UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

FABRICIO CÉSAR TOFANI

# ANÁLISE DOS PROGRAMAS DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO DO BRASIL A PARTIR DE GRAFOS DE COLABORAÇÃO

Ouro Preto
2026

FABRICIO CÉSAR TOFANI

# ANÁLISE DOS PROGRAMAS DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO DO BRASIL A PARTIR DE GRAFOS DE COLABORAÇÃO

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Vander Luis de Souza Freitas
Coorientador: Prof. Dr. Eduardo José da Silva Luz

Ouro Preto
2026

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

T644a Tofani, Fabricio César.
Análise dos programas de pós-graduação em Ciência da Computação do Brasil a partir de grafos de colaboração. [manuscrito] / Fabricio César Tofani. - 2026.
30 f.: il.: color., gráf., tab..

Orientador: Prof. Dr. Vander Luiz de Souza Freitas.
Coorientador: Prof. Dr. Eduardo José da Silva Luz.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Biológicas. Graduação em Ciência da Computação .

1. Redes de computadores - Cooperação. 2. Teoria dos grafos. 3. Ciência. 4. Banco de dados bibliográficos. 5. Bibliometria. I. Freitas, Vander Luiz de Souza. II. Luz, Eduardo José da Silva. III. Universidade Federal de Ouro Preto. IV. Título.

CDU 681.3.04.2

Bibliotecário(a) Responsável: Renata Mara de Almeida - CRB-6: 2507

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

# FOLHA DE APROVAÇÃO

**Fabricio Cesar Tofani**

**Grafos de colaboração em publicações de professores credenciados em programas de Pós-Graduação em Ciência da Computação no Brasil**

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 24 de Julho de 2026.

Membros da banca

Vander Luis de Souza Freitas (Orientador) - Doutor - Universidade Federal de Ouro Preto
Eduardo José da Silva Luz (Coorientador) - Doutor - Universidade Federal de Ouro Preto
Gladston Juliano Prates Moreira (Examinador) - Doutor - Universidade Federal de Ouro Preto
Fernando Henrique Oliveira Duarte (Examinador) - Mestre - Universidade Federal de Ouro Preto

Vander Luis de Souza Freitas, Orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 31/07/2026.

Documento assinado eletronicamente por **Vander Luis de Souza Freitas**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 31/07/2026, às 15:35, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **1142659** e o código CRC **C3AC54B5**.



R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591692 - www.ufop.br

# Resumo

No Brasil, a cada quatro anos, a CAPES avalia os Programas de Pós-Graduação (PPGs) das diversas instituições de ensino brasileiras e lhes atribui uma nota correspondente ao seu desempenho, baseada em uma gama de fatores como produção de artigos, número de docentes e discentes, entre outros. Esta monografia avalia se apenas a produção de artigos e a colaboração dos autores são suficientes para explicar as notas dos programas de Pós-Graduação em Ciência da Computação. Para tal, constrói-se uma rede de colaboração a partir de uma base de dados montada com informações obtidas no relatório da CAPES de 2022 e por meio de consultas à OpenAlex, uma base de dados de autores, artigos e instituições relacionadas à pesquisa científica, na qual os nós representam autores e as arestas representam a coautoria em um mesmo artigo. A partir dessa rede computam-se diversas métricas, como grau, PageRank e coeficiente de clusterização, além da detecção de comunidades. Os resultados mostram que a estrutura de colaboração guarda relação com as notas: no nível do programa, a predição classifica cerca de 91% dos programas dentro de uma nota da real, e a separação entre programas medianos e de excelência atinge cerca de 80% de acurácia. Observa-se ainda que a produtividade dos docentes e a topologia da colaboração são os fatores mais associados às notas, enquanto ponderar as arestas por citações ou produtividade tende a diluir esse sinal.

**Palavras-chave**: Redes de colaboração, Ciência das redes, *Science of science*, OpenAlex, Dados bibliométricos.

# Abstract

In Brazil, every four years, CAPES evaluates the Postgraduate Programs (PPGs) of the various Brazilian educational institutions and assigns them a grade corresponding to their performance, based on a range of factors such as article production, number of faculty and students, among others. This monograph evaluates whether article production and author collaboration alone are enough to explain the grades of Postgraduate programs in Computer Science. To this end, a collaboration network is built from a database assembled with information obtained from the 2022 CAPES report and from queries to OpenAlex, a database of authors, articles and institutions related to scientific research, in which nodes represent authors and edges represent co-authorship of the same article. From this network several metrics are computed, such as degree, PageRank and clustering coefficient, in addition to community detection. The results show that the collaboration structure is related to the grades: at the program level, the prediction classifies about 91% of the programs within one grade of the actual one, and the separation between average and excellence programs reaches about 80% accuracy. It is also observed that faculty productivity and the collaboration topology are the factors most associated with the grades, while weighting edges by citations or productivity tends to dilute this signal.

**Keywords**: Collaboration networks, Network science, Science of Science, OpenAlex, Bibliometric data.

# Lista de Figuras

# Lista de Tabelas

# Lista de Abreviaturas e Siglas

| | |
|---|---|
| API | *Application Programming Interface* (Interface de Programação de Aplicações) |
| CAPES | Coordenação de Aperfeiçoamento de Pessoal de Nível Superior |
| JSON | *JavaScript Object Notation* |
| MAE | Erro Absoluto Médio (do inglês *Mean Absolute Error*) |
| NetSci | *Network Science* (Ciência das Redes) |
| NMI | *Normalized Mutual Information* (Informação Mútua Normalizada) |
| PPG | Programa de Pós-Graduação |
| UFOP | Universidade Federal de Ouro Preto |

# Sumário

# 1 Introdução

A Ciência é uma das muitas áreas de interesse da humanidade, em que podemos explicar os diversos fenômenos físicos, biológicos e sociais que permeiam a nossa realidade, englobando uma grande quantidade de áreas distintas do conhecimento. Dado isto, é inevitável que com o avançar das contribuições ao longo do tempo surja a necessidade de se estudar a própria área da ciência, desde seu histórico até a forma como as entidades científicas interagem entre si. A Ciência pode ser descrita como uma rede complexa e auto-organizada de projetos, autores, instituições e temas que interagem entre si e evoluem continuamente (FORTUNATO et al., 2018).

Um dos desafios dessa área de estudos sobre a própria ciência é lidar com a quantidade de dados acadêmicos gerados pelas diversas pesquisas, em que nas épocas mais antigas não se existia uma forma eficiente de armazenar todo esse conhecimento. Atualmente, com o crescimento das novas tecnologias e com a alta disponibilidade de dados podemos ter armazenados vários detalhes de pesquisas científicas como artigos, autores e instituições guardados digitalmente, o que fornece uma grande quantidade de oportunidades para entender e explorar as interações entre as pesquisas e os pesquisadores. Nos dias atuais podemos fazer uma simples pesquisa em ferramentas como o *Google Scholar*[1] e ter acesso a um grande número de artigos e ter acessos a vários metadados desses artigos.

Uma das bases digitais que contém detalhes sobre as interações entre as entidades científicas é a biblioteca digital OpenAlex Priem, Piwowar e Orr (2022). A OpenAlex é um catálogo de dados *Open source* com centenas de milhões de entidades e bilhões de interações entre elas. Esses dados podem ser acessados via a OpenAlex[2] API (*Application Programming Interface*) em que os dados podem ser filtrados pelas queries de busca e são retornados no formato JSON. Existem vários tipos de entidades na OpenAlex, mas para este estudo são necessárias apenas três: autores (*authors*), instituições (*institutions*) e publicações (*works*).

No Brasil as universidades têm Programas de Pós-Graduação (PPG) que ofertam cursos de Mestrado e Doutorado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)[3] a qual faz uma análise quadrienal dos Programas de Pós-Graduação no Brasil observando características desses programas. Como resultado dessa análise cada Programa recebe uma nota de 1 a 7 conforme ilustra a Tabela 1.1. Apenas programas com nota 3 ou maior tem permitido o seu funcionamento. Programas com notas 6 e 7 são PPGs considerados de excelência internacional para todos os cursos.

[1] <https://scholar.google.com/>
[2] <https://docs.openalex.org/>
[3] <https://www.gov.br/capes/pt-br>

Tabela 1.1 – Notas da avaliação dos PPGs.

| Nota | Significado |
|---|---|
| 7 e 6 | Programas com excelência internacional para doutorado. |
| 5 | Programas com excelência nacional.<br>Nota máxima para programas que oferecem apenas o mestrado. |
| 4 | Programas com bom desempenho porém ainda não com excelência.<br>Nota mínima para doutorados. |
| 3 | Programas que atendem os requisitos mínimos de qualidade.<br>Nota mínima para mestrados. |
| 1 e 2 | Autorizações de funcionamento canceladas.<br>Não há o reconhecimento dos cursos de<br>mestrado e/ou doutorado oferecidos pelo programa. |

## 1.1 Justificativa

A cada quatro anos a CAPES reavalia os PPGs. Os critérios de avaliação são diversos e extensos. Esta monografia avalia se apenas a produção bibliográfica por si só é capaz de explicar as notas dos Programas de Pós-Graduação acadêmicos em Ciência da Computação.

## 1.2 Objetivos

O objetivo geral desta monografia é relacionar as métricas de uma rede de colaboração científica entre os professores das instituições públicas e privadas incluídas nos PPGs em Ciência da Computação no Brasil (com foco apenas nos programas stricto-sensu, ou seja, programas voltados para a formação de mestres e doutores, excluindo os programas de especialização) e as notas destes programas aos quais eles são credenciados. Os objetivos específicos são listados a seguir:

- Montar uma base de dados com a maioria dos professores de universidades inclusas nos PPGs de computação relacionando as informações contidas nos currículos Lattes com os dados da OpenAlex;
- Implementar uma rede de colaboração representando as interações entre esses autores na publicação de artigos;
- Relacionar as métricas encontradas nessas redes com as notas dos PPGs.

