mas com uma distribuição de impacto mais ampla.

- Eigenvector Centrality: Similar ao gráfico anterior, o impacto varia dependendo do valor

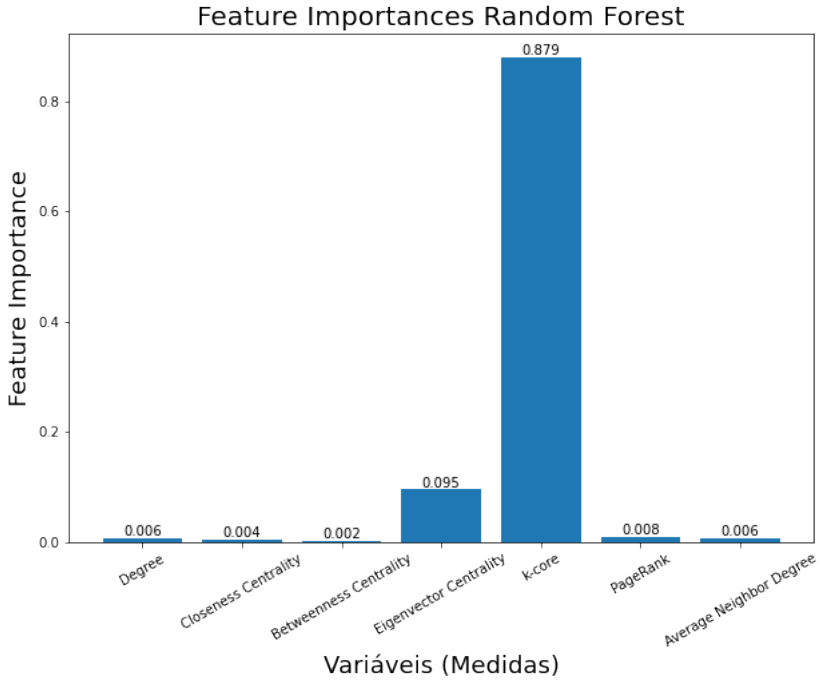
da característica, mas aqui a dispersão é mais concentrada, sugerindo que a maioria dos impactos são pequenos e próximos de zero. Continua sendo uma característica influente, mas com menos variabilidade comparada ao k-core.

- Degree: O impacto é distribuído tanto em valores positivos quanto negativos, dependendo dos valores da característica. Mantém-se como uma característica de relevância moderada no modelo.
- Average Neighbor Degree e PageRank: Ambas apresentam impactos variáveis, mas geralmente menores, com a maioria dos pontos próximos a zero. Isso sugere que esses atributos não têm um impacto tão grande quanto os outros. Estão mais abaixo na lista, indicando menor importância em comparação com outras características.
- Betweenness Centrality: Esta característica parece ter o menor impacto de todas, com a maioria dos pontos concentrados em torno de zero. Está no final da lista, sugerindo que é a característica menos relevante no modelo.

Portanto, k-core e Closeness Centrality são as características mais influentes, com o k-core mostrando um impacto positivo mais forte à medida que o valor da característica aumenta.

