UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

ALEX JUNIOR NUNES DA SILVA

ANÁLISE DAS REDES BRASILEIRAS DE COAUTORIA NOS PROGRAMAS DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO POR MEIO DE MEDIDAS TOPOLÓGICAS

DISSERTAÇÃO DE MESTRADO

CORNÉLIO PROCÓPIO

2019

ALEX JUNIOR NUNES DA SILVA

ANÁLISE DAS REDES BRASILEIRAS DE COAUTORIA NOS PROGRAMAS DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO POR MEIO DE MEDIDAS TOPOLÓGICAS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática da Universidade Tecnológica Federal do Paraná – UTFPR como requisito parcial para a obtenção do título de "Mestre em Informática".

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Orientador: Prof. Dr. Fabrício Martins Lopes

Coorientador: Prof. Dr. Jesús Pascual Mena-Chalco

CORNÉLIO PROCÓPIO

2019

Dados Internacionais de Catalogação na Publicação

S586 Silva, Alex Junior Nunes da

Análise das redes brasileiras de coautoria nos programas de pós-graduação em ciência da computação por meio de medidas topológicas / Alex Junior Nunes da Silva. – 2019.

102 f. : il. color. ; 31 cm.

Orientador: Fabrício Martins Lopes.

Coorientador: Jesús Pascual Mena-Chalco.

Dissertação (Mestrado) – Universidade Tecnológica Federal do Paraná. Programa de Pós-Graduação em Informática, Cornélio Procópio, 2019.

Bibliografia: p. 91-99.

1. Coautoria. 2. Universidades e faculdades - Pós-graduação. 3. Bibliometria. 4. Informática – Dissertações. I. Lopes, Fabrício Martins, orient. II. Mena-Chalco, Jesús Pascual, coorient. III. Universidade Tecnológica Federal do Paraná. Programa de Pós-Graduação em Informática. IV. Título.

CDD (22. ed.) 004

Biblioteca da UTFPR - Câmpus Cornélio Procópio

Bibliotecário/Documentalista responsável:
Romeu Righetti de Araujo – CRB-9/1676

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Câmpus Cornélio Procópio
Programa de Pós-Graduação em Informática

Título da Dissertação Nº 66:

"ANÁLISE DAS REDES BRASILEIRAS DE COAUTORIA NOS PROGRAMAS DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO POR MEIO DE MEDIDAS TOPOLÓGICAS".

por

Alex Junior Nunes da Silva

Orientador: **Prof. Dr. Fabricio Martins Lopes**
Co-orientador: **Prof. Dr. Jesús Pascual Mena Chalco**

Esta dissertação foi apresentada como requisito parcial à obtenção do grau de MESTRE EM INFORMÁTICA – Área de Concentração: Computação Aplicada, pelo Programa de Pós-Graduação em Informática – PPGI – da Universidade Tecnológica Federal do Paraná – UTFPR – Câmpus Cornélio Procópio, às 14h00 do dia 15 de agosto de 2019. O trabalho foi _____ pela Banca Examinadora, composta pelos professores:

_____
Prof. Dr. Fabricio Martins Lopes
(Presidente – UTFPR-CP)

_____
Prof. Dr. André Yoshiaki Kashiwabara
(UTFPR-CP)

_____
Prof. Dr. Sylvio Barbon Junior
(UEL)

Visto da coordenação:

_____
Danilo Sipoli Sanches
Coordenador do Programa de Pós-Graduação em Informática
UTFPR Câmpus Cornélio Procópio

A Folha de Aprovação assinada encontra-se na Coordenação do Programa.

Dedico este trabalho a Deus,
Ele merece tudo.

AGRADECIMENTOS

Certamente essa seção não irá atender a todas as pessoas que participaram dessa significativa etapa de minha vida. Assim sendo, desde já peço perdão àquelas que não estão diretamente citadas entre essas palavras, mas podem estar certas que fazem parte do meu pensamento e de minha gratidão.

Agradeço primeiramente a Deus, pois sem Ele eu nada seria. Dele por Ele e para Ele são todas as coisas.

Agradeço ao meu orientador Prof. Dr. Fabrício Martins Lopes, pela sabedoria, conhecimento, paciência e cuidado com que me guiou nesta trajetória. Também ao meu coorientador, Prof. Dr. Jesús Pascual Mena-Chalco por sempre se disponibilizar, compartilhar da suas experiências, conhecimentos e dar importantes contribuições a este trabalho.

Estendo meus agradecimentos ao amigo Matheus Montanini que auxiliou diretamente neste projeto, aos meus colegas de sala, aos meus colegas de trabalho que me auxiliaram em todo o período deste trabalho.

A Secretaria do Curso e os demais departamentos correlatos, pela cooperação.

Gostaria de deixar registrado também, o meu reconhecimento à minha família, pois acredito que sem o apoio deles seria muito difícil vencer este desafio, em especial à minha mãe, Maria de Fátima por todos os ensinamentos, conselhos e cuidado.

Agradeço também a minha namorada Larisse da Silva Souza por todo o suporte e compreensão em meus momentos de ausência.

Não posso deixar de fora dos agradecimentos os professores que participaram das minhas bancas e que colaboraram grandemente com apontamentos, sugestões e correções para este trabalho, são eles: Dr. Sylvio Barbon Junior, Dr. André Yoshiaki Kashiwabara e Dr. Willian Massami Watanabe.

Enfim, a todos os que por algum motivo contribuíram para a realização desta pesquisa.

"Não vos amoldeis às estruturas deste mundo,
mas transformai-vos pela renovação da mente,
a fim de distinguir qual é a vontade de Deus:
o que é bom, o que Lhe é agradável, o que é perfeito."
(Bíblia Sagrada, Romanos 12, 2)

RESUMO

SILVA, Alex Junior Nunes da. **Análise das Redes Brasileiras de Coautoria nos Programas de Pós-Graduação em Ciência da Computação por Meio de Medidas Topológicas.** 2019. 103 f. Dissertação de mestrado - Programa de Pós-Graduação em Informática, Universidade Tecnológica Federal do Paraná, Cornélio Procópio, 2019.

A análise das redes sociais tem se tornado uma área de grande atenção e foco nos últimos anos, pois com ela, podem ser observados padrões de comportamento entre seus componentes, bem como suas interações podem ser estudadas. Redes de coautoria, são um exemplo de rede social, na qual um pesquisador passa a ter uma ligação com outro pesquisador quando ambos compartilham a coautoria em um artigo publicado. A partir das redes formadas por essas ligações, medidas topológicas podem ser aplicadas para investigar padrões, classificar e prever os seu comportamento. Nesse trabalho, foi realizada a análise dos programas Brasileiros de pós-graduação em Ciência da Computação, para tal, foram extraídos os currículos acadêmicos da Plataforma Lattes, mapeadas as conexões entre os pesquisadores e geradas a representação das redes por meio de grafos. A partir desses grafos foram extraídas diversas medidas topológicas para compor um vetor de características para a respectiva classificação. Nesse sentido, esse trabalho propõe um índice quantitativo para medir a produtividade dos programas a partir da Média de Pesquisadores por Publicações. Além disso, são propostos três índices qualitativos de colaboração acadêmica: o Índice de Primeiro Autor, o Índice de Colaboração e o Índice de Senioridade, os quais analisam a posição em que um autor participa em uma publicação. As medidas extraídas e as medidas propostas foram analisadas levando em consideração a avaliação realizada periodicamente aos programas (Nota CAPES) para validar suas efetividades. A análise foi realizada considerando abordagens de classificação como *Random Forest*, seleção de características como *Best First* e também medidas de correlação. Os resultados gerados indicam grande relevância e identificam padrões de comportamento entre os programas que podem justificar a classificação realizada pela CAPES.

Palavras-chave: Coautoria. Pós-Graduação. Bibliometria. Redes Complexas. Medidas Topológicas. Reconhecimento de Padrões. Índice de Senioridade.

ABSTRACT

SILVA, Alex Junior Nunes da. **Analysis of Brazilian Co-authoship Networks in Graduate Programs in Computer Science through Topological Measurements.** 2019. 103 f. Masters dissertation - Programa de Pós-Graduação em Informática, Universidade Tecnológica Federal do Paraná, Cornélio Procópio, 2019.

The analysis of social networks has become an area of great attention and focus in recent years, as it can be observed the behavior between its components, as well as their interactions. Co-authoring networks are an example of a social network, in which a researcher has a connection with another researcher as they share co-authoring in published articles, on these networks, topological measures can be applied to investigate patterns, classify and predict their behavior. In this work, the analysis of the Brazilian postgraduate programs in Computer Science was carried out. To this end, academic curricula were extracted from the Lattes Platform, mapping the connections between researchers and generating the representation of connections through graphs. It was adopted several topological measurements in order to evaluate the graphs. A quantitative index "Average Researchers per Publications" to measure program productivity was proposed, three qualitative indices of academic collaboration were also proposed, called "First Author Index", "Collaboration Index", and "Seniority Index" which analyze the author position in a publication and gives it a rating. The analyzed measurements were compared with the government's periodic evaluation of the programs (CAPES Note) to validate their effectiveness. For comparison, classification approaches such as Random Forest, feature selection such as Best First and correlation were adopted. The results indicate a high relevance and identify patterns of behavior among the programs that explains the government's evalution of the graduate programs.

Keywords: Co-authorship. Postgraduate studies. Bibliometrics. Complex networks. Topological measurements. Pattern Recognition. Seniority Index.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AutoML	*Automated Machine Learning*
BI	*Business Intelligence*
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CE	*Comunity Edition*
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CSV	*Comma-Separated Values*
DBPL	*Digital Bibliography & Library Project*
DER	Diagrama de Entidade e Relacionamento
EE	*Enterprise Edition*
ETL	*Extraction, Transformantion and Load*
ID	Número Identificador
IES	Instituição de Ensino Superior
ISSN	*International Standard Serial Number*
MBA	*Master Business Administration*
NC	Nota CAPES
PDI	*Pentaho Data Integration*
SBS	*Sequential Backward Selection*
SFFS	*Sequential Forward Floating Selection*
SFS	*Sequential Forward Selection*
SGBD	Sistema de Gerenciamento de Banco de Dados
SNPG	Sistema Nacional de Pós-Graduação
SQL	*Structured Query Language*
XML	*Extensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

A teoria dos grafos foi criada em 1736 pelo matemático Leonhard Euler com o intuito de solucionar um problema na cidade de Königsberg localizada na antiga Prússia. Nessa cidade os habitantes indagavam, se havia a possibilidade de atravessar todas as 7 pontes que sobrepunham o rio Prejel sem passar por uma mesma ponte mais de uma vez (BARABÁSI, 2009; EULER, 1736).

Euler resolveu o problema interpretando as pontes como a estrutura de um grafo, ou seja, um conjunto nós (vértices) ligados por *links* (arestas), sendo que as faixas de terras eram denominadas como nós e as pontes que as conectavam eram chamadas de *links*. Ele provou de maneira simples, não ser possível atravessar todas as pontes sem passar por uma mesma ponte mais de uma vez, pois os nós com o número ímpar de *links* deveriam ser os nós de partida e/ou de chegada, portanto, não poderiam existir mais de 2 nós com a quantidade ímpar de *links*, os demais nós no meio do percurso deveriam ter um número par de links, todavia, existiam 4 nós com *links* ímpares, provando ser impossível a hipótese anteriormente levantada (BARABÁSI, 2009; EULER, 1736; GABARDO, 2015).

Com as contribuições de Euler na teoria dos grafos, vários matemáticos desenvolveram algumas soluções de problemas do mundo real modelando-os em forma de grafos. Grafos podem ser análogos à redes, de encontro a esse cenário, cria-se então a teoria das redes que são geralmente um agrupamento de grafos e essas são largamente utilizadas para entender e classificar várias redes naturais e/ou artificiais do mundo real como, por exemplo, na biologia, na qual redes são utilizadas para identificar genes ou proteínas importantes em um ambiente biológico, o modelo de redes também é encontrado na grande rede de computadores, a internet (BARABÁSI, 2009). Nesse contexto, destacam-se também as redes sociais, que são um caso particular de redes de interações entre indivíduos que compartilham informações entre si e estão inseridos em um determinado contexto.

De acordo com Barabási (2009) o fenômeno de redes sociais pode ser exemplificado com o cristianismo, que foi um movimento religioso criado pelo seu líder Jesus de Nazaré sendo que o mesmo possuía na época ideias adversas aos judeus e as autoridades romanas o que poderia fazer com que não houvesse sucesso no crescimento do movimento, contudo, um discípulo seu chamado Paulo, disseminou as mensagens dentre vários cercos sociais observando sempre as principais comunidades contemporâneas a ele e percorrendo cerca de 16 mil quilômetros em 12 anos. Atualmente mais de 2 bilhões de pessoas se declaram cristãs, cerca de 31,4% da população mundial (CIA, 2018), Barabási afirma que Paulo foi o primeiro caixeiro-viajante do cristianismo e utilizou-se do conceito de redes sociais para difundir suas ideias e crenças.

No cenário atual, é elevado o crescimento das redes sociais, o Facebook[1] possuía em março de 2019, uma média de 2,38 bilhões de usuários ativos (FACEBOOK, 2019). Um indivíduo pode possuir um registro em mais de um serviço de rede social e com base nas informações movimentadas pelo mesmo, é possível utilizar ferramentas de análise de dados para mapear seus comportamentos revelando seus interesses, hábitos e padrões de consumo, sendo que de posse dessas informações as empresas podem oferecer serviços e produtos de maneira mais assertiva aos seus clientes e clientes em potenciais (GABARDO, 2015; JAMJUNTRA et al., 2017).

Atualmente, pelo menos duas das dez maiores empresas mundiais de tecnologia, a Google e a Facebook, faturam bilhões de dólares todos os anos analisando o produto "informação" (FACEBOOK, 2018; VIEW; APRIL; GOOG, 2018; WINKLER; BARR, 2014), isso fornece uma visão da importância de se extrair dados de redes sociais.

Quando o assunto é a análise de redes, frequentemente cita-se as redes complexas, que são redes compostas por elementos característicos, sendo que elas possuem comunidades de vértices (nós) com um grande número de conexões entre si, a distribuição de graus das suas arestas, comumente obedecem uma lei de potência e a forma como seus elementos se agrupam podem obedecer a alguns modelos que fora observados por pesquisadores. Um exemplo de rede complexa são as redes sociais (GABARDO, 2015).

Assim como as redes sociais particulares, um outro modelo de redes que pode ser estudado, é o das redes sociais de colaborações acadêmicas, essas, podem representar a interação entre os diferentes pesquisadores, de distintos lugares do país, portanto, é possível identificar as colaborações entre projetos de pesquisa e trabalhos científicos. Com a análise dessas redes é possível vislumbrar o elo de conexão entre 2 ou mais pesquisadores, entre os programas de pós-graduação ou entre as instituições de ensino superior (BORDIN; GONÇALVES; TODESCO, 2014; LOPES et al., 2011; NEWMAN, 2001a; NEWMAN, 2001b).

Com as análises das redes de colaborações acadêmicas podem ser inferidas métricas nas redes e seu comportamento pode ser previsto. Por conseguinte, pode-se observar, por exemplo, os padrões de interações entre os programas de pós-graduação. Esses padrões presentes nas redes podem servir de subsídio à tomada de decisões, pois a observação deles pode levar ao entendimento de como ocorrem as colaborações entre os pesquisadores, grupos de pesquisas, entre pesquisadores de diferentes regiões, entre os programas de pós-graduação, dentre outros (BORDIN; GONÇALVES; TODESCO, 2014; LOPES et al., 2011; MENA-CHALCO et al., 2014; NEWMAN, 2001a; NEWMAN, 2001b).

Os programas de pós-graduação no Brasil são regulados e avaliados pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), sendo que o órgão possui

[1] Plataforma que oferece o mais popular serviço particular de rede social. No qual um usuário pode, de maneiras heterogêneas, interagir com outros usuários.

sistemas de informação internos nos quais os dados coletados são analisados com objetivo de avaliar e acompanhar os programas, pesquisadores e instituições. Após a avaliação dos programas a CAPES os classifica, divulgando publicamente as notas dessa avaliação. Logo, de acordo com a análise realizada pela CAPES, é atribuída uma nota de qualidade ao programa avaliado (GATTI et al., 2003; HORTA; MORAES, 2005).

Pertinente a esse contexto, é proposto neste trabalho a geração e análise de redes sociais acadêmicas com o objetivo de investigar padrões de colaboração entre pesquisadores e entre programas de pós-graduação do Brasil, buscando associar as notas de avaliação recebidas pela CAPES aos padrões topológicos das redes de interações.

1.1 Problemas e Premissas

Fora observado que, com relação às publicações, não existem métricas efetivamente claras para que o coordenador de um programa de pós-graduação possa explorar e com isso tomar decisões adequadas sobre qual caminho ele deve seguir com o intuito de melhorar os indicadores e, por conseguinte melhorar a classificação do programa junto a CAPES, sendo que sem essas métricas de maneira clara os coordenadores podem usar métodos intuitivos o que pode gerar erros, análises espúrias e o processo torna-se trabalhoso.

A contribuição externa entre os programas de pós-graduação, isto é, os pesquisadores de um programa que realizam projetos de pesquisa com os pesquisadores de outros programas, pode ser um fator importante para avaliar o processo de evolução de um programa, porém obter essa informação pode não ser um procedimento trivial, pois, é elevado o número de publicações, não existe uma base de dados aberta e unificada com os metadados de todas as publicações e não existe uma ferramenta computacional específica para essa análise (ALVES; YANASSE; SOMA, 2011; VANZ, 2009).

Um dado que pode ser observado é a ordem de citação na autoria dos projetos de pesquisas, essa ordem pode ter diversas regras, como por exemplo, a ordem alfabética dos nomes dos pesquisadores envolvidos no projeto. Todavia, considerando a área de Ciência da Computação, comumente é adotada a ordem com o pesquisador principal, sendo aquele que é responsável pela execução da pesquisa e/ou o que mais contribuiu para a pesquisa é o primeiro autor, já os demais coautores podem ser corresponsáveis por tarefas de menor esforço dentro da pesquisa. É comum também o pesquisador mais experiente, o responsável pelo projeto de pesquisas ou o orientador do pesquisador principal ser o último coautor citado nos trabalhos. Essa ordem de citação pode indicar importantes padrões, desse modo, é interessante observá-los, sendo que, não existe uma ferramenta que extraia esses dados e os exiba de maneira direta (ALVARENGA, 2007; HILÁRIO; GRÁCIO; GUIMARÃES, 2018; NEWMAN, 2001a; NEWMAN, 2001b; PETROIANU, 2002).

1.2 Levantamento de Hipóteses

São apresentadas abaixo algumas hipóteses a serem respondidas com o resultado deste trabalho, essa hipóteses foram levantadas com base no estudo dos trabalhos anteriores.

1. Redes de programas com diferentes notas de avaliação CAPES possuem diferença significativa em sua estrutura e dinâmica;
2. Programas com maiores notas CAPES possuem maior colaboração interna (intraprogramas) que os demais;
3. Programas com notas CAPES mais baixas possuem um maior Índice de Primeiro Autor;
4. Pesquisadores de programas com Nota CAPES intermediárias possuem maior Índice de Colaboração
5. Programas com notas CAPES mais elevadas possuem um grande número de pesquisadores com alto Índice de Senioridade.

1.3 Objetivos

1.3.1 Objetivo Geral

Tem-se como objetivo deste trabalho a identificação de padrões de comportamento em redes de coautoria acadêmica, por meio da análise de métricas topológicas das redes de programas brasileiros de pós-graduação em Ciência da Computação bem como a proposição de novas métricas de colaboração acadêmica.

1.3.2 Objetivos Específicos

Para o alcance do objetivo acima descrito, é necessária a conclusão das seguintes etapas:

- Extração e cruzamento dos dados;
- Aplicação e análise das métricas de redes complexas existentes;
- Aplicação e análise das novas métricas propostas;
- Comparação entre as métricas e a avaliação CAPES.

1.4 Justificativa

A identificação de padrões topológicos nas redes podem fornecer *insigths* sobre como as mesmas são organizadas, com isso, esses padrões podem ser estudados e medidas sobre eles podem ser aplicadas, essas por sua vez, podem definir as redes e até mesmo classificá-las. Observar os padrões nessas redes pode de maneira informatizada, auxiliar na escolha de revisores mais apropriados para eventos e/ou periódicos.

As informações geradas como resultado deste trabalho podem ser de grande importância para a gestão das instituições de ensino superior (IES), sendo que os dados também poderão ser utilizados pelos pesquisadores para subsidiar suas tomadas de decisões no planejamento de suas colaborações, desenvolvimento de ações nos programas de pós-graduação e melhoria qualitativa junto a CAPES, desenvolvimento de políticas públicas no fomento a pesquisas, entre outras.

Com essas informações, as IES e os pesquisadores passariam a ter uma visão consolidada da interação entre: os pesquisadores, os programas e as IES.

1.5 Organização do Texto

Este documento está organizado de forma em que no Capítulo 2 é realizada uma fundamentação teórica com o intuito de definir minimamente os elementos necessários para a total compreensão do itens abordados no projeto, bem como os trabalhos que são relacionados a esse. No Capítulo 3 são descritos os passos realizados para a obtenção e processamento dos dados que serão utilizados. No Capítulo 4, é apresentado o método utilizado neste trabalho e que fora aplicado sobre os dados extraídos. O Capítulo 5 está composto com os resultados alcançados com a aplicação do método e também a interpretação dos mesmos aplicado ao trabalho. O capítulo 6 conta com o desfecho realizado nesse trabalho e a indicação de possíveis trabalhos futuros. Finalmente nos Apêndices A e B são exibidos respectivamente o Diagrama de Entidade e Relacionamento do banco de dados gerado e uma descrição detalhada das fontes de dados utilizadas.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados os principais conceitos utilizados neste trabalho, de forma sucinta exibimos as definições dos conceitos de grafos, redes complexas, seleção de características e métricas acadêmicas, assim como trabalhos relacionados ao tema deste projeto.

2.1 Grafos

Os grafos são abstrações matemáticas que podem ser usados na modelagem de muitos problemas reais de diversas áreas (AHO; HOPCROFT; ULLMAN, 1974; BONDY; MURTY, 1976). Seu uso é de grande abrangência por possuir uma estrutura simples que permite modelar sistemas sociais e de comunicações (LANCICHINETTI et al., 2010; ROSA, 2011). Ele é usado para representar elementos, denominados vértices ou nós e suas ligações denominadas arestas, *links* ou ainda *edges*. Pode-se definir um grafo como: $\mathbf{G(V,E)}$, em que $\mathbf{G}$ é um grafo com um conjunto $\mathbf{V = \{v_1, v_2, v_3, ..., v_n\}}$ de vértices conectados por um conjunto $\mathbf{E = \{e_1, e_2, e_3, ..., e_n\}}$ de arestas.

A Figura 1 exibe um exemplo básico de um grafo de cinco vértices (1, 2, 3, 4 e 5) conectados por arestas de igual peso e não direcionadas.

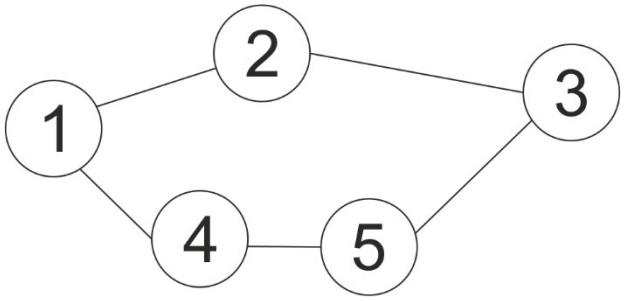

Figura 1 – Grafo de exemplo composto por 5 vértices.

Fonte: Autoria própria

Os vértices do grafo são compostos por elementos dos quais podem receber um código de identificação (*ID*), ou também podem carregar outras informações de acordo com o contexto onde está sendo aplicado. Uma métrica comumente utilizada sobre os vértices é o seu grau, que é um número que representa a quantidade de arestas que são conectadas a esse vértice.

As arestas que conectam os vértices, podem possuir direções específicas, sendo que a estes grafos dá-se o nome de grafo direcionado ou grafo dirigido. Assim como a direção, uma aresta pode possuir um peso ou então um custo associado.

A Figura 2 ilustra um exemplo de um grafo de cinco vértices (1, 2, 3, 4 e 5) conectados por arestas direcionadas e com peso (custo). Com uma análise visual é possível inferir que não há possibilidade de percorrer um caminho do vértice 2 para a vértice 1, pois não existe esse caminho, todavia, o contrário é possível, pode-se inferir também que o custo do caminho entre o vértice 4 até o 5 é maior do que o custo do caminho entre o vértice 1 até o 2.

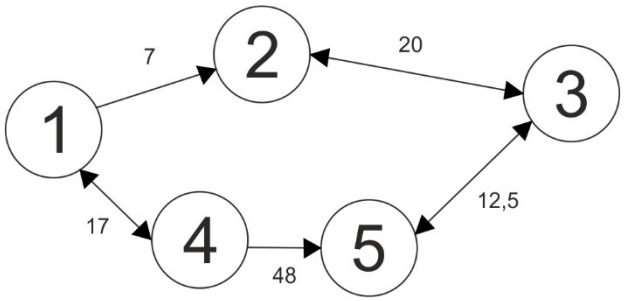

Figura 2 – Grafo de 5 vértices com arestas dirigidas e com peso.

Fonte: Autoria própria

2.2 Redes Complexas

Em algumas situações os grafos podem ser caracterizados como redes, isto é, um grupo de elementos conectados a outros elementos por meio de uniões específicas, sendo que essas redes podem ser concretas como, por exemplo, a internet, os sistemas de voo e as redes do cérebro humano (neurais); ou ainda podem ser abstratas como por exemplo as redes de interações entre indivíduos (BOCCALETTI et al., 2006).

As redes reais possuem propriedades que foram observadas ao longo do tempo por pesquisadores que afirmam que existem características dispostas nelas, que são características de conectividade que não seguem padrões aleatórios, suas estruturas são irregulares complexas e evoluem dinamicamente ao longo do tempo (BOCCALETTI et al., 2006). Redes do mundo real podem possuir vértices com grande número de conexões (*hubs*), podem também existir comunidades de vértices e a distribuição de graus (quantidade de conexões de um vértice) podem seguir lei de potência matemática (BARABÁSI; BONABEAU, 2003; COSTA et al., 2007). Dessas observações tem-se a teoria das redes complexas sendo que uma grande quantidade de sistemas da natureza e da sociedade podem ser representados por meio de redes complexas (ALBERT; BARABÁSI, 2002; LOPES, 2011).

Em redes do mundo real, observa-se que existem padrões de organização e estruturação, e esses padrões ainda que são aplicados para cada tipo de rede ou classe de redes, determinam a conectividade nas redes e influenciam na dinâmica das informações trafegadas em seu interior (COSTA et al., 2007). Nas seguintes seções, são explicados os modelos comumente estudados para entender o comportamento dessas redes, sendo que elas apresentam topologias e características que as diferem.

2.2.1 Modelo de Erdös-Renyi

O modelo de redes aleatório criado por Paul Erdös e Alfréd Rényi pode-se considerar como o mais simples de redes complexas (COSTA et al., 2007; LOPES, 2011), nele existe uma probabilidade $\boldsymbol{0 < p < 1}$ de que haja a conexão entre 2 vértices da rede. Nesse modelo a distribuição de frequências dos graus (número de conexões por vértice) da rede obedece uma distribuição muito próxima a Poisson (COSTA et al., 2007; GABARDO, 2015), sendo ela:

$$P(k) = \frac{e^{-(k)}(k)^k}{k!}. \tag{2.1}$$

Por conseguinte, esse modelo de redes também pode ser encontrado na literatura com o nome de *Poisson random graphs* (BOCCALETTI et al., 2006).

2.2.2 Modelo de Barabási-Albert

Albert-Lázló e Réka Albert conceberam um método de geração de redes (BARABASI; ALBERT, 1999) em que para as arestas serem conectadas observa-se o grau dos vértices e os que possuem um maior grau têm uma maior possibilidade de receber as novas conexões. Esse paradigma é conhecido como modelo de ligação ou anexação preferencial, essa particularidade é denominada ricos ficam mais ricos (*rich-get-richer*) sendo que nesse modelo de rede a distribuição de frequência dos graus seguem uma lei de potência. Redes sem escala (*scale-free*) também é comumente chamada essas redes, uma vez que, os vértices com maior êxito recebem vantagens com relação aos demais (GABARDO, 2015).

A probabilidade *(PB)* de um novo vértice v_i conectar-se a um vértice v_j já existente é proporcional ao grau k_j do vértice existente v_j, conforme demonstrado em:

$$PB(v_\mathrm{i} \rightarrow v_\mathrm{j}) = \frac{k_\mathrm{j}}{\sum_\mathrm{u} k_\mathrm{u}}. \tag{2.2}$$

2.2.3 Modelo de Watts e Strogatz

No modelo proposto por Duncan J. Watts e Steven H. Strogatz, o coeficiente de *clustering* é elevado e o comprimento médio entre os caminhos é pequeno. Esse modelo utiliza o clássico paradigma de mundo pequeno (*Small-world*), também conhecido como a "teoria dos seis graus de separação", no qual através de um estudo científico, realizado pelo pesquisador Stanley Milgram no ano de 1967, foi demonstrado que duas pessoas diferentes possuem amigos (conexão) em comum passando por no máximo seis outras pessoas. Por conseguinte, seria possível enviar um recado por uma "pessoa A" e passando por no máximo seis pessoas diferentes esse recado poderia chegar até "uma pessoa B" (BARABÁSI, 2009; GABARDO, 2015; WATTS; STROGATZ, 1998).

Esse modelo de redes possui características semelhantes ao modelo proposto por Erdös e Rényi, todavia, nesse são gerados pequenos grupos de três vértices conectados entre si (tríades), sendo que esse modelo é bem próximo das redes do mundo real (GABARDO, 2015).

2.3 Métrica de Redes

Nessa seção são apresentadas algumas das métricas quantitativas comumente aplicadas em grafos e redes complexas por distintos algoritmos.

2.3.1 Ordem e Tamanho

A ordem de um grafo é um parâmetro quantitativo que quando extraído, expressa a quantidade de vértices que formam o grafo. Essa métrica pode fornecer dados de grande relevância pois ela exibe trivialmente a dimensão do grafo analisado, sendo possível observar e comparar diferentes redes usando essa medida.

O tamanho de um grafo é a soma da ordem (número de vértices) com o número total de arestas. Embora essa medida seja pouco complexa, ela pode, assim como a ordem do grafo, exibir dados de grande relevância no estudo e análise de redes/grafos.