## 1.3 Estrutura da Monografia

**Capítulo 1:** Introdução.

**Capítulo 2:** Revisão Bibliográfica/ Embasamento Teórico (com o referencial teórico e trabalhos relacionados).

**Capítulo 3:** Metodologia ou Desenvolvimento (material e métodos).

**Capítulo 4:** Resultados e Discussões.

**Capítulo 5:** Considerações Finais.

Esta monografia integra um projeto mais amplo de análise de dados bibliográficos de professores credenciados em Programas de Pós-Graduação em Computação no Brasil, no qual a base de dados foi construída e os dados foram analisados sob diferentes perspectivas, incluindo *deep-learning*, *embeddings*, ciência das redes e mineração de dados.

# 2 Revisão Bibliográfica

Neste capítulo discutem-se e apresentam-se definições sobre *Network Science* (ou Ciência das Redes) abreviada como **NetSci** e, para isso, é preciso compreender vários conceitos sobre grafos e sobre Redes de Colaboração, bem como trabalhos relacionados a essa área. Mas primeiramente é importante definir do que se trata a *Network Science* de um modo geral. O campo da *Network Science* surge a partir de uma necessidade de se compreender sistemas complexos e intrincados que surgem à medida que as sociedades vão avançando social e tecnologicamente, tornando a área algo extremamente recente. Essa recência se dá por dois fatores: conseguirmos representar grandes redes com ferramentas tecnológicas e a universalidade das características das redes que representam diferentes sistemas como a WWW, iterações entre proteínas, a internet, redes sociais, redes de transporte e muitas outras. Sendo assim podemos definir a NetSci como uma forma de representar estruturas complexas e as interações entre seus componentes (BARABASI, 2016, chapter 1).

## 2.1 Trabalhos Relacionados

O termo ***Science of Science*** (Ciência da Ciência) foi cunhado por Fortunato et al. (2018) para descrever a área que estuda a própria publicação de artigos científicos e a sua relação com diversas circunstâncias, sejam elas autores, instituições, temas ou áreas do conhecimento. No artigo, os autores abordam características dos estudos sobre publicações científicas feitos sobre diversas bases de dados ao longo dos anos, que só se tornaram possíveis devido aos avanços tecnológicos no armazenamento de dados. Várias análises sobre dados de outros estudos são feitas e algumas conclusões são tiradas de forma mais geral, como a disparidade de gênero entre as publicações ou o fato de o momento da carreira de um cientista pouco se correlacionar com quando ele obterá seu artigo mais citado.

No contexto deste estudo, analisa-se como os diferentes autores interagem entre si e, para isso, é preciso compreender como são os modelos de contratação e retenção de professores dentro das universidades, além de compreender como professores não necessariamente atuam nas universidades em que obtiveram os seus doutorados. Essa análise é feita por Wapman et al. (2022) utilizando as universidades dos Estados Unidos como objeto de estudo. Nesse estudo foi utilizada uma base de dados de recrutamento de acadêmicos de diversas universidades dos Estados Unidos a partir de diversas áreas distintas dentro da última década (2011-2020). A partir desses dados foi então criada uma rede em que os nós representam universidades e as arestas representam indivíduos que fizeram doutorado em uma universidade e trabalham em outra. O estudo encontrou diversas características dessa rede como o fato de que muitos professores são auto-contratados nas instituições em que obtiveram seus doutorados e a desigualdade de formação

de novos professores, sendo poucas universidades as formadoras da grande maioria de novos professores.

Outro tipo de análise interessante feita nesse contexto é a questão da mobilidade e os impactos negativos da limitação da mobilidade dos e do isolacionismo internacional dos acadêmicos atualmente. Sugimoto et al. (2017) faz um estudo sobre esse tema por meio de uma análise da base de dados *Web of Science*[1], que é uma base amplamente utilizada em pesquisas acadêmicas, em que foram analisados mais de 14 milhões de artigos científicos publicados entre 2008 e 2015. O estudo revela que 96% dos acadêmicos são afiliados a um único país, enquanto apenas 4% são pesquisadores cuja carreira inclui afiliações em mais de um país. Eles apontam que os cientistas mais móveis são mais citados em geral do que os cientistas restritos a apenas um país com grandes variações dependendo da região comparada. Além disso o artigo cita impactos de eventos recentes como o *Brexit* que acarreta na saída de pesquisadores de instituições britânicas devido as incertezas em relação ao financiamento científico não mais atrelado à União Europeia. Outras tendências como a migração de cientistas da Europa e da Ásia para a América do Norte também são notadas.

Existem estudos também referentes á redes de colaboração da área científica. Newman (2001) analisa as estruturas das redes e o comportamento das comunidades de autores de artigos. Nesse estudo dois cientistas são conectados caso eles sejam autores de um mesmo artigo. Newman encontra várias características dessas redes como o fato de que elas formam pequenos agrupamentos (*clusters*) dentro das comunidades de cientistas além de mostrar alguns aspectos como as diferenças entre as comunidades científicas como as de física e de biomedicina. O artigo encontrou uma grande quantidade de resultados como por exemplo o fato de que dois autores podem ter uma conexão entre si por meio de outros autores, fazendo com que a rede seja altamente conectada. Além disso as quantidades de artigos escritos por autores, numero de autores por artigo e quantidade de colaboradores que cada autor possui são relacionados por meio de uma lei de potência.

Barabási et al. (2002) investiga de como funciona a co-autoria dentro do ambiente científico e como ela evolui ao longo do tempo em alguns dos campos do conhecimento. Para este estudo a base de dados utilizada continha os títulos e autores de todos os artigos científicos publicados mais relevantes do período entre 1991 e 1998 e se limitaram às áreas da matemática e neurociência. Essas áreas foram selecionadas principalmente porque são áreas com características distintas no quesito de colaboração, sendo a neurociência altamente colaborativa e a matemática sendo mais individual. A rede foi construída com apenas um tipo de nó que representa os autores, sendo que as arestas entre esses autores representam co-autoria de artigos. Nesse estudo, várias informações importantes foram encontradas a respeito dessas redes sendo elas: o fato de que existem poucos nós da rede que possuem um alto grau de conexão enquanto a maioria dos nós são pouco conectados, o que gera uma quantidade significativa de clusters centrados em alguns

[1] <https://www.webofscience.com/wos>

indivíduos; Além disso há uma tendência de crescimento do grau do grafo como um todo a medida que o tempo passa já que autores com muitas colaborações tendem a colaborar com mais artigos.

Em outro estudo relacionado, Moody (2004) explora como as redes de colaboração de pesquisas científicas se desenvolveram ao longo de 36 anos. O estudo usa dados obtidos numa base de dados chamada *Sociological Abstracts* a qual cobre todas as publicações na área de sociologia. No estudo foi construída uma rede de colaboração a partir desses dados, a qual possuía dois tipos de nós: os artigos e os autores, em que cada autor que contribuiu com um determinado artigo possuía uma aresta conectada a aquele artigo independente de serem ou não co-autores. Essa rede foi implementada utilizando diagramas de visualização de redes e os estudos foram feitos usando modelos estatísticos para extrair diversas informações sobre a rede. O estudo encontrou algumas tendências relacionais desses artigos dentro desse intervalo de tempo como o fato de que artigos possuírem mais co-autores se tornou mais comum ao longo dos anos, possivelmente devido a necessidades financeiras, diferenças de nível de treinamento dos autores entre as disciplinas e a divisão de trabalho dos cientistas em geral.

No âmbito mais específico da Ciência da Computação temos estudos que analisam redes de colaboração referentes a autores dessa área. Franceschet (2011) utilizou uma base de dados chamada *The DLBP Computer Science Bibliography* que continha 1.6 milhões de entradas que continham diversos artigos científicos da área da Ciência da Computação, como por exemplo o *Journal of the American Society for Information Science and Technology, Scientometrics*, e o *Journal of Informetrics*. A base de dados é uma das mais utilizadas por pesquisadores da área de pesquisa relacionada à informática e computação devido a sua alta precisão e pode ser considerada uma predecessora da OpenAlex. No estudo foram incluídas publicações de 1936 a 2008, a partir da qual foi construído uma rede representada por um grafo bipartido com dois tipos de nós similares ao de estudos citados anteriormente em que uma aresta existe entre um nó autor e um nó artigo se aquele autor contribuiu com aquele artigo. Esse estudo conseguiu encontrar diversas tendências na área da Ciência da Computação, tais como: há uma grande expansão da área ao decorrer do tempo, observa-se um aumento da quantidade de autores ativos e também que a área da computação se encontra em um ponto intermediário no quesito de colaboração em um mesmo artigo.

A OpenAlex é uma base de dados relativamente recente e ainda em crescimento mas já existem alguns estudos que a utilizam como fonte de dados. Para fazer uma relação dos acadêmicos com suas presenças nas redes sociais, Mongeon, Bowman e Costas (2023) fazem uma análise de perfis de pesquisadores na rede social X (antigamente chamada de *Twitter*). O estudo teve como objetivo ligar autores existentes na OpenAlex com seus *usernames* do X. Uma das dificuldades que o estudo teve foi comparar os nomes dos cientistas a seus nomes de usuário já que nem sempre os usuários colocam seu próprio nome nas seus identificadores em redes como o X, isso sem contar o fato de que muitos não utilizam caracteres presentes no alfabeto

latino. Além disso nem sempre as contas presentes no ORCID daquele pesquisador são contas válidas. O estudo então gerou uma base de dados a partir de um modelo comparativo entre vários dados obtidos via OpenAlex, ORCID e o X, porém dentro das limitações de *usernames* citadas anteriormente.