3.2.2.3 *OpenFlights*

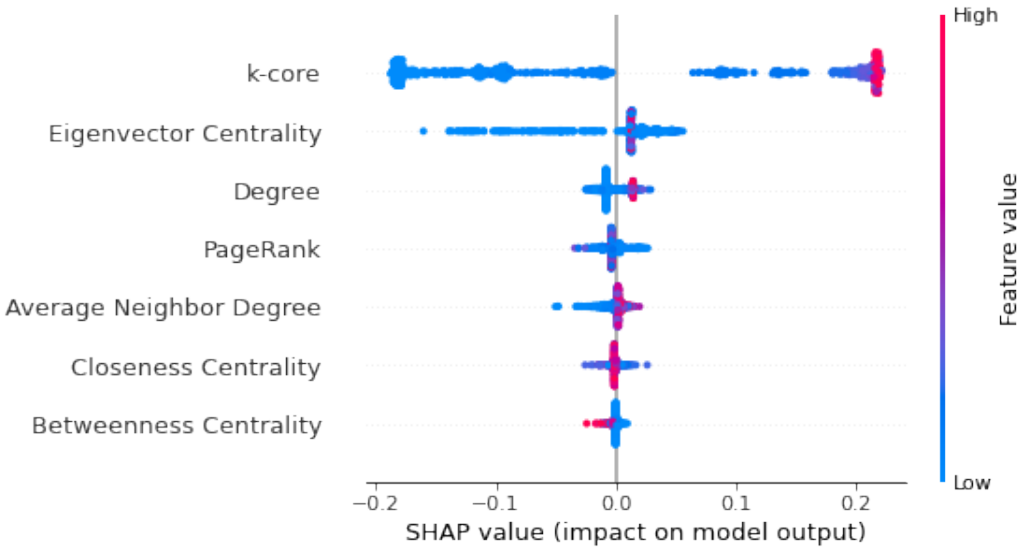
Figura 17 – Features importance



Na figura 17 fica evidente que a medida k-core (0.879) é a característica mais determinante no modelo para se prever a fração final de recuperados. A medida Eingenvector Centrality (0.095) também possui a sua importância, embora menos densa.

Em contrapartida, Degree (0.006), Closeness Centrality (0.004), Betweenness Centrality (0.002), PageRank (0.008), Average Neighbor Degree (0.006) não contribuem significativamente para as previsões do modelo.

Figura 18 – SHAP Value



Neste summary plot dos valores SHAP (figura 18), o impacto das características no modelo de machine learning é apresentado de forma similar aos gráficos anteriores, mas com algumas diferenças importantes nos valores e na distribuição dos impactos. Vamos analisar cada um dos pontos principais:

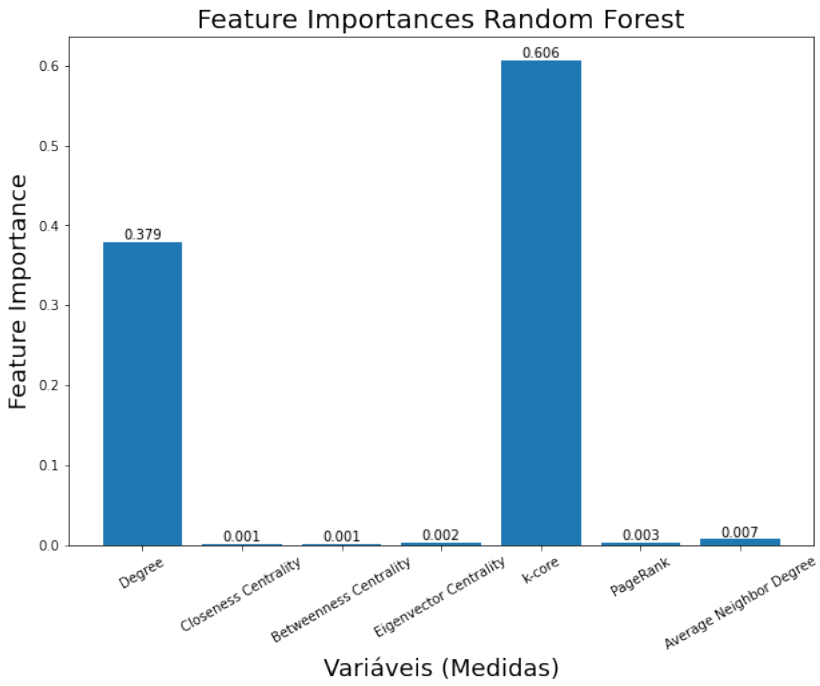
- k-core: Esta característica tem um impacto significativo no modelo, com valores SHAP que podem ser tanto positivos quanto negativos. Notavelmente, os valores altos de k-core (vermelho) tendem a ter um impacto positivo expressivo, enquanto valores baixos (azul) têm um impacto mais variável, incluindo valores negativos significativos.
- Closeness Centrality e Betweenness Centrality: Ambas as características têm impactos bastante pequenos, com a maioria dos valores de SHAP concentrados em torno de zero, sugerindo que elas têm uma influência mínima no modelo.
- Eigenvector Centrality: Tem uma distribuição mais concentrada em torno de zero, mas ainda apresenta impactos significativos tanto positivos quanto negativos, dependendo do valor da característica. Valores altos de Eigenvector Centrality (vermelho) tendem a contribuir positivamente para o modelo.
- Degree: Mostra uma distribuição de impacto mais centrada em torno de zero, com alguns valores extremos que podem ter um impacto positivo ou negativo significativo. Isso sugere que o impacto de Degree no modelo é mais variável.