2.3.2 Grau e grau médio

O grau (*degree*, ou ainda *degree centrality*) ou conectividade de um grafo diz respeito a quantidade de conexões diretas que um vértice possui. No caso dos grafos dirigidos, esses graus podem ser divididos em graus de entrada (*indegree*) ou em graus de saída (*outdegree*) (BOCCALETTI et al., 2006). Essa medida pode ter grande importância, pois através dela pode-se verificar a importância que cada nó possui na rede, sendo que, os vértices com muitas conexões (graus) podem ser considerados vértices de grande importância para a rede, já os vértices sem conexões podem ser irrelevantes a mesma (LOPES, 2011).

Outra medida comumente calculada é o grau médio da rede (*average degree*), que calcula-se somando o grau de todos os vértices e dividi-se pelo total de vértices da rede. Para calcular o grau médio de uma rede $\langle k \rangle$, deve-se atentar ao fato que uma mesma aresta está conectada a dois vértices diferentes, portanto essa medida pode ser obtida por:

$$\langle k \rangle = \frac{2E}{V} \tag{2.3}$$

em que **E** são as arestas que são somadas 2 vezes e **V** os vértices do grafo (GABARDO, 2015).

É comum também calcular e exibir a distribuição de graus (*degree distribution*) da rede, sendo que ela informa a quantidade de grafos com uma determinada quantidade de grau o que torna simples o processo de avaliação a disseminação dos graus na rede.

2.3.3 Densidade

A densidade de um grafo está relacionada com a sua ordem e seu tamanho, ou seja, é proporção entre a quantidade de arestas e vértices, através dessa medida é possível definir se um grafo é denso caso ele possua uma grande quantidade de arestas para uma quantidade de vértices; ou ainda se o grafo é esparso caso contrário. Para calcular a densidade dividi-se o número de arestas pelo número total de arestas possíveis da rede (GABARDO, 2015; MENA-CHALCO et al., 2014).

2.3.4 Coeficiente de aglomeração

Um *cluster* é definido como um grupo de vértices fortemente conectados entre si, portanto uma comunidade de vértices, observa-se que em muitas redes do mundo real e principalmente nas redes sociais, os vértices tendem a criarem grupos com grande quantidade de conexões entre si (HOLLAND; LEINHARDT, 1971). O coeficiente de aglomeração (coeficiente de *cluster*, agrupamento, transitividade ou ainda *clustering coefficient*), indica a probabilidade de que dois vértices adjacentes estejam conectados a outro, esse coeficiente pode obedecer dois grupos, o coeficiente de aglomeração local ou o coeficiente de aglomeração global, sendo que o primeiro é uma medida específica de um vértice e seu objetivo é determinar a densidade das arestas estabelecidas entre os vizinhos desse nó, já no segundo a medida é a disposição da rede em existir conjuntos de elementos, portanto uma visão geral do agrupamento da rede (BOCCALETTI et al., 2006; GABARDO, 2015; MENA-CHALCO et al., 2014).

O coeficiente de *cluster* local indica a tendência que um nó tem de formar um subgrafo com todas os vértices conectados entre si (Clique). Um exemplo de Clique são os trios, que são grupos de 3 vértices fortemente conectados entre si (triângulos). O cálculo do coeficiente de *cluster* local atende a seguinte equação:

$$C_{\mathrm{i}} = \frac{2n_{\mathrm{i}}}{k_{\mathrm{i}}(k_{\mathrm{i}} - 1)}, \tag{2.4}$$

em que o coeficiente de aglomeração local C_i do vértice i, n_i denota a quantidade de arestas entre os k_i vértices de i até esse determinado vértice (BOCCALETTI et al., 2006).

O coeficiente de *cluster* global mensura a tendência da rede para desenvolver grupos. O cálculo consiste na quantidade de trios conectados entre si na rede (triângulos), conforme:

$$C = \frac{3 \times \text{número de triângulos no grafo}}{\text{número de trios conectados no grafo}} \tag{2.5}$$

sendo que no coeficiente de *cluster* global C divide-se três vezes cada triângulo pois cada triângulo corresponde a três trios, pelo número de trio de vértices conectados onde um único vértice é o principal e os outros dois são de acompanhamento (BOCCALETTI et al., 2006).

O valor de grau de um nó pode variar de zero onde esse não possui nenhuma conexão com outro nó, sendo classificado como nó isolado, ou o valor máximo que é a quantidade total de vértices da rede subtraindo um que é o nó observado, caso o nó possua esse valor significa que ele está conectado diretamente a todos os outros vértices da rede, sendo ele um importante *cluster* (MENA-CHALCO et al., 2014).

2.3.5 Centralidade

Em redes sociais alguns vértices podem possuir maior importância que outros, para representar essa importância usa-se as medidas de centralidade (FREEMAN, 1977; FREEMAN, 1978). Ao observar as medidas de centralidade é possível identificar quem são os indivíduos de maior influência na rede, uma vez que esses indivíduos possuem um maior controle do fluxo de informações do que os demais, geralmente esses indivíduos também são importantes para as redes, porque caso eles sejam removidos os caminhos mais curtos podem ser perdidos levando assim a um maior aumento no tempo de tráfego de uma informação na rede (NEWMAN, 2001b; WASSERMAN; FAUST, 1994).

As medidas de centralidade são dividas em 2 categorias, centralidade local e centralidade global, sendo que na local analisa-se a relevância de um vértice com relação à sua vizinhança; já na centralidade global a análise é realizada com base na relevância de um vértice pela totalidade da rede. Como exemplo de centralidade local tem a medida centralidade de grau (*degree centrality*), já na centralidade global têm-se a centralidade de intermediação (*betweenness centrality*) e a centralidade de proximidade (*closeness centrality*). Na centralidade de intermediação Freeman (1978) propôs a atribuição de

importância para um vértice de acordo com o fluxo de informações que passam através dele, para isso, analisa-se os vértices que estão no menor "caminho" de conexão entre outros 2 vértices, agindo como uma ponte de informações. Na centralidade de proximidade Freeman (1978) propôs uma medida global de centralidade levando em consideração a proximidade dos pontos, sendo assim essa medida avalia a proximidade de um vértice com os demais vértices da rede, a definição dela é a soma das menores distância (geodésica) entre o vértice e todos os demais vértices da rede.

2.4 Vulnerabilidade em Redes

Um ponto com grande importância a ser observado e analisado em ambientes de redes complexas é a vulnerabilidade de uma rede (ALBERT; BARABÁSI, 2002; KOVÁCS; BARABÁSI, 2015), para isso analisa-se a importância de cada vértice dela e os classifica como altamente importantes ou pouco importantes para a sua organização. Entende-se que um vértice possui grande importância para a estrutura da rede se caso ele seja removido essa estrutura sofrerá grandes alterações, o impacto dessas alterações estão atrelados a forma como os caminhos são dispostos entres as arestas. Existem diferentes formas de estudar a vulnerabilidade de redes, pode-se por exemplo avaliar as interrupções de uma rede logo após a remoção de um componente e considerar toda a reação em cascata, como também pode-se avaliar sem uma abordagem em cascata, embora ambas tenham em comum a análise da importância de um vértice (índice de centralidade) os resultados entre as duas abordagens podem ser diferentes entre si (LATORA; MARCHIORI, 2005).

2.4.1 O fenômeno *rich-club* em redes

Em algumas redes complexas é observado o fenômeno *rich-club*, que são comunidades de vértices com grande número de conexões (*hubs*) conectados entre si, dessa maneira, o próprio nome do fenômeno trivialmente o caracteriza pois existe um "clube de vértices ricos em conexões" (ZHOU; MONDRAGON, 2004).

É possível extrair uma métrica, o coeficiente *rich-club*, que mede a proporção em que os *hubs* de uma rede são conectados entre si, esse coeficiente pode servir como avaliação da robustez de uma rede, pois quanto maior o valor, mais fortemente conectados, o que indica que caso um desses *hubs* seja removido, menor será o impacto na estrutura da rede (COLIZZA et al., 2006; MCAULEY; COSTA; CAETANO, 2007; OPSAHL et al., 2008; ZHOU; MONDRAGON, 2004).

2.5 Classificação e Seleção de Características

Para classificar objetos pode-se usar técnicas de reconhecimento de padrões, essas técnicas permitem encontrar propriedades nesses objetos e agrupá-los em conjuntos ou classes de acordo com suas com características iguais ou semelhantes (CHEN; HAN; YU, 1996; MACQUEEN, 1967; MICHIE; SPIEGELHALTER; TAYLOR, 1994; RUSSELL; NORVIG, 2002; SYMEONIDIS; MITKAS, 2005; ZHENG et al., 2018).

Os métodos de classificação são comumente subdivididos em 3 grupos: a classificação supervisionada, onde existe uma coluna de dados com os valores conhecidos (rótulos) e o objetivo é fazer com que o modelo aprenda as regras para mapear os outros dados de entrada para que sejam separados em classes de acordo como o rótulo informado; a classificação não-supervisionada, onde não se conhece o rótulo das classes e o modelo precisa entender os relacionamentos entre os dados sem que haja uma supervisão; há também o aprendizado por reforço onde o modelo realiza ações que são avaliadas, e um retorno com uma recompensa ou uma punição é realizado com base na ação desenvolvida (CHEN; HAN; YU, 1996; MICHIE; SPIEGELHALTER; TAYLOR, 1994; RUSSELL; NORVIG, 2002; SYMEONIDIS; MITKAS, 2005).

Como os processos de classificação utilizam características/atributos presentes nos dados para serem executados, as caraterísticas têm um importante papel nesse contexto, por conseguinte, existem estudos e linhas de pesquisas voltadas para a seleção de características, todavia, como não é o escopo principal deste trabalho, não serão apresentadas de maneira detalhada o estudo da seleção de características.

É importante observar que existem algoritmos de seleção de características, bem como existem os de extração de características, sendo que no primeiro são observadas as amostras (objetos) e são analisadas quais são as características que melhor representam e separam os diferentes objetos; já na extração de características, são criadas novas características combinando características iniciais (CAMPOS, 2001; LOPES, 2011; WEBB, 2002). Alguns autores podem utilizar o termo Seleção de Atributos (*Attribute Selection*), que é um sinônimo para tratar a Seleção de Características.

2.5.1 Busca Sequencial para Frente (SFS)

O algoritmo de Busca Sequencial para Frente (*Sequential Forward Selection* ou SFS) é um algoritmo do tipo *wrapper*, ou seja, a medida de desempenho de previsão do processo de aprendizagem do mesmo vem do resultado da sua função critério, isso auxilia na verificação quanto a qualidade das variáveis utilizadas (GUYON; ELISSEEFF, 2003; LOPES, 2011).

O SFS é um algoritmo que inicia com um conjunto vazio e adiciona recursos selecionados por uma função de avaliação, o algoritmo vai adicionando essas características a cada

iteração, cada característica junto a anterior constitui o melhor grupo de características, portanto o algoritmo melhora a cada iteração, até que uma condição de parada seja satisfeita.

Existe também o algoritmo de Busca Sequencial para Trás (*Sequential Backward Selection* ou SBS) que possui seu funcionamento muito próximo ao SFS, todavia, no SBS ao invés de ser adicionada a caraterística mais relevante, nele a menos relevante é removida a cada iteração, essa abordagem é conhecida como *top-down* (LOPES, 2011; PUDIL; NOVOVIČOVÁ; KITTLER, 1994). Para este projeto o SFS será utilizado com sua abordagem *bottom-up*.

2.5.2 Busca Sequencial Flutuante para Frente (SFFS)

O algoritmo SFS pode apresentar uma propriedade adversa denominada efeito *nesting*, que ocorre porque uma característica que não faz parte da solução ótima é inserida no resultado do algoritmo e a mesma permanece, levando o algoritmo a uma solução sub-ótima (SOMOL et al., 1999). Isso pode acontecer porque duas características em conjuntos pode classificar um conjunto corretamente, todavia, uma dessas características em particular pode ser ruins nessa classificação (LOPES, 2011).

Para evitar o efeito *nesting*, pode-se utilizar o algoritmo de Busca Sequencial Flutuante para Frente (*Sequential Forward Floating Selection* ou SFFS), esse algoritmo pode evitar esse problema porque com ele é possível inserir as características no conjunto de classificação de maneira flutuante (LOPES, 2011; PUDIL; NOVOVIČOVÁ; KITTLER, 1994).

O SFFS tem grande eficiência computacional e chega muito próximo a uma solução ótima, existem adaptações dele que entregam melhores resultados, entretanto, o custo computacional necessário também cresce o que muitas vezes os torna inviável e mesmo que eles entregam melhores resultados, não conseguem impedir em sua totalidade o efeito *nesting*.

O funcionamento do SFFS inicia com um quadro vazio ($k=0$) e vai aplicando-se o algoritmo SFS até que o quadro resposta seja igual a dois ($k=2$), quando o agrupamento torna-se maior que 2 ($k>2$) aplica-se então o algoritmo SBS removendo características menos relevantes. O algoritmo SFFS segue dessa forma até o momento em que encontra uma condição de parada, sendo que os melhores resultados de cada iteração são armazenados e dentre eles o melhor é a resposta final do algoritmo.

A Figura 3 exibe o fluxograma do funcionamento básico do algoritmo SFFS, onde K é o tamanho do grupo com a solução naquele instante e d especifica a condição de parada do algoritmo, portanto, o tamanho da solução final.

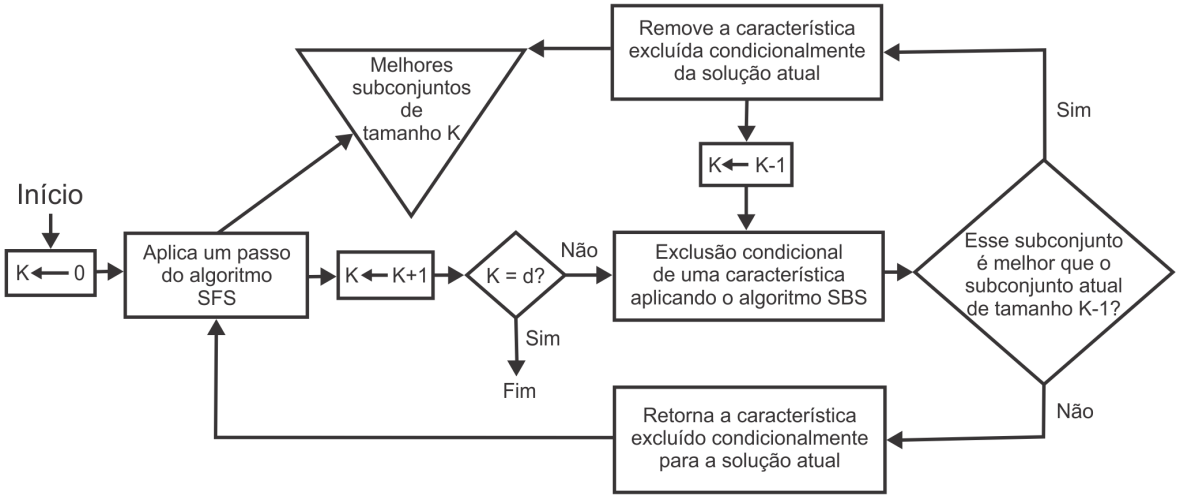

Figura 3 – Fluxograma simplificado do funcionamento do SFFS.

Fonte: (LOPES, 2011), originalmente adaptado de (SOMOL et al., 1999)

2.6 Entropia

Embora a ideia de entropia tenha sido inserida em 1865 na área de termodinâmica (CLAUSIUS, 1879), ela também é utilizada na área da Teoria da Informação, sendo que nesse segmento ela pode demonstrar a quantidade de informações presentes em um local, bem como classificar o nível de desordem de um agrupamento de informações (BISHOP, 1995; SHANNON, 1948).

Shannon e Weaver (1963) definem uma forma de entropia denominada Entropia de Shannon, ela exibe a probabilidade do acontecimento do evento P(x), sendo que X é uma variável aleatória que pode ter um valor do conjunto discreto. Dessa forma, segue a equação 2.4 onde uma média dos logaritmos das possíveis chances de ocorrências *x(log(P(x))* com ponderamento das suas probabilidades P(x) e assumindo que 0 x $log(0) = 0$:

$$H(X) = -\sum_{x \in X} P(x) \log P(x), \text{ sendo que: } \sum_{x \in X} P(x) = 1. \tag{2.6}$$

Dessa maneira, quanto mais elevada a entropia, mais elevado o nível de incerteza de prever o valor, portanto a entropia é a grandeza de incerteza de uma variante.

Partindo desse conceito, a entropia paralela entre duas variáveis diferentes (X, Y) define-se como:

$$H(X, Y) = -\sum_{x \in X, y \in Y} P(x, y) \log P(x, y), \tag{2.7}$$

sendo que *P(x,y)* corresponde a probabilidade conjunta das duas variáveis (LOPES, 2011).

2.7 Aprendizado de Máquina Automatizado (*AutoML*)

O aprendizado de máquina (*Machine Learning* em inglês) é um campo da inteligência artificial onde são treinados modelos que, através da experiência em seus acertos e erros passam a aprender regras que posteriormente podem ser utilizadas para tomar decisões e/ou realizar previsões sem a necessidade de uma programação explícita (MITCHELL, 1997; SAMUEL, 1959).

Um conceito que pode ser aplicado no contexto de aprendizado de máquina é o Aprendizado de Máquina Automatizado, normalmente citado pela sigla *AutoML*, proveniente do termo em inglês *Automated Machine Learning*. Esse, é aplicado porque desenvolver modelos de aprendizado de máquina pode ser um processo de grande complexidade, exigir um grande esforço e tempo, dessa forma, o *AutoML* automatiza o processo permitindo que o usuário não necessite ter grandes conhecimentos nos algoritmos e faz com que o tempo seja empregado em outros pontos de um projeto (GUYON et al., 2015; HU; HUANG, 2017; LI et al., 2019).

O *AutoML* executa, com base em técnicas de otimização, uma série de algoritmos de seleção de características em conjunto com algoritmos de classificação e regressão sobre os dados e avalia o resultado retornado de cada execução. Assim como os hyperparâmetros dos algoritmos de aprendizado são ajustados para maximizar o acerto nas classificações e diminuir os erros, o *AutoML* realiza testes com diferentes algoritmos para mostrar ao usuário qual seria a melhor configuração para aquele conjunto de dados sem a supervisão humana (GUYON et al., 2015; HU; HUANG, 2017; LI et al., 2019; THORNTON et al., 2012).

2.8 Coeficiente de Correlação de Postos de Spearman

Variáveis quantitativas são atributos de elementos que podem ser expressos por meio de valores numéricos sejam eles valores finitos (discretos) ou intervalos de valores (contínuos); já as variáveis qualitativas são as que não podem ser exibidas numericamente e geralmente seus valores são categorizados para exibirem qualidades dos elementos.

Ao analisar um conjunto de variáveis quantitativas é possível verificar qual a correlação entre elas, em estatística uma medida linear de grande utilização é o Coeficiente de Correlação de Postos de Spearman (também chamado de Coeficiente de Spearman) que é uma medida proposta por Charles Spearman (SPEARMAN, 1904) sendo a mais antiga estatística baseada em postos (*ranking*) (BAUER, 2007), essa medida analisa qual a intensidade em que pode ser descrita a relação entre uma hipotética variável X e uma

variável Y usando a função monótona, sendo que essa função mantém ou inverte sua relação de ordem, dessa forma, uma variável X pode ser diretamente proporcional a uma variável Y quando o valor de ambas crescem conjuntamente, ou ela pode ser inversamente proporcional sendo que nesse caso quando o valor da variável X cresce o valor da variável Y diminui, porém essa relação entre elas possui uma correlação que é exibida pelo coeficiente de Spearman.

2.9 Medidas de Similaridade entre Cadeias de Texto

Encontrar similaridade entre *Strings* é uma abordagem largamente requerida em diversos sistemas de informação, pois com ela é possível, por exemplo, detectar plágios em textos, pesquisar palavras em páginas da internet, comparar diversos textos, dentre outras várias aplicações. Em muitas linguagens de programação a comparação entre *Strings* não pode ser feita da mesma maneira trivial em que se comparam dois números por exemplo, portanto, são necessárias estratégias específicas para encontrar correspondência entre ambas as *Strings* (SAVIĆ; IVANOVIĆ; JAIN, 2019; ZHANG; HU; BIAN, 2017).

Uma estratégia comumente utilizada para comparar duas *Strings* é a "distância de edição (*edit distance*)" que é uma forma de quantificar o quão divergente são duas *Strings*, para isso, conta-se o mínimo de operações (inserções, exclusões ou substituições) necessárias para transformar uma "*String* α" em uma "*String* β". (JAMJUNTRA et al., 2017; KUMAR; VIBHA; VENUGOPAL, 2016; ZHANG; HU; BIAN, 2017). Vários algoritmos de medidas de distância podem ser aplicados na comparação entre *Strings* (SAVIĆ; IVANOVIĆ; JAIN, 2019), todavia, neste projeto foi utilizada a distância Levenshtein, pois é um algoritmo clássico (LEVENSHTEIN, 1966), bastante usado na literatura e o mesmo é apropriado para o uso neste trabalho, sendo que o foco deste trabalho não é comparar os diferentes algoritmos de medidas de distância.

Um termo que pode ser utilizado para indicar uma comparação entre 2 cadeias de texto é o termo em inglês "*match*", que pode referir-se ao quão correspondente é uma *String* de outra.

2.10 Programas de Pós Graduação

Os programas do Sistema Nacional de Pós-Graduação (SNPG) no Brasil, são divididos em 2 grupos diferentes, o *lato sensu*: que são programas de especialização que entregam um certificado de conclusão e não um diploma ao aluno concluinte, sendo que os cursos com a titulação *Master Business Administration* (MBA) fazem parte desse grupo de programas; e também os programas *stricto sensu*, são programas de doutorado acadêmico ou mestrado que pode adotar a modalidade acadêmica ou profissional. Os programas

stricto sensu normalmente fazem seleção dos seus candidatos através da divulgação de um edital, sendo que ao concluir essa modalidade de programa o aluno recebe um diploma com a respectiva titulação alcançada (EDUCAÇÃO, 2018).

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) declara que os cursos de pós-graduação vêm crescendo no pais, sendo que nos anos de 2013 a 2016, esse crescimento registrou a taxa de 25% no número de programas. No mesmo período os programas de Mestrado Profissional possuíram um aumento de 77% (CAPES, 2017a), portanto, com base nesses dados de crescimento dos programas, são necessárias ferramentas computacionais para avaliar suas evoluções e comportamentos (CAPES, 2017b).

2.11 Métricas Sobre Atuação Acadêmica

Pode-se entender como métricas acadêmicas os indicadores referentes ao ensino, à pesquisa e à extensão, essas métricas nem sempre são fáceis de serem obtidas, todavia, são de grande importância para a gestão acadêmica. Pode-se mensurar dados como: a quantidade de orientações (trabalhos de graduação, estágios curriculares obrigatórios, especializações, mestrados, doutorados); publicações; quantidade de projetos externos às IES; consultorias em empresas; bancas de apresentações e defesas de trabalhos.

No quesito pesquisa, já existem alguns indicadores mais formais pois, para classificar as revistas científicas existentes no Brasil, a CAPES, desenvolveu o Qualis, que é um índice qualitativo aplicado às produções científicas em periódicos e em conferências no caso da Ciência da Computação. O Qualis é dividido em estratos que categorizam as revistas científicas de acordo com critérios específicos da área de pesquisa na qual ela atua (OLIVEIRA et al., 2015). O Qualis pode ser utilizado como uma métrica acadêmica, podendo de acordo com a classificação (estrato) Qualis, avaliar um programa (BARATA, 2016).

Para avaliar os programas de pós-graduação no Brasil a CAPES utiliza sete níveis indo de 1 até o 7. Programas que recebem nota 1 ou 2 não são recomendados pela CAPES. A nota 3 é a mínima recomendada pela CAPES e, normalmente, são atribuídas a programas recentes que possuem apenas o mestrado. As notas 6 e 7 são atribuídas aos programas que possuem seu desempenho comparado aos melhores programas no mundo, com colaborações e inserção internacional. Para realizar essas avaliações a CAPES utiliza-se da análise de metodologia específica e os programas são examinados por especialistas, nessas análises a CAPES considera o intervalo de três anos (triênios) (LOPES et al., 2011), todavia, recentemente foi adotado o período avaliativo de quatro anos (quadrienal).

2.12 Autoria Acadêmica

Um trabalho acadêmico pode ter a colaboração de diversos coautores em seu desenvolvimento, sendo que geralmente cada autor ou coautor tem um papel no trabalho e com a contribuição de todos o resultado final é atingido (ALVARENGA, 2007; BORDIN; GONÇALVES; TODESCO, 2014; HILÁRIO; GRÁCIO; GUIMARÃES, 2018; PETROIANU, 2002).

Todos que tiveram uma contribuição em um trabalho, podem ser citados como autores e/ou coautores e a ordem em que são citados depende da área e da instituição pela qual o trabalho fora publicado, existem áreas onde a convenção exige que os autores sejam citados em ordem alfabética, porém, esse tipo de modelo de citação pode trazer discriminações para pesquisadores nos quais seus sobrenomes começam com letras presentes mais no final do alfabeto, bem como podem favorecer os pesquisadores com sobrenomes que iniciam-se com letras mais no começo do alfabeto. Outro ponto a ser analisado é que nos trabalhos com mais de três autores, a abreviação "*et al.*" é usada para representá-los, portanto, isso poderia prejudicar um autor apenas pelo seu sobrenome, ainda que o mesmo tenha desenvolvido um maior esforço no projeto.

A Comissão de Integridade de Pesquisa do CNPq publicou um relatório técnico (CHAVES et al., 2011) com 21 diretrizes para a boa conduta na produção acadêmica no Brasil, destaca-se então a diretriz 19 que define que todos os autores em um trabalho possuem responsabilidades sobre ele, todavia, ao primeiro autor a responsabilidade é integral, sendo que nos demais ela é parcial conforme suas colaborações. Com isso, pode-se observar que o CNPq também reconhece que a ordem de citação em um trabalho possui diferentes responsabilidades sobre o mesmo.

2.13 Plataforma Lattes

A Plataforma Lattes é composta por sistemas de informação e seu nome presta homenagem ao físico Césare Mansueto Giulio Lattes, ela é gerenciada pelo Instituto Stela, em parceria com o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), que é responsável por agrupar bases de dados com informações sobre perfis acadêmicos de pesquisadores brasileiros, grupos de pesquisas e sobre as instituições de ensino superior do país (GUEDES, 2001; PAPER; CATARINA, 2012). Os dados disponibilizados na plataforma são públicos e ela tem como principal elemento o Currículo Lattes (ALVES; YANASSE; SOMA, 2012).

Os dados são acessados através de páginas e portais Web, portanto, a interação desses dados com os de outros sistemas de informação pode não ser uma tarefa trivial.

O Currículo Lattes é atualmente a principal base de dados de currículos acadêmicos

do país (MENA-CHALCO et al., 2014). Tem o objetivo de centralizar e padronizar informações e registros de vida pregressa e atual da comunidade científica brasileira, onde arquiva, gerencia e disponibiliza atualmente dados de mais de 6 milhões de perfis de pesquisadores, discentes e docentes, que em sua maioria atuam no Brasil (ALVES; YANASSE; SOMA, 2011). Uma motivação para que essa base esteja sempre atualizada é que os editais das agências de fomento utilizam esses dados para disponibilizar recursos para pesquisadores e grupos de pesquisadores. Ainda, essa base também é utilizada para colher dados para o reconhecimento de cursos de graduação e a avaliação de programas de pós-graduação (ALVES et al., 2015; FERRAZ; QUONIAM; ALVARES, 2014).

2.14 Plataforma Sucupira

O Sucupira [1], é um sistema de informação criado para extrair informações do Currículo Lattes e apresentá-las aos usuários finais. O objetivo da criação do sistema é que ele seja a base de dados referência para o SNPG, para isso, os dados do mesmo devem ser disponibilizados em tempo real (*streaming*) informando todos os lançamentos da CAPES (CAPES, 2014).

Nessa plataforma são criados relatórios de pesquisadores relacionando-os de acordo com áreas de pesquisas, regiões e publicações. Através do estudo de (ALVES; YANASSE; SOMA, 2011) pode-se constatar a eficiência do sistema para a integração de dados do Currículo Lattes, todavia o sistema Sucupira não permite personalizações e caso seja necessário ter uma visão diferente das redes ele não pode ser utilizado.

2.15 Ferramentas Mais Relevantes Para Este Trabalho

Nessa seção são apresentados as ferramentas mais importantes para este projeto, as que possuem pouca utilização ou uma relevância secundária serão apresentadas no Capítulo de materiais, porém as dessa seção são as que foram significativas para os resultados deste projeto.

2.15.1 Suíte Pentaho

O Pentaho é um pacote de sistemas de informações pertencente a empresa Hitachi Vantara que é utilizado em inteligência de negócios (*Business Intelligence* ou BI). É uma solução que conta com 2 versões: uma gratuita, *open-source* (versão *Comunnity Edition* ou CE) e outra versão paga (versão *Enterprise Edition* ou EE), o que difere ambas as versões é que na versão paga, o usuário passa a ter suporte técnico diretamente da Pentaho, bem como ela conta também com algumas ferramentas exclusivas de análises de *Dashboards*.

[1] Pode ser acessado em: < https://sucupira.capes.gov.br/sucupira/ >. Último acesso em: 13 fev. 2018.

Desenvolvida com a linguagem de programação Java, a ferramenta é uma das mais usadas e com maior reputação dentre as soluções de BI existentes (MARINHEIRO; BERNARDINO, 2013).

O Pentaho Data Integration (PDI), também conhecido como Kettle, é o módulo de *Extraction, Transformation and Load* (ETL) da suíte Pentaho. Portanto ele é o responsável por toda a manipulação e estruturação dos dados que as outras aplicações do Pentaho irão utilizar (MUSSA et al., 2018).

Embora o PDI seja parte da suíte Pentaho, ele pode trabalhar sem a necessidade das outras aplicações (*stand-alone*). Dessa maneira, ele pode ser utilizado para realizar os processos de ETL de maneira descomplicada (MARINHEIRO; BERNARDINO, 2013).

O PDI trabalha com dois tipos de atividades, as transformações (*transformations*), que são uma série linear de operações aplicadas sobre os dados; e os trabalhos (*jobs*), que são os gerenciamentos das transformações. Os trabalhos controlam o fluxo de dados entre as transformações e os possíveis erros que possam acontecer nessas transformações. Tanto nas transformações quanto nos trabalhos, o PDI utiliza componentes visuais para separar e identificar cada estágio de uma atividade, dessa forma, cada componente tem sua característica e é responsável por uma etapa.