## 2.2 Fundamentação Teórica

Nesta seção abordam-se alguns dos elementos cruciais para o entendimento do que foi feito. É importante ter o conhecimento do que são grafos, de como são representados e as suas diversas características. Além disso também precisamos definir o que são Redes de Colaboração.

### 2.2.1 Grafo

Um grafo ou rede é uma abstração que funciona como uma maneira de representar como os elementos de um sistema interagem uns com os outros dentro daquele contexto. Um grafo pode ser usado para representar diversos tipos de sistemas, por exemplo, um grupo de pessoas e as amizades entre elas. Um grafo é composto de basicamente dois elementos:

- Nó (rede) ou Vértice (grafo): Representação de um elemento que compõe um sistema. No exemplo acima as pessoas seriam representadas pelos nós.
- Link (rede) ou Aresta (grafo): Interações entre os nós da rede. No exemplo se duas pessoas possuem uma amizade elas teriam uma aresta entre si.

Matematicamente representamos um grafo como:

$$G(V, E)$$

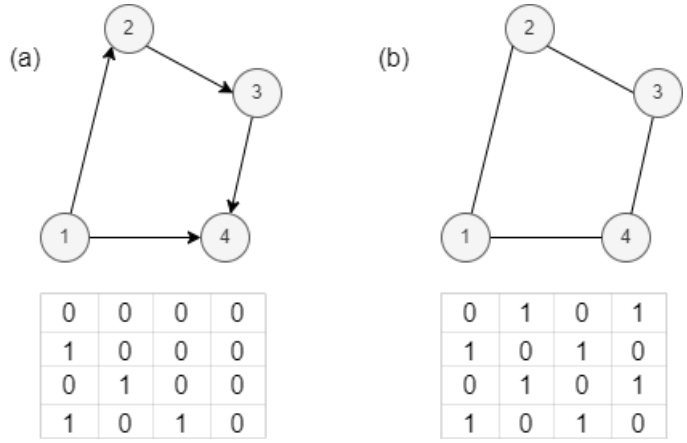
em que $V$ representa os vértices e $E$ representam as arestas (*edges*).

Um grafo pode ser direcionado ou não direcionado. Os grafos direcionados tem arestas que representam interações que possuem uma direção entre si, ou seja, dados dois vértices A e B, uma aresta pode sair de A para B ou de B para A em que essas interações representam estados diferentes. Por exemplo uma representação de ligações telefônicas precisa de uma rede direcionada já que se eu ligar para um amigo não significa a mesma coisa de esse amigo ligar para mim. Geralmente as arestas direcionados são representadas por uma seta e nesses grafos são chamadas de arcos. Já os grafos não direcionados tem arestas que representam casos onde não há diferença entre de onde saem as interações. Um exemplo é um grafo de relacionamentos românticos entre pessoas. Se A se relaciona com B, B se relaciona com A da mesma forma (BARABASI, 2016).

O grau $k_i$ de um determinado vértice $i$ em um grafo determina com quantos outros vértices daquele grafo ele está conectado. Num grafo direcionado vértices podem ter graus de entrada e graus de saída (BARABASI, 2016).

#### 2.2.1.1 Representação

Uma das maneiras de se representar um grafo é por meio de uma matriz de adjacências como mostrado na Figura 2.1.




| | | | |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 |
| 1 | 0 | 1 | 0 |

| | | | |
|---|---|---|---|
| 0 | 1 | 0 | 1 |
| 1 | 0 | 1 | 0 |
| 0 | 1 | 0 | 1 |
| 1 | 0 | 1 | 0 |

Figura 2.1 – Representação de um grafo direcionado (a) e de um grafo não direcionado (b) como uma matriz de adjacências. Fonte: feito pelo autor utilizando diagrams.io <https://app.diagrams.net/>.

A matriz de adjacências de uma rede direcionada com $N$ nós é uma matriz de $N$ linhas e $N$ colunas em que seus elementos são:

- $A_{ij} = 1$ caso exista um arco do vértice $i$ para o vértice $j$
- $A_{ij} = 0$ caso não haja conexão entre os vértices $i$ e $j$

O grau de um vértice $i$ representado em uma matriz pode ser encontrado pela soma dos $N$ valores da linha $i$ ou da coluna $i$ em um grafo não direcionado:

$$k_i = \sum_{j=1}^{N} A_{ij} = \sum_{i=1}^{N} A_{ji}.$$

Já em um grafo direcionado podemos encontrar o grau de entrada e de saída de um vértice olhando as linhas para a entrada e a coluna para a saída (BARABASI, 2016):

$$k_{i(in)} = \sum_{j=1}^{N} A_{ji},$$

$$K_{i(out)} = \sum_{j=1}^{N} A_{ji}.$$

#### 2.2.1.2 Grafos ponderados

Até então nas definições de grafos, vimos apenas grafos que possuem arestas simples, que indicam simplesmente a existência de conexões. No entanto as conexões podem ter mais informações adicionais que uma aresta simples não pode representar. Nessas situações fazemos o uso de arestas ponderadas as quais formam um grafo ponderado. Nos grafos ponderados cada conexão $(i, j)$ possui um peso $w$ associado que é representado por um valor escalar. Numa rede que representa a conexão de linhas de energia elétrica por exemplo, o peso de cada aresta pode representar a quantidade de corrente elétrica que flui de um vértice a outro. Um grafo simples tem todos os pesos das arestas iguais a 1. Na Matriz de Adjacências representamos cada aresta pelo seu peso $A_{ij} = w_{ij}$ (BARABASI, 2016).

#### 2.2.1.3 Conectividade e Clusterização

Uma característica a se avaliar em um determinado grupo de nós de uma rede é a sua conectividade. Dois nós de um grafo não-direcionado são ditos conectados quando existe um caminho entre esses dois nós, ou seja, eles estão ligados entre si por meio de um conjunto de arestas e nós. Os nós conectados não necessariamente precisam ter uma aresta conectando os dois diretamente, apenas precisa ser possível chegar de um para o outro percorrendo arestas. Um grafo $G$ é considerado conectado se para todo par de nós de $G$ existe algum caminho entre esses dois nós. Se existir ao menos um par de nós em $G$ em que não se consegue criar um caminho entre eles, dizemos que $G$ é um grafo desconectado. Caso um grafo tenha dois componentes (conjuntos de nós e arestas) ligados entre si por uma única aresta $m$ dizemos que $m$ é uma ponte, em que a remoção desta aresta faz com que o grafo se torne desconectado. Definir a conectividade de um grafo é uma tarefa difícil já que é necessário analisar todos os pares de nós (BARABASI, 2016).

O coeficiente de clusterização é uma métrica para calcular o quão conectada é a vizinhança de um determinado nó. Para um nó $i$ de grau $k_i$ o coeficiente de clusterização se dá por

$$C_i = \frac{2L_i}{k_i(k_i - 1)},$$

em que $L_i$ representa a quantidade de arestas entre os $k_i$ vizinhos do nó $i$. O resultado dessa equação é um valor entre 0 e 1 que indica a probabilidade de que dois dos nós da vizinhança de $i$ se conectam uns com os outros. Por exemplo, se $C_i = 0$ indica que os vizinhos de $i$ não se conectam entre si enquanto $C_i = 1$ indica que os vizinhos de $i$ formam um grafo completo entre si (clique) como explicitado na Figura 2.2. O grau médio de clusterização de um grafo representa o coeficiente médio de cada nó do grafo e é dado por

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

em que $N$ representa os vértices e o $C_i$ o coeficiente local desses vértices. Embora esse valor seja definido para grafos não-direcionados também existem definições genéricas que funcionam para grafos direcionados e/ou ponderados (BARABASI, 2016).

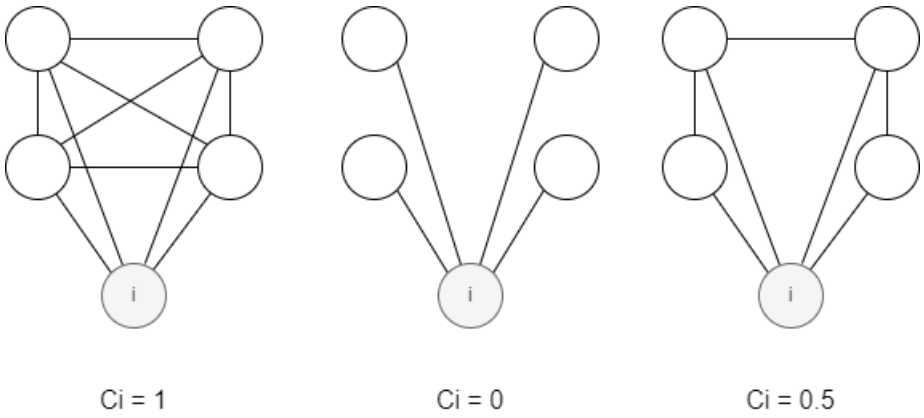



Figura 2.2 – Representação do coeficiente de clusterização de um nó $i$ baseado na sua vizinhança. O valor de $C_i$ indica a probabilidade de que um par de nós da vizinhança de $i$ forme um subgrafo conectado. Fonte: adaptado de (BARABASI, 2016) criado pelo autor utilzando diagrams.io <https://app.diagrams.net/>

#### 2.2.1.4 Centralidade

Além do grau, existem outras medidas que buscam quantificar a importância ou a influência de um nó dentro da rede, chamadas de medidas de centralidade. A própria contagem de conexões de um nó, o seu grau, já é uma medida de centralidade, conhecida como centralidade de grau: nós com muitas conexões tendem a ser mais importantes. No entanto, nem sempre a quantidade de conexões reflete bem a importância de um nó, pois estar conectado a poucos nós muito influentes pode ser mais relevante do que estar conectado a muitos nós pouco influentes.