- Average Neighbor Degree: A característica apresenta uma distribuição equilibrada em torno de zero, com impactos majoritariamente pequenos, tanto positivos quanto negativos.
- PageRank: A maioria dos valores de SHAP para PageRank está concentrada perto de zero, com uma ligeira tendência positiva para valores mais altos (vermelho). O impacto geral é relativamente pequeno.

Em resumo, k-core é a característica mais importante e tem um impacto substancial no modelo, especialmente quando seus valores são altos. Já a Eigenvector Centrality também é influente, mas com uma variação de impacto menor.

3.2.2.4 US airports

Figura 19 – Features importance

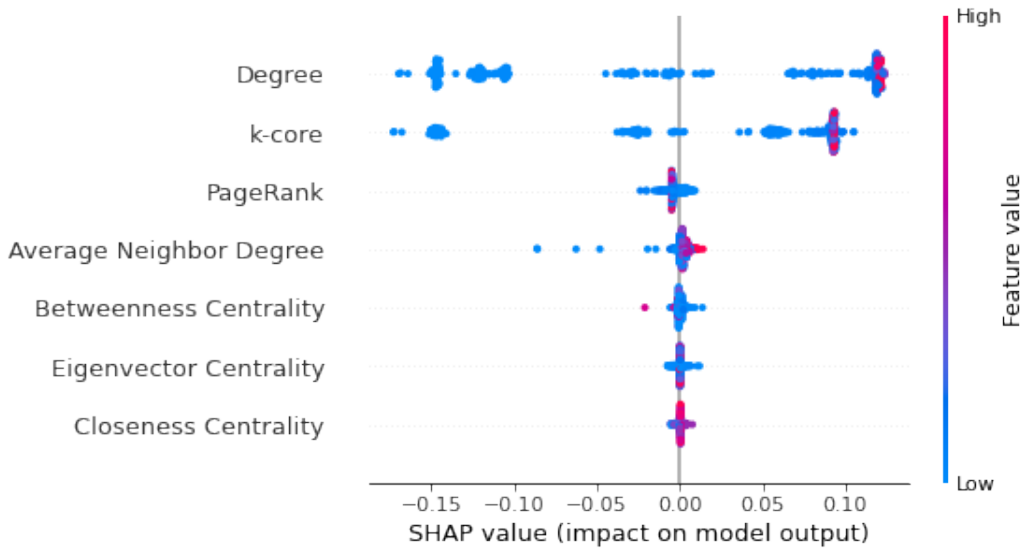


Para a rede US airports, k-core (0.606) e Degree (0.379) são as mais significativas, enquanto as restantes (Closeness Centrality (0.001), Betweenness Centrality (0.001), Eigenvector Centrality (0.002), PageRank (0.003) e Average Neighbor Degree (0.007)) são pouco relevantes para as previsões do modelo.

Isso indica que o grau dos nós e o grupo de nós altamente interligados são os aspectos que influenciam no modelo para a previsão da fração final de recuperados.