A Figura 4 ilustra um exemplo de transformação no PDI, onde é possível observar que cada etapa é encapsulada em um elemento denominado *step*, sendo que em cada um deles é realizada uma tarefa sobre os dados e o fluxo de dados segue a direção informada pelas setas.

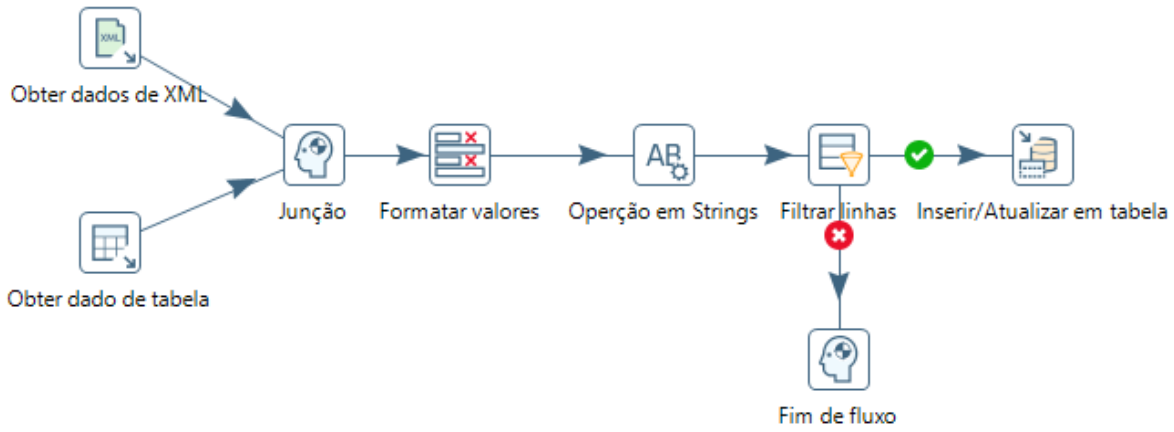

Figura 4 – Exemplo de uma Transformação no PDI.

Fonte: Autoria própria

2.15.2 O WEKA

O WEKA (acrônimo para *Waikato Environment Knowledge Analysis*) é uma ferramenta de código aberto, criada na Universidade de Waikato na Nova Zelândia sobre a linguagem de programação Java e com o intuito de reunir diferentes implementações de algoritmos de aprendizado de máquina, sendo que o WEKA também fornece soluções para o pré-processamento, além de soluções para a visualização de conjuntos de dados (FRANK; HALL; WITTEN, 2017; HALL et al., 2009; WITTEN; FRANK; HALL, 2011).

Ele foi criado porque os pesquisadores perceberam que existiam diferentes modelos de aprendizado disponíveis sendo que cada um era executado em uma linguagem de programação diferente, sobre plataformas diferentes e formato de dados de entrada e saída diferentes o que tornava suas implementações trabalhosas e dificultava o processo de comparação entre os modelos e algoritmos. Partindo dessa ideia eles desenvolveram uma "caixa de ferramentas" para o cientistas de dados, sendo que essa ferramenta é utilizada tanto para pesquisadores quanto para profissionais, pois nela foram implementados vários métodos para resolver todos os problemas padrões de mineração de dados, como: classificação, regressão, agrupamento, regras de associação e seleção de atributos, além dos algoritmos de pré-processamento e das ferramentas de visualizações conforme mencionado anteriormente (FRANK; HALL; WITTEN, 2017; HALL et al., 2009).

O WEKA unifica em um ambiente gráfico, todos os algoritmos acima citados e padroniza a forma como eles são aplicados aos conjuntos de dados, bem como a maneira como seus parâmetros são alterados. O WEKA facilita também a entrada de dados, sendo que essa, é realizada por meio de uma matriz relacional no formato padrão da ferramenta (ARFF) e essa pode ser convertida à partir de arquivos JSONs, CSVs ou ainda por consultas em banco de dados (WITTEN; FRANK; HALL, 2011).

Uma vantagem na utilização do WEKA é que podem ser realizado a limpeza e adequação de dados, aplicados vários modelos e realizado vários treinamentos, aprendizados e avaliações sem que o usuário digite uma linha de código sequer, pois com o ambiente gráfico esse processo é facilitado, todavia, caso o usuário queira trabalhar com o ambiente de comandos, também é possível o que torna a ferramenta mais robusta e diversificada. Usuários também podem implementar algoritmos que não existem na ferramenta e disponibilizá-los por meio de um portal (*Package Manager*) para que outros usuários possam utilizá-los.

O WEKA também conta com o Auto-WEKA que é uma ferramenta de AutoML, sendo que quando utilizada o processo de seleção do algoritmo de seleção de atributos, modelo de classificação, aprendizagem e seus hyper-parâmetros são configurados de maneira automatizada (KOTTHOFF; LEYTON-BROWN, 2016; THORNTON et al., 2013).

2.16 Trabalhos Relacionados

Newman publicou dois artigos nos quais foram analisadas as publicações entre os anos de 1995 e 1999 nas áreas de física, pesquisa biomédica e ciência da computação e criou redes sociais para analisar cada uma dessas áreas, ele analisou algumas métricas nessas redes como a média de pesquisadores por publicação, a média de publicações por pesquisadores, coeficiente de aglomeração, presença e tamanho do componente gigante, medidas de distância entre pesquisadores e ainda medidas de centralidade nas redes (NEWMAN, 2001a; NEWMAN, 2001b).

Ao analisar os dados gerados nestes trabalhos, Newman pôde observar que as distribuições das quantidades de autores por trabalhos e trabalhos por autores segue uma lei de potência matemática, ele observou também que todas as redes analisadas possuía um componente conectado com a maior parte dos elementos da rede (componente gigante), também observou que as estruturas de redes possuem diferenças entre si, outro ponto observado é que as redes analisadas possuem pequenas distâncias entre os pares de autores, o que as classificam como redes de mundo pequeno, ele também mostrou que, para a maioria dos autores, a maior parte dos caminhos entre eles e outros pesquisadores, passam por apenas um ou dois outros pesquisadores.

Em (LOPES et al., 2011) os autores desenvolvem um processo para avaliar os programas de pós-graduação no Brasil, para isso, eles analisam as redes de coautorias dos pesquisadores de Ciência da Computação e aplicam sobre elas 2 medidas de qualidade, para classificar os programas, propostas pelos autores, a primeira é a "eficiência social", que baseia-se na necessidade de comportamento colaborativo e também na não existência de indivíduos "socialmente ineficientes", a segunda é a análise de "maior autovalor", que infere qualidade baseando-se na necessidade de um grande número de bons pesquisadores e também na grande densidade de colaborações dentro do programa.

O escopo deste trabalho está apenas na análise interna dos programas, portanto, caso seja necessária a análise com base nas colaborações externas (entre programas) é necessária uma extensão deste projeto ou uma outra abordagem. O conjunto de dados utilizado neste trabalho foi obtido através da *Digital Bibliography & Library Project* (*DBLP*) e foram considerados 732 pesquisadores dos 27 programas brasileiros de pós-graduação em Ciência da Computação, isso pode ser um fator negativo, pois caso uma publicação não seja indexada por essa biblioteca a publicação ficará de fora dessa análise.

A conclusão deste trabalho foi que ao analisar as redes de coautoria pode-se concluir que nos programas principais os autores possuem uma grande quantidade de colaboração entre si. Foi realizada também uma comparação com a classificação CAPES nos programas e os índices propostos pelos pesquisadores são correlatos à analise da CAPES.

Em (MENA-CHALCO et al., 2014) os autores evidenciam a grande importância

em analisar redes de coautorias para verificar padrões de interações entre pesquisadores e grupos de pesquisadores, os autores também corroboram a alta relevância em utilizar-se a plataforma Lattes como fonte de extração de dados.

Nesse trabalho são extraídos os dados e geradas as redes de coautoria dos pesquisadores das oito grandes áreas de conhecimento que separam os programas brasileiros, sendo que o principal intuito é analisar qual a estrutura bem como a dinâmica das redes de coautoria entre o programa (intraprograma) e entre outros programas (interprograma) de todas as áreas de conhecimento, os autores usaram uma janela temporal de 3 anos, baseando-se no período de avaliação CAPES. Na fase de montagem do conjunto de dados os autores chamam a atenção a um caso em que foi observado um único pesquisador, com um grande nível de interação, que possuía 340.000 conexões com outros pesquisadores, essa informação poderia ser melhor detalhada, por exemplo dessas conexões, em quais publicações o pesquisador foi primeiro e segundo autor? Em quais publicações ele foi o último? Como esse não era o foco do trabalho os autores não responderam essas perguntas, mas poderia ser uma extensão deste trabalho.

Ao analisar as redes os autores concluíram que cada grande área possui uma estrutura de interação diferente confirmando assim uma hipótese levantada no início do trabalho. Foram extraídas 10 medidas topológicas de cada uma das redes e através dessas medidas pode-se observar que existem grandes diferenças entre as áreas e medidas como comprimento médio entre as redes variaram de entre 4.6031 a 8.1501, é possível observar também que a área de Ciências da Saúde possui muitos coautores e uma grande interação entre ela se comparado com outras áreas.

Os autores concluem também que a quantidade de coautoria cresce apressadamente no decorrer do tempo e o crescimento está correlato ao número de vértices e arestas presentes na rede. De acordo com os dados gerados, no geral as redes dobram de tamanho para cada triênio, os resultados também deslumbraram que algumas redes não possuem propriedades de mundo pequeno sendo presente apenas nas áreas de linguística, letras e artes, pois as mesmas alta transitividade e um número meno de comprimento médio entre os caminhos, porém essa caracterização só é observada nos primeiros períodos de avaliação, pois nos últimos a topologia das redes não atendem esse padrão. Foi observado também que as áreas de ciências biológicas e humanas possuem um comportamento estável, diferente das áreas de ciências agrárias, ciências humanas, ciências da saúde, linguística, letras e arte que possuem um comportamento oscilante, que pode indicar que essas áreas possuem dependências de outras áreas do conhecimento.

Este trabalho foi realizado com dados coletados em maio de 2011, portanto, pode ser que as informações desse trabalho descrito em (MENA-CHALCO et al., 2014) podem ter diferenças das informações levantadas para este projeto.

Em (BORDIN; GONÇALVES; TODESCO, 2014) os autores analisaram a produção

bibliográfica por meio das redes de coautoria do programa de Engenharia e Gestão do Conhecimento da Universidade Federal de Santa Catarina entre os anos de 2005 e 2012, para isso, eles extraíram os dados através dos relatórios de produção bibliográfica da CAPES, analisando artigos publicados em livros e revistas e trabalhos em eventos, a correção dos nomes dos autores e abreviações foi realizada manualmente o que gerou uma alta demanda de tempo. Como resultado dessa publicação os autores exibiram alguns resultados interessantes sobre o programa analisado, eles analisaram o número de publicações, quais delas são de quais tipos, qual o percentual de produções individuais e a densidade da rede.

Eles utilizaram também uma métrica que é o número médio de autores por produção bibliográfica onde eles dividiram o total de pesquisadores (vértices) de um programa pelo total de publicações (arestas), sendo que quanto mais produtivo um programa, maior será esse indicador, sendo que os autores concluíram que esse índice manteve estável no decorrer do período analisado.

Algumas medidas de redes complexas foram também analisadas tais como: número de vértices, componentes, componentes gigante, distância média, centralidade de grau, centralidade de intermediação e centralidade de proximidade. Com essas medidas os autores puderam classificar e exibir os autores com maior centralidade de grau, portanto, os autores mais influentes em cada programa. Com os resultados do trabalho eles concluíram também que no primeiro ano do programa o percentual de publicações individuais foi o mais elevado, pois o programa era novo e não havia grande quantidade de trabalhos com coautoria.

Apesar de analisar as medidas os autores não fizeram nenhum tipo de análise qualitativa das publicações e não fizeram comparação com a avaliação da CAPES ou com qualquer outra avaliação.

Em (BORDONS et al., 2015) os autores analisaram a estrutura das redes de coautoria das áreas de Nanociência, Farmacologia e Estatística no período de 2006 à 2008 na Espanha, a performance individual dos pesquisadores foi observada usando como base o Índice-G e a fonte de dados foram artigos acadêmicos das respectivas áreas. Eles observaram que a estrutura da rede de Farmacologia e Nanociência são semelhantes e mais densas do que a rede de Estatística que por sua vez é mais fragmentada e possui um número menor de conexões, sendo assim os grupos de pesquisas de Estatísticas são menores ou seus pesquisadores trabalham mais individualmente do que nas outras duas áreas. Com o resultado dessa pesquisa os autores conseguiram confirmar que os autores com maior quantidade de colaborações (alta distribuição de graus) ou com ligações mais fortes com seus coautores são mais propensos a possuir um maior Índice-G. Os resultados apontaram também para o fato de que o coeficiente de aglomeração apresenta associação negativa com o Índice-G.

Em (DIGIAMPIETRI et al., 2014) os autores apresentam diferentes formas de

visualizar o desempenho dos programas, classificá-los de acordo com cada perspectiva levantada e também combiná-los, ele também mostra a correlação entre as diferentes métricas, além de discutir como os programas interagem. Os autores avaliam os programas de duas formas diferentes: a primeira sobre a produtividade de um programa baseado em índices bibliométricos (SJR, fator de impacto JCR e Qualis), mostrando o número de citações, Índice-G e Índice-H das máquinas de busca *Microsoft Academic Search* e *Google Scholar*; a segunda forma de avaliação leva em conta as redes sociais acadêmicas, onde são criados grafos em que os vértices são os programas e as arestas a interação entre esses programas, nessa avaliação o período considerado está entre 2004 e 2009 e os dados são extraídos do Currículo Lattes. Neste trabalho é possível verificar através das medidas de produtividade a evolução dos programas, observa-se também que os *rankings* de produtividade podem ter grande variação dependendo da medida utilizada. Com os resultados do trabalho, os autores concluem que os programas com os melhores índices topológicos são os mais produtivos.

Os estudos anteriormente citados são de grande relevância para o contexto deste trabalho, todavia, nenhum deles realiza a mesma análise deste trabalho, pois alguns analisam bases de dados diferentes (BORDIN; GONÇALVES; TODESCO, 2014; LOPES et al., 2011; NEWMAN, 2001a; NEWMAN, 2001b), outros analisam a mesma base de dados, porém a análise é feita por visões distintas (DIGIAMPIETRI et al., 2014) e até em outros períodos (MENA-CHALCO et al., 2014), sendo que os resultados alcançados diferem-se dos propostos neste trabalho.

3 CONJUNTO DE DADOS

Nessa seção são apresentados os conjuntos de dados utilizados e os algoritmos aplicados sobre eles.

3.1 Gerenciamento dos Conjuntos de Dados

A aquisição, processamento e o armazenamento dos dados foram divididos em 3 sub-etapas cujos fluxos são demonstrados no diagrama presente na Figura 5, onde cada linha representa cada uma das 3 sub-etapas, cada coluna representa qual a fonte de dados utilizada, qual o processo realizado e qual o artefato gerado como resultado desse processo, por fim, todo o processamento é arquivado no banco de dados. Nessa Figura, as nuvens identificam que os dados são acessados através da *internet*, os processos estão descritos nas caixas, os arquivos são representados por pequenas páginas e o banco de dados representado por uma pilha de círculos.

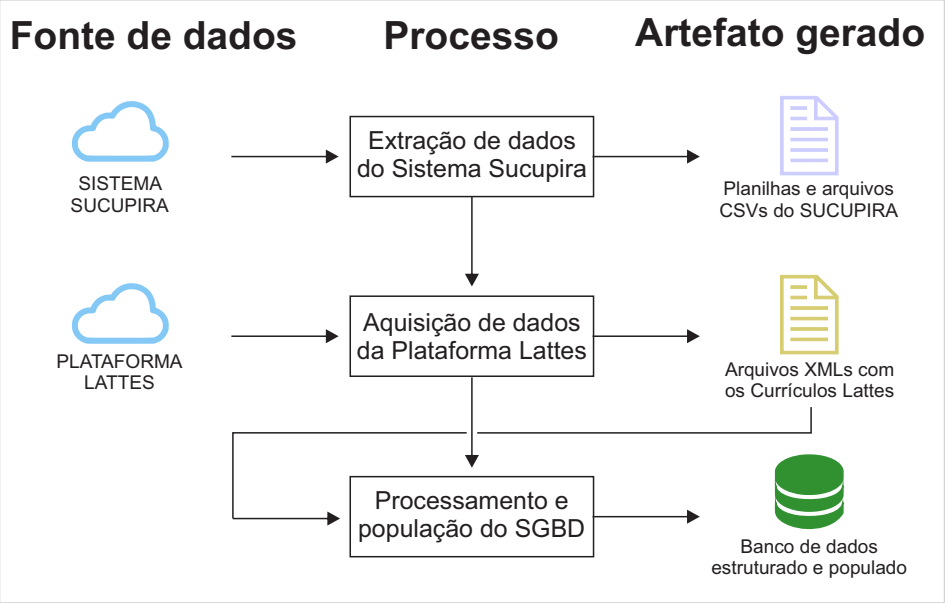

Figura 5 – Fluxo geral do processamento dos dados.

Fonte: Autoria própria

3.2 Etapa 1 – Aquisição de dados

A Figura 6 exibe um diagrama com o fluxo do processo de aquisição de dados do Portal CAPES (sistema Sucupira) e da Plataforma Lattes, na coluna à esquerda são dispostas as fontes de dados utilizadas; na coluna central, são exibidos os processos

realizados em cada sub etapa; já na coluna à direita é onde são exibidos cada um dos artefatos gerado por cada processo dessa etapa, sendo que o artefato final é um diretório com todos os Currículos Lattes anteriormente filtrados e baixados em formato XML.

As fontes de dados descritas nessa etapa do projeto possuem um código de classificação e estão catalogadas no Apêndice B deste trabalho.

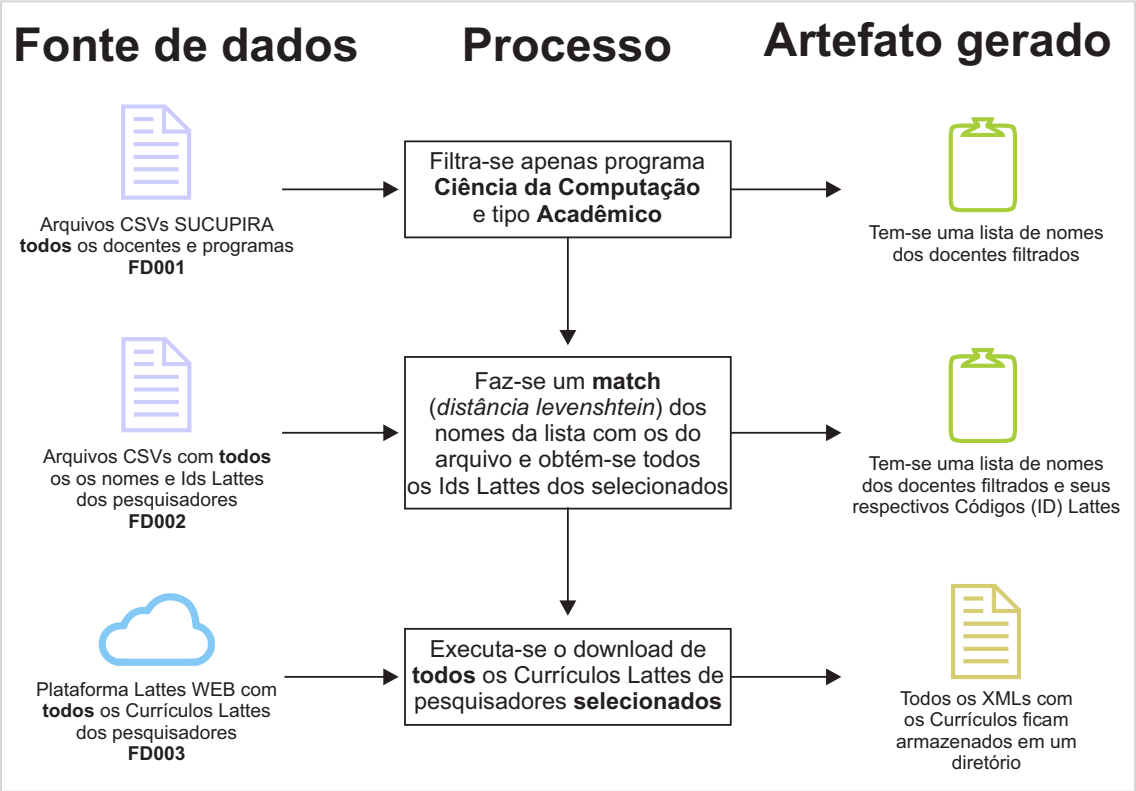

Figura 6 – Fluxo da etapa de aquisição de dados do Portal CAPES.

Fonte: Autoria própria

A fonte de dados inicial, foi obtida por meio de um arquivo de dados CSV com 89.255 registros, extraído do sistema de informação Sucupira. Nessa fonte estão presentes vários dados de extrema importância ao projeto, como por exemplo: o nome dos docentes, quais os programas em que eles estão vinculados, quais as instituições as quais eles pertencem, dentre outros dados.

Com os dados dessa fonte foi possível realizar os primeiros filtros, que foram os programas, para aparecer somente pesquisadores vinculados à Ciência da Computação e também o tipo de programa, para exibir somente os programas acadêmicos.

No pseudocódigo Algoritmo 1 é exibido o algoritmo utilizado para realizar os filtros nos dados.

A Tabela 1 demonstra os filtros e os dados resultantes, sendo que por meio dela

Algoritmo 1: Filtra registros

```
Input: Objeto x com todos os registros da planilha.
Output: Apenas os registros filtrados.
if x.programa é "CIÊNCIA DA COMPUTAÇÃO" then
    if x.tipo é "ACADÊMICO" then
        seleciona o registro;
    end
end
```

é possível observar que dos 89.255 registros iniciais o número foi reduzido para apenas 1.644, ou seja, mais de 98% de redução com os filtros aplicados.

Tabela 1 – Filtros aplicados sobre a planilha inicial.

Campo	**Valor**	**Registros resultantes**
PROGRAMA	CIÊNCIA DA COMPUTAÇÃO	1.875
TIPO	ACADÊMICO	1.644

Conforme mencionado anteriormente, essa fonte de dados colabora com diversas informações de extrema relevância, todavia, quando deseja-se outros dados não presentes nela, deve ser utilizada uma abordagem diferente. No caso deste projeto, é de suma importância ter-se os Currículos Lattes dos pesquisadores, e para obtê-los é necessário o Código Lattes, que é uma informação numérica e única (ID) que identifica cada pesquisador cadastrado na Plataforma Lattes.

Para obter o código Lattes dos pesquisadores desejados, utilizou-se um arquivo de dados CSV onde são informados 4.555.770 registros contendo o nome completo do pesquisador e seu respectivo Código Lattes, os detalhes de como foi gerado esse arquivo podem ser encontrados em (MENA-CHALCO et al., 2014). Os 1.644 dados extraídos do CSV do CAPES foram então cruzados com todos os registros do arquivo CSV da Plataforma Lattes identificando 1.808 registros.

Um dos desafios de construir uma rede de coautoria está na etapa de extração dos dados de uma fonte, pois comumente usa-se o nome de um autor para identificá-lo, todavia, os dados podem conter falhas como por exemplo nomes estarem escritos com caracteres diferentes, sem acentos, ou ainda podem haver a presença de homônimos em alguns casos, sendo que nesse, dois pesquisadores diferentes podem ser identificados como o mesmo apenas por conta do nome em comum.

No contexto deste projeto, para realizar o cruzamento dos dados de ambas as fontes, foi utilizado o campo com o nome do pesquisador, comparando por meio da medida de distância de Levenshtein, usando um limiar de proximidade onde apenas *Strings* com o valor de operações i, onde: $1 <= i$, assim foi possível identificar nomes muito próximos, que

se porventura tivessem um caractere equivocadamente cadastrado, ainda sim fosse possível a identificação. As *Strings* foram todas normalizadas antes de realizar a comparação, eliminando assim todos os acentos e caracteres especiais que poderiam afetar no resultado do algoritmo de Levenshtein.

Observa-se também que o processo resultante gerou um número maior (1.808 registros) que em uma das fontes (1.644 registros), que não deveria ocorrer como resultado de um cruzamento de dados. Esse número maior foi gerado porque foram encontrados 310 homônimos, ou seja, diferentes pesquisadores com o mesmo nome ou com cadastros duplicados de um mesmo pesquisador dentro da Plataforma Lattes. Para esses dados identificados como homônimos foi observado a área de atuação do pesquisador, eliminando assim falsos vínculos. A Tabela 2 exibe um resumo da quantidade de dados de cada uma das fontes.

Tabela 2 – Quantidade de dados de cada fonte.

Descrição da fonte de dados	**Quantidade de registros (tuplas)**
CSV com todos os códigos Lattes	4.555.770
CSV do SUCUPIRA com os dados do portal CAPES	1.644
Quantidade de registros encontrados	1.808
Quantidade de dados com nomes homônimos	310

O processo de cruzamento de dados foi realizado utilizando-se a ferramenta Pentaho Data Integration (PDI), pois com os procedimentos disponíveis nele foi possível realizar o processo sem a necessidade de escrever grandes quantidades de linha de código, otimizando assim essa etapa.

3.3 Etapa 2 – Processamento e População

Após a aquisição dos dados do Currículos Lattes, inicia-se o processo de população dos dados no sistema gerenciador de banco de dados (SGBD), para isso, foi criado um banco de dados utilizando o SGBD Postrgres, o Diagrama de Entidade e Relacionamento (DER) pode ser visualizado no Apêndice A deste documento.

A primeira sub-etapa dessa fase é a inserção dos dados na tabela PESQUISADORES do SGBD, para isso, todos os XMLs são lidos e são selecionados os dados pessoais e de endereço dos pesquisadores. É realizado um processo de comparação do nome completo do pesquisador e seu código Lattes, caso esse pesquisador já possua algum registro nessa tabela, seus dados são apenas atualizados, caso contrário, é inserido um novo registro.

A segunda sub-etapa é responsável por popular a tabela EXTRATO_QUALIS do SGBD, nessa tabela, são listadas todas as conferências, um código único que as identificam

International Standard Serial Number (ISSN), seu título e seu estrato Qualis. A fonte de dados dessa etapa, é um documento de planilha eletrônica, obtido no portal da (CAPES, 2017c) com os dados de todas as classificações das publicações realizadas na área de ciência da computação atualizadas até o ano de 2016. Esses dados servirão para avaliações futuras de programas.

Na terceira sub-etapa as publicações que não são artigos serão inseridas no SGBD, porém, nessa etapa serão filtrados os dados para que somente sejam inseridas as produções bibliográficas que sejam trabalhos em eventos e que a sua natureza seja "completo", conforme demonstrado no pseudocódigo Algoritmo 2.

Algoritmo 2: Filtra publicações

Input: Objeto p com todos as publicações da planilha.
Output: Apenas as publicações completas.
if *p.natureza é "COMPLETO"* **then**
  seleciona o registro;
end

Nessa etapa não há qualquer tipo de vínculo da publicação com os pesquisadores, é apenas realizada a inserção delas na tabela PUBLICACOES, todo o processo de relacionamento entre pesquisador/publicação será feito adiante. Como a tabela de publicações é genérica para qualquer tipo de publicação é inserido então uma *Flag* numérica na coluna "tipo", com o valor 2 para identificar posteriormente que todos os registros na tabela com o valor 2 de tipo, trata-se de uma publicação completa em um evento.

A sub-etapa seguinte é onde insere-se os dados referentes a publicações de artigos, esses dados são igualmente inseridos na tabela de PUBLICACOES o que irá diferir dos outros dados é que nesses, adiciona-se o valor numérico 1 como *flag* na coluna "tipo" do SGBD, possibilitando assim a aplicação de filtros posteriormente. Os dados nessa fase são filtrados para que sejam salvos apenas os dados de publicações de artigos com a natureza "completo", o valor do ISSN por diversas vezes não é armazenado com uma formatação padronizada nos XMLs do Lattes, dessa forma, nessa transformação é realizado um pré-processamento responsável por deixar esse valor dentro de um padrão antes de inseri-lo no banco de dados do projeto. Nessa etapa também é feito o primeiro relacionamento, onde o ISSN é utilizado para relacionar os dados da tabela de PUBLICACOES com os dados da tabela EXTRATO_QUALIS, dessa forma, é possível ver a classificação Qualis de cada uma das publicações de artigos.

Nessa sexta sub-etapa é onde realiza-se um cruzamento de informações, no qual, através dos XMLs com informações do Lattes verifica-se quais são os pesquisadores de cada uma das publicações e cria-se uma ligação entre eles. Para isso, compara-se o título de cada trabalho do XML com todos os títulos de trabalhos armazenados no banco de dados do projeto, essa comparação é realizada usando o algoritmo de distância de Levenhstein e

os títulos com um limiar de proximidade onde apenas *Strings* com o valor de operações *i*, onde: *2 <= i*, são escolhidos, compara-se então o nome completo do pesquisador com o nome vinculado a publicação, o algoritmo e suas configurações seguem o mesmo do exemplo anterior.

Ao final dessa etapa tem-se o relacionamento entre as publicações e os pesquisadores, portanto, essa é uma etapa primordial ao projeto, porém, como trata-se de um processo de comparação entre *strings*, é exigido um grande tempo computacional para sua conclusão, portanto, para um outro conjunto maior de dados talvez seja necessária uma outra abordagem baseado em soluções de *Big Data* e/ou *e-Science*.

A sub-etapa de número sete, é muito próxima a etapa anterior (seis), diferindo apenas que ao invés de comparar os dados de publicações, nessa usa-se os dados de publicações de artigos. Todavia, os algoritmos, as configurações e a estrutura é mesma da sub-etapa seis.

Na sub-etapa oito a fonte de dados utilizada é o arquivo CSV do portal da CAPES outrora utilizado na Etapa 1, dessa fonte, são extraídas todas as instituições de ensino superior e armazenadas na tabela INSTITUICAO_ENSINO_SUPERIOR do banco de dados do projeto.

A sub-etapa nove é altamente semelhante à anterior (oito), sendo que a única diferença entre elas é que nessa os dados são as áreas de avaliação ao invés das IES, e a tabela de destino é a AREA_AVALIACAO.