A centralidade de autovetor (*eigenvector centrality*) captura essa ideia ao definir a importância de um nó de forma recursiva: um nó é tão mais central quanto mais centrais forem os seus vizinhos. Formalmente, a centralidade $x_i$ de um nó $i$ é proporcional à soma das centralidades dos seus vizinhos, o que corresponde ao autovetor da matriz de adjacências associado ao seu maior autovalor (BARABASI, 2016).

O PageRank é uma variação da centralidade de autovetor, originalmente proposto para ordenar páginas da web em função da sua relevância. Ele modela um caminhante aleatório que percorre a rede seguindo as arestas e, ocasionalmente, salta para um nó qualquer; a centralidade de um nó corresponde à frequência com que esse caminhante o visita. Dessa forma, um nó recebe importância dos nós conectados a ele, ponderada pela importância desses nós (BARABASI, 2016).

Há ainda medidas de centralidade baseadas em caminhos mínimos, como a intermediação (*betweenness*), que mede o quanto um nó atua como ponte nos caminhos mais curtos entre os demais, e a proximidade (*closeness*), que mede o quão perto um nó está de todos os outros. Essas medidas são custosas de calcular em redes muito grandes, pois exigem o cálculo dos caminhos mínimos entre todos os pares de nós (BARABASI, 2016).

#### 2.2.1.5 Comunidades e Modularidade

Uma característica importante de muitas redes reais é a presença de comunidades, que são grupos de nós densamente conectados entre si e com relativamente poucas conexões com o restante da rede. Em uma rede de colaboração científica, por exemplo, uma comunidade pode corresponder a um grupo de pesquisadores que colaboram frequentemente uns com os outros (BARABASI, 2016).

Para avaliar o quão boa é uma divisão da rede em comunidades utiliza-se a modularidade, denotada por $Q$. A modularidade compara a fração de arestas que ficam dentro das comunidades com a fração que seria esperada caso as arestas fossem distribuídas aleatoriamente, preservando o grau de cada nó. Ela é dada por

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

em que $m$ é o número de arestas, $A_{ij}$ é o peso da aresta entre $i$ e $j$, $k_i$ é o grau do nó $i$, $c_i$ é a comunidade a que $i$ pertence e $\delta(c_i, c_j)$ vale 1 quando $i$ e $j$ estão na mesma comunidade e 0 caso contrário. Valores mais altos de $Q$ indicam uma estrutura de comunidades mais bem definida (BARABASI, 2016).

Encontrar a partição que maximiza a modularidade é um problema computacionalmente custoso, de modo que se utilizam métodos heurísticos. Um dos mais difundidos é o método de Louvain (BLONDEL et al., 2008), um algoritmo guloso e hierárquico capaz de processar redes com milhões de nós em tempo reduzido. O método alterna duas etapas: primeiro cada nó é movido para a comunidade de um vizinho sempre que isso aumenta a modularidade; em seguida, cada comunidade é agregada em um único nó, e o processo se repete sobre a rede reduzida. O algoritmo aceita um parâmetro de resolução, que controla a granularidade das comunidades encontradas, produzindo comunidades menores e mais numerosas para valores maiores.

### 2.2.2 Redes de colaboração (*Collaboration Networks*)

A partir dos conceitos de grafos definidos na subseção anterior podemos compreender a definição de uma rede de colaboração. O estudo das redes de colaboração consiste então de uma análise sobre a estrutura, comportamento e a evolução das interações dessa rede. Uma rede de colaboração pode ser composta por pessoas ou recursos, por exemplo, e pode ter elementos mistos. As redes de colaboração com seu crescimento constante nos diversos tipos de situações, geram alguns subtipos de redes. Nesse sentido podemos dizer que uma rede de colaboração nada mais é do que uma representação de entidades autônomas que colaboram entre si para alcançar um objetivo específico e que estão em constante evolução (CAMARINHA-MATOS; AFSARMANESH, 2004). Alguns exemplos são:

- ***Virtual Enterprise (VE)***: Uma aliança temporária entre duas ou mais empresas que buscam

compartilhar experiências e habilidades para melhor responder oportunidades de mercado. Essa cooperação é suportada por redes computacionais.

- ***Virtual Organization (VO)***: Similar ao VE porém nem todos os elementos dessa rede são necessariamente empresas lucrativas e sim algum tipo de organização auxiliar.
- ***VO Breeding Environment (VBE***: Representam uma associação entre diferentes organizações e suas instituições relacionadas. Essas organizações podem ser cooperativas ou competitivas dependendo do contexto e isso influi em como as instituições intermediárias se relacionam com elas.
- ***e-science***: referente a redes de colaboração na área científica com registros dinâmicos de autores, artigos e instituições.

Um dos problemas dessa área das redes de colaboração é a falta de definições formais dos elementos que a compõem por meio de teorias formais, paradigmas e ferramentas para construir os modelos representativos dessas redes, o que sem sombra de dúvidas dificulta as interações de diferentes autores em relação a essa área (CAMARINHA-MATOS; AFSARMANESH, 2004).

Nesta monografia utiliza-se uma rede de colaboração simples que envolve os autores e co-autores de trabalhos encontrados na OpenAlex. Na rede os nós representam cada autor individual enquanto um link representa se esses dois autores colaboraram em algum mesmo artigo, em que maiores quantidades de colaborações entre um par de autores implica em um maior peso da aresta, funcionando de uma maneira similar a rede descrita no trabalho de Newman (2001). Para tal feito, uma rede do tipo *e-science* é um ponto de partida inicial a se considerar para a implementação da rede.

### 2.2.3 Regressão Logística

Os métodos de regressão são muito populares em análise de dados. A regressão logística, assim como vários outros modelos estatísticos, tem como objetivo encontrar a melhor representação da interação entre uma variável de saída dependente de uma resposta e uma variável independente que pode dar uma explicação sobre a saída. A regressão logística trata de problemas em que os dados de saída são binários, representando estados como verdadeiro ou falso ou pertencimento ou não-pertencimento, o que difere da regressão linear por exemplo que lida com dados de saída contínuos, ou seja, a regressão logística é utilizada para classificar dados em valores categóricos (JR; LEMESHOW; STURDIVANT, 2013).

Uma possibilidade para a predição de probabilidade de um valor de entrada ser positivo seria utilizar uma equação linear do tipo $p(x) = ax + b$. Fazendo essa predição rapidamente notamos que alguns dos valores probabilísticos de $p(x)$ são menores que 0 para alguns valores falsos e são maiores que 1 para alguns valores verdadeiros, o que viola o intervalo aceitável de probabilidades que está contido entre $0$ e $1$. Para que não haja esse problema é necessário modelar

a função para calcular $p(x)$ de forma a limita-la ao intervalo desejado. Na regressão logística é usada a função logística que é dada por:

$$p(x) = \frac{e^{ax+b}}{1 + e^{ax+b}}$$

Com isso garantimos que $p(x)$ fica contida entre um intervalo de 0 e 1 e os dados ficam distribuídos como uma curva S que define claramente uma divisão entre as duas classes de dados, como mostrado na Figura 2.3.

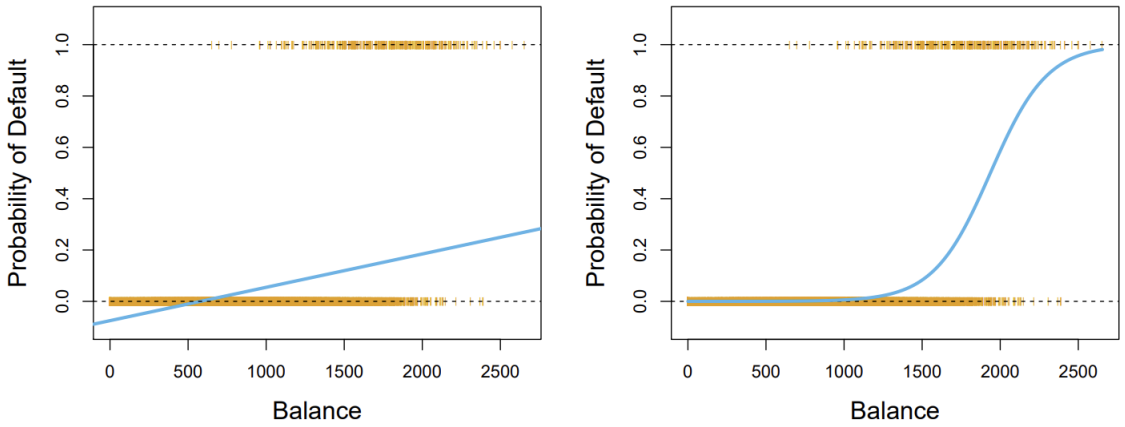



Figura 2.3 – Comparação da regressão linear com a logística para a divisão de dados categóricos. Fonte: (JR; LEMESHOW; STURDIVANT, 2013)

A Chance (do inglês *odd*) é um valor que obtemos a partir de manipulações na fórmula e pode ter um valor entre 0 e $+\infty$ e implicam em baixas e altas probabilidades respectivamente. Ao calcularmos o logaritmo natural de ambas as partes da função chegamos no Logit que é uma função que consegue mapear os valores das probabilidades para valores reais

$$\ln\left(\frac{p(x)}{1 - p(x)}\right) = ax + b.$$

Como a relação com a variável $x$ não é uma linha reta, o coeficiente $a$ de inclinação não corresponde a uma mudança em $p(x)$ relacionada com o aumento de $x$ em uma unidade e sim depende do valor atual de $x$.