Por meio da figura 20, pode-se interpretar que:

Figura 20 – SHAP Value



- k-core: Tem uma distribuição ampla de valores SHAP, com impactos tanto positivos quanto negativos, mas geralmente com uma tendência de impacto positivo, especialmente para valores mais altos.
- Betweenness Centrality: A maioria dos valores SHAP para esta característica está concentrada em torno de zero, sugerindo que tem pouco impacto nas previsões do modelo, com algumas pequenas variações negativas.
- Eigenvector Centrality e Closeness Centrality: Essas características mostram pouca variação nos valores SHAP, com a maioria dos impactos próximos a zero. Isso indica que têm um impacto mínimo nas previsões do modelo.
- Degree: Aparece como a característica mais importante no modelo, com uma distribuição de valores SHAP que pode ser tanto positiva quanto negativa. Valores altos de Degree (vermelho) tendem a ter um impacto positivo considerável no modelo, enquanto valores baixos (azul) podem ter um impacto negativo ou muito próximo de zero.
- Average Neighbor Degree: A maioria dos impactos está próxima de zero, com uma ligeira tendência para impactos negativos, especialmente para valores mais baixos.
- PageRank: Apresenta uma distribuição mais concentrada em torno de zero, mas com alguns valores que indicam impacto positivo, principalmente quando a característica possui valores altos.

Desse modo, pode-se concluir que Degree e k-core são as características mais influentes no modelo, com Degree tendo um impacto ligeiramente maior, enquanto a PageRank tem uma importância moderada, mas ainda contribui de maneira significativa em alguns casos. Já Average Neighbor Degree, Betweenness Centrality, Eigenvector Centrality, e Closeness Centrality têm

menos impacto, com a maioria dos valores SHAP perto de zero, indicando que têm uma influência mínima no modelo.

CAPÍTULO

4

CONCLUSÕES

Nossas análises revelaram a importância das medidas de centralidade na previsão da fração final de recuperados. Através de correlações de Pearson, observamos que Closeness Centrality e k-core apresentaram as correlações mais fortes e positivas com a fração final de recuperados em quase todas as redes estudadas, sugerindo que indivíduos bem conectados ou próximos tendem a ter um impacto positivo na recuperação. Por outro lado, Betweenness Centrality e Average Neighbour Degree foram as medidas com menores correlações em praticamente todas as redes, o que nos indica que essas medidas não são medidas tão boas para identificar os principais propagadores em redes de aeroportos. Ainda, é válido destacar que dependendo do valor de β escolhido, a correlação das centralidades com o número de infectados muda. O valor escolhido de $\beta = 0.3$ representa uma relação positiva, mas de magnitude moderada.

Além disso, utilizamos um modelo de regressão linear e *Random Forest* para avaliar a importância relativa de cada medida de centralidade na previsão da fração final de recuperados. E os resultados corroboram em sua maioria com os valores das correlações, onde as medidas Betweenness Centrality e Average Neighbor Degree apresentam as menores importâncias. Assim como k-core e Closeness Centrality sempre desempenham altas importâncias em ambos os métodos.

A análise dos resultados obtidos nas regressões lineares para as diferentes redes (Netscience, Air traffic control, OpenFlights e US airports) nos permitiu compreender o papel das medidas de centralidade na dinâmica de recuperação, refletida pela fração final de indivíduos recuperados.

Na rede "Netscience", Degree e Closeness Centrality têm coeficientes positivos significativos, indicando que indivíduos mais centrais ou próximos na rede favorecem a recuperação. Por outro lado, Betweenness Centrality, embora positiva, tem um coeficiente muito próximo de zero, sugerindo impacto limitado. Eigenvector Centrality e k-core também têm coeficientes positivos, mas menores, enquanto PageRank apresenta um coeficiente negativo considerável, indicando

que maior importância na rede pode prejudicar a recuperação. A Average Neighbor Degree também tem um coeficiente negativo com impacto ligeiramente negativo. Em resumo, Degree e Closeness Centrality são os principais impulsionadores da recuperação, enquanto PageRank parece ter efeito adverso.