No processo com a sub-etapa dez, os programas de pós-graduação são carregados de duas fontes de dados, sendo uma o arquivo CSV utilizado na Etapa 1 e outra uma planilha eletrônica publicada no portal da CAPES com os programas nacionais de pós-graduação.

Os dados de ambas as fontes são pré-processados para que não haja repetição entre as instituições, são realizados então dois cruzamentos de informações, no primeiro, é observado o a área de avaliação do programa e vinculado ao registro correspondente na tabela AREA_AVALIACAO; já no segundo é comparada a instituição de ensino com as cadastradas na tabela INSTITUICAO_ENSINO_SUPERIOR e vinculado o relacionamento, para executar ambas as comparações foi usado as mesmas parametrizações e o mesmo algoritmo de comparação entre *strings* das sub-etapas anteriores.

Na sub-etapa 11 do projeto os docentes são vinculados aos programas, para isso, usa-se a fonte de dados da Etapa 1 juntamente com uma planilha eletrônica coletada do portal da CAPES onde todos os docentes e seus vínculos com os programas são listados. Essa é a fase que exige um maior tempo computacional no processamento, pois trata-se de uma série de comparações de distância entre *Strings*.

3.3.1 Tratamento de duplicidades

Uma etapa importante foi o processo de limpeza dos dados e um problema encontrado estava na inserção dos registros de publicações, pois como trata-se de uma informação inserida manualmente na plataforma Lattes pelos pesquisadores, algumas inconsistências foram encontradas tais como: publicações com mesmo título mas com datas (anos) diferentes, publicações com o mesmo título mas em eventos diferentes. Ao observar os dados encontramos um padrão bastante comum no que diz respeito às publicações, observou-se que geralmente as datas diferenciavam-se em cerca de 1 ano, observou-se também que os eventos diferentes na verdade eram o mesmo evento porém eram escritos de maneira diferentes como por exemplo em um pesquisador A o evento figurado era escrito "Seminário Brasileiro de Redes Complexas", já o pesquisador B que era coautor no mesmo trabalho descrevia "2019 Seminário Brasileiro de Redes Complexas (SBRC)", como a comparação é feita por *String*, ajustou-se o algoritmo para que ele identificasse as 2 Strings desse exemplo como a mesma *String*, por conseguinte, o mesmo evento. Esses ajustes aumentaram a acurácia dos dados e preveniu para que não houvesse nenhuma publicação duplicada na tabela, pois esses publicações duplicadas poderiam gerar análises espúrias em processos subjacentes.

4 PROCEDIMENTOS METODOLÓGICOS

Nesta seção são apresentados os procedimentos utilizados neste projeto.

4.1 Processo

A Figura 7 exibe o fluxo de trabalho (*workflow*) geral do projeto sendo que na coluna mais à esquerda são exibidas as fontes de dados (entradas) das quais cada processo irá alimentar-se, na coluna ao centro é exibido cada um dos processos e o resultado ao final de cada um deles (saídas) é demonstrado na coluna mais à direita, Artefato gerado. Os processos são dependentes entre si, pois pode-se observar que o artefato gerado como resultado em um processo serve como fonte de dados em outro processo, portanto, os mesmos devem seguir a ordem conforme demonstrada nesse fluxo.

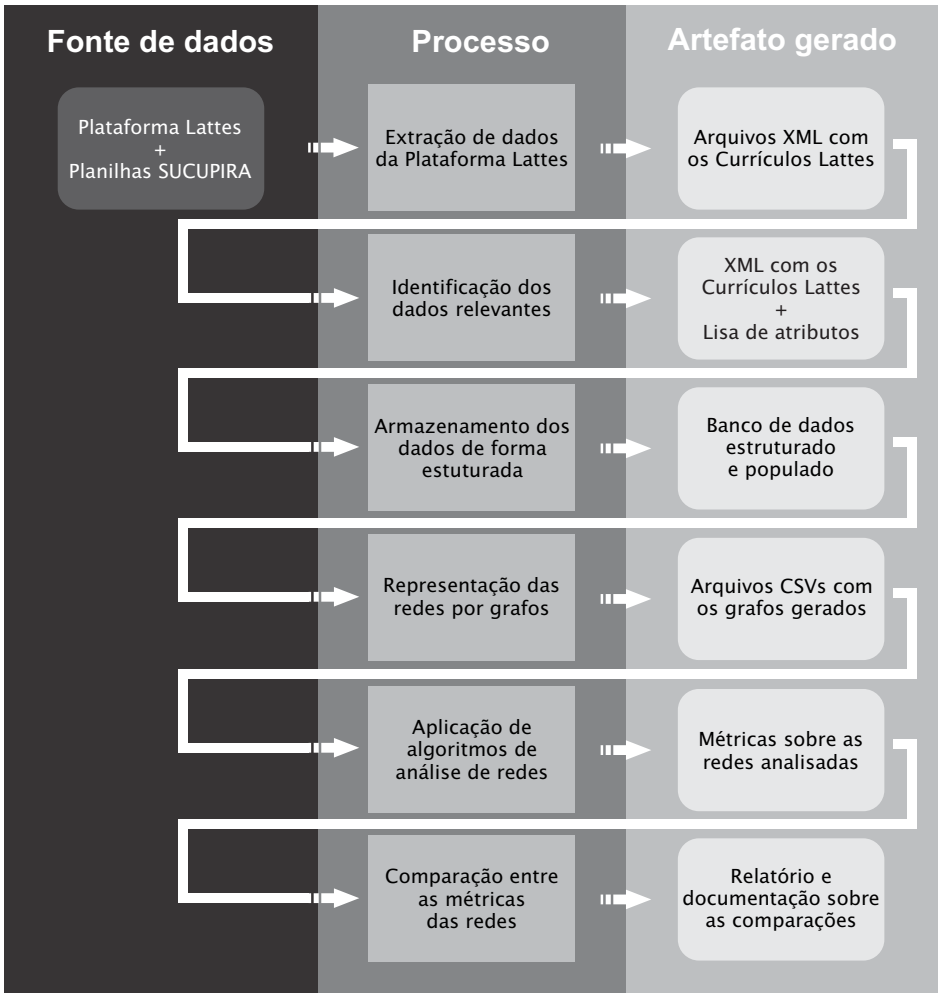

Figura 7 – Fluxo de trabalho geral do projeto.

Fonte: Autoria própria

4.2 Prospecção de Dados Acadêmicos

Os dados acadêmicos são disponibilizados por meio do portal brasileiro de dados abertos[1]. Cada um desses dados pode fornecer diversas fontes de informações que, quando cruzadas com outras fontes, entregam uma série de conhecimento.

Para este projeto, tem-se a necessidade de capturar esses dados e estruturá-los de uma maneira em que seja facilitada sua posterior utilização. Para isso os dados devem ser baixados por meio dos portais WEB e armazenados em uma estrutura adequada. Uma estrutura comumente utilizada e que pode ser condizente com essa demanda é a de um SGBD.

4.3 Identificação de Dados Relevantes ao Projeto

Para a identificação dos dados relevantes ao projeto analisa-se cada coluna de dado proveniente da planilhas dos portais WEB. Dados redundantes são ignorados bem como novos dados podem ser criados cruzando dois dados diferentes.

4.4 Agrupamento dos programas

Os programas neste projeto são agrupados de acordo com sua nota de avaliação CAPES, sendo que os programas com notas inferior a 3 são descartados desse projeto, haja visto que a CAPES não recomenda manter essa nota. Para a maior parte das análises consideramos 5 classes de programas, que vão de programas com a Nota CAPES 3 até os programas com a Nota CAPES 7, todavia, em determinados momentos agrupamos os programas em apenas 3 classes sendo: os programas de nota baixa, que são os programas com Notas CAPES 3 e 4 onde os denominamos como “Programas C”; os programas de nota intermediária que são os programas de nota 5, que chamamos “Programas B” e por último os programas de notas altas que são os programas de Nota CAPES 6 e 7 que chamamos de “Programas A”. Uma quantidade menor de classes mostrou-se mais assertiva em alguns algoritmos.

4.5 Índices de produções

Neste projeto define-se indicadores para estabelecer a ordem de importância dos autores em publicações, estes serão estabelecidos levando em consideração a quantidade de publicações de um determinado pesquisador e em qual posição o nome desse pesquisador encontra-se inserido nessas publicações. Para isso assume-se o cenário onde o primeiro nome informado em uma publicação, pertence ao autor principal do trabalho, portanto,

[1] Pode ser acessado em http://dados.gov.br/

aquele que empregou um maior esforço ao mesmo, definiu os materiais, métodos e objetivos dele e realizou a análise final dos resultados; já o último nome informado é do pesquisador responsável pela orientação da pesquisa e/ou pelo projeto, o pesquisador nessa posição possui maior senioridade; os pesquisadores intermediários são aqueles que contribuíram para o trabalho de diferentes maneiras, porém não se encaixam nas outras classificações.

A ordem em que esses pesquisadores são citados é informada no Currículo Lattes e ao extrair as informações do mesmo tem-se o campo "ORDEM-DE-AUTORIA" onde cada autor e coautor é citado e classificado no trabalho. Abaixo são descritos especificamente cada um dos índices e como é realizado o cálculo para sua obtenção, esses índices são gerados considerando os 3 períodos de avaliação CAPES, portanto, cada período possuí seu respectivo índice.

A motivação da criação desses índices é que nos trabalhos observados, não encontrou-se um indicador que qualificasse as colaborações em forma de coautoria em publicações, embora alguns estudos realizem a análise da quantidade de publicações em coautorias, não se leva em consideração que a ordem da citação pode revelar padrões importantes a serem analisados e detalhar a coautoria com base na forma como foi realizada a colaboração, e com base nesses indicadores traçar perfis inerentes aos pesquisadores e/ou programas.

4.5.1 Índice de Primeiro Autor

Conforme citado anteriormente, esse indicador exibe o percentual de publicações onde o autor é o primeiro nome a ser citado.

O Índice de Primeiro Autor de um pesquisador F_p é definido através do resultado da razão entre o total de publicações onde seu nome é o primeiro, por todas as suas publicações, conforme a equação abaixo.

$$F_{\mathrm{p}} = \frac{\text{número de publicações como primeiro autor}}{\text{número total de publicações}} \tag{4.1}$$

4.5.2 Índice de Colaboração

O Índice de Colaboração de um pesquisador I_p exibe a quantidade de vezes onde um pesquisador é citado fora das extremidades (primeiro ou último) em um trabalho, é definido como a quantidade de vezes onde ele é como intermediário, dividido pela quantidade total das publicações do mesmo, conforme a equação abaixo.

$$I_{\mathrm{p}} = \frac{\text{número de publicações como autor intermediário}}{\text{número total de publicações}} \tag{4.2}$$

4.5.3 Índice de senioridade

O Índice de Senioridade S_p é definido através do resultado da razão entre o total de publicações citado como último autor por todas as publicações de um pesquisador, conforme a equação abaixo.

$$S_{\mathrm{p}} = \frac{\text{número de publicações como último autor}}{\text{número total de publicações}} \tag{4.3}$$

Parte-se da hipótese que programas mais bem avaliados, com notas CAPES entre 6 e 7 são os programas onde grande parte de seus pesquisadores possuem maior Índice de senioridade, dessa forma, analisando os dados neste projeto, essa hipótese será respondida.

4.6 Média de Pesquisadores por Publicação

Grossman e Arbor (1995) observaram que o número médio de autores para cada artigo na matemática sofrera um aumento naqueles últimos 60 anos, onde em média era 1 pesquisador para cada artigo, foi para 1,5. Newman (2001b), realizou estudos sobre a quantidade média de autores por artigo e também a quantidade média de artigos para cada autor entre os anos de 1995 e 1999, eles encontraram alguns padrões nas redes analisadas como por exemplo, um padrão de lei de potência nos dados bem como a presença de componentes gigantes nessas redes. Todavia, embora essas análises foram de grande relevância, elas foram realizadas com base nos pesquisadores e não fora realizada uma análise dos programas, partindo desse cenário propomos a Média de Pesquisadores por Publicação (MPP) é uma medida que embora seja bem simples, sua interpretação pode ser de grande utilidade, ela é o resultado da divisão do número de vértices pelo número de arestas, com ela pode-se avaliar se um programa é em média mais eficiente que outro, partindo do conceito que programa mais eficiente é aquele que possui menor quantidade de pesquisadores (vértices) e maior quantidade de publicações (arestas), nessa abordagem quanto mais próximo de 0, mais eficiente é o programa. Abaixo é definida a fórmula para obtenção da MPP, deve-se observar que é uma medida de média, portanto em alguns momentos esse valores pode ser decimais.

$$MPP = \frac{\text{número total de pesquisadores (vértices)}}{\text{número total de publicações (arestas)}} \tag{4.4}$$

Essa medida é quantitativa, portanto, para sua análise deve-se levar em consideração que o intuito dela é apresentar a quantidade de pesquisadores de um programa e se a quantidade de publicações acompanham esse valor, com ela, pode-se avaliar se os programas possuem muito pesquisadores, mas pouca publicação, ou o inverso. O cálculo

dessa medida como acima definido, não contabiliza a média individual de colaboradores de cada publicação e exibe um resumo, pois se assim fosse feita, ela poderia ser uma medida qualitativa que avalia quantos autores em média uma publicação possui, porém esse não é o objetivo de análise deste projeto.

4.7 Organização do Grafo de Colaboração Acadêmica

Sendo um grafo **G(V,E)** onde **V** são os vértices e **E** o conjunto de arestas, para este projeto define-se como um grafo de colaboração acadêmica sendo os vértices os pesquisadores, e as arestas as publicações que eles possuem em colaboração.

A Figura 8 exibe esses grafos, através dela é possível observar que o pesquisador A e o Pesquisador C possui poucos projetos de pesquisa em colaboração, porém, o Pesquisador E e o Pesquisador F possui uma grande quantidade de projetos de pesquisa em colaboração.

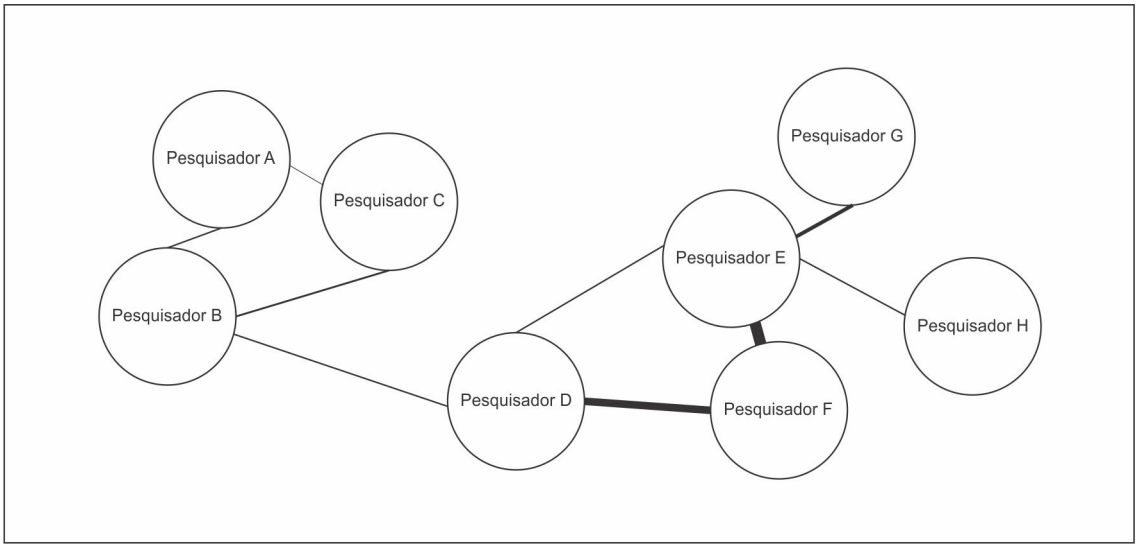

Figura 8 – Representação dos grafos com os pesquisadores.

Fonte: Autoria própria

4.8 Definição das Matrizes de Características

Após a geração das redes, são realizadas consultas utilizando a Linguagem de Consulta Estruturada (*Structured Query Language* ou SQL) no SGBD extraindo as características topológicas dos grafos, elas irão compor vetores de características $\overrightarrow{V}_c$, sendo que cada rede formará um vetor, portanto para os grafos: G_1, G_2, G_3, ..., G_n, teremos os vetores: $\overrightarrow{V}_1$, $\overrightarrow{V}_2$, $\overrightarrow{V}_3$, ..., $\overrightarrow{V}_n$. O vetor de características é definido por:

$$V_{\mathrm{c}} = \begin{bmatrix} C_1 \\ C_2 \\ C_3 \\ ... \\ C_{\mathrm{n}} \end{bmatrix}$$

onde, *C(n)* é cada uma das características topológica desse grafo. Após a geração dos vetores de características os mesmos serão agrupados em uma matriz de características $\mathrm{M_c}$, sendo que nessa matriz, cada linha será o dado de um vetor e suas colunas serão cada uma das características e por último uma coluna com a nota CAPES (NC) daquele grupo, portanto define-se a matriz como:

$$M_{\mathrm{c}} = (\sum \overrightarrow{V}_{\mathrm{c}}) * (\sum C + 1) \tag{4.5}$$

dando origem a:

$$M_{\mathrm{c}} = \begin{bmatrix} C_1 & C_2 & C_3 & ... & C_{\mathrm{n}} & NC \\ C_1 & C_2 & C_3 & ... & C_{\mathrm{n}} & NC \\ C_1 & C_2 & C_3 & ... & C_{\mathrm{n}} & NC \end{bmatrix}$$

4.9 Classificação e Seleção de Características

Com o intuito de avaliar a eficácia dos indicadores usados essa é uma etapa de elevada importância dentro deste projeto, para realizá-la utiliza-se a ferramenta WEKA, é então testados os diferentes modelos de seleção de atributos bem como os de classificação supervisionada.

Após a geração das matrizes de características a fase a ser aplicada é a de seleção de características, para esse fim irá ser utilizado um sistema que é o resultado do trabalho descrito em (LOPES; MARTINS; CESAR, 2008), onde os autores criam uma *interface* gráfica que pode ser executada em diversas plataformas e sistema operacionais, pois ela é criada sobre a tecnologia Java e seu código é aberto (*open source*).

O sistema DimReduction[2] desenvolvido pelos autores foi criado para abordagens de problemas de bioinformática, todavia, ele pode ser estendido como é o caso deste projeto. Esse sistema conta com alguns algoritmos de seleção de características, funções critérios e ferramentas de visualizações gráficas tais como gráficos de dispersão e coordenadas paralelas. Muito embora o sistema conta com diferentes algoritmos de seleção de características, nesse projeto irá ser utilizado o de Busca Sequencial Flutuante para Frente (SFFS). A função critério utilizada será a Entropia Condicional Média, sendo que algumas estratégias de

[2] Pode ser acessado em: https://sourceforge.net/projects/dimreduction/

estimativas de erro também serão aplicadas como a penalização de em instâncias raramente observadas, conforme também descrito em (LOPES; MARTINS; CESAR, 2008).

A etapa de classificação supervisionada é a etapa definida como mais importante neste projeto, uma vez que junto a ela podem ser associadas técnicas de seleção de características como uma etapa anterior, fazendo com que os resultados obtidos sejam de grande acurácia. Para realizar essa etapa o modelo referência utilizado foi o de Floresta Aleatória (*Random Forest*) que é um algoritmo de classificação supervisionada não-linear baseado em árvore de decisão criado por Ho (1995), a escolha desse modelo dá-se ao fato do mesmo ser amplamente utilizado na literatura e seus resultados já serem validados por outros projetos anteriores, bem como a boa adaptação aos dados desse trabalho. Muito embora os principais testes tenham sido executados no *Random Forest*, outros modelos também serão utilizados e quando aplicados, serão informados.

A implementação do *Random Forest* utilizada neste trabalho é a realizada por Breiman (2001). Breiman (1996b) observou que no momento em que são construídas as árvores para o processo de classificação, cerca de 37% dos dados de de treinamento são perdidos pois os preditores que utilizam agregação *bootstrap* (*bagging*) usam apenas 2/3 dos dados (BREIMAN, 1996a), dessa forma, ele criou uma forma de usar os dados que não foram usados na construção da árvore para testá-la e mensurar o erro de predição que melhora o desempenho da previsão e evita a necessidade de um conjunto de dados de validação independente, essa mensuração é denominada *out-of-bag-error* também chamada de *out-of-bag-estimate*, que em resumo é uma estimativa interna de erro de uma floresta aleatória no momento em que ela está sendo formada.

4.10 Medidas Utilizadas

Neste projeto foram utilizadas algumas medidas topológicas sobre as redes representadas, abaixo são informadas essas medidas e detalhadas. Muito embora tenham sido usadas um grupo de medidas, algumas não trouxeram resultados minimamente relevantes para o conjunto de dados utilizado, portanto, essas foram descartadas e não serão apresentadas nessa seção.

Algumas funções retornam como resultado um vetor com um valor específico para cada vértice, como por exemplo uma função que exibe qual é a conectividade de um vértice x na rede, essa função analisa cada vértice e retorna um valor específico para cada, porém, muitas vezes deseja-se analisar a rede como um todo e não analisar uma granularidade menor (cada vértice). Uma estratégia abordada neste projeto foi utilizar 3 formas de medida para se obter uma estatística básica da função, sendo elas: o mínimo, que informa o valor mínimo encontrado no vetor; um valor máximo, que exibe sempre o maior valor encontrado; e um valor médio que trata-se de uma média aritmética simples de todos os

valores do vetor. Com isso tem-se a intenção de analisar cada medida de maneira global na rede.

4.10.1 Pacote NetSwan

Nessa subseção são descritos os algoritmos e medidas utilizadas neste projeto provenientes do pacote *NetSwan* (LHOMME, 2015) desenvolvido sobre a linguagem R.

- *SwanCloseness* - Calcula o impacto na proximidade *(Closeness)* quando é removido um vértice da rede, para isso, calcula-se a mudança na soma do inverso das distâncias entre todos os pares de vértices ao excluir um vértice x.
- A função *SwanConectvity*, calcula a perda de conectividade quando é removido um vértice da rede, o resultado desse cálculo mensura a diminuição no número de relacionamentos entre cada vértice da rede quando um vértice ou vários são removidos.
- A função *SwanCombinatory*, realiza o cálculo da vulnerabilidade de uma rede e qual é sua resistência à remoção de vértices seja por falhas aleatórias ou ataques intencionais. Esse algoritmo remove vértice por vértice da rede primeiramente de maneira aleatória, posteriormente na ordem decrescente de seu grau de carga *(betweenness)*, por último ele faz um cenário em cascata recauculando as cargas depois de cada vértice removido.
- A função *SwanEfficiency*, verifica a alteração na soma das distâncias entre todos os pares de vértices ao excluir um determinado vértice.

4.10.2 Pacote brainGraph

Nessa subseção são descritos os algoritmos e medidas utilizadas neste projeto provenientes do pacote *brainGraph* (WATSON, 2019) desenvolvido sobre a linguagem R. O pacote foi criado com o objetivo de ser utilizado em redes cerebrais, sendo que diversas medidas levam em consideração o mapeamento do cérebro, todavia, algumas medidas podem ser estendidas e utilizadas em outros contextos como é o caso desse projeto.

- A função *efficiency* calcula qual é a eficiência de cada vértice da rede, a função matemática que a define é expressa abaixo, sendo que a eficiência E_{nodal} do vértice i presente no grafo G, onde: d_{ij} é o menor comprimento de caminho entre os vértices i e j e N é o total de vértices do grafo.

$$E_{\text{nodal}}(i) = \frac{1}{N-1} \sum_{j \in G} \frac{1}{d_{\text{ij}}} \tag{4.6}$$

- A função *vulnerability* realiza o cálculo da vulnerabilidade dos vértices da rede (LATORA; MARCHIORI, 2005), essa vulnerabilidade pode ser considerada como a queda proporcional na eficiência global da rede quando um vértice específico é removido. A vulnerabilidade é considerada como máxima através de todos os vértices. Como essa função retorna um vetor de valores, utiliza-se os valores estatísticos anteriormente definidos.

- A função *communicability* realiza o calculo da comunicabilidade de uma rede, para isso, essa medida considera todos os caminhos possíveis entre pares de vértices, sejam eles os caminhos mais curtos ou não.

- A função *coeff_var* calcula o coeficiente de variação de uma rede, sua fórmula é expressa abaixo, sendo que o coeficiente de variação do vetor x *CV(x)* é o desvio padrão *sd(x)* dividido pela média *mean(x)*.

$$CV(x) = \frac{sd(x)}{mean(x)} \tag{4.7}$$

- A função *rich_club_coeff* calcula o coeficiente de *Rich Club* da rede.

4.10.3 Pacote iGraph

Nessa subseção são descritos os algoritmos e funções da biblioteca iGraph (CSARDI; NEPUSZ, 2006; CSARDI, 2019). O iGraph é uma biblioteca bastante utilizada na linguagem R e fornece diversas medidas topológicas que quando aplicadas em grafos entregam uma série de informações que podem possuir grande relevância para possíveis análises.

As medidas utilizadas do pacote são:

- O Número Total de Vértices e Arestas na rede, que são a quantidade de elementos presentes na rede, sendo que a definição desses grafos é apresentada em 4.7 Organização do Grafo de Colaboração Acadêmica.

- A Quantidade de Vértices isolados na rede é a contagem de elementos que possuem um grau de conectividade nulo, ou seja, os vértices que não possuem nenhuma conexão com outros vértices no grafo. A razão entre esse número de vértices isolados e o total de vértices da rede, geram a medida Percentual de Vértices Isolados.

- O Grau *(degree)* do grafo revela o número de arestas que um determinado vértice possuí. A média aritmética desses valores exibe o Grau Médio da rede, conforme definido anteriormente em 2.3.2 Grau e grau médio.

- O Coeficiente de aglomeração *(cluster)* exibe o grau de tendência em que os vértices do grafo tem a se agrupar e formar grupos de comunidades *(clusters)*, conforme definido em 2.3.4 Coeficiente de Aglomeração.

- O Coeficiente de Assortatividade *(assortativity coefficient)* expressa a tendência de vértices se conectarem a outros vértices similares ou diferentes em algum aspecto, como o grau, por exemplo. Os valores variam entre -1 e 1, sendo que -1 são os vértices que se conectam unicamente com elementos diferentes e 1, os vértices se conectam exclusivamente com elementos similares.

- O Caminho Médio da rede exibe o número médio de arestas entre duas vértices.

- O Diâmetro da Rede informa qual é a maior distância entre os pares de quaisquer vértices da rede.

- A medida de Densidade da Rede informa a razão entre o número de conexões entre os vértices do grafo e o número de possíveis conexões, conforme definido em 2.3.3 Densidade.

- A Centralidade de Grau *(degree centrality)* mede o quão importante é um vértice na rede levando em conta as conexões inerentes a ele. Essa medida caracteriza-se sendo que quanto mais conexões nó vértice, maior será seu grau de centralidade, portanto, assume-se que esse nó possui maior importância (FREEMAN, 1978).

- O índice de *Betweenness Centrality* também mensura a importância de um vértice na rede, porém essa medida leva em consideração o número de caminhos mínimos entre dois nós escolhidos de maneira aleatória sendo que o elemento em foco tem que estar entre eles.

- A medida *Closeness Centrality* bem como as medidas anteriores avalia a importância de um vértice baseando-se na distância de todos os outros elementos do grafo.

- O *Eigenvector Centrality* avalia a importância de um vértice levando em consideração além da quantidade de conexões (grau), também a qualidade dessas conexões, dessa forma influencia na medida se o vértice está conectado a outro(s) vértice(s) com um grande número de conexões.

5 RESULTADOS

Neste capítulo são apresentados os resultados obtidos ao considerar como fonte de dados os currículos Lattes, conforme detalhado nos capítulos anteriores. Todos os dados utilizados neste trabalho estão dentro de 3 períodos de avaliação da CAPES, sendo eles o Período 1 que vai do ano de 2007 à 2009; o Período 2 de 2010 à 2012 e; o Período 3 do ano de 2013 ao ano de 2016. Em todas as análises serão considerados os 3 períodos como um único indo de 2007 à 2016 e nos casos onde cada período será analisado em específico será destacado e informado.

5.1 Pré-processamento

Todas medidas de cada característica da rede foram normalizadas para ser exibidos somente números dentro do intervalo entre 0 e 1 e o intuito da normalização é que as medidas não influenciem os algoritmos pela diferença de grandeza entre elas. A equação abaixo exibe a fórmula utilizada para a normalização onde o valor normalizado de x (VN) é o resultado do valor de x menos o valor mínimo da medida ($VMin$) dividido pelo valor máximo da medida, ($VMax$) menos o valor mínimo da medida($VMin$).

$$VN(x) = \frac{(x - VMin)}{(VMax - VMin)} \tag{5.1}$$

Algumas medidas dos pacotes de análise de vulnerabilidade das redes retornavam vetores de dados, essas medidas foram então transformamos em 3 medidas para cada programa, sendo elas: o mínimo, que exibe o menor valor do vetor, o máximo com o valor mais elevado e a média, que é uma média aritmética simples dos valores do vetor, como por exemplo a medida de vulnerabilidade *"Vulnerability"* do pacote brainGraph, que terá o minVulnerability, o avgVulnerability e o maxVulnerability que são respectivamente os valores mínimos, médios e máximos do vetor de análise de vulnerabilidade retornado pela algoritmo.

Os programas possuem tamanhos diferentes, com isso, as medidas podem gerar valores que podem levar a análises espúrias se não for levada em consideração esse tamanho, por exemplo, a medida de caminho médio é completamente dependente do tamanho do programa, pois programas maiores certamente terão maiores medidas de caminho médio que programas menores, porém, deve ser levado em consideração o tamanho do programa e com o objetivo de resolver a questão da diferença entre o tamanho, fizemos uma ponderação utilizando o número total de arestas ou vértices da rede, dependendo de qual medida é analisada, dessa forma, no exemplo acima dividimos o valor do caminho médio pelo número

de arestas, como os dados estão normalizados no mesmo intervalo de valores esse processo é possível, dessa forma o tamanho do programa não influenciará na medida analisada.

5.2 Índices propostos

São exibidos abaixo os resultados dos índices que foram propostos neste projeto, sendo eles os 3 Índices de Colaboração e a Média de Pesquisadores por Publicação.