Os coeficientes $a$ e $b$ são desconhecidos e precisam ser estimados por meio dos dados de treinamento. Enquanto na regressão linear é utilizado a estimativa dos mínimos quadrados $R^2$, para a regressão logística utilizamos o *Maximum Likelihood Estimation* já que se busca estimar o valor dos coeficientes a fim de que a probabilidade predita de um dado corresponda o máximo ao seu estado atual, ou seja, se o valor de $p(x)$ seja próximo de zero para dados categorizados como *False* e próximo de 1 para dados *True* (JAMES et al., 2013).

A regressão logística descrita até aqui trata de problemas binários, com apenas duas classes possíveis. Quando a variável de saída pode assumir mais de duas categorias, como as

notas dos programas que variam de 3 a 7, utiliza-se a extensão multinomial da regressão logística. Nessa formulação, uma das classes é tomada como referência e o modelo estima, para cada uma das demais classes, um conjunto próprio de coeficientes que descreve o logaritmo da chance daquela classe em relação à classe de referência. Assim, para um problema com $K$ classes, são ajustados $K - 1$ conjuntos de coeficientes, e a probabilidade predita para cada classe é obtida normalizando essas chances de forma que somem 1. Alternativamente, o problema pode ser decomposto em uma série de classificadores binários, cada um separando uma classe das demais (JR; LEMESHOW; STURDIVANT, 2013).

### 2.2.4 Medidas de Associação entre Variáveis Categóricas

Para verificar se a estrutura de comunidades da rede guarda relação com as notas dos programas é necessário medir a associação entre duas variáveis categóricas: a comunidade a que cada autor pertence e a nota do seu programa. Duas medidas são utilizadas para isso nesta pesquisa, ambas variando entre 0, quando não há associação alguma, e 1, quando a associação é total.

O $V$ de Cramér é uma medida baseada no teste qui-quadrado ($\chi^2$), que avalia se a distribuição conjunta de duas variáveis categóricas se afasta do que seria esperado caso elas fossem independentes. A partir da estatística $\chi^2$, o $V$ de Cramér a normaliza para o intervalo entre 0 e 1, sendo dado por

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1,\, c-1)}},$$

em que $n$ é o número total de observações e $r$ e $c$ são, respectivamente, o número de categorias de cada variável (CRAMÉR, 1946).

A informação mútua normalizada (*Normalized Mutual Information*, NMI) é uma medida oriunda da teoria da informação, bastante utilizada para comparar duas partições de um mesmo conjunto de elementos. A informação mútua quantifica o quanto conhecer a comunidade de um elemento reduz a incerteza sobre a sua nota, e vice-versa; a normalização a leva ao intervalo entre 0 e 1, em que 0 indica partições independentes e 1 indica partições idênticas a menos dos rótulos (BARABASI, 2016).

# 3 Desenvolvimento

Neste capítulo descreve-se todo o processo de construção da rede de colaboração, bem como o processo de busca e preenchimento da base de dados utilizada. Além disso, apresentam-se algumas informações mais específicas sobre a OpenAlex, a fim de mostrar como a ferramenta foi utilizada ao longo do processo de desenvolvimento.

## 3.1 Base de Dados

O projeto consiste de duas bases de dados distintas a partir das quais se constrói a rede de colaboração. Essas bases são preenchidas de acordo com consultas feitas na OpenAlex, que são a busca pelas instituições de ensino, pelos autores e pelos trabalhos dos autores. Cada uma dessas buscas tem uma chamada de API relacionada a ela, exemplificadas a seguir.

Inicialmente foi necessária a busca pelas instituições de ensino que são cadastradas nos PPGs a fim de encontrar qual o identificador de cada uma na OpenAlex. A ideia era buscar as universidades pelo nome ou sigla utilizando o parâmetro "*search*" na chamada de API da instituição. Com isso recebemos como resposta todas as universidades que atendem o parâmetro passado e assim conseguimos obter qual o Identificados daquela universidade. Na Figura 3.1 temos um exemplo de busca por instituição via parâmetro <https://api.openalex.org/institutions?search=ufop> em que buscamos pela sigla "ufop" e temos como resposta um array de resultados. Em cada resultado o valor id retornado é a url correspondente a instituição, em que o identificador é a cadeia de caracteres logo após a url base.


```
"results": [
    {
        "id": "https://openalex.org/I10824318",
        "ror": "https://ror.org/056s65p46",
        "display_name": "Universidade Federal de Ouro Preto",
        "relevance_score": 14170.968,
        "country_code": "BR",
        "type": "education",
        "lineage": [
            "https://openalex.org/I10824318"
        ],
        "homepage_url": "http://www.ufop.br/",
        "image_url": "https://commons.wikimedia.org/w/index.php?
            title=Special:Redirect/file/UFOP%20logo.gif"
```


Figura 3.1 – Exemplo de consulta de instituição - no caso a Universidade Federal de Ouro Preto. Fonte: busca na OpenAlex feita pelo próprio autor por meio do Postman.

Com os dados das instituições em mãos, conseguimos conferir se os IDs das instituições encontrados nos resultados dos autores batem com as instituições destes. Na planilha de autores inicial, havia somente os dados do PPG e os nomes dos docentes de cada instituição listada. Para cada linha era necessário buscar pelo currículo lattes daquele autor, e preencher mais uma série de informações baseadas nesse currículo, descritas mais adiante. A partir dessas informações obtidas do lattes conseguimos buscar e conferir dados desses professores na OpenAlex. Quando fazemos uma busca na chamada de api de autores, por um autor pelo seu nome por exemplo, temos como resposta uma lista de possíveis autores que atendem a aquele nome integralmente ou parcialmente. Coube a nós analisar individualmente os resultados das chamadas e encontrar o autor que mais condiz com os valores do lattes. Da mesma forma que nas instituições, o id do autor está presente no resultado. Após isso podemos consultar por esse identificador e obtemos uma série de informações relevantes como mostra a Figura 3.2.

Um cuidado importante nessa etapa é a desambiguação de autores. Um mesmo nome pode corresponder a vários pesquisadores distintos e, ao contrário, um mesmo pesquisador pode ter mais de um perfil na OpenAlex ou ter trabalhos atribuídos incorretamente. Esse é um problema conhecido e relevante em bases bibliométricas, inclusive na própria utilização da OpenAlex pela CAPES para levantar a produção dos programas. Para mitigá-lo, a correspondência entre cada docente e o seu identificador na OpenAlex foi conferida manualmente, cruzando as informações do currículo Lattes (instituição, área de atuação, orientador e histórico de publicações) com os dados retornados pela API. Cada correspondência foi registrada na base com um indicador de confiabilidade, de forma que os casos duvidosos pudessem ser revistos ou descartados. Ainda assim, a desambiguação permanece uma fonte potencial de ruído nos resultados.

```
{
    "id": "https://openalex.org/A5069290754",
    "orcid": null,
    "display_name": "Carl Sagan",
    "display_name_alternatives": [ ...
    ],
    "works_count": 776,
    "cited_by_count": 20793,
    "summary_stats": { ...
    },
    "ids": { ...
    },
    "affiliations": [ ...
    ],
    "last_known_institution": { ...
    },
    "last_known_institutions": [ ...
    ],
    "x_concepts": [ ...
    ],
    "counts_by_year": [ ...
    ],
    "works_api_url": "https://api.openalex.org/works?filter=author.id:A5069290754",
    "updated_date": "2024-01-23T13:29:51.887146",
    "created_date": "2023-07-21"
}
```


Figura 3.2 – Exemplo de consulta de autor - no caso Carl Sagan. Fonte: busca na OpenAlex feita pelo próprio autor por meio do Postman.

A partir das consultas e dos dados contidos na OpenAlex, a nossa equipe de estudo conseguiu montar uma base de dados de docentes das universidades pertencentes aos PPGs além de uma base auxiliar com os dados dos PPGs. Os dados foram inseridos manualmente pelos alunos a partir das consultas individuais. As colunas da base são descritas a seguir:

- ppg_codigo: O código do PPG definido pela CAPES.
- ppg_nome: O nome do curso (Ciência da computação ou só Computação).
- ppg_nota: A nota do programa como (descrito na Tabela 1.1, com exceção do valor abaixo de 3).
- institution_id: O id da instituição obtido via OpenAlex.
- ies_sigla: Sigla da instituição da universidade.
- nome_docente: Nome do autor (Docente).
- doutorado_ano: Ano em que o autor concluiu o doutorado
- regime_trabalho: Regime de trabalho do autor (Integral, Dedicação exclusiva, etc)
- carga_horaria: Horas semanais de dedicação ao doutorado
- link_do_lattes: url refrente ao currículo lattes do autor

- author_id: id do autor advindo da OpenAlex
- bolsista_produtividade: campo booleano que define se o autor é ou não um bolsista de produtividade.
- extrato_bolsa_produtividade: tipo de bolsa de produtividade do autor caso ele tenha.
- doutorado_institution_id: id da instituição do autor, obtido da OpenAlex.
- doutorado_institution_name: nome da instituição do autor.
- doutorado_ppg_codigo: código do PPG daquele curso que o autor faz doutorado.
- doutorado_supervisor_id: id do orientador de doutorado do autor obtido na OpenAlex.
- doutorado_supervisor_name: nome do orientador.