Na rede "Air traffic control", os resultados corroboraram a importância de Degree e Closeness Centrality, cujos coeficientes positivos reforçam o papel da conectividade direta e proximidade física na recuperação. No entanto, a Betweenness Centrality e a Eigenvector Centrality mostraram coeficientes negativos, sugerindo que, ao considerar a rede como um todo, o impacto de indivíduos intermediários ou de alta posição estratégica pode ser limitado. O PageRank também apresentou um coeficiente negativo, o que sugere que a centralidade percebida pela importância dos nós na rede não necessariamente promove a recuperação.

Para a rede "OpenFlights", destacam-se os fortes coeficientes positivos de Closeness Centrality e k-core, reforçando a relevância dessas métricas na recuperação. Embora outras métricas como Degree e Eigenvector Centrality também tenham mostrado valores positivos, o impacto foi menos pronunciado. O PageRank, inicialmente indicativo de uma correlação positiva, perdeu sua relevância ao ser analisado na regressão, evidenciando que, neste contexto, sua importância é reduzida em comparação a outras variáveis.

Na rede "US airports", o coeficiente mais forte foi para Eigenvector Centrality, sugerindo que a posição estratégica, mais do que a simples conectividade ou proximidade, exerce maior influência na recuperação. Além disso, as métricas Closeness Centrality e k-core também se mostraram significativas, corroborando os resultados observados em redes anteriores. Já o impacto de Degree, Betweenness Centrality, PageRank e Average Neighbor Degree foi mais limitado, com coeficientes baixos, refletindo uma influência reduzida na recuperação, especialmente em relação à Average Neighbor Degree, cujos resultados diferiram da correlação inicial.

Em síntese, os coeficientes de regressão indicam que as medidas de centralidade de Degree e Closeness Centrality têm um papel consistente e significativo na recuperação em diversas redes, com k-core e Eigenvector Centrality também apresentando impactos positivos, embora mais moderados. Em contrapartida, o PageRank e a Average Neighbor Degree mostraram-se menos relevantes ou até adversos, dependendo da rede analisada.

Ainda, é válido destacar que calcular o erro no contexto de um modelo de regressão linear, usando testes como o teste t usado, é essencial para avaliar a qualidade e confiabilidade do modelo estatístico. Esses achados oferecem importantes contribuições para a compreensão do papel das métricas de centralidade na análise de redes complexas e podem servir de base para a tomada de decisões em contextos em que a propagação e a recuperação desempenham papéis cruciais.

Já no *Random Forest*, a análise dos resultados obtidos para as redes "Netscience", "Air traffic control", "OpenFlights"e "US airports"proporciona uma visão abrangente sobre a

importância relativa das diferentes métricas de centralidade para a previsão da fração final de recuperados. De modo geral, as *feature importances* e os valores SHAP ajudam a identificar as variáveis mais influentes em cada rede, contribuindo para a compreensão das dinâmicas de propagação e recuperação.

Na rede "Netscience", as métricas Eigenvector Centrality, Closeness Centrality e Average Neighbor Degree se destacam com as maiores importâncias, indicando que essas características são cruciais para o modelo de Random Forest prever a fração final de recuperados. A Eigenvector Centrality tem um impacto predominante negativo, mas à medida que seus valores aumentam, o impacto tende a ser menos negativo. Degree, PageRank e Betweenness Centrality, embora importantes, possuem importâncias menores, sugerindo um papel menos significativo na previsão da recuperação.

Para a rede "Air traffic control", as métricas Closeness Centrality e k-core emergem como as mais determinantes, com impactos positivos mais pronunciados, especialmente para valores altos dessas características. Em comparação, Degree e Betweenness Centrality apresentam contribuições menores, enquanto PageRank tem impacto quase nulo no modelo. A análise SHAP revela uma variabilidade considerável no impacto de Closeness Centrality, com valores tanto positivos quanto negativos, o que pode refletir a complexidade da proximidade física na rede.