5.2.1 Índices de colaboração

Os índices propostos neste projeto foram gerados e suas análises realizadas. Os programas foram agrupados de acordo com sua avaliação CAPES e a Figura 9 exibe a média dos índices de cada grupo, nas linhas são exibidas as médias dos Índices de Senioridade, dos Índices de Primeiro Autor e dos Índidces de Colaboração. É possível observar que a medida em que as notas de avaliação da CAPES aumenta, os Índices de Senioridade aumentam conjuntamente, tendo um comportamento linear o que confirma a Hipótese 2 anteriormente levantada, observa-se também que os índices de Colaboração e Senioridade são bem próximos nos programas de Notas 6 e 7. É visível que o Índice de Senioridade está em seu menor valor nos programas NOTA3 e o maior nos programas NOTA7, bem como o inverso ocorre com o Índice de Primeiro Autor, portanto ambos os índices apresentam uma diferença significativa em seus respectivos extremos, pode-se também afirmar que em todos os programas o maior Índice é o de Colaboração exceto em programas de NOTA6, onde o maior Índice é o de Senioridade.

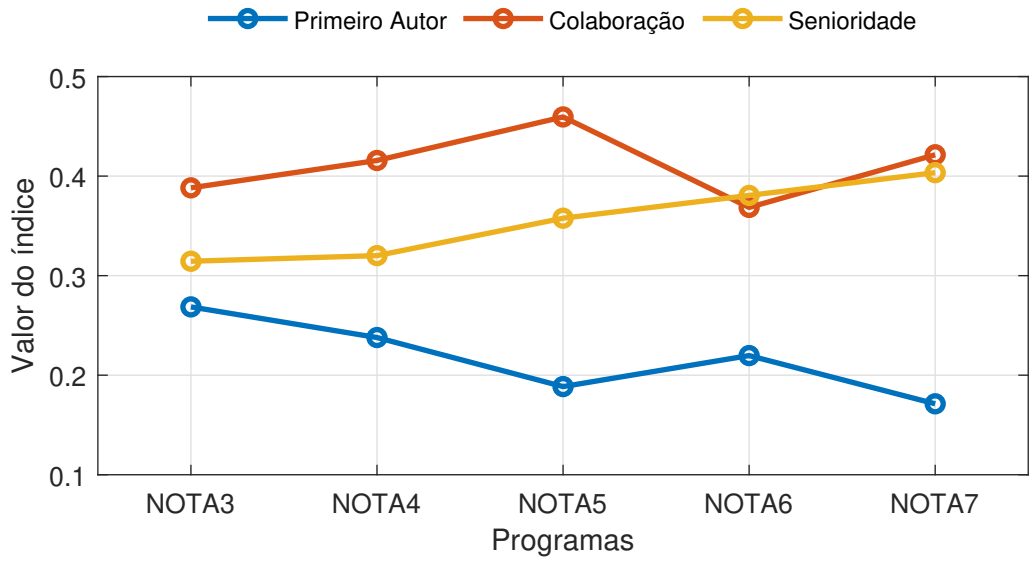

Figura 9 – Índices propostos.

Fonte: Autoria própria

Também foi realizada a avaliação dos Índices de colaboração, porém com o agrupamento dos programas usando as três classes (Programas A, Programas B e Programas C), conforme definido na Seção 4.4, a Figura 10 exibe a avaliação da média das medidas onde é possível verificar que o Índice de Primeiro Autor nos Programas A e Programas B é praticamente o mesmo, assim como os Índices de Colaboração entre os Programas A e Programas C, também observa-se que os Programas A possuem valores muito próximos no Índice de Colaboração e Índice de Senioridade. Nessa análise é possível verificar ainda a linearidade em que é elevado o Índice de Senioridade conforme aumenta a classificação dos programas.

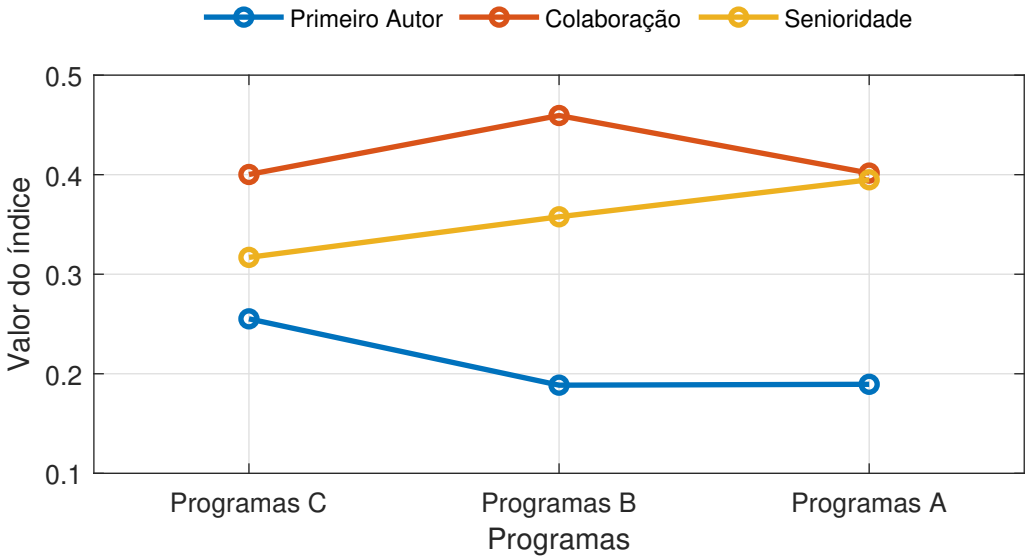

Figura 10 – Índices de Colaboração aplicado à 3 classes.

Fonte: Autoria própria

5.3 Média de Pesquisadores por Publicações

Com o intuito de observar a produtividade de um programa foi extraído o percentual de pesquisadores por publicações das redes e ele apresentou resultados de grande relevância, sendo que com o resultado dos dados gerados, é possível observar que conforme há um aumento na avaliação CAPES dos programas, há também uma tendência de diminuição na medida proposta, confirmando assim que os programas menos bem avaliados possuem menor produtividade no tocante exclusivamente à quantidade de publicações. A Figura 11 demonstra a tendência de queda na média de pesquisadores por publicações conforme o aumento da Nota CAPES, com exceção dos programas NOTA6 que possui um comportamento um pouco atípico, confirmando a maior produtividade de programas mais bem avaliados conforme dito acima.

Em várias análises neste trabalho será utilizado esse modelo de gráfico conhecido

como diagrama de caixa ou *boxplot*, onde na barra inferior (mais próxima do eixo x) são exibidos os valores mínimos, a barra logo acima exibe o valor do primeiro quartil, a barra ao centro o valor da mediana, na barra superior a ela o valor do terceiro quartil e na barra mais superior de todas são expressos os valores máximos do dado analisado, os valores discrepantes (*outliers*) são representados por pequenos círculos fora das barras anteriormente descritas.

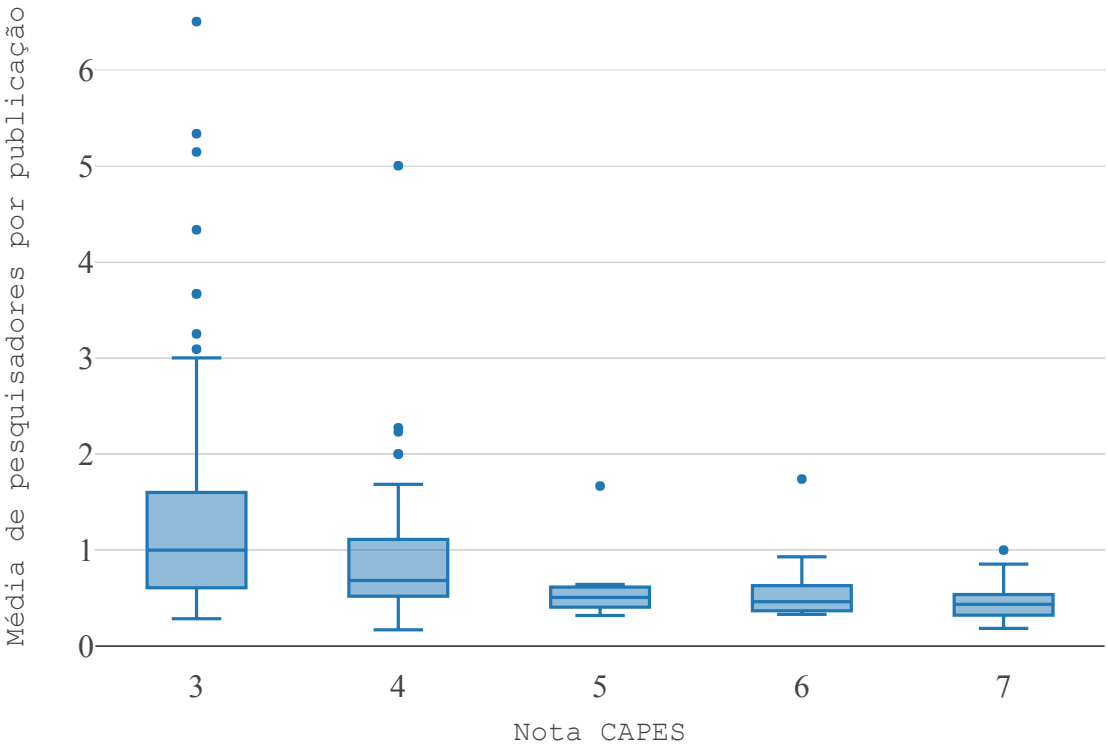

Figura 11 – Média de Pesquisadores por Publicações.

Fonte: Autoria própria

5.4 Seleção de Atributos

Nessa etapa aplicamos sobre os dados gerados modelos de seleção de atributos (*FS*) com o objetivo de selecionar as características mais importantes para a classificação supervisionada dos dados. O algoritmo usado como avaliador de atributos foi o CfsSubsetEval, que é um algoritmo de correlação baseado na seleção de característica (HALL, 1999). O método de buscas utilizado foi o melhor-primeiro (*Best First*), que é um método de busca global que avalia qual a melhor opção para ser escolhida atendendo a uma função heurística F(n)=h(n). Essa junção de algoritmos de avaliação e função de busca elencou 9 atributos como os melhores atributos para a classificação dos dados com base na Nota CAPES, executamos também uma validação cruzada utilizando como parâmetro 10

amostras, para entendermos dos atributos quais são os mais relevantes. Os resultados da validação cruzada, são expressos na Figura 12, onde das 10 amostras, qual o percentual de uso de cada atributo na execução. A interpretação de cada um desses atributos será detalhada posteriormente.

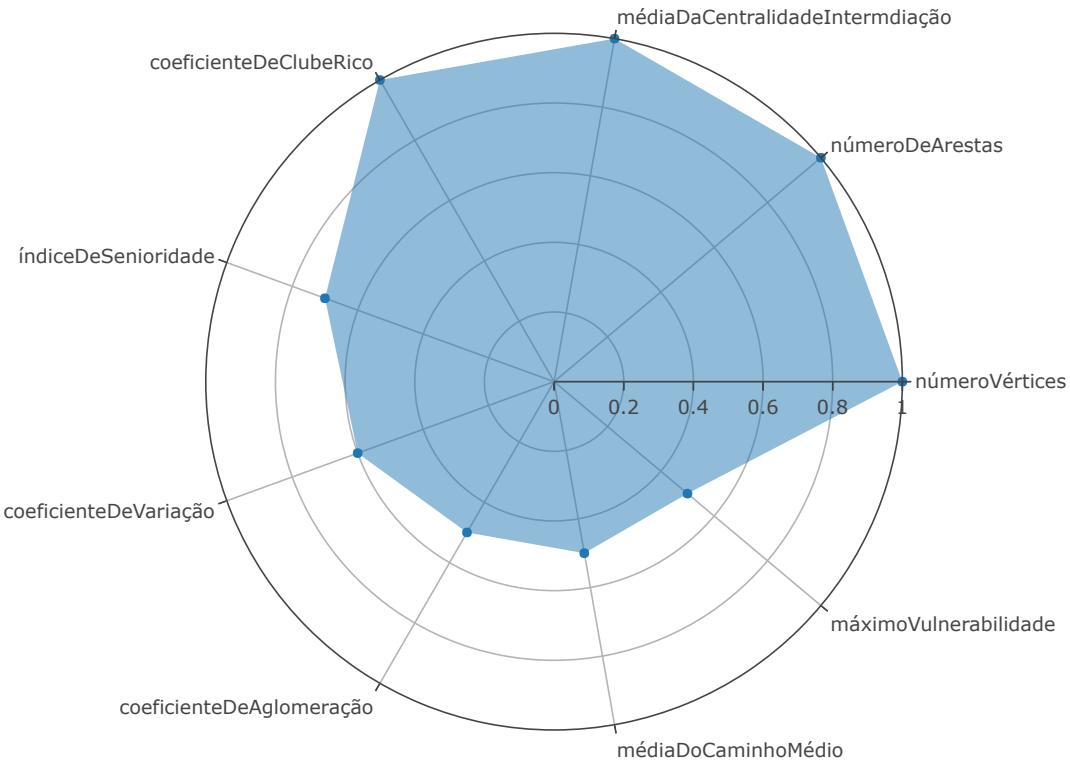

Figura 12 – Características mais relevantes escolhidas pelo algoritmo de Seleção de Atributos.

Fonte: Autoria própria

5.5 Algoritmo de Classificação *Random Forest*

Após o processamento dos dados, aplicou-se o algoritmo *Random Forest* com as configurações padrões da ferramenta WEKA. O mecanismo de teste utilizado foi o de validação cruzada *k-fold* (*cross-validation*) com uma quantidade de 10 dobras (*folds*) considerando os dados totais, portanto, garante-se que uma parte dos dados é usada na validação do modelo, a Figura 13 ilustra o processo de validação cruzada, onde é possível observar que a cada execução uma parte dos dados é utilizada para validar o modelo e outra parte (*k-1*) é utilizada para treinar o modelo, sendo nesse exemplo, 1 parte usada para a validação e 9 para o treinamento.

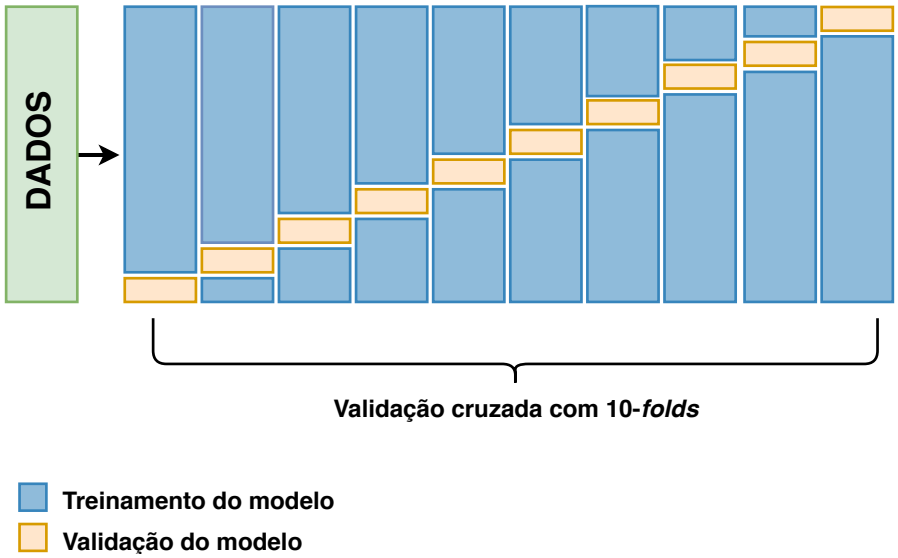

Figura 13 – Exemplo de validação cruzada *k-fold* onde o k é igual a 10.

Fonte: Autoria própria

Os dados retornados pelo modelo são as amostras classificadas, o valor de erro e, o mais importante para este projeto, são exibidos os valores de importância de cada atributo (*Random Forest Importance*), ou seja, o peso de cada atributo no processo de aprendizagem do modelo, sendo que com essa informação pode se definir quais são os atributos que melhor definem esses dados, uma vez que atributos com maior importância na definição da Nota CAPES, são os atributos com maior relevância. Para esse modelo, as medidas de maior relevância são as que diz respeito à quantidade de vértices (número de vértice, número de vértices isolados e percentual de vértices isolados), à quantidade de arestas (número de arestas, coeficiente de aglomeração, diâmetro da rede e densidade da rede), uma medida de semelhança de vértices (coeficiente de assortatividade), e também a Média de Pesquisadores de Publicações foi pontuada como de grande relevância para a classificação do modelo.

É exibido na Figura 14 um gráfico com o resumo dos dados retornados pelo *Random Forest* onde é possível verificar os atributos que possuem maior importância na definição da Nota CAPES, os valores presentes no gráfico diz respeito à importância de cada atributo com base na redução média de impurezas, portanto os atributos com maior pureza utilizados na classificação. Essas medidas e suas interpretações nos dados deste projeto serão explicadas e detalhadas em capítulos posteriores, mas é possível observar que, além de outras, a medida proposta da Média de Pesquisadores por Publicação (MPP) possui grande relevância para esse modelo de classificação.

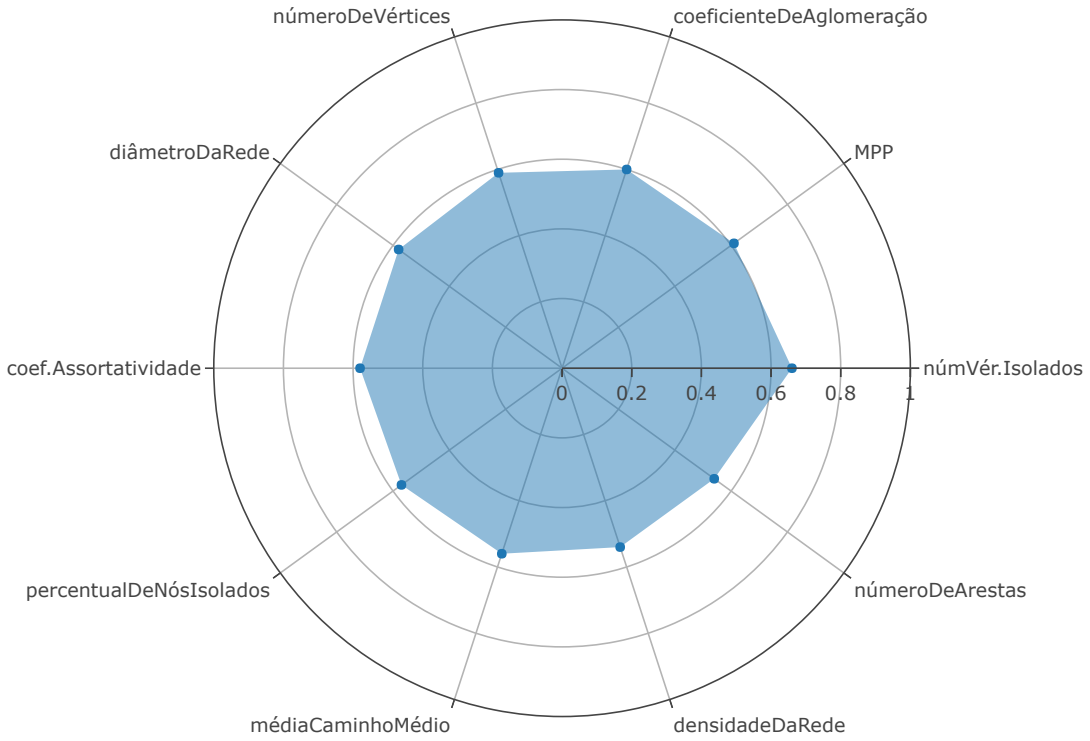

Figura 14 – Importância de cada característica para o Random Forest.

Fonte: Autoria própria

A Tabela 5.5 exibe os dados do *out-of-bag-estimates* onde é possível verificar a previsão de classes corretamente classificadas, os erros pertinentes e o coeficiente de Kappa (COHEN, 1968).

Tabela 3 – Resumo do *out-of-bag-estimates.*

Descrição	**Valor ou percentual**
Instâncias classificadas corretamente	57,310 %
Instâncias classificadas incorretamente	42,690 %
Coeficiente de Kappa	0,359
Erro absoluto médio	0,205
Raiz quadrada de erro médio	0,329

Os dados estatísticos de relações entre o resultado do modelo por cada classe é exibido na Tabela 5.5. Comumente são exibidos os valores de precisão e de revocação juntos, porém nessa Tabela é possível observar que a coluna de precisão é exibida, porém a coluna de revocação (*recall*) está oculta porque seus valores são exatamente o mesmos da taxa de Verdadeiro Positivo (VP). São exibidos então os valores referentes: a taxa de verdadeiros positivos (VP); falsos positivos (FP); a precisão do modelo, onde é possível observar que os programas das extremidades (NOTA3 e NOTA7) foram mais precisos que

os demais na classificação; a *F-measure* que combina a precisão e a revocação e também é conhecida como *F1-score*; o MCC que é o coeficiente de correlação de Matthews (1975), e por último os valores referentes a área da curva COR também chamada pelo termo em inglês *ROC Curve* que é a representação por meio de um gráfico dos verdadeiros positivos X os falsos positivos. Na última linha são exibidas as médias aritméticas de cada uma das colunas.

Tabela 4 – Acurácia do modelo em cada classe.

Taxa VP	**Taxa FP**	**Precisão**	***F-Measure***	**MCC**	**Área ROC**	**Classe**
0,720	0,250	0,692	0,706	0,468	0,828	NOTA3
0,534	0,283	0,492	0,512	0,247	0,701	NOTA4
0,071	0,032	0,167	0,100	0,059	0,741	NOTA5
0,222	0,056	0,182	0,200	0,152	0,838	NOTA6
0,533	0,032	0,615	0,571	0,535	0,929	NOTA7
0,561	0,214	0,548	0,552	0,349	0,787	-

5.6 Algoritmo de Busca SFFS

Nessa etapa foi aplicado sobre os dados o algoritmo definido em 2.5.2 Busca Sequencial Flutuante para Frente (SFFS), sendo que os dados foram previamente normalizados e a função critério utilizado foi a Entropia definido em 2.6 Entropia. Para realizar a classificação foi configurado para gerar conjuntos com no máximo as 3 características mais relevantes para correlacionar com a Nota CAPES.

A função fora executada utilizando a abordagem *Holdout* sendo que 80% do conjunto de dados fora utilizado para a estimação dos parâmetros (treinamento) do modelo e 20% para o teste (validação), os resultados são expressos na Figura 15 sendo que consideramos somente as 10 melhores características que definem a Nota CAPES.

É possível observar que 6 características são de grande relevância para a convergência do algoritmo, portanto essa característica possuem grande correlação com a Nota CAPES, algumas características são comum para outros algoritmos e todas elas serão melhor detalhadas em capítulos posteriores.

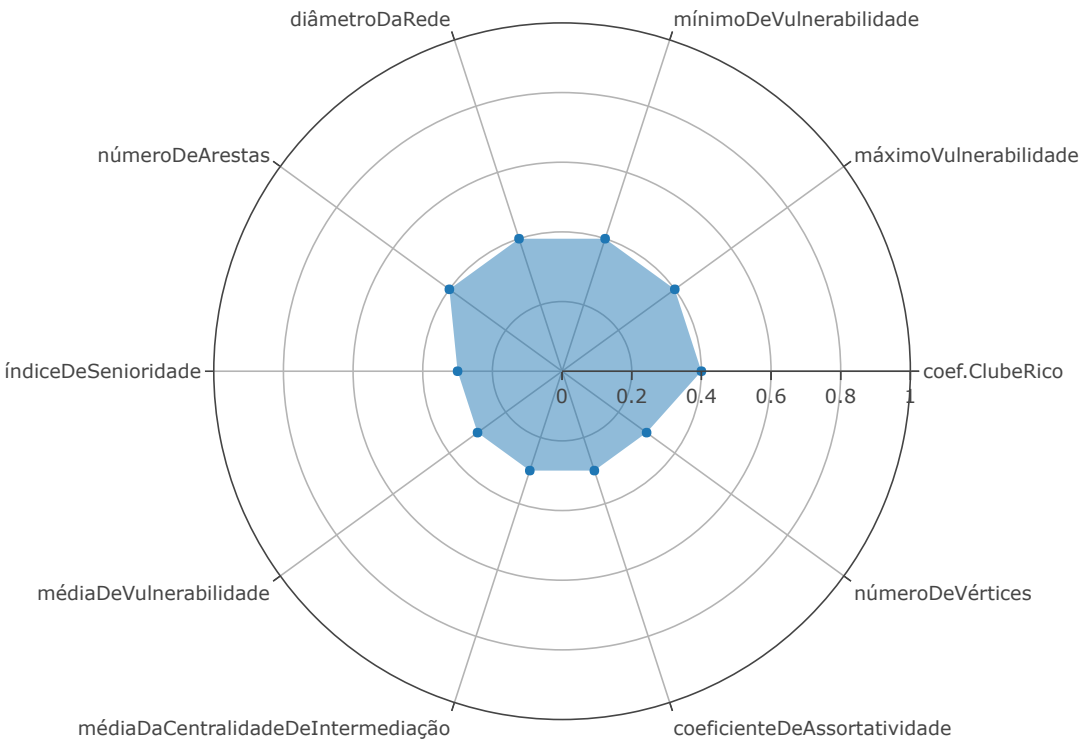

Figura 15 – Frequência de características mais relevantes para o SFFS.

Fonte: Autoria própria

5.7 Coeficiente de Correlação de Spearman

Nessa fase, aplicamos os modelos estatísticos para medir o coeficiente de correlação de Spearman, com o intuito de verificar qual a correlação de cada medida com a Nota CAPES, para calcular esse coeficiente todos os dados precisam ser numéricos, porém neste projeto tratamos a Nota CAPES como um informação nominal, pois os algoritmos de classificação podem realizar análises espúrias ao receber dados numéricos como rótulo da classe, porém nessa etapa os dados foram mantidos numéricos.

Os resultados da execução dessas medidas e os coeficientes de correlação são expressos na Figura 16 que exibe as 10 medidas com maior correlação e qual a correlação das mesmas, os dados que geraram os gráficos são os números absolutos, portanto, neste gráfico não é levada em consideração se a medida é diretamente correlata a Nota CAPES ou se ela é inversamente correlata, sendo que essa análise, bem como o detalhamento de cada uma dessas medidas será realizado em capítulos posteriores.

Embora cada uma das medidas serão discutidas posteriormente é possível ver que algumas medidas de vulnerabilidade de redes aparecem como tendo grande importância na classificação da Nota CAPES, isso mostra que essas medidas possuem uma avaliação

interessante para observar os padrões em redes de coautoria.

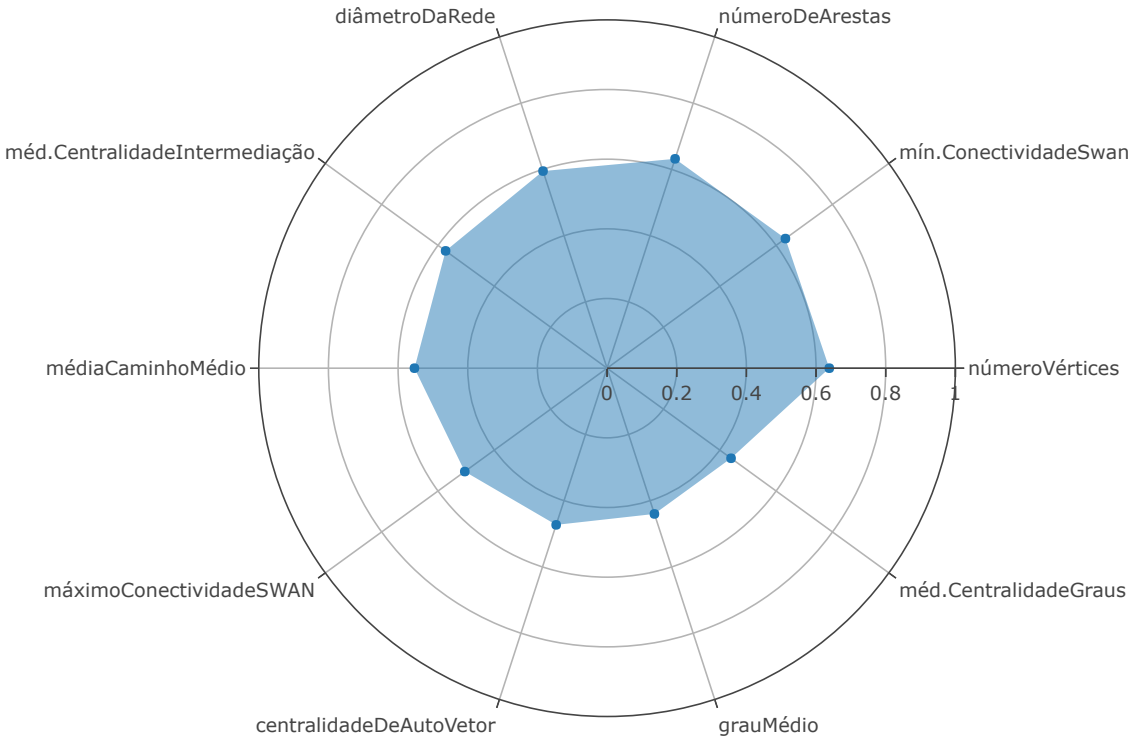

Figura 16 – Percentual de características mais relevantes para o Coef. de Spearman.

Fonte: Autoria própria

5.8 Aprendizado de Máquina Automatizado (*AutoML*)

Embora o conceito de *AutoML* seja bastante amplo e pode abordar muitas etapas e estratégias com recursividade, otimização de hyperparâmetros e seleção de modelos, de maneira simplificada aplicou-se sobre os dados uma das abordagens comuns de *AutoML*, que realizou um *tuning* com o intuito de verificar de maneira automatizada qual o melhor algoritmo de classificação supervisionada bem como os parâmetros do mesmo para que os dados fossem classificados de maneira otimizada.

Essa abordagem conseguiu configurar o algoritmo J48 (originalmente criado por Ross Quinlan e chamado de C4.5 (SALZBERG, 1994)), que é uma árvore de decisão e o mesmo obteve uma taxa de acerto de 77,78% para os dados com 5 classes (Notas CAPES de 3 à 7) e 85,97% para os dados separados por 3 classes, sendo que essa acurácia fortalece a concepção da utilização das medidas topológicas para a definição da Nota CAPES.

A Tabela 5.8 exibe a matriz de confusão que o algoritmo J48 gerou quando realizou a classificação, com as 171 amostras é possível observar que o algoritmo realiza a classificação

com grande acurácia na maioria das classes, observa-se que as classes com maior quantidade de amostras são as que o algoritmo melhor classifica.