Cada autor tem um campo extra chamado "works_api_url" que consiste em uma outra chamada de api que busca dentro dos trabalhos quais deles possuem o autor passado como parâmetro de url como autor do artigo. Com isso podemos ter detalhes sobre trabalhos específicos. Na Figura 3.3 temos um exemplo de resposta da consulta de trabalhos com os autores em destaque. O array "authorships" é utilizado para montar a rede de colaborações, já que com ele conseguimos obter todos os identificadores dos autores que colaboraram na pesquisa.


```
s,
"authorships": [
    {
        "author_position": "first",
        "author": {
            "id": "https://openalex.org/A5040062779",
            "display_name": "Christopher F. Chyba",
            "orcid": "https://orcid.org/0000-0002-6757-4522"
        },
        "institutions": [
            {
                "id": "https://openalex.org/I205783295",
                "display_name": "Cornell University",
                "ror": "https://ror.org/05bnh6r87",
                "country_code": "US",
                "type": "education",
                "lineage": [
                    "https://openalex.org/I205783295"
                ]
            }
        ],
        "countries": [
            "US"
        ],
        "is_corresponding": false,
        "raw_author_name": "Christopher Chyba",
        "raw_affiliation_string": "Laboratory for Planetary Studies Cornell University, Ithaca, USA",
        "raw_affiliation_strings": [
            "Laboratory for Planetary Studies Cornell University, Ithaca, USA"
        ]
    },
```


Figura 3.3 – Exemplo de consulta de trabalhos de um autor. Fonte: busca na OpenAlex feita pelo próprio autor por meio do Postman.

Estes são os dados mais importantes para a construção da rede de colaboração. Na rede cada nó pode ser representado por um identificador de autor e a partir do array “authorships” conseguimos definir as arestas entre esses autores, e inclusive podemos adicionar pesos às arestas dependendo de quantas vezes um par de autores colaborou em um mesmo artigo. Com esses dados podemos correlacionar notas dos PPGs e outros fatores no quão influentes determinados autores são.

Como não necessariamente o número de interações entre autores seja suficiente para explicar a importância de uma interação, uma das potenciais ideias é montar diversas versões de redes em que os pesos das arestas variam de acordo com certos critérios como a relevância dos *journals* em que os artigos foram publicados, números de citações, entre outros. A partir dessas diferentes redes podemos montar uma rede final com a combinação linear dessas redes.

Para a construção da rede foi utilizada a linguagem Python. O acesso à OpenAlex é feito por meio da biblioteca *pyalex*[1], e todas as respostas da API (autores, trabalhos e *sources*) são persistidas em um banco de dados local em SQLite, que funciona como *cache*: cada entidade é consultada na API uma única vez e reaproveitada nas execuções seguintes, o que evita requisições desnecessárias, respeita os limites de uso da OpenAlex e torna os experimentos reprodutíveis sem depender de novas coletas. A construção e a manipulação das redes utilizam a biblioteca NetworkX[2], e cada rede gerada é persistida em disco no formato Parquet para reutilização. Para a detecção de comunidades e para a classificação é utilizada a biblioteca scikit-learn[3].

## 3.2 Construção das Redes de Colaboração

A partir da lista de professores (*seeds*) obtida da base de dados, para cada professor são recuperados todos os seus trabalhos publicados na OpenAlex. Para cada trabalho, o conjunto de coautores (obtido do *array* “authorships”) forma um clique: todo par de autores que assina um mesmo artigo recebe uma aresta. Os nós representam autores e as arestas representam a coautoria. Foi adotado o recorte temporal de trabalhos publicados até 31 de dezembro de 2022, de forma a alinhar a rede à avaliação quadrienal da CAPES referente ao período. Como os coautores incluem também pesquisadores que não pertencem aos PPGs, a rede resultante é significativamente maior do que o conjunto de professores.

Conforme antecipado, uma única definição de peso pode não capturar toda a importância de uma colaboração. Por isso foram construídas diferentes versões da rede, todas com a mesma topologia (mesmos nós e arestas), variando apenas o peso $w$ das arestas:

- **Coautoria (contagem)**: $w$ igual ao número de vezes que o par de autores colaborou, seguindo a definição de Newman (2001).

[1] <https://github.com/J535D165/pyalex>
[2] <https://networkx.org/>
[3] <https://scikit-learn.org/stable/>

- **Coautoria simples**: rede binária, com $w = 1$ para toda aresta.
- **Citações dos artigos**: $w$ igual à média do número de citações dos trabalhos compartilhados pelo par.
- **Produtividade do autor**: $w$ igual à média do número de trabalhos (*works-count*) dos dois autores.
- **Citações do autor**: $w$ igual à média do total de citações dos dois autores.

As três primeiras redes derivam diretamente da varredura dos trabalhos, enquanto as duas últimas são obtidas reponderando a rede de coautoria a partir de atributos já disponíveis nos nós, sem novas consultas à API. Uma sexta rede, ponderada pela relevância do *journal* (índice-h da fonte de publicação), foi prevista mas não é analisada nesta monografia por exigir a coleta adicional das entidades *source* da OpenAlex. Cada aresta armazena, além do peso $w$, o seu inverso $1/w$, utilizado como distância nas métricas baseadas em caminhos mínimos, já que nelas o peso representa importância e não distância.

Por fim, foi construída uma rede combinada, que agrega as demais em uma única representação por meio de uma combinação linear dos pesos. Como as redes têm a mesma topologia mas pesos em escalas muito diferentes (contagens, citações e produtividade variam de unidades a milhares), uma soma direta seria dominada pela rede de maior magnitude. Para evitar isso, o peso de cada rede é primeiro normalizado pela sua maior aresta, passando ao intervalo entre 0 e 1, e a rede combinada recebe a média desses pesos normalizados. Dessa forma cada rede contribui igualmente para o peso final.

## 3.3 Métricas de Rede

Sobre cada rede são calculadas métricas que caracterizam a posição de cada autor na estrutura de colaboração. Foram utilizadas as métricas de **grau** (número e soma dos pesos das colaborações), **PageRank**, **coeficiente de clusterização** (Seção 2) e **centralidade de autovetor** (*eigenvector*), que mede a importância de um nó em função da importância de seus vizinhos. As métricas de intermediação (*betweenness*) e proximidade (*closeness*), embora implementadas, não foram calculadas sobre as redes completas por terem custo proporcional ao produto do número de nós pelo número de arestas, o que as torna impraticáveis na escala obtida. As quatro métricas efetivamente calculadas (grau, PageRank, clusterização e centralidade de autovetor) foram obtidas para cada autor em todas as redes e, junto com os atributos de produtividade, compõem o vetor de *features* utilizado na classificação (Seção 3.5).

## 3.4 Detecção de Comunidades

Para identificar grupos de autores densamente conectados entre si foi utilizado o método de Louvain, que busca uma partição da rede maximizando a modularidade. A escolha desse método se deve principalmente à sua eficiência: ele particiona redes com milhões de arestas em poucos segundos, o que é essencial na escala deste trabalho (mais de um milhão de arestas), enquanto alternativas exatas ou baseadas em modularidade gulosa tradicional se mostraram inviáveis nesse porte. Existem outros algoritmos de detecção de comunidades, como abordagens espectrais, hierárquicas e de otimização multiobjetivo, capazes de capturar estruturas que a maximização gulosa da modularidade nem sempre alcança; a comparação com esses métodos permanece como possibilidade de trabalho futuro. O método aceita um parâmetro de resolução: valores maiores produzem mais comunidades, cada uma menor. Foram testados diferentes valores de resolução a fim de observar como a granularidade das comunidades se relaciona com as notas dos programas. Nas redes cujas arestas podem ter peso zero (citações de artigos ou de autores sem citações), foi aplicada uma suavização somando uma unidade a cada peso, pois o método interpretaria um peso zero como ausência de aresta, fragmentando a rede em milhares de comunidades unitárias.

Cada comunidade é então caracterizada pela distribuição das notas dos programas dos professores que a compõem (coautores externos, sem nota, são ignorados). Como medida da associação entre a partição em comunidades e as notas dos PPGs, além da pureza das comunidades foram utilizadas a informação mútua normalizada (NMI) e o $V$ de Cramér, ambas variando entre 0 (nenhuma associação) e 1 (associação total).

## 3.5 Classificação das Notas dos Programas

Por fim, buscou-se predizer a nota do PPG a partir das características de rede e de produtividade dos professores, por meio de regressão logística. Para cada professor foi montado um vetor de *features* reunindo as métricas de rede das diferentes redes e atributos de produtividade obtidos da OpenAlex (número de trabalhos, total de citações, índice-h e índice-i10). Como a nota assume cinco categorias (de 3 a 7), foi empregada a extensão multinomial da regressão logística apresentada na Seção 2, que generaliza a formulação binária para múltiplas classes. As *features* foram padronizadas antes do ajuste.

A classificação foi conduzida em dois níveis. No **nível do autor**, cada professor é uma amostra rotulada com a nota do seu programa. No **nível do programa**, as *features* dos professores são agregadas por PPG pela média simples: para cada programa, cada característica, tanto as métricas de rede quanto os atributos de produtividade, recebe a média dos valores dos docentes credenciados nele, e o programa se torna uma amostra rotulada com a sua nota, nível em que ela de fato é definida. Um professor credenciado em mais de um programa contribui para a média de cada um deles. O desempenho foi avaliado por validação cruzada e comparado a um classificador

*baseline* que sempre prediz a nota mais frequente. No nível do autor, em que há muitas amostras, utilizou-se validação cruzada estratificada em cinco partições. No nível do programa, em que há apenas algumas dezenas de amostras e notas pouco frequentes, adotou-se a validação cruzada *leave-one-out*, que a cada iteração deixa um único programa de fora e treina com os demais, aproveitando ao máximo os dados disponíveis. Como as notas são desbalanceadas, com algumas muito mais frequentes que outras, a acurácia isolada pode ser enganosa (um modelo que sempre prevê a nota mais comum já obtém acurácia razoável); por isso reporta-se também a medida $F_1$ macro, que dá peso igual a todas as classes e penaliza o modelo que ignora as notas menos frequentes. Como a nota é ainda ordinal, reportam-se também o acerto dentro de $\pm 1$ nota e o erro absoluto médio (MAE).