Na rede "OpenFlights", a métrica k-core tem um impacto substancial, sendo a característica mais importante para a previsão da recuperação. Eigenvector Centrality, apesar de sua importância, apresenta uma variação de impacto mais concentrada, sugerindo uma relação mais complexa com a fração de recuperados. As demais métricas, como Degree, Closeness Centrality, e PageRank, têm impactos menores, refletindo sua menor relevância para a previsão da recuperação.

Por fim, na rede "US airports", tanto k-core quanto Degree se destacam como as variáveis mais influentes, com o Degree mostrando um impacto ligeiramente mais significativo. As outras características, como Closeness Centrality, Betweenness Centrality, Eigenvector Centrality e PageRank, apresentam impactos muito pequenos, com a maioria dos valores SHAP próximos de zero, indicando que essas métricas têm uma influência mínima no modelo.

Em resumo, os resultados mostram que, embora as métricas de Degree, Closeness Centrality, k-core e Eigenvector Centrality sejam geralmente as mais relevantes, sua importância varia consideravelmente de acordo com a rede analisada. k-core surge como uma característica central em várias redes, enquanto Betweenness Centrality e PageRank demonstram uma contribuição mais modesta. Essa análise fornece uma compreensão detalhada de como diferentes medidas de centralidade influenciam a dinâmica de recuperação, nos permitindo compreender a modelagem de redes complexas e a tomada de decisões em contextos de propagação e recuperação.

Em síntese, a identificação de propagadores influentes em redes complexas fornece uma base sólida para aprimorar a tomada de decisões em sistemas reais, especialmente em contextos

que envolvem propagação dinâmica. No caso de uma doença infecciosa, por exemplo, saber quais indivíduos ou localidades exercem papel central na disseminação permite direcionar campanhas de vacinação, ações de contenção e alocação de recursos de forma mais eficiente. Em vez de medidas generalizadas, é possível atuar de maneira específica sobre os nós mais críticos da rede, reduzindo significativamente o alcance e a velocidade da transmissão. Essa abordagem baseada na estrutura da rede potencializa a eficácia das intervenções e reforça a importância de incorporar análises topológicas na formulação de políticas públicas e estratégias operacionais.

REFERÊNCIAS

ADAMIC, L.; HUBERMAN, B. The web's hidden order. **Commun. ACM**, v. 44, p. 55–59, 09 2001. Citado na página 10.

ARRUDA, G. F. de; BARBIERI, A. L.; RODRIGUEZ, P. M.; MORENO, Y.; COSTA, L. da F.; RODRIGUES, F. A. **The role of centrality for the identification of influential spreaders in complex networks**. 2014. Disponível em: <https://arxiv.org/abs/1404.4528>. Citado na página 15.

BAÑOS, R. A.; BORGE-HOLTHOEFER, J.; MORENO, Y. The role of hidden influentials in the diffusion of online information cascades. **EPJ Data Science**, v. 2, n. 1, p. 6, Jul 2013. ISSN 2193-1127. Disponível em: <https://doi.org/10.1140/epjds18>. Citado na página 10.

BARABÁSI, A.-L. **Network Science**. Cambridge, UK: Cambridge University Press, 2018. ISBN 9781107076266. Disponível em: <http://networksciencebook.com/>. Citado nas páginas 9 e 13.

BARABáSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **Science**, v. 286, n. 5439, p. 509–512, 1999. Disponível em: <https://www.science.org/doi/abs/10.1126/science.286.5439.509>. Citado nas páginas 15, 18 e 20.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001. ISSN 1573-0565. Disponível em: <https://doi.org/10.1023/A:1010933404324>. Citado na página 43.

COHEN, R.; HAVLIN, S. **Complex Networks: Structure, Robustness and Function**. [S.l.: s.n.], 2010. ISBN 978-0521841566. Citado na página 14.