Tabela 5 – Matriz de confusão do AutoWEKA.

a	**b**	**c**	**d**	**e**	**<− classificado como**
65	10	0	0	0	a = NOTA3
10	47	1	0	0	b = NOTA4
0	6	7	1	0	c = NOTA5
0	0	1	7	1	d = NOTA6
2	2	3	1	7	e = NOTA7

5.9 Correlação Entre os Resultados dos Algoritmos

Como foram executados diferentes algoritmos e diferentes paradigmas de classificação e reconhecimento de padrões, resultados diferentes foram alcançados, porém algumas medidas foram comuns para mais de um algoritmo, dessa forma nessa seção apresenta-se quais foram as medidas mais relevantes para cada algoritmo e com isso qual medida é mais importante para o conjunto de dados. Na Tabela 5.9 as medidas utilizadas pelo algoritmo de Seleção de Características (SC) são todas listadas pois essa abordagem foi utilizada como a principal nesse projeto, já as medidas retornadas pelo algoritmo *Random Forest* (RF), Coeficiente de Correlação de Postos de Spearman (SPER) e o *Sequential Forward Floating Search* (SFFS), só apenas exibidas as 10 medidas que possuem maior relevância para cada um respeitando o critério que essa medida deve estar em algum outro algoritmo.

Tabela 6 – Características Mais Relevantes.

Medida/Algoritmo	**SC**	**RF**	**SPER**	**SFFS**	**Total**
númeroArestas	S	S	S	S	4
númeroVértices	S	S	S	S	4
caminhoMédio	S	S	S	N	3
coeficienteAglomeração	S	S	N	N	2
médiaCentralidadeIntermediação	S	N	S	N	2
maxVulnerabilidade	S	N	N	N	2
coeficienteAssortatividade	N	S	N	S	2
diâmetroRede	N	S	S	N	2
índiceSenioridade	S	N	N	S	2
coeficienteClubeRico	S	N	N	N	1
coeficienteVariação	S	N	N	N	1
numVérticesIsolados	N	S	N	N	1
percentualVérticesIsoladas	N	S	N	N	1
médiaCentralidadeProximidade	N	S	N	N	1

5.10 Interpretação de cada Atributo no Contexto deste Projeto

Nesta seção é explicado o significado no contexto deste projeto de cada atributo escolhido como sendo de grande relevância para cada um dos algoritmos. Nessa seção é apenas realizada a interpretação de cada medida, pois as definições de cada uma é anteriormente realizada na seção 4.10. Os valores exibidos nos gráficos são os valores médios de cada grupo de programas, portanto essas informações são menos sensíveis a situações onde um único programa é destoante dos demais.

5.10.1 Número de vértices

Este atributo no contexto deste projeto representa a quantidade de autores que cada rede (programa) possui, ele diz respeito ao tamanho da rede e é um dos atributos que mais tem correlação com a Nota CAPES, conforme a Figura 17 pode-se observar que os programas com menores notas de avaliação CAPES também possuem uma menor média no número de vértices, portanto, com base nestes dados é correto afirmar que programas maiores (com relação a quantidade de pesquisadores) são programas mais bem avaliado, essa afirmativa tem grande coerência, pois programas com maiores notas, recebem mais verbas e podem aumentar a quantidade de pesquisadores. Observa-se que os programas de NOTA6 estão um pouco fora da tendência padrão e acabam diferindo um pouco do previsto para os demais.

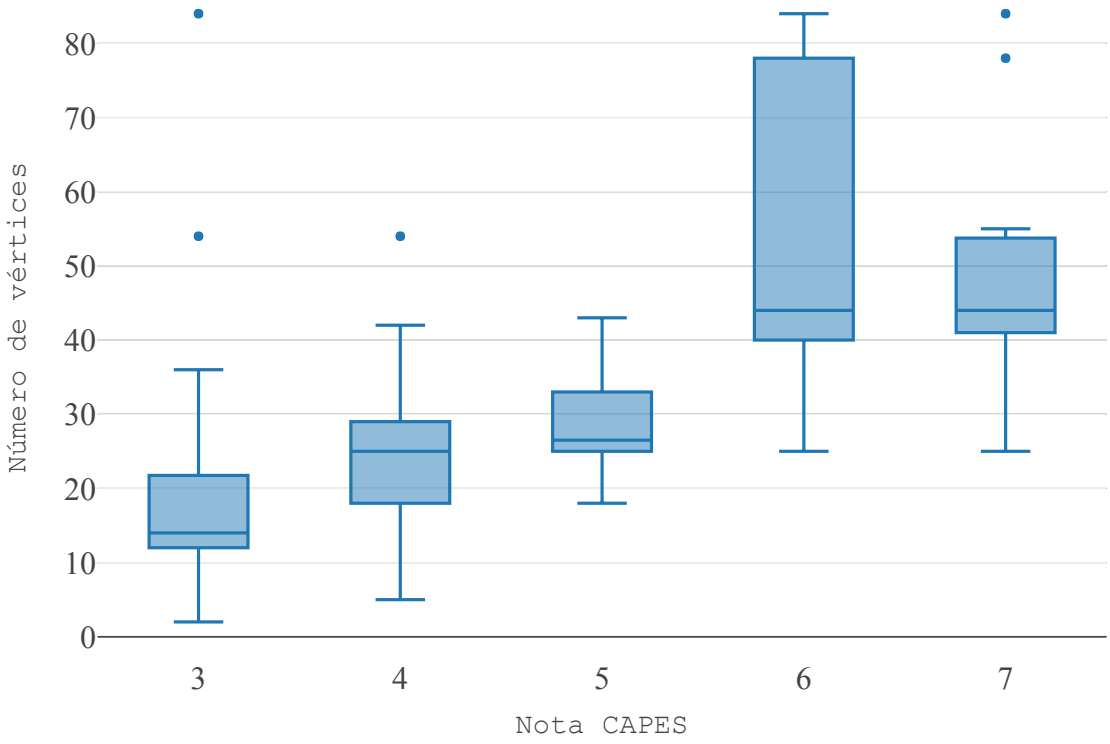

Figura 17 – Média do número de vértices em cada grupo de avaliação CAPES.

Fonte: Autoria própria

5.10.2 Número de arestas

Este atributo revela a quantidade de publicações em coautoria dentro de um programa, as arestas são a ponte de ligação entre os vértices, portanto, o ponto de conexão entre os autores, essa medida também é uma das medidas mais relevantes para a análise dos programas e possui um elevado nível de correlação com a Nota CAPES, como essa medida diz respeito ao tamanho dos caminhos da rede, outras medidas como caminho médio estão bem correlacionadas com ela. Programas com um maior valor nessa medida são os programas que possuem mais quantidade de publicações em coautoria, portanto, essa medida pode ser significativa na qualificação dos programas. A Figura 18 mostra a média do número de arestas em cada grupo de avaliação CAPES, onde é possível ver a tendência em que quanto maior a avaliação do programa, maior a média de publicações em coautoria.

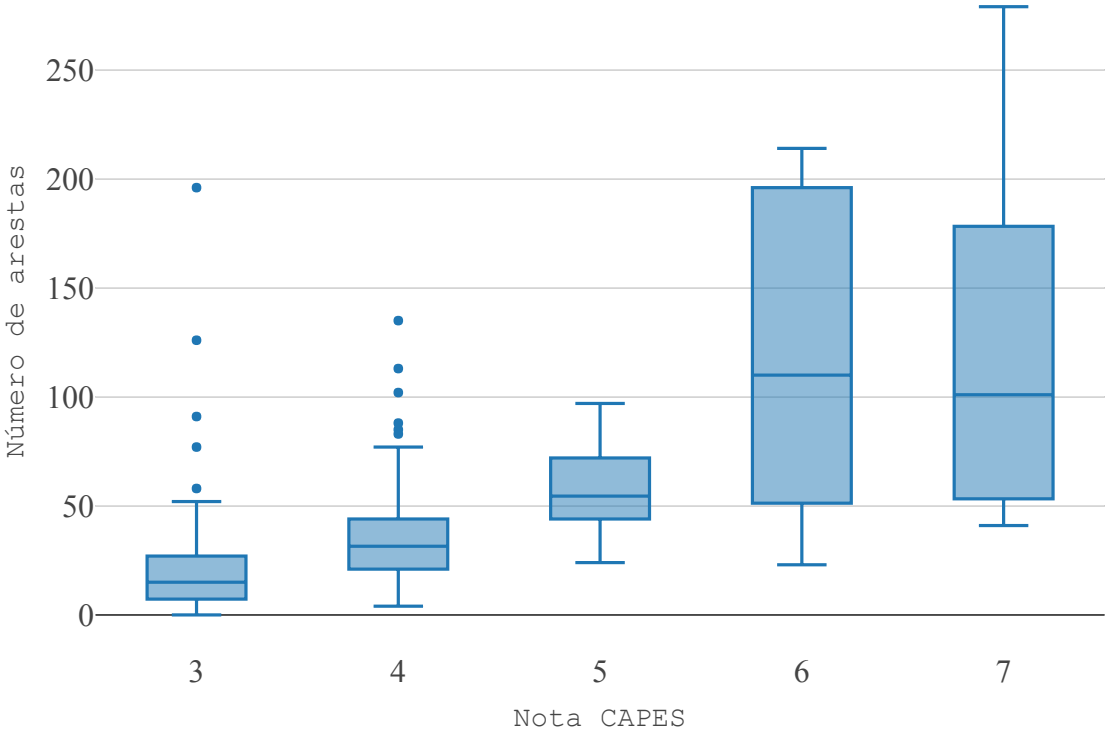

Figura 18 – Média do número de arestas em cada grupo de avaliação CAPES.

Fonte: Autoria própria

Uma análise que podemos fazer com base nas medidas de média de vértices e média de arestas dos programas é ver que embora os programas de NOTA6 possuam maior quantidade de pesquisadores (vértices), esses programas não possuem maior quantidade de publicações (arestas) que os programas de NOTA7, o que pode significar que programas de NOTA7 embora possuem em média menos pesquisadores que programas de NOTA6 ainda sim possuem maior quantidade de publicações em coautoria, ainda que essa medida não leva em consideração a qualidade das publicações mas já é possível observar este padrão, a Figura 19 representa essas afirmações.

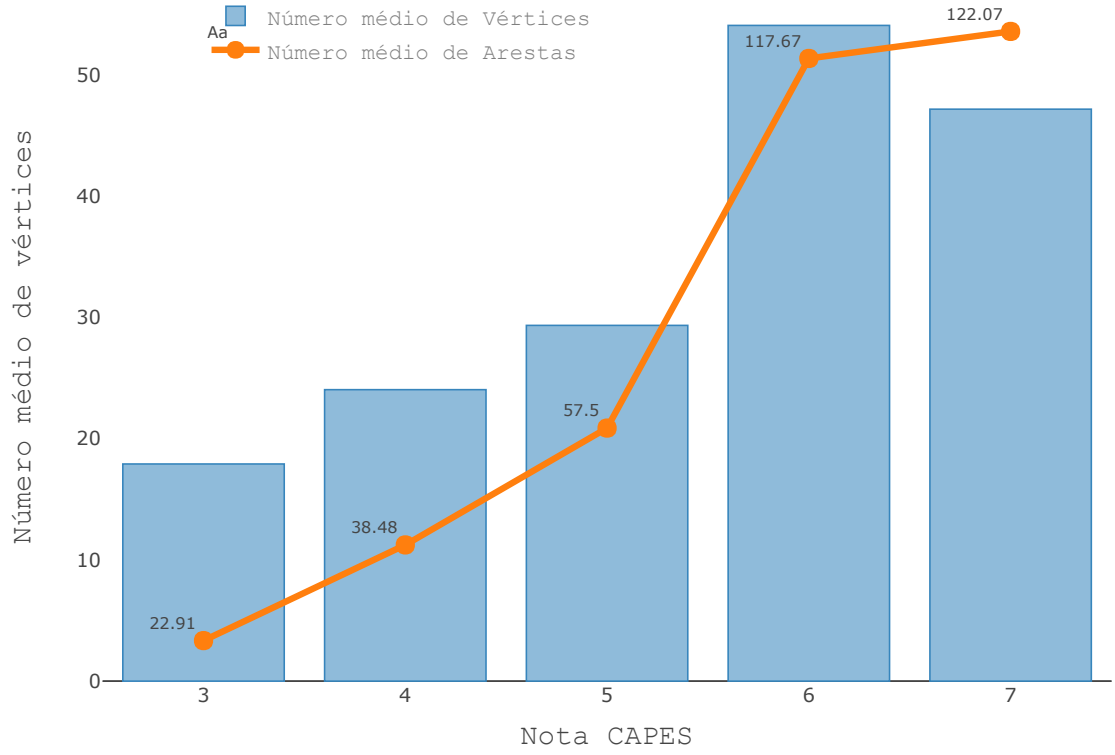

Figura 19 – Média do número de arestas e de vértices de cada grupo de avaliação CAPES.

Fonte: Autoria própria

5.10.3 Caminho Médio

A medida de Caminho Médio (*Average Path Length*) exibe o número médio de etapas nos menores caminhos para todos os pares possíveis de vértices do grafo, é uma medida da eficiência da informação ou do transporte de massa em uma rede. Pode-se analisar essa medida como o número médio de pesquisadores que um vértice α precisa interagir para alcançar o vértice β, partindo do conceito que α e β não possuem uma conexão direta.

No contexto deste projeto, quanto maior o valor desse caminho médio, mais difícil é a colaboração entre todos os pesquisadores do programa, o resultado dessa medida possui grande correlação com as medidas de número de vértices e o número de arestas, pois são elas que determinam o comprimento do caminho. Observa-se na Figura 20 que os programas mais bem avaliados possuem um maior índice de caminho médio, pois como são programas com maiores quantidades de pesquisadores e de publicações, a média dos caminhos será maior, todavia essa análise pode estar incorreta pois com ela não se sabe exatamente se um programa possui maior valor de caminho médio que outro programa pelo fato dela não estar ponderada.

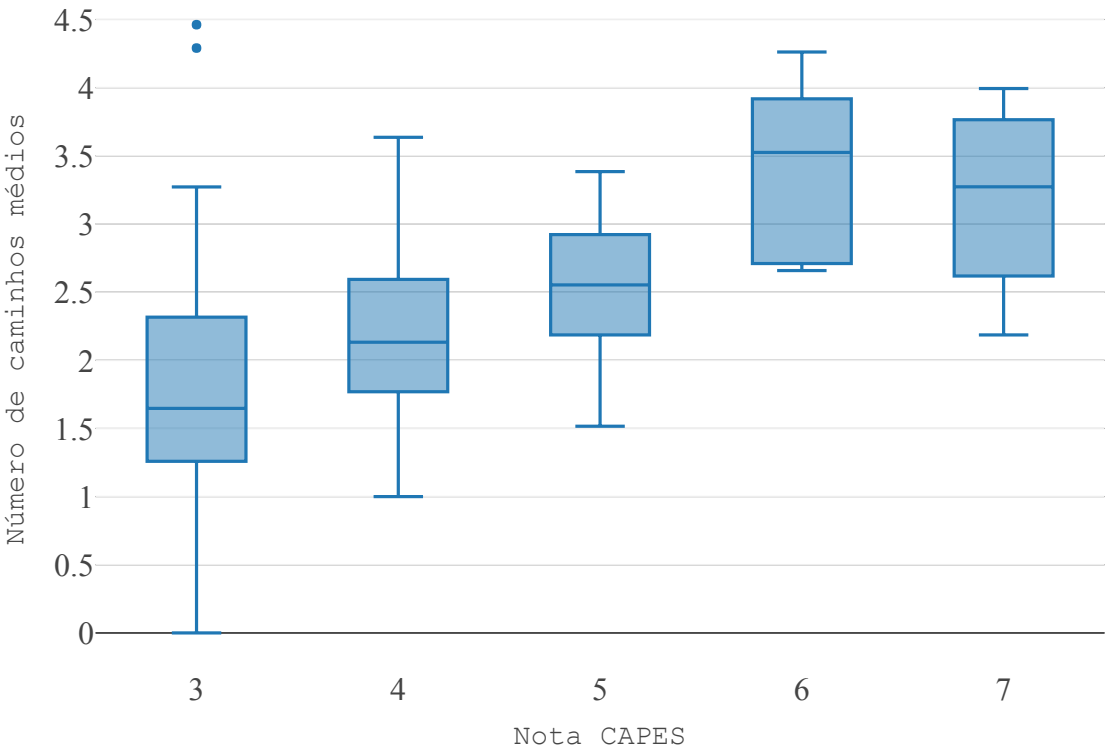

Figura 20 – Caminho Médio

Fonte: Autoria própria

Para contornar a diferença entre o tamanho dos programas foi realizada uma ponderação e com os dados ponderados, têm-se então uma interpretação completamente inversa da anterior, pois nessa nova abordagem os programas com maiores notas CAPES possuem menor caminho médio, portanto, os pesquisadores são conectados mais rapidamente com outros pesquisadores distintos, as médias dos caminhos médios são exibidas na Figura 21.

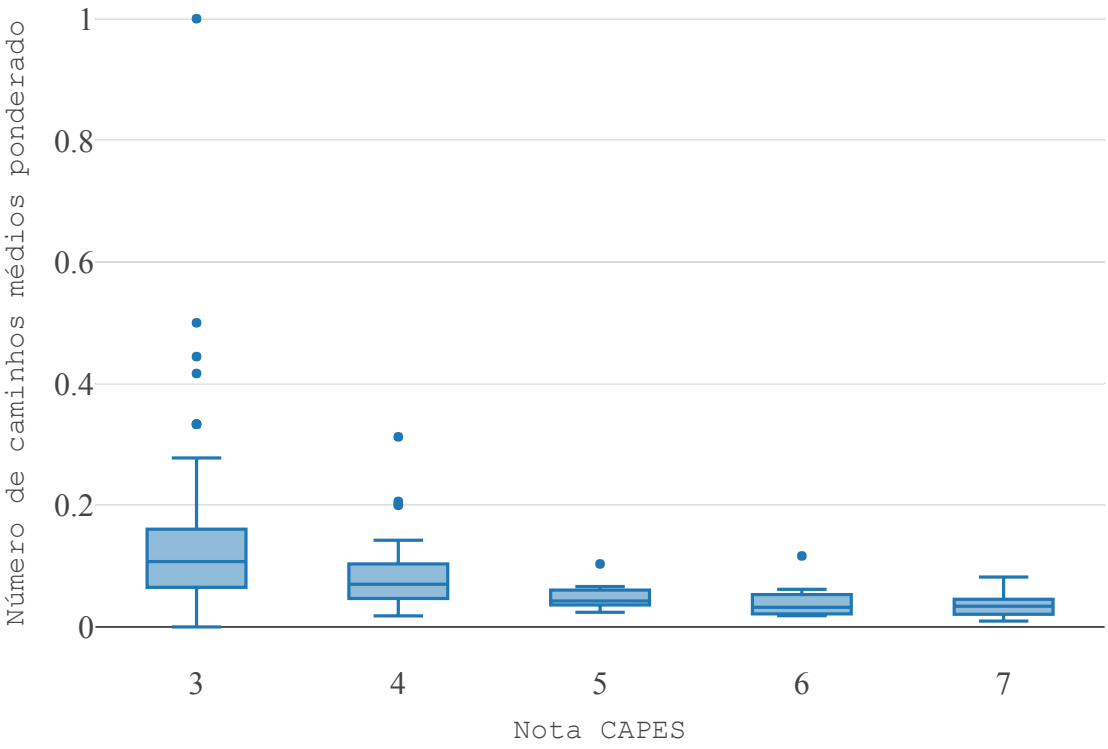

Figura 21 – Caminho Médio Ponderado

Fonte: Autoria própria

5.10.4 Coeficiente de Aglomeração

O Coeficiente de Aglomeração está diretamente relacionado à quantidade de pesquisadores com grande número de conexões com outros pesquisadores, esses pesquisadores são considerados como *hubs*, e esses *hubs* podem ser significativos para a estrutura dos programas. Como essa medida analisa a presença de elementos fortemente conectados o que pode ser comparado a "comunidades" nas redes, quanto maior o seu valor, interpreta-se que maior é a quantidade de pesquisadores com características em comum que interagem entre si.

Com os dados gerados como resultado deste projeto, não é possível observar uma tendência padrão, conforme exposto na Figura 22, entretanto, essa medida pode ser observada como uma medida completamente dependente do tamanho de cada rede (vértices e arestas), portanto, calcula-se o coeficiente de aglomeração médio (GABARDO, 2015) ponderando com o número de vértices para que o tamanho das redes não influencie na média do coeficiente de aglomeração.

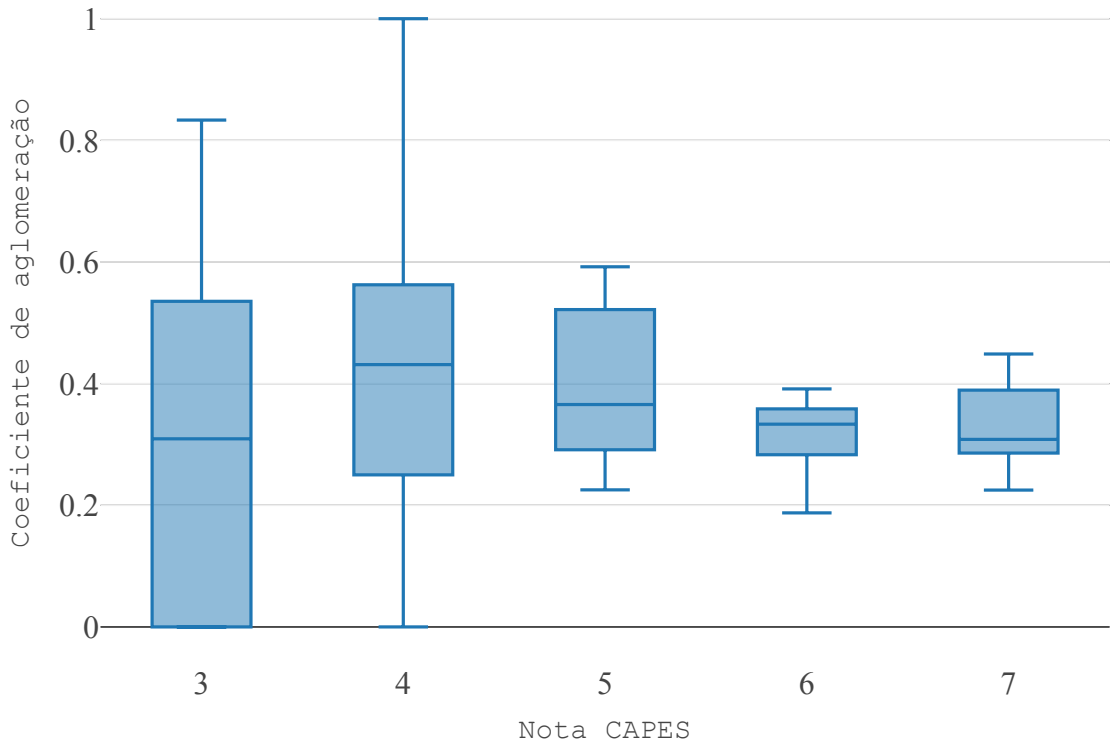

Figura 22 – Coeficiente de Aglomeração

Fonte: Autoria própria

A Figura 23 exibe os dados resultantes sendo que com essa ponderação, é possível então verificar uma tendência padrão de decaimento do coeficiente de aglomeração conforme o aumento da avaliação CAPES dos programas, a interpretação desses dados é que os programas mais bem avaliados possuem menor quantidade de *hubs*, portanto, possuem menor quantidade de comunidades de pesquisadores que possuem semelhança entre si, o que pode concluir que existe maior heterogeneidade entre os autores, sendo que os programas menos bem avaliados são mais semelhantes a modelos de rede de mundo pequeno, conforme definido em na Seção 2.2.3.

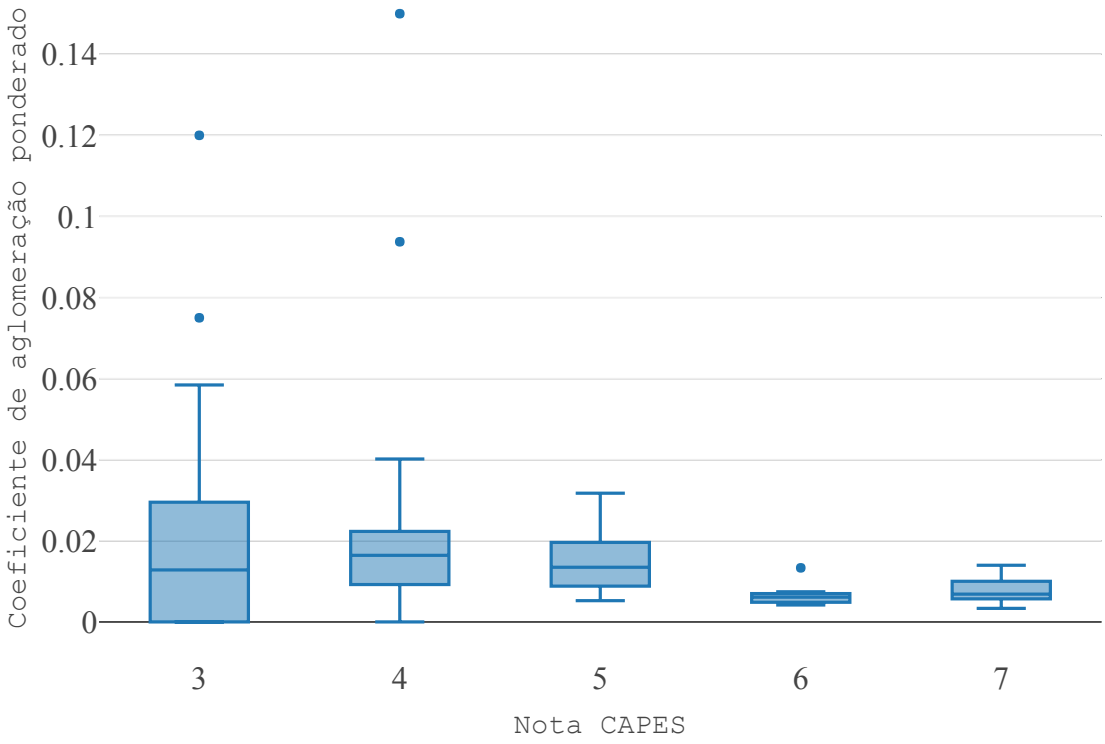

Figura 23 – Média do Coeficiente de Aglomeração Ponderado

Fonte: Autoria própria

5.10.5 Média de Centralidade de Intermediação

A medida de centralidade de intermediação (*betweenness centrality*), diz respeito à quantidade de conexões que passam por um vértice, ela leva em consideração a centralidade de um vértice com relação ao caminho mais curto de outros vértices. No contexto deste trabalho quanto maior o valor da centralidade de intermediação, mais um pesquisador participa direta ou indiretamente das publicações total desse programa, portanto, mais influente e importante é o pesquisador no programa, como esse valor é individual para cada vértice, cada programa gera um vetor, coletamos então a média desse vetor que é o valor comparado à Nota CAPES pelas diferentes abordagens. A Figura 24 exibe a média desses valores médios de centralidade de intermediação nos programas, é possível então verificar que quanto maior a Nota CAPES de um programa, maior a média de centralidade de intermediação dele, portanto, programas mais bem avaliados possuem a característica de possuírem mais pesquisadores que atuam como "pontes" conectando a outros pesquisadores mediante a coautoria em publicações, dessa forma, programa menos bem avaliados possuem menos pesquisadores influentes, sendo assim mais dependentes dos vértices com maior centralidade de intermediação. Essa medida não teve a necessidade de ser ponderada uma vez que a alteração na tendência foi muito pouca.

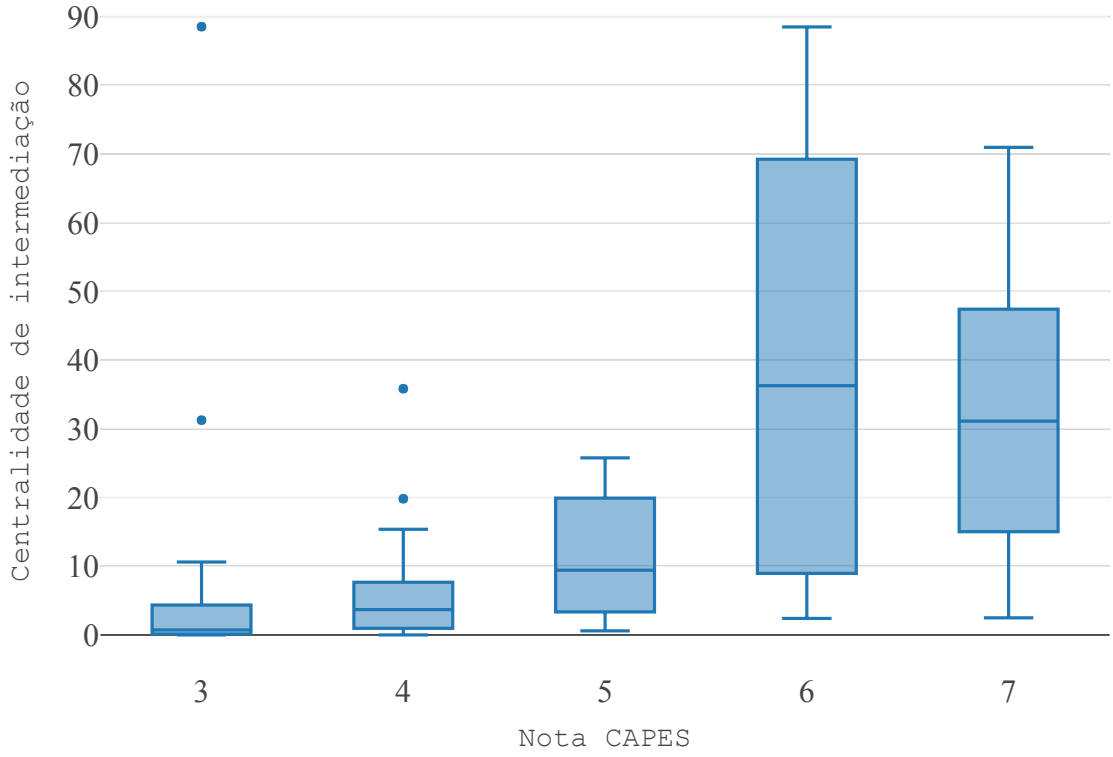

Figura 24 – Média das Centralidades de Intermediações

Fonte: Autoria própria

5.10.6 Diâmetro da Rede

O diâmetro da rede informa qual é a maior distância geodésica do grafo, ou seja, qual é o maior dos menores caminhos entre cada par de vértices, dessa forma no contexto deste projeto, essa medida indica o quão distantes estão os pesquisadores em um programa, sendo que quanto maior o valor dela, mais distante está um pesquisador X de um pesquisador Y. Ao observar os dados gerados (Figura 25), é possível verificar que a medida cresce proporcional à Nota CAPES, todavia, como essa medida trata de distância e os programas possuem tamanhos diferentes, pondera-se a medida com o número total de arestas do programa, obtendo então os novos resultados adversos aos anteriores que concluem que quando maior a avaliação CAPES dos programas, menos distantes estão seus pesquisadores quando se leva em consideração as publicações em coautoria entre os mesmos, conforme é demonstrado na Figura 26, por mais que os programas possuem maior quantidade de pesquisadores e publicações, ainda sim a rede é mais concentrada.