Duas variações foram exploradas na classificação. A primeira diz respeito ao balanceamento das classes: como as notas não ocorrem com a mesma frequência, é possível ponderar cada classe pelo inverso da sua frequência, o que favorece o acerto nas notas raras ao custo da acurácia global. No nível do programa, em que os dados são escassos, essa ponderação mostrou-se contraproducente, e os melhores resultados foram obtidos sem ela. A segunda variação é uma formulação binária do alvo, na qual as notas são agrupadas em dois patamares, de um lado as notas 3 e 4 e de outro as notas de 5 a 7, buscando uma tarefa mais robusta diante do número reduzido de programas e da baixa frequência de algumas notas.

# 4 Resultados

A base de dados de docentes foi concluída para esta versão final, reunindo pouco mais de 1.500 professores credenciados nos PPGs de Ciência da Computação, distribuídos em 62 instituições distintas. Durante a coleta, cada docente foi associado ao seu perfil na OpenAlex por meio do cruzamento com o currículo Lattes; os registros cuja correspondência não pôde ser confirmada com segurança foram descartados, para não introduzir ruído por erros de desambiguação. As análises de classificação apresentadas adiante consideram os 1.464 professores para os quais foi possível obter tanto as métricas de rede quanto a nota do programa. Esta seção apresenta primeiro alguns números gerais da base e, em seguida, os resultados da rede de colaboração, da detecção de comunidades e da classificação das notas.

Tabela 4.1 – Números de Professores e PPGs por nota de PPG.

| Nota | Num. de professores | Num. de PPGs |
|---|---|---|
| 7 | 371 | 8 |
| 6 | 132 | 4 |
| 5 | 281 | 12 |
| 4 | 483 | 27 |
| 3 | 240 | 15 |

Em uma primeira análise temos, como mostrado na Tabela 4.1 a quantidade de professores por nota do PPG. Inicialmente já é notável que as universidades com notas mais altas concentram uma maior quantidade de professores por instituição, com os programas de nota 7 e 6 acumulando uma média aproximada de 46 e 33 docentes respectivamente. Isso se dá possivelmente por serem instituições maiores e de mais prestígio. Além disso universidades com mais docentes tendem a ter uma maior nota em geral.

Tabela 4.2 – Professores com orientadores de universidades fora dos PPGs.

| Nota | Num. de professores |
|---|---|
| 7 | 195 |
| 6 | 45 |
| 5 | 74 |
| 4 | 79 |
| 3 | 29 |

Outra métrica interessante é que cerca de 195 autores de programas nota 7 possuem orientadores de universidades que estão fora dos PPGs, ou seja, possivelmente universidades internacionais, o que é bem acima dos valores para as outras notas como mostra a Tabela 4.2.

A seguir são apresentados os resultados obtidos a partir da rede de colaboração construída, da detecção de comunidades e da classificação das notas dos programas.

## 4.1 A Rede de Colaboração Construída

Aplicando a metodologia da Seção 3.2 sobre os professores dos PPGs, considerando os trabalhos publicados até o final de 2022, obteve-se uma rede de coautoria com 107.115 nós (autores) e 1.015.773 arestas (colaborações). A maior parte dos nós corresponde a coautores externos aos programas, uma vez que a varredura dos trabalhos alcança todos os colaboradores dos professores. Entre os próprios professores dos PPGs há 8.847 arestas de coautoria, evidenciando um volume expressivo de colaboração direta entre os docentes dos programas.

Sobre essa topologia foram geradas as cinco versões da rede descritas na Seção 3.2, que se diferenciam apenas pelos pesos das arestas, e sobre cada uma foram calculadas as métricas de rede da Seção 3.3.

## 4.2 Comunidades e Notas dos PPGs

A Tabela 4.3 apresenta, para cada rede, a associação entre a partição em comunidades (obtida pelo método de Louvain com resolução 2,0) e as notas dos programas. A pureza mede a fração da nota predominante em cada comunidade, ponderada pelo número de professores; o *lift* é a razão entre essa pureza e a pureza de referência de 0,323, obtida ao se atribuir a todos a nota mais comum.

Tabela 4.3 – Associação entre comunidades (Louvain, resolução 2,0) e as notas dos PPGs, por rede. Pureza de referência (*baseline*) $= 0{,}323$.

| **Rede** | **Comunidades** | **Pureza** | ***Lift*** | **NMI** | $V$ **de Cramér** |
|---|---|---|---|---|---|
| Coautoria simples | 175 | 0,545 | 1,69 | 0,168 | 0,524 |
| Coautoria (contagem) | 166 | 0,535 | 1,66 | 0,157 | 0,505 |
| Produtividade do autor | 156 | 0,471 | 1,46 | 0,126 | 0,442 |
| Citações do autor | 160 | 0,456 | 1,41 | 0,111 | 0,405 |
| Citações dos artigos | 114 | 0,449 | 1,39 | 0,115 | 0,393 |

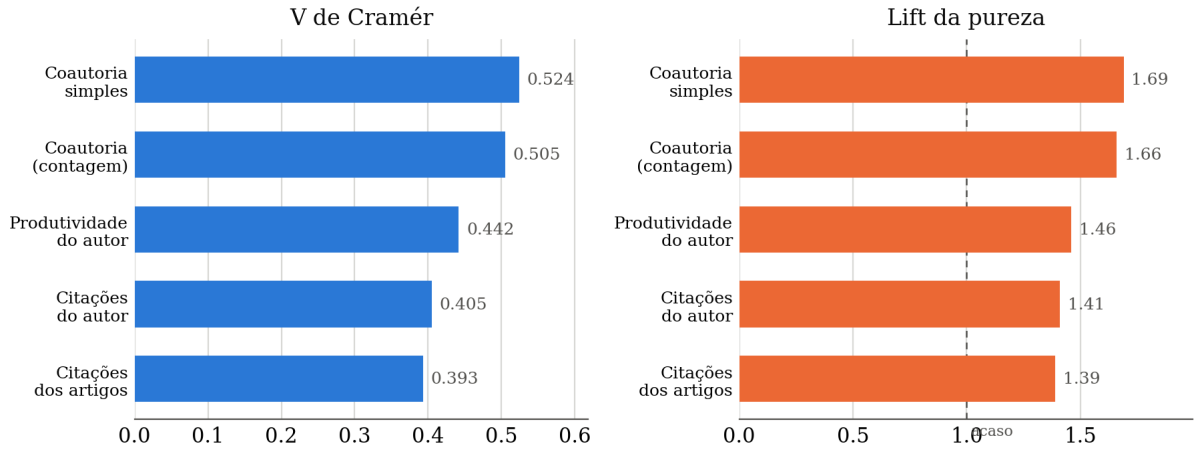



Figura 4.1 – Associação entre as comunidades detectadas e as notas dos PPGs, por rede, segundo o $V$ de Cramér e o *lift* da pureza sobre a referência. A linha tracejada indica o *lift* igual a 1, esperado ao acaso. Fonte: elaborada pelo autor.

A Figura 4.1 apresenta graficamente essa comparação. Todas as redes apresentam *lift* superior a 1, indicando que as comunidades agrupam professores de notas semelhantes acima do que seria esperado ao acaso. A associação é moderada, com o $V$ de Cramér chegando a 0,52 na melhor rede. Nota-se que as redes de topologia pura (coautoria simples e por contagem) refletem melhor as notas do que as redes ponderadas por citações ou produtividade, o que sugere que *quem colabora com quem* carrega mais informação sobre a nota do programa do que a magnitude das colaborações. Um padrão recorrente é a formação de comunidades dominadas por programas de nota 7, indicando que os docentes de programas de excelência tendem a colaborar densamente entre si.

## 4.3 Classificação das Notas

A Tabela 4.4 resume o desempenho da regressão logística na predição da nota do PPG, nos dois níveis descritos na Seção 3.5. Os valores entre parênteses correspondem ao *baseline* (predizer sempre a nota mais frequente).

Tabela 4.4 – Desempenho da regressão logística na predição da nota, com o *baseline* entre parênteses.

| **Métrica** | **Nível do autor** | **Nível do programa** |
|---|---|---|
| Amostras | 1.464 professores | 66 programas |
| Acurácia | 0,396 (0,322) | 0,515 (0,409) |
| $F_1$ macro | 0,256 | 0,394 |
| Acerto $\pm 1$ nota | 0,704 (0,665) | 0,909 (0,818) |
| MAE | 1,103 (1,262) | 0,621 (0,894) |

No nível do autor o sinal é fraco: o modelo supera o *baseline*, mas prevê a nota exata em menos da metade dos casos. Isso é esperado, pois a nota é atribuída ao programa, e não ao indivíduo, havendo grande variação entre professores de um mesmo PPG. Ao agregar as *features* por programa, nível em que a nota é de fato definida, o desempenho melhora de forma expressiva: cerca de 91% dos programas são classificados a no máximo uma nota de distância do valor real (acerto dentro de $\pm 1$), com erro médio de 0,62 nota. Como a nota é ordinal, o modelo, mesmo quando erra, tende a errar por pouco, confundindo notas adjacentes.

Quanto às *features* mais relevantes, no nível do autor destaca-se a centralidade de autovetor na rede de coautoria, seguida de indicadores de produtividade. No nível do programa, ilustrado na Figura 4.2, predominam os indicadores de produtividade agregada (índice-i10, índice-h e total de citações), complementados por métricas de rede como o coeficiente de clusterização e o PageRank. Esse resultado é coerente com o objetivo desta pesquisa: a estrutura de colaboração, combinada à produtividade dos docentes, carrega informação relevante sobre a nota do programa, ainda que não a determine completamente.