KEMPE, D.; KLEINBERG, J.; TARDOS, E. Maximizing the spread of influence through a social network. In: **Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2003. (KDD '03), p. 137–146. ISBN 1581137370. Disponível em: <https://doi.org/10.1145/956750.956769>. Citado na página 10.

KITSAK, M.; GALLOS, L. K.; HAVLIN, S.; LILJEROS, F.; MUCHNIK, L.; STANLEY, H. E.; MAKSE, H. A. Identification of influential spreaders in complex networks. **Nature Physics**, v. 6, n. 11, p. 888–893, Nov 2010. ISSN 1745-2481. Disponível em: <https://doi.org/10.1038/nphys1746>. Citado na página 10.

LESKOVEC, J.; KRAUSE, A.; GUESTRIN, C.; FALOUTSOS, C.; VANBRIESEN, J.; GLANCE, N. Cost-effective outbreak detection in networks. In: **Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2007. (KDD '07), p. 420–429. ISBN 9781595936097. Disponível em: <https://doi.org/10.1145/1281192.1281239>. Citado na página 10.

MEYER, C. D. Matrix analysis and applied linear algebra. In: . [s.n.], 2000. Disponível em: <https://api.semanticscholar.org/CorpusID:118122532>. Citado na página 18.

MILO, R.; ITZKOVITZ, S.; KASHTAN, N.; LEVITT, R.; SHEN-ORR, S.; AYZENSHTAT, I.; SHEFFER, M.; ALON, U. Superfamilies of evolved and designed networks. **Science**, v. 303, n. 5663, p. 1538–1542, 2004. Disponível em: <https://www.science.org/doi/abs/10.1126/science.1089167>. Citado na página 15.

NEWMAN, M. **Networks: An Introduction**. Oxford University Press, 2010. ISBN 9780199206650. Disponível em: <https://doi.org/10.1093/acprof:oso/9780199206650.001.0001>. Citado nas páginas 13, 16, 17 e 19.

NEWMAN, M.; GIRVAN, M. Finding and evaluating community structure in networks. **Physical review. E, Statistical, nonlinear, and soft matter physics**, v. 69, p. 026113, 03 2004. Citado na página 15.

NEWMAN, M. E. J. The structure of scientific collaboration networks. **Proceedings of the National Academy of Sciences**, Proceedings of the National Academy of Sciences, v. 98, n. 2, p. 404–409, jan. 2001. ISSN 1091-6490. Disponível em: <http://dx.doi.org/10.1073/pnas.98.2.404>. Citado na página 22.

PASTOR-SATORRAS, R.; CASTELLANO, C.; MIEGHEM, P. V.; VESPIGNANI, A. Epidemic processes in complex networks. **Rev. Mod. Phys.**, American Physical Society, v. 87, p. 925–979, Aug 2015. Disponível em: <https://link.aps.org/doi/10.1103/RevModPhys.87.925>. Citado nas páginas 9 e 19.

RODRIGUES, F. A.; PERON, T.; CONNAUGHTON, C.; KURTHS, J.; MORENO, Y. A machine learning approach to predicting dynamical observables from network structure. **Proceedings of the Royal Society A**, v. 481, n. 20240435, 2025. Citado na página 15.

SEIDMAN, S. B. Network structure and minimum degree. **Social Networks**, v. 5, n. 3, p. 269–287, 1983. ISSN 0378-8733. Disponível em: <https://www.sciencedirect.com/science/article/pii/037887338390028X>. Citado na página 17.

SHAPLEY, L. S. 17. a value for n-person games. In: ____. **Contributions to the Theory of Games, Volume II**. Princeton: Princeton University Press, 1953. p. 307–318. ISBN 9781400881970. Disponível em: <https://doi.org/10.1515/9781400881970-018>. Citado na página 44.

WATTS, D. J.; STROGATZ, S. H. **Collective dynamics of 'small-world' networks.** [S.l.], 1998. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/9623998/>. Citado nas páginas 13, 15 e 19.


ICMC
USP
SÃO CARLOS