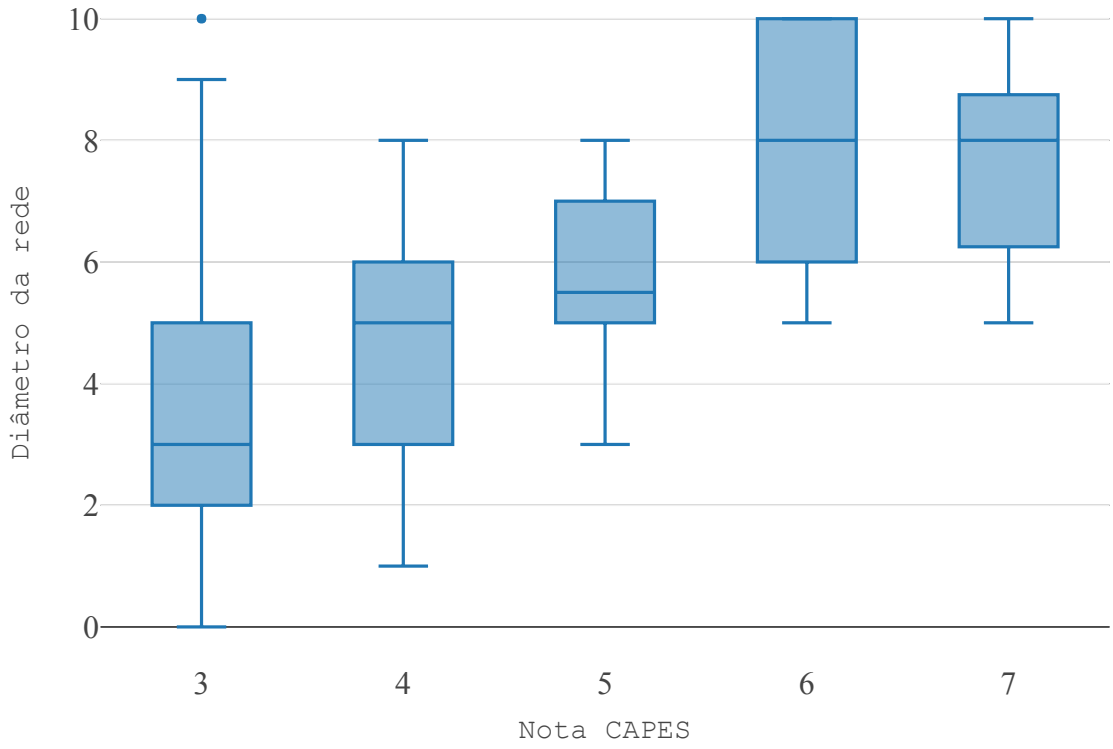

Figura 25 – Média do Diâmetro da Rede

Fonte: Autoria própria

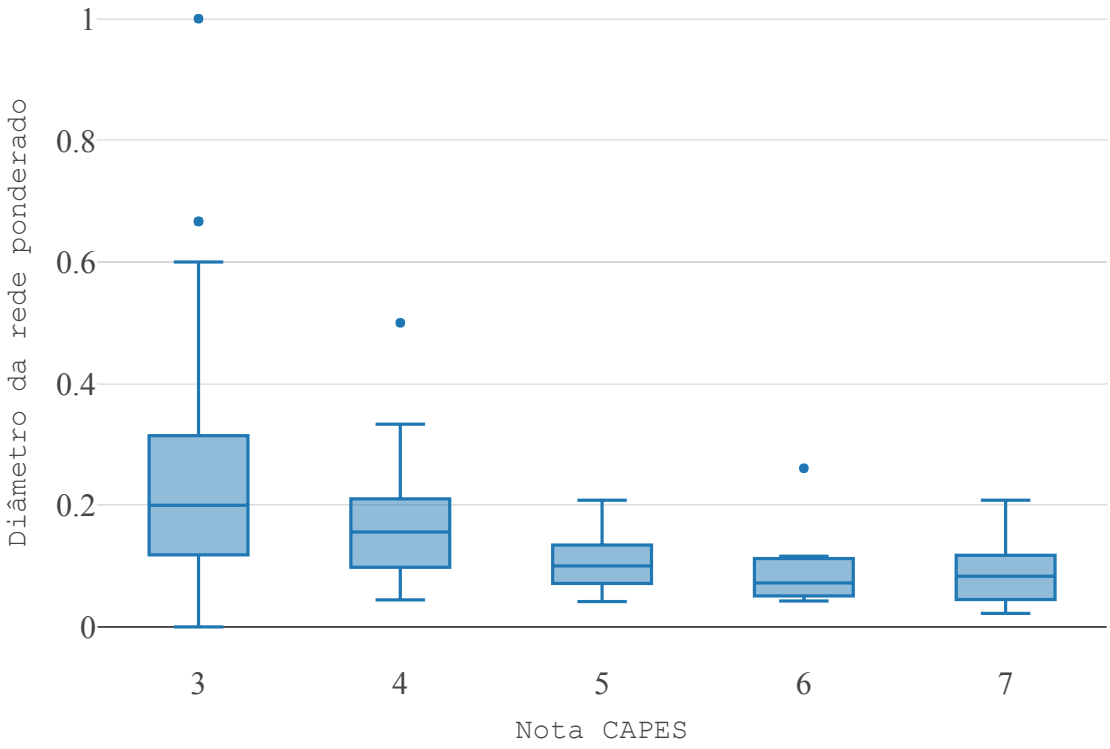

Figura 26 – Média do Diâmetro da Rede Ponderado

Fonte: Autoria própria

5.10.7 Coeficiente de Assortatividade

O coeficiente de assortatividade é uma métrica que avalia a semelhança entre os vértices da rede e qual a tendência que esses vértices semelhantes possuem em se conectar. No contexto deste projeto essa medida avalia o quão parecidos são os pesquisadores de um programa e qual a tendência que esses pesquisadores possuem em serem coautores em um mesmo projeto de pesquisa, para medir a semelhança desses pesquisadores o algoritmo analisa alguns valores do vértice, como por exemplo o seu grau e/ou a importância (centralidade) dele e consegue comparar com outros pesquisadores vizinhos. Na Figura 27, são exibidos os valores médios de coeficiente de assortatividade de cada grupo de programas avaliados pela CAPES, ao observar este gráfico torna-se difícil encontrar uma tendência padrão nos dados, porém na Figura 28 os dados são ponderados com a quantidade média de vértices de cada grupo, sendo que assim é possível observar uma tendência padrão de descensão neste coeficiente conforme há uma ascensão na avaliação CAPES, com esses dados pode-se afirmar que conforme os programas crescem na avaliação CAPES os seus respectivos pesquisadores são menos semelhantes.

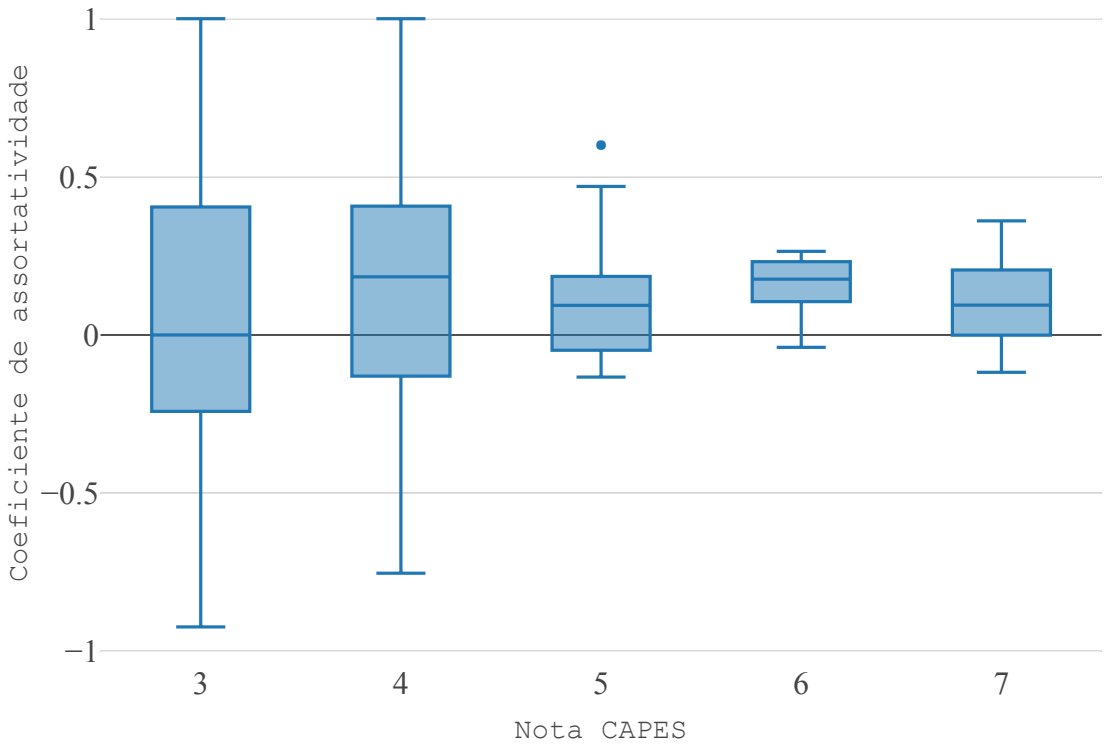

Figura 27 – Média do Coeficiente de Assortatividade

Fonte: Autoria própria

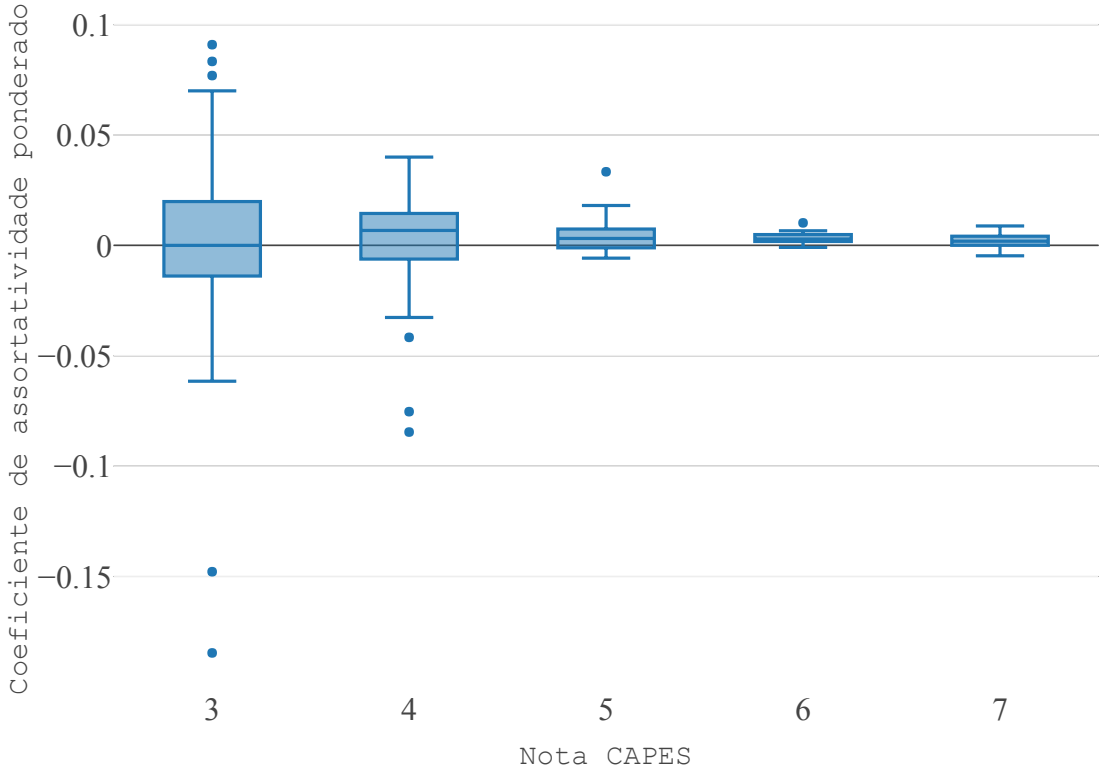

Figura 28 – Média do Coeficiente de Assortatividade Ponderado pelo Número de Vértices

Fonte: Autoria própria

5.10.8 Coeficiente de Clube Rico

O coeficiente de clube rico (*rich club*) é uma medida de vulnerabilidade e exibe a quantidade de grupo de *hubs* existentes na rede. Se um programa possuir vários pesquisadores com grande centralidade conectados de maneira em que os mesmos formam um grupo, esse programa terá um alto valor no coeficiente de clube rico, porém esse clube rico pode ser um ponto vulnerável da rede, pois caso a mesmo sofra um ataque exatamente nesse ponto, os danos serão maximizados, dessa forma, um programa com menor valor desse coeficiente pode ser um programa onde os pesquisadores de grande influência estão melhor espalhados pela rede, tornando-as assim mais robustas.

Essa medida pode ser confundida muitas vezes com o coeficiente de aglomeração, mas na verdade o que as diferem é que nessa medida (clube rico) são analisados os grupos de *hubs* existentes na rede; já na de aglomeração analisa-se apenas os *hubs* isoladamente, ainda que os mesmos não formam grupos entre si.

A interpretação que temos ao analisar os dados resultantes deste projeto é que os programas melhor avaliados pelas CAPES possuem menor valor de coeficiente de clube rico, portanto, esses programas possuem uma menor quantidade de grupo de *hubs*, essa medida possui também correlação com o coeficiente de aglomeração (*cluster*), pois ela

basicamente contabiliza qual a quantidade de grupo de *clusters* na rede. Na Figura 29 estão apresentadas as médias dos coeficientes de clube rico de cada categoria de avaliação CAPES, corroborando com as afirmações acima descritas, nessa medida não exibimos os valores ponderados porque quando aplicada a ponderação, a tendência dos dados mantém-se a mesma mudando apenas a grandeza dos números.

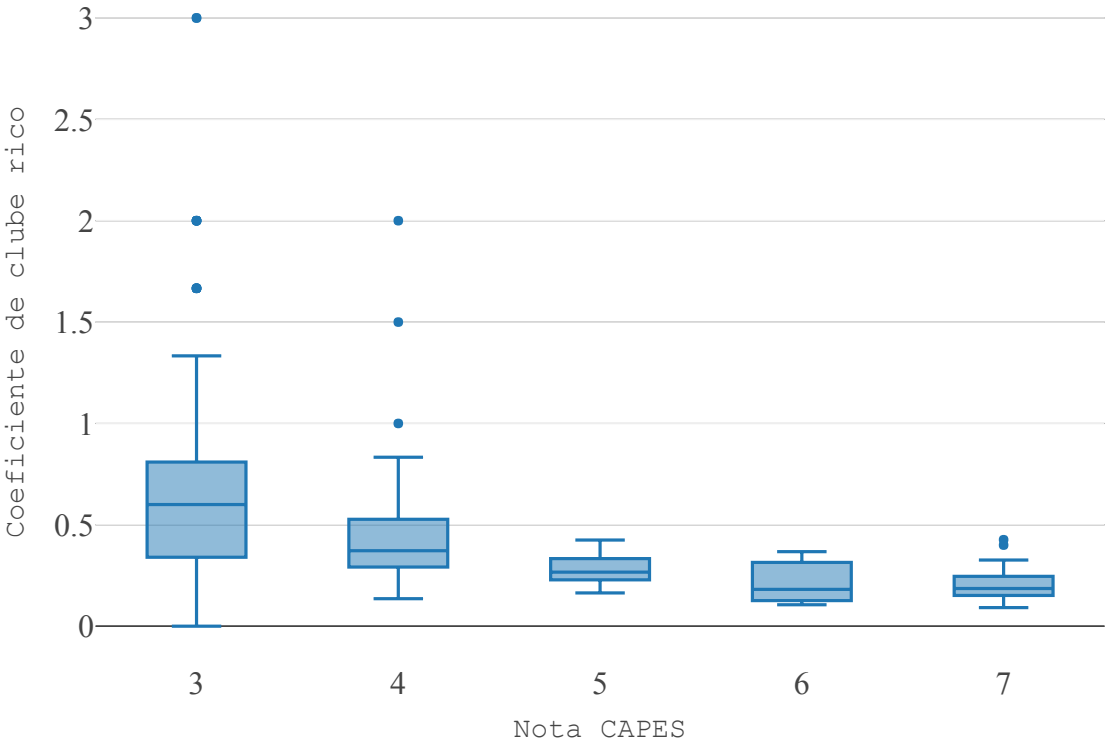

Figura 29 – Média do Coeficiente de Clube Rico

Fonte: Autoria própria

5.10.9 Vulnerabilidade

Essa função examina a vulnerabilidade da rede, no contexto deste projeto, uma rede com maior vulnerabilidade é uma rede onde caso um pesquisador aleatório seja removido a estrutura da rede sofrerá grandes alterações, pois um pesquisador com grande influência na rede pode ser um ponto de vulnerabilidade para essa rede. Com os dados resultantes deste trabalho é possível afirmar que programas com maiores avaliações CAPES possuem menor vulnerabilidade, isso é, programas mais bem avaliados são menos dependentes de um pesquisador específico, sendo que caso o mesmo seja removido do programa, esse programa continuará com sua estrutura levemente alterada, tornando-os programas mais estáveis e sólidos do que os programas menos bem avaliado pela Nota CAPES, sendo que esses possuem grande dependência de um grupo de pesquisador que torna um ponto de vulnerabilidade ao programa.

Conforme explicado anteriormente, essa medida analisa individualmente cada vértice e retorna o seu valor de vulnerabilidade, sendo assim, essa medida retornou um vetor com os valores de vulnerabilidade de toda a rede, desse vetor fora extraídos os valores mínimos, os valores médio e o valor máximo do mesmo, sendo que esse valor máximo foi o que os algoritmos apontaram como tendo maior relevância na classificação junto a Nota CAPES, a média desses valores está representada na Figura 30, onde é possível observar a tendência de queda da vulnerabilidade ante o aumento da classificação CAPES dos programas.

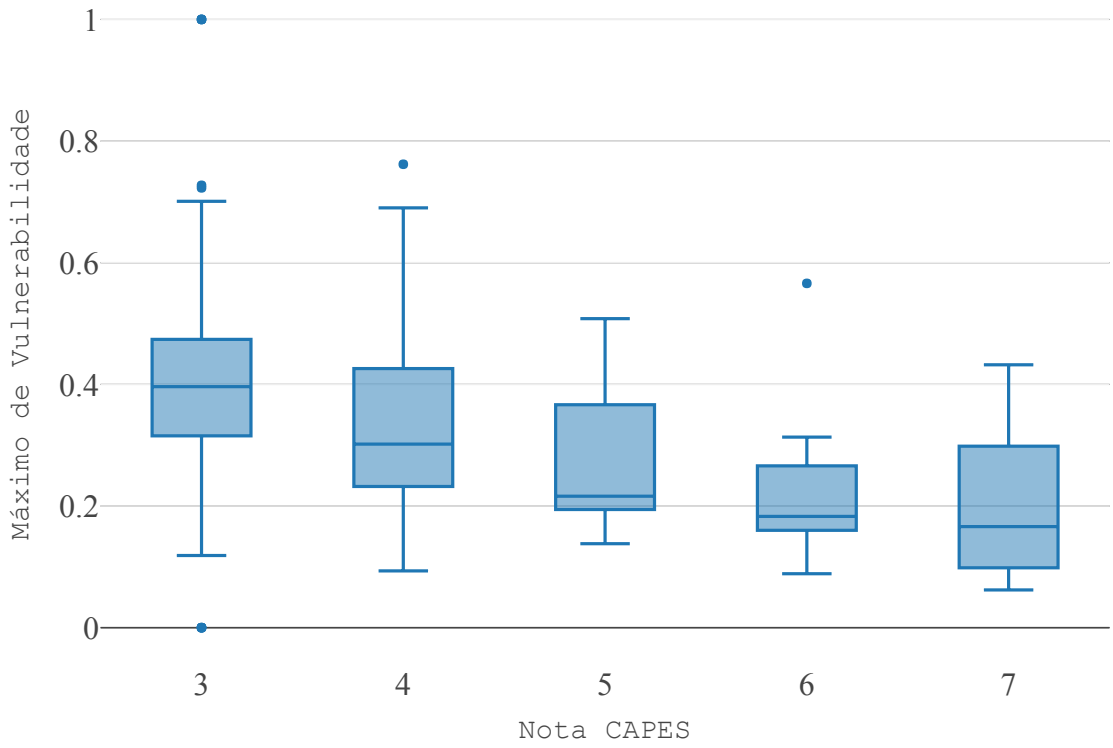

Figura 30 – Média do Máximo de Vulnerabilidade em cada Programa

Fonte: Autoria própria

5.10.10 Coeficiente de Variação

O coeficiente de variação é o oposto do coeficiente de assortatividade, pois ele demonstra qual a variação da rede levando em consideração os atributos de cada vértice. Os dados resultantes neste projeto reafirma que programas com maior Nota CAPES, possuem maior variação entre os pesquisadores, conforme expresso na Figura 31, exceto entre os programas de NOTA6 e os programas de NOTA7 onde os de menor nota possuem maior variação.

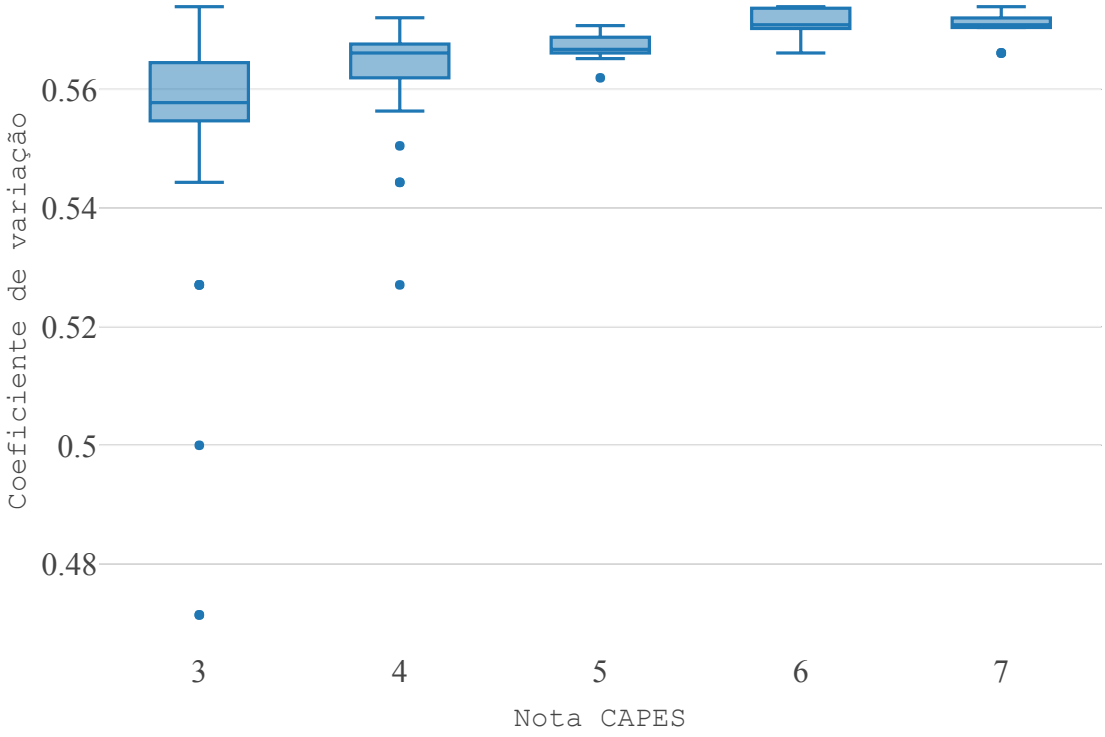

Figura 31 – Média do Coeficiente de Variação de cada Programa.

Fonte: Autoria própria

5.11 Análises temporais

Nessa seção são expostos os resultados das análises dos programas/medidas levando em consideração o tempo das 3 janelas de avaliação CAPES, para isso, filtramos os dados para separar cada grupo de avaliação CAPES (NOTA3, NOTA4, ..., NOTA7) separamos também cada um dos 3 períodos de avaliação CAPES, com isso obtemos uma média de cada medida e podemos comparar para verificar a evolução ao longo do tempo.

5.11.1 Características de Maior Relevância

Como neste projeto trabalhamos com uma grande quantidade de medidas, para realizar a análise temporal utilizamos somente as 9 medidas de maior importância na definição da Nota CAPES de acordo com o modelo de seleção de características, conforme exposto na Seção 5.4. Na Figura 32 são expressas as análises temporais de cada uma das 9 medidas, onde cada medida está em um gráfico de linhas e dentro de cada gráfico são separados os grupos de avaliação CAPES de acordo com a cor de cada linha, ainda dentro do gráfico cada coluna corresponde a um período de avaliação CAPES sendo eles P1 (2007 à 2009), P2 (2010 à 2012) e P3 (2013 à 2016).

As medidas são explicadas abaixo:

- pode se observar que o número de vértices representando o número de pesquisadores diminui nos programas de NOTA3 e NOTA6, nos demais ele cresce ao longo dos períodos, a mesma tendência aplica-se ao número de arestas (quantidade de publicações);
- o coeficiente de aglomeração só cresce nos programas de NOTA4;
- o caminho médio segue a mesma tendência padrão do número de vértices e arestas porém de maneira mais sutil com exceção dos programas de NOTA4 que possuem ascensão nessa medida;
- a centralidade de intermediação é bem próxima ao caminho médio porém a queda dos programas de NOTA6 e a ascensão dos programas de NOTA7 é muito mais acentuada;
- o coeficiente de aglomeração obedece um padrão em quase todos os programas onde o mesmo vem decaindo ao longo do tempo, com exceção dos programas de NOTA6 que do período 1 para o período 2 tem uma ascensão;
- o coeficiente de clube rico nos programas de NOTA3 e NOTA7 tem a tendência de decair, já nos programas de NOTA6 tem a tendência de ascensão nos programas de NOTA4 ele tem uma queda do período 1 para o período 2, porém tem uma ascensão do 2 para o 3 e por último os programas NOTA5 tem um crescimento do período 1 para o período 2 e depois tem estabilidade do 2 para o 3;
- o coeficiente de variação de todos os programas é elevado por ele fica bem claro a separação entre os programas todos os programas têm uma queda ao longo do tempo com exceção dos programas de NOTA7 que têm uma ascensão ao longo do tempo;
- o índice de senioridade tem ascensão em todos os programas ao longo do tempo, com exceção dos programas de NOTA6 que do período 2 para o período 3 tem uma queda.

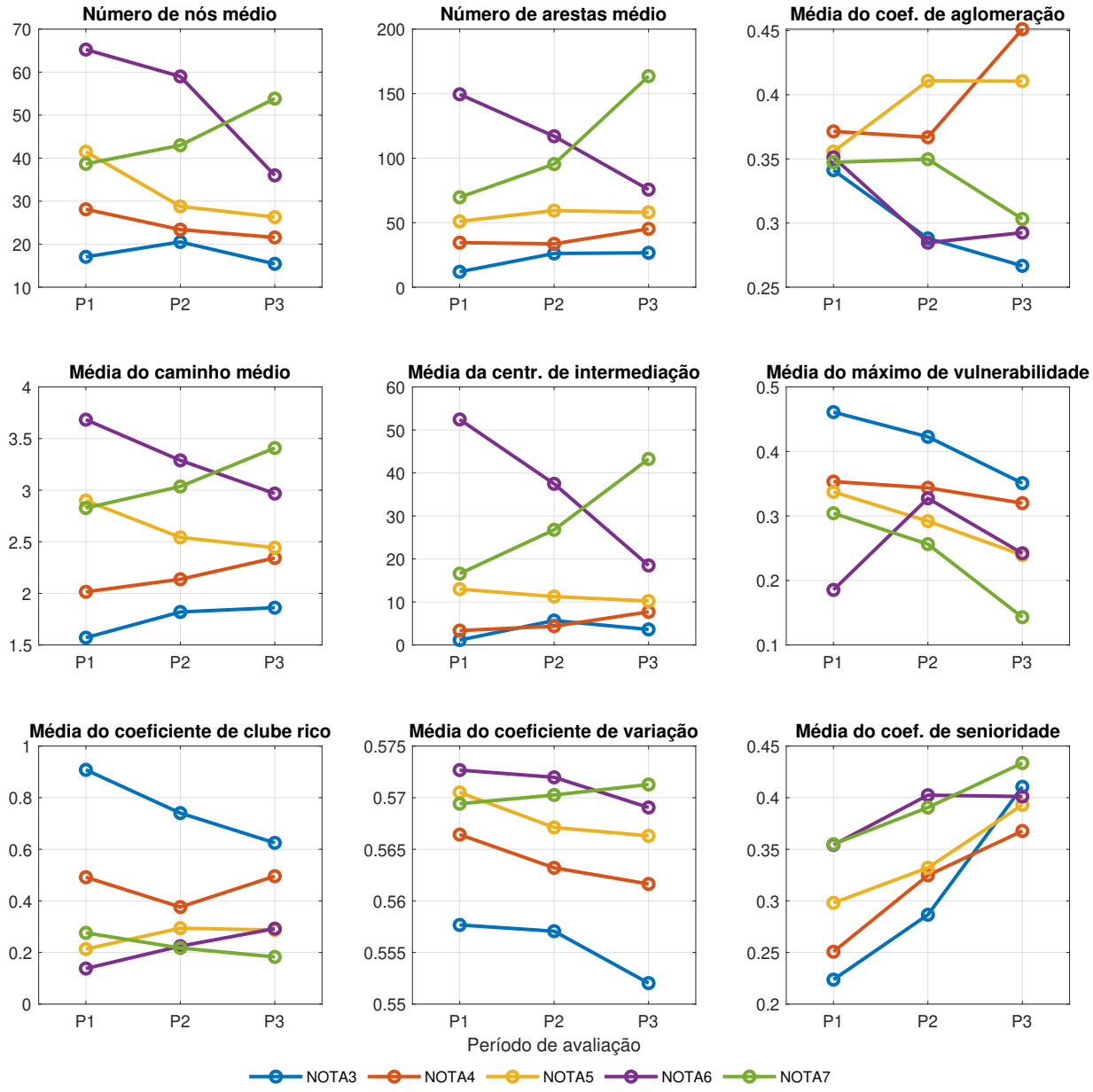

Figura 32 – Análises temporais de cada medida de grande relevância ao longo dos 3 períodos CAPES.