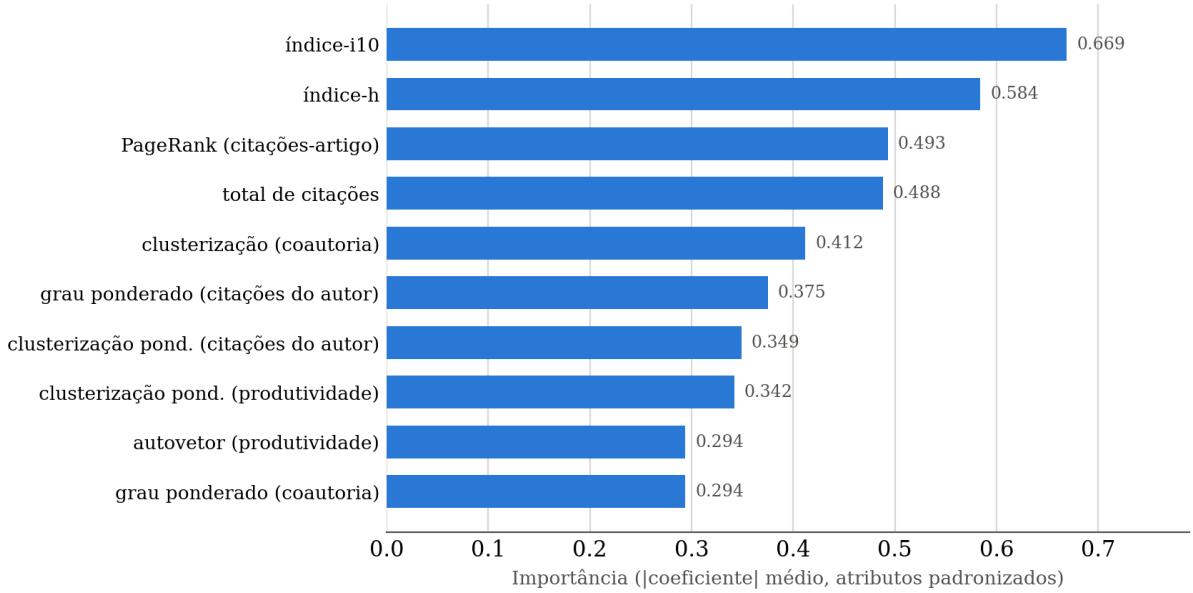



Figura 4.2 – Importância das dez *features* mais relevantes na classificação da nota no nível do programa, medida pelo valor absoluto médio dos coeficientes da regressão logística sobre atributos padronizados. Fonte: elaborada pelo autor.

Considerando o pequeno número de programas e a presença de notas pouco frequentes, avaliou-se também uma formulação binária do problema, na qual os programas são separados em dois grupos: os de notas 3 e 4, considerados medianos, contra os de notas de 5 a 7, considerados bons ou de excelência. Nessa formulação o modelo alcança acurácia de 0,803, contra 0,636 do *baseline*, e medida $F_1$ macro de 0,789, classificando corretamente 35 dos 42 programas do primeiro grupo e 18 dos 24 do segundo. A separação entre programas medianos e programas de

destaque mostra-se, portanto, bem mais nítida do que a predição da nota exata, o que reforça que as características de rede e de produtividade capturam sobretudo a distinção entre patamares de qualidade.

Também se avaliou o efeito de incluir a rede combinada descrita na Seção 3.2 entre as *features* do modelo no nível do programa. Ao contrário do esperado, a rede combinada não melhorou a predição: a acurácia caiu de 0,515 para 0,439, o acerto dentro de $\pm 1$ nota passou de 0,909 para 0,864 e o erro médio subiu de 0,62 para 0,74. Como todas as redes compartilham a mesma topologia, a rede combinada acrescenta poucas *features* realmente novas, e essas não trazem informação adicional relevante; com apenas 66 amostras, o acréscimo de variáveis pouco informativas favorece o sobreajuste e degrada o desempenho. O resultado é coerente com o observado na análise de comunidades, em que a topologia pura e os indicadores de produtividade se mostraram mais associados às notas do que as ponderações por citação ou produtividade. Combinar todas as redes, portanto, tende a diluir o sinal em vez de reforçá-lo.

Deve-se ressaltar que a análise no nível do programa envolve apenas 66 amostras, com algumas notas pouco representadas (a nota 6, por exemplo, aparece em somente 4 programas), o que confere caráter exploratório a esses resultados.

# 5 Considerações Finais

Esta monografia investigou se as notas atribuídas pela CAPES aos programas de pós-graduação em Ciência da Computação do Brasil guardam relação com a forma como os seus docentes colaboram entre si na produção científica. Para responder a essa questão, foi construída uma rede de colaboração a partir dos dados da OpenAlex, reunindo os professores credenciados nesses programas e os seus coautores, sobre a qual foram calculadas métricas de rede e aplicadas técnicas de detecção de comunidades e de classificação das notas.

Os resultados indicam que a estrutura de colaboração entre os docentes de fato guarda relação com a nota dos programas, ainda que essa relação seja parcial e não determinística. No nível do autor individual o sinal é fraco, o que era esperado, uma vez que a nota é atribuída ao programa e não ao pesquisador. Ao agregar as características por programa, no entanto, o desempenho da predição melhora de forma expressiva: cerca de 91% dos programas são classificados a no máximo uma nota de distância do valor real, e a separação entre programas medianos e programas de excelência atinge cerca de 80% de acurácia.

Entre os fatores mais associados às notas destacam-se a produtividade agregada dos docentes e a posição que eles ocupam na estrutura de colaboração. A análise de comunidades reforça esse resultado: as comunidades detectadas apresentam associação moderada com as notas, e observa-se que os docentes de programas de excelência tendem a colaborar densamente entre si. Um resultado relevante e um tanto contraintuitivo é que a topologia pura da colaboração, ou seja, quem colabora com quem, mostrou-se mais associada às notas do que as redes ponderadas por citações ou produtividade. Na mesma direção, a rede que combina linearmente todas as versões não superou as redes individuais, o que sugere que combinar os diferentes critérios tende a diluir o sinal em vez de reforçá-lo.

Deve-se ressaltar o caráter exploratório destes resultados, dado que a análise no nível do programa envolve um número reduzido de amostras, com algumas notas pouco representadas. Ainda assim, a pesquisa evidencia que existe relação mensurável entre a estrutura de colaboração dos docentes e a avaliação dos programas. Como possível desdobramento futuro, tais indicadores poderiam compor, de forma complementar, os critérios de avaliação dos programas de pós-graduação.

# Referências


BARABASI, A. L. *Network Science*. Cambridge University Press, 2016. Disponível em: <http://networksciencebook.com/>.

BARABáSI, A.-L.; JEONG, H.; NÉDA, Z.; RAVASZ, E.; SCHUBERT, A.; VICSEK, T. Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, Elsevier, v. 311, n. 3-4, p. 590–614, 2002.

BLONDEL, V. D.; GUILLAUME, J.-L.; LAMBIOTTE, R.; LEFEBVRE, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, IOP Publishing, v. 2008, n. 10, p. P10008, 2008.

CAMARINHA-MATOS, L. M.; AFSARMANESH, H. The emerging discipline of collaborative networks. In: CAMARINHA-MATOS, L. M. (Ed.). *Virtual Enterprises and Collaborative Networks*. Boston, MA: Springer US, 2004. p. 3–16. ISBN 978-1-4020-8139-2.

CRAMÉR, H. *Mathematical Methods of Statistics*. [S.l.]: Princeton University Press, 1946.

FORTUNATO, S.; BERGSTROM, C. T.; BÖRNER, K.; EVANS, J. A.; HELBING, D.; MILOJEVIĆ, S.; PETERSEN, A. M.; RADICCHI, F.; SINATRA, R.; UZZI, B. et al. Science of science. *Science*, American Association for the Advancement of Science, v. 359, n. 6379, p. eaao0185, 2018.

FRANCESCHET, M. Collaboration in computer science: A network science approach. *Journal of the American Society for Information Science and Technology*, Wiley Online Library, v. 62, n. 10, p. 1992–2012, 2011.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. et al. *An introduction to statistical learning*. [S.l.]: Springer, 2013. v. 112.

JR, D. W. H.; LEMESHOW, S.; STURDIVANT, R. X. *Applied logistic regression*. [S.l.]: John Wiley & Sons, 2013. v. 398.

MONGEON, P.; BOWMAN, T. D.; COSTAS, R. An open data set of scholars on twitter. *Quantitative Science Studies*, MIT Press, v. 4, n. 2, p. 314–324, 2023.

MOODY, J. The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American sociological review*, Sage Publications Sage CA: Los Angeles, CA, v. 69, n. 2, p. 213–238, 2004.

NEWMAN, M. E. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, National Acad Sciences, v. 98, n. 2, p. 404–409, 2001.

PRIEM, J.; PIWOWAR, H.; ORR, R. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*, 2022. Disponível em: <https://arxiv.org/abs/2205.01833>.

SUGIMOTO, C. R.; ROBINSON-GARCIA, N.; MURRAY, D. S.; YEGROS-YEGROS, A.; COSTAS, R.; LARIVIèRE, V. Scientists have most impact when they're free to move. *Nature*, v. 550, n. 7674, p. 29–31, October 2017. ISSN 1476-4687. Disponível em: <https://doi.org/10.1038/550029a>.

WAPMAN, K. H.; ZHANG, S.; CLAUSET, A.; LARREMORE, D. B. Quantifying hierarchy and dynamics in us faculty hiring and retention. *Nature*, Nature Publishing Group UK London, v. 610, n. 7930, p. 120–127, 2022.