Fonte: Autoria própria

5.11.2 Índices Propostos de Colaboração

Os índices de colaboração que foram propostos neste projeto também foram analisados ao longo dos 3 períodos de avaliação CAPES e os resultados foram que: o Índice de Primeiro Autor tem uma tendência de queda em todos os programas com exceção dos programas de NOTA6 que teve uma ascensão do período 1 para o 2; o Índice de colaboração tem ascensão do período 1 para o 2 exceto para os programas NOTA6 que é o inverso, do período 2 para o 3 tem queda ou mantém se estável exceto os programas

NOTA6 que tem grande ascensão, nesse índice o destaque está para os programas de NOTA5 que do período 1 para o período 2 tem uma ascensão proeminente; o Índice de Senioridade possui uma tendência padrão de ascensão em todos os programas com exceção dos programas NOTA6 que possuí uma leve decaída do período 2 para o período 3, essas tendências são expressas na Figura 33 onde é possível verificar todas as linhas com os grupos de programas e as colunas com os índices/períodos.

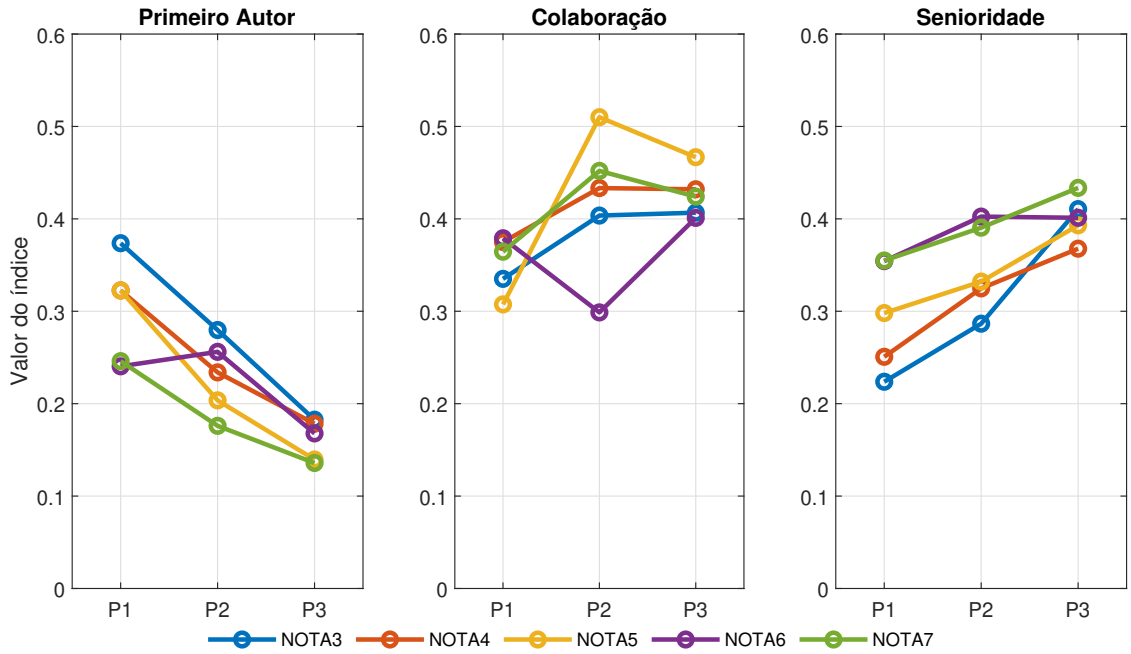

Figura 33 – Média dos Índices Propostos ao Longo dos 3 Períodos de Avaliação CAPES.

Fonte: Autoria própria

5.12 Discussões

No Capítulo 1.2 foram levantadas algumas hipóteses e nesta seção as mesmas são listadas a fim de discutir sobre sua validação no contexto dos programas brasileiros em Ciência da Computação.

1. **Redes de programas com diferentes notas de avaliação CAPES possuem diferença significativa em sua estrutura e dinâmica:**

   Hipótese confirmada, pois observou-se que em todas as medidas relevantes é diferenciado os valores conforme a classificação CAPES dos programas.

2. **Programas com maiores notas CAPES possuem maior colaboração interna (intraprogramas) que os demais:**

   Hipótese confirmada, pois como é possível observar nos resultados, programas mais bem avaliados possuem maior quantidade de arestas, mesmo ponderando a medida

e levando em consideração a quantidade de pesquisadores, é possível observar que esses programas possuem menor valor na média de pesquisadores por publicações o que confirma a hipótese levantada.

3. **Programas com notas CAPES mais baixas possuem um maior Índice de Primeiro Autor:**

   Hipótese confirmada, como é possível observar na Figura 9, quanto menor a classificação dos programas, maior é o valor desse índice, o que indica que os programas com a menor avaliação possuem mais integrantes que são citados como primeiro autores nas publicações;

4. **Pesquisadores de programas com Nota CAPES intermediárias possuem maior Índice de Colaboração:**

   Hipótese confirmada, como observa-se nos resultados deste projeto, os maiores valores do Índice de Colaboração estão em programas intermediários, sendo que os programas NOTA5 destaca-se dos demais como tendo o valor mais elevado.

5. **Programas com notas CAPES mais elevadas possuem um grande número de pesquisadores com alto Índice de Senioridade:**

   Hipótese confirmada, pois o Índice de Senioridade possui alta correlação com a Nota CAPES, sendo que o mesmo foi apontado por vários modelos de seleção de características como um atributo de grande importância e grande correspondência com a Nota CAPES.

6 CONCLUSÕES E TRABALHOS FUTUROS

Analisamos dados de 1.644 pesquisadores agrupados em 62 programas de pós-graduação em Ciência da Computação no Brasil em um período de 10 anos (2007 à 2016), aplicamos 38 medidas de análises topológicas de rede juntamente com 4 novas medidas propostas e podemos observar que as medidas estudadas podem ser significativas na análise dos programas. Medidas de quantidade e centralidade de vértices, são as medidas mais relevantes na classificação de programas usando como base a nota de avaliação da CAPES, medidas de vulnerabilidade também podem oferecer dados de grande relevância para a avaliação desses programas, além disso, medidas que dizem respeito ao tamanho das redes/caminhos também foram importantes neste projeto, essas medidas reunidas podem classificar os programas com uma alta taxa de acurácia.

Foi proposta uma medida quantitativa que analisou a média de pesquisadores para cada publicação nos programas, com isso foi possível observar e traçar um padrão de comportamento onde, programas mais bem avaliados qualitativamente possuem menor média de pesquisadores por publicação, dessa forma, pode-se afirmar que esses programas são mais produtivos que os demais, uma vez que a quantidade individual de publicações é mais elevada que os demais programas.

Dos três índices de colaboração propostos neste trabalho o Índice de Senioridade destacou-se oferecendo uma alta afinidade com a Nota CAPES, com isso é possível concluir que programas com maior Nota CAPES possuem maior número de pesquisadores que são listados, por último em publicações, indicando que esses pesquisadores possuem maior quantidade de participação como orientadores em projetos de pesquisa. Embora os demais índices não tenha a mesma linearidade do Índice de Senioridade, eles também são importantes para analisar os programas e podem fornecer uma boa visão do padrão de comportamento dos programas, uma vez que cada índice foi relevante para um determinado grupo de programas, pois enquanto o Índice de Senioridade teve seus picos em programas de maior avaliação; o Índice de Colaboração atingiu o ápice em programas de avaliação intermediária, o que indica que nesses programas os pesquisadores são mais citados como colaboradores nos projetos; o Índice de Primeiro Autor teve seu maior valor em programas de menor avaliação, o que indica que nesses programas grande parte dos pesquisadores aparecem como primeiro autor.

Os programas possuem um comportamento semelhante nas análises, todavia os grupo de programas classificados com a NOTA6 é o que mais difere dos demais sendo que por diversas vezes onde todos os outros tinham tendência de ascensão em alguma medida topológica ele tinha declínio e o contrário também, portanto, este grupo de programas possui um comportamento atípico aos demais.

Foi observado que todas as medidas utilizadas tiveram variações significativas ao longo dos três períodos de avaliação CAPES, o que indica que os programas são dinâmicos e por mais que um programa não possua alteração na nota CAPES dentre destes três períodos, ainda sim esse programa possui alterações em suas redes de colaboração.

Pode-se então concluir que programas mais bem avaliados em comparação com os menos avaliados possuem: maior quantidade de autores; maior quantidade de publicações; menor caminho médio, isto é os pesquisadores estão mais próximos aos outros; menor quantidade de pesquisadores com grande número de conexão, conectados diretamente entre si; menor quantidade de comunidades de pesquisadores com grande número de conexões entre si; maior quantidade de pesquisadores centrais, isto é, pesquisadores que conectam outros pesquisadores (influentes); maior proximidade entre os pesquisadores, portanto, um pesquisador pode estar à poucos outros pesquisadores de distância de outros; menor semelhança entre o perfil dos pesquisadores; maior variação entre as medidas dos pesquisadores; menor vulnerabilidade, portanto, se um pesquisador for removido o impacto no programa é menor; menor quantidade média de autores por publicações, portanto maior produtividade do programa; maior valor de Índice de Senioridade; menor valor de Índice de Primeiro Autor.

As análises realizadas como resultado deste projeto podem ser significativas na classificação dos programas, os coordenadores podem usar essas informações para avaliar os pontos em que podem haver melhorias em seus programas e orientado aos dados buscar estratégias para a evolução dos mesmos e a ascensão na avaliação da CAPES.

A metodologia deste projeto pode ser estendida e aplicada em outras áreas, com isso, pode-se realizar comparações entre as áreas observando se os padrões se mantém ou se há diferença entre eles.

Foi observado também que os programas de NOTA6 possuem um comportamento atípico se comparado com os demais programas, como não é o escopo deste trabalho um trabalho futuro pode analisar esses programas e entender o que faz esses programas diferentes dos demais.

Neste projeto os esforços foram concentrados em dados de publicações de 2 tipos, todavia, outros tipos de produções podem ser avaliadas, como por exemplo, participações em bancas, correção de artigos, orientações de projetos dentre outras medidas também informadas no Currículo Lattes.

REFERÊNCIAS

AHO, A. V.; HOPCROFT, J. E.; ULLMAN, J. D. *The Design and Analysis of Computer Programs*. 1st. ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1974. 450 p. ISBN 9781475737998. Citado na página 21.

ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. *Reviews of Modern Physics*, American Physical Society, v. 74, n. 1, p. 47–97, jan 2002. Disponível em: <https://link.aps.org/doi/10.1103/RevModPhys.74.47>. Citado 2 vezes nas páginas 22 e 27.

ALVARENGA, P. J. L. *Um Estudo Sobre Referências Bibliográficas Na Área De Ciência Da Computação*. Tese (Master Degree) — Universidade Federal de Minas Gerais, 2007. Citado 2 vezes nas páginas 18 e 34.

ALVES, A. D.; YANASSE, H. H.; SOMA, N. Y. SUCUPIRA: A system for Information extraction of the Lattes Platform to identify academic social networks. *6th Iberian Conference on Information Systems and Technologies (CISTI 2011)*, p. 1–6, 2011. Citado 2 vezes nas páginas 18 e 35.

ALVES, A. D.; YANASSE, H. H.; SOMA, N. Y. LattesMiner: uma linguagem de domínio específico para extração automática de informações da Plataforma Lattes. *XII Workshop de Computação Aplicada -WORCAP 2012*, 2012. Disponível em: <http://mtc-m18.sid.inpe.br/col/sid.inpe.br/mtc-m18/2013/01.15.16.10/doc/worcap2012{_}submission{_}61-AlexandreD.Alv>. Citado na página 34.

ALVES, A. D. et al. Mapeamento de Competências Tecnológicas: Um Sistema de Informação para auxiliar o processo de decisão da Petrobras baseado na Plataforma Lattes. In: *Information Systems and Technologies (CISTI), 2015 10th Iberian Conference on*. [S.l.: s.n.], 2015. p. 1–6. Citado na página 35.

BARABÁSI, A.-L. Linked: A nova Ciência dos Networks. *São Paulo: Leopardo*, p. 256, 2009. Citado 2 vezes nas páginas 16 e 24.

BARABASI, A.-L.; ALBERT, R. Emergence os Scaling in Random Networks. *Science*, v. 286, n. October, p. 509–512, 1999. Disponível em: <http://barabasi.com/f/67.pdf>. Citado na página 23.

BARABÁSI, A.-L.; BONABEAU, E. Scale-Free Networks. *Scientific American*, v. 288, n. 5, p. 60–69, 2003. ISSN 0036-8733. Disponível em: <http://www.nature.com/doifinder/10.1038/scientificamerican0503-60>. Citado na página 22.

BARATA, R. d. C. B. Dez coisas que você deveria saber sobre o Qualis. *Revista Brasileira de Pós Graduação*, v. 13, n. 1, p. 13–40, 2016. ISSN 2358-2332. Citado na página 33.

BAUER, L. *Estimação do coeficiente de correlação de spearman ponderado*. Tese (Master Degree) — Universidade Federal do Rio Grande do Sul, 2007. Disponível em: <http://hdl.handle.net/10183/11499>. Citado na página 31.

BISHOP, C. M. *Neural Networks for Pattern Recognition*. Clarendon Press, 1995. 482 p. (Advanced Texts in Econometrics). ISBN 9780198538646. Disponível em: <https://books.google.com.br/books?id=-aAwQO{_}-r>. Citado na página 30.

BOCCALETTI, S. et al. Complex networks: Structure and dynamics. *Physics Reports*, v. 424, n. 4-5, p. 175–308, 2006. ISSN 03701573. Citado 5 vezes nas páginas 22, 23, 24, 25 e 26.

BONDY, J. A.; MURTY, U. S. R. *Graph Theory With Applications*. 1. ed. Ontario: Springer Publishing Company, Incorporated, 1976. 270 p. ISBN 0333226941. Citado na página 21.

BORDIN, A. S.; GONÇALVES, A. L.; TODESCO, J. L. Análise da colaboração científica departamental através de redes de coautoria. *Perspectivas em Ciência da Informação*, v. 19, n. 2, p. 37–52, jun 2014. ISSN 1413-9936. Disponível em: <http://www.scielo.br/scielo.php?script=sci{_}arttext{&}pid=S1413-99362014000200004{&}lng=p>. Citado 4 vezes nas páginas 17, 34, 39 e 41.

BORDONS, M. et al. The relationship between the research performance of scientists and their position in co-authorship networks in three fields. *Journal of Informetrics*, Elsevier Ltd, v. 9, n. 1, p. 135–144, jan 2015. ISSN 17511577. Disponível em: <http://dx.doi.org/10.1016/j.joi.2014.12.001https://linkinghub.elsevier.com/retrieve/pii/S1751157714001138>. Citado na página 40.

BREIMAN, L. Bagging predictors. *Machine Learning*, v. 24, n. 2, p. 123–140, aug 1996. ISSN 0885-6125. Disponível em: <http://link.springer.com/10.1007/BF00058655>. Citado na página 55.

BREIMAN, L. *Out-of-Bag Estimation*. Berkeley, CA: [s.n.], 1996. Disponível em: <https://www.stat.berkeley.edu/{~}breiman/OOBestimation.p>. Citado na página 55.

BREIMAN, L. Random Forests. *Machine Learning*, Cambridge, v. 45, n. 1, p. 5–32, 2001. ISSN 1573-0565. Disponível em: <https://doi.org/10.1023/A:1010933404324>. Citado na página 55.

CAMPOS, T. E. de. *Técnicas de seleção de características com aplicações em reconhecimento de faces*. 160 p. Tese (Doutorado) — Universidade de São Paulo, São Paulo, may 2001. Disponível em: <http://www.teses.usp.br/teses/disponiveis/45/45134/tde-23112001-085134/>. Citado na página 28.

CAPES. *Plataforma Sucupira*. 2014. Disponível em: <http://www.capes.gov.br/avaliacao/plataforma-sucupira>. Citado na página 35.

CAPES. *Avaliação da CAPES aponta crescimento da pós-graduação brasileira*. 2017. Citado na página 33.

CAPES. *Avaliação da CAPES aponta crescimento da pós-graduação brasileira*. 2017. Disponível em: <http://www.capes.gov.br/sala-de-imprensa/noticias/8558-avaliacao-da-capes-aponta-crescimento-da-pos-graduacao-brasileira>. Citado na página 33.

CAPES. *Avaliação da CAPES aponta crescimento da pós-graduação brasileira*. 2017. Disponível em: <http://www.capes.gov.br/sala-de-imprensa/noticias/8558-avaliacao-da-capes-aponta-crescimento-da-pos-graduacao-brasileira>. Citado na página 46.

CHAVES, A. S. et al. *Relatório da Comissão de Integridade de Pesquisa do CNPq*. [S.l.], 2011. 7 p. Disponível em: <http://www.cnpq.br/documents/10157/a8927840-2b8f-43b9-8962-5a2ccfa74dda>. Citado na página 34.

CHEN, M.-S.; HAN, J.; YU, P. S. Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, v. 8, n. 6, p. 866–883, 1996. ISSN 10414347. Disponível em: <http://ieeexplore.ieee.org/document/553155/>. Citado na página 28.

CIA, C. I. A. *The World Factbook*. 2018. Disponível em: <https://www.cia.gov/library/publications/the-world-factbook/fields/2122.html>. Citado na página 16.

CLAUSIUS, R. J. E. *The Mechanical Theory Of Heat*. London: Kessinger Publishing, LLC, 1879. 424 p. ISBN 1437414001. Citado na página 30.

COHEN, J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, v. 70, n. 4, p. 213–220, 1968. ISSN 1939-1455. Disponível em: <http://doi.apa.org/getdoi.cfm?doi=10.1037/h0026256>. Citado na página 65.

COLIZZA, V. et al. Detecting rich-club ordering in complex networks. *Nature Physics*, v. 2, n. 2, p. 110–115, feb 2006. ISSN 1745-2473. Disponível em: <http://arxiv.org/abs/physics/0602134http://dx.doi.org/10.1038/nphys209http://www.nature.com/articles/nphys209>. Citado na página 27.

COSTA, L. d. F. et al. Characterization of complex networks: A survey of measurements. *Advances in Physics*, v. 56, n. 1, p. 167–242, 2007. ISSN 0001-8732. Disponível em: <http://dx.doi.org/10.1080/00018730601170527>. Citado 2 vezes nas páginas 22 e 23.

CSARDI, G. *iGraph: network analysis library*. [S.l.], 2019. R package version 1.0.1. Disponível em: <https://github.com/igraph/rigraph>. Citado na página 57.

CSARDI, G.; NEPUSZ, T. The igraph software package for complex network research. *InterJournal*, Complex Systems, p. 1695, 2006. Disponível em: <http://igraph.org>. Citado na página 57.

DIGIAMPIETRI, L. A. et al. BraX-Ray: An X-Ray of the Brazilian Computer Science Graduate Programs. *PLoS ONE*, v. 9, n. 4, p. e94541, apr 2014. ISSN 1932-6203. Disponível em: <https://dx.plos.org/10.1371/journal.pone.0094541>. Citado 2 vezes nas páginas 40 e 41.

EDUCAÇÃO, M. da. *Qual a diferença entre pós-graduação lato sensu e stricto sensu?* 2018. Disponível em: <http://portal.mec.gov.br/component/content/article?id=13072:qual-a-diferenca-entre-pos-graduacao-lato-sensu-e-stricto-sensu>. Citado na página 33.

EULER, L. *Solutio problematis ad geometrian situs pertinentis*. 1736. 128–140 p. Citado na página 16.

FACEBOOK. *Facebook - Facebook Reports First Quarter 2018 Results*. 2018. Disponível em: <https://investor.fb.com/investor-news/press-release-details/2018/Facebook-Reports-First-Quarter-2018-Results/default.aspx>. Citado na página 17.

FACEBOOK. *Company Info | Facebook Newsroom*. 2019. Disponível em: <https://newsroom.fb.com/company-info/>. Citado na página 17.

FERRAZ, R. R. N.; QUONIAM, L.; ALVARES, L. M. A. D. R. Avaliação de redes multidisciplinares com a ferramenta scriptlattes: os casos da nanotecnologia, da dengue e de um programa de pós-graduação Stricto Sensu em Administração. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação*, v. 19, n. 40, p. 67, 2014. ISSN 1518-2924. Disponível em: <https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2014v19n40p67>. Citado na página 35.

FRANK, E.; HALL, M. A.; WITTEN, I. H. The WEKA workbench. In: *Data Mining*. Elsevier, 2017. p. 553–571. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/B9780128042915000246>. Citado na página 37.

FREEMAN, L. C. A Set of Measures of Centrality Based on Betweenness. *Sociometry*, v. 40, n. 1, p. 35, mar 1977. ISSN 00380431. Disponível em: <https://www.jstor.org/stable/3033543?origin=crossref>. Citado na página 26.

FREEMAN, L. C. Centrality in Social Networks Conceptual Clarification. *Social Networks*, v. 1, n. 1968, p. 215–239, 1978. Citado 3 vezes nas páginas 26, 27 e 58.

GABARDO, A. C. *Análise de Redes Sociais - Uma Visão Computacional*. 1. ed. São Paulo: Novatec, 2015. 144 p. ISBN 978-85-7522-417-5. Citado 6 vezes nas páginas 16, 17, 23, 24, 25 e 75.

GATTI, B. et al. O modelo de avaliação da CAPES. *Revista Brasileira de Educação*, n. 22, p. 137–144, apr 2003. ISSN 1413-2478. Disponível em: <http://www.scielo.br/scielo.php?script=sci{_}arttext{&}pid=S1413-24782003000100012{&}lng=p>. Citado na página 18.

GROSSMAN, J. W.; ARBOR, A. On a Portion of the Well-Known Collaboration Graph. *Congressus Numerantium*, v. 108, 1995. Citado na página 52.

GUEDES, C. A. *Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Currículo Lattes: Perguntas e Respostas.* 2001. 1–4 p. Disponível em: <http://www.pucrs.campus2.br/manuais/dicas{_}lattes.> Citado na página 34.

GUYON, I. et al. Design of the 2015 ChaLearn AutoML challenge. In: *2015 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2015. p. 1–8. ISBN 978-1-4799-1960-4. Disponível em: <http://ieeexplore.ieee.org/document/7280767/>. Citado na página 31.

GUYON, I.; ELISSEEFF, A. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research (JMLR)*, v. 3, n. 3, p. 1157–1182, 2003. ISSN 00032670. Citado na página 28.

HALL, M. et al. The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, v. 11, n. 1, p. 10, nov 2009. ISSN 19310145. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/14733285.2016.1271943http://portal.acm.org/citation.cfm?doid=1656274.1656278>. Citado na página 37.

HALL, M. A. *Correlation-based Feature Selection for Machine Learning*. Tese (Doutorado) — The University of Waikato, 1999. Citado na página 62.

HILÁRIO, C. M.; GRÁCIO, M. C. C.; GUIMARÃES, J. A. C. Aspectos éticos da coautoria em publicações científicas. *Em Questão*, v. 24, n. 2, p. 12, apr 2018. ISSN 1808-5245. Disponível em: <http://seer.ufrgs.br/index.php/EmQuestao/article/view/76312>. Citado 2 vezes nas páginas 18 e 34.

HO, T. K. Random decision forests. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. IEEE Comput. Soc. Press, 1995. v. 1, p. 278–282. ISBN 0-8186-7128-9. Disponível em: <http://ieeexplore.ieee.org/document/598994/>. Citado na página 55.

HOLLAND, P. W.; LEINHARDT, S. Transitivity in Structural Models of Small Groups. *Comparative Group Studies*, v. 2, n. 2, p. 107–124, may 1971. ISSN 0010-4108. Disponível em: <http://journals.sagepub.com/doi/10.1177/104649647100200201>. Citado na página 25.

HORTA, J. S. B.; MORAES, M. C. M. de. O sistema CAPES de avaliação da pós-graduação: da área de educação à grande área de ciências humanas. *Revista Brasileira de Educação*, n. 30, p. 95–116, dec 2005. ISSN 1413-2478. Disponível em: <http://www.scielo.br/scielo.php?script=sci{_}arttext{&}pid=S1413-24782005000300008{&}lng=p>. Citado na página 18.

HU, Y.-j.; HUANG, S.-w. Challenges of automated machine learning on causal impact analytics for policy evaluation. In: *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)*. IEEE, 2017. p. 1–6. ISBN 978-1-5090-6710-7. Disponível em: <http://ieeexplore.ieee.org/document/8343571/>. Citado na página 31.

JAMJUNTRA, L. et al. Social network user identification. *2017 9th International Conference on Knowledge and Smart Technology (KST)*, p. 132–137, 2017. Disponível em: <http://ieeexplore.ieee.org/document/7886120/>. Citado 2 vezes nas páginas 17 e 32.

KOTTHOFF, L.; LEYTON-BROWN, K. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *Journal of Machine Learning Research*, v. 17, p. 1–5, 2016. Citado na página 37.

KOVÁCS, I. A.; BARABÁSI, A.-L. Destruction perfected. *Nature*, v. 524, n. 7563, p. 38–39, 2015. ISSN 0028-0836. Disponível em: <http://www.nature.com/articles/524038a>. Citado na página 27.

KUMAR, B. T. H.; VIBHA, L.; VENUGOPAL, K. R. Web page access prediction using hierarchical clustering based on modified levenshtein distance and higher order Markov model. *2016 IEEE Region 10 Symposium (TENSYMP)*, p. 1–6, 2016. Disponível em: <http://ieeexplore.ieee.org/document/7519368/>. Citado na página 32.

LANCICHINETTI, A. et al. Characterizing the Community Structure of Complex Networks. *PLoS ONE*, v. 5, n. 8, p. e11976, 2010. ISSN 1932-6203. Disponível em: <http://dx.plos.org/10.1371/journal.pone.0011976>. Citado na página 21.

LATORA, V.; MARCHIORI, M. Vulnerability and protection of infrastructure networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, v. 71, n. 1, p. 1–4, 2005. ISSN 15393755. Citado 2 vezes nas páginas 27 e 57.

LEVENSHTEIN, V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, v. 10, n. 8, p. 707–710, 1966. ISSN 00385689. Citado na página 32.

LHOMME, S. *NetSwan: Network Strengths and Weaknesses Analysis*. [S.l.], 2015. R package version 0.1. Disponível em: <https://github.com/cran/NetSwan>. Citado na página 56.

LI, Z. et al. A Blockchain and AutoML Approach for Open and Automated Customer Service. *IEEE Transactions on Industrial Informatics*, IEEE, v. 15, n. 6, p. 3642–3651, jun 2019. ISSN 1551-3203. Disponível em: <https://ieeexplore.ieee.org/document/8649758/>. Citado na página 31.

LOPES, F.; MARTINS, D.; CESAR, R. M. Feature selection environment for genomic applications. *BMC Bioinformatics*, v. 9, n. 1, p. 451, 2008. ISSN 1471-2105. Disponível em: <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-451>. Citado 2 vezes nas páginas 54 e 55.

LOPES, G. R. et al. Ranking Strategy for Graduate Programs Evaluation. *Proceedings of 7th International Conference on Information Technology and Application*, n. Icita, p. 59–64, 2011. Citado 4 vezes nas páginas 17, 33, 38 e 41.

LOPES, M. F. *Redes complexas de expressão gênica: síntese, identificação, análise e aplicações*. 110 p. Tese (Doutorado) — Universidade de São Paulo, 2011. Citado 7 vezes nas páginas 22, 23, 24, 28, 29, 30 e 31.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*. [s.n.], 1967. p. 281—-297. ISSN 1059-9495. Disponível em: <http://link.springer.com/10.1007/s11665-016-2173-6>. Citado na página 28.

MARINHEIRO, A.; BERNARDINO, J. Analysis of Open Source Business Intelligence Suites Análise de Suites Open Source Business Intelligence. 2013. Citado na página 36.

MATTHEWS, B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, v. 405, n. 2, p. 442–451, oct 1975. ISSN 00052795. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/0005279575901099>. Citado na página 66.

MCAULEY, J. J.; COSTA, L. d. F.; CAETANO, T. S. Rich-club phenomenon across complex network hierarchies. *Applied Physics Letters*, v. 91, n. 8, p. 084103, aug 2007. ISSN 0003-6951. Disponível em: <http://aip.scitation.org/doi/10.1063/1.2773951>. Citado na página 27.

MENA-CHALCO, J. P. et al. Brazilian Bibliometric Coauthorship Networks. *Journal of the Association for Information Science and Technology*, v. 65, n. 7, p. 1424–1445, 2014. ISSN 19335954. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.23010>. Citado 9 vezes nas páginas 17, 25, 26, 35, 38, 39, 41, 44 e 103.

MICHIE, D.; SPIEGELHALTER, D. J.; TAYLOR, C. C. *Machine Learning, Neural and Statistical Classification*. Upper Saddle River, NJ, USA: Prentice Hall, 1994. 289 p. Citado na página 28.

MILGRAM, S. The Small-World Problem. *Psychology Today*, v. 1, n. 1, p. 61–67, 1967. ISSN 00134252. Citado na página 24.

MITCHELL, T. M. *Machine Learning*. McGraw-Hill, 1997. 414 p. ISBN 9780071154673. Disponível em: <https://books.google.com.br/books?id=EoYBngEACAAJ>. Citado na página 31.

MUSSA, M. d. S. et al. Business Intelligence in Education: An Application of Pentaho Software. *Revista Produção e Desenvolvimento*, v. 4, p. 29–41, 2018. Citado na página 36.

NEWMAN, M. E. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, v. 64, n. 1, p. 7, 2001. ISSN 1063651X. Citado 4 vezes nas páginas 17, 18, 38 e 41.

NEWMAN, M. E. J. Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, v. 64, n. 1, p. 016131, jun 2001. ISSN 1063-651X. Disponível em: <https://link.aps.org/doi/10.1103/PhysRevE.64.016131>. Citado 6 vezes nas páginas 17, 18, 26, 38, 41 e 52.

OLIVEIRA, A. B. et al. Comparação entre o Qualis/Capes e os índices H e G: o caso do portal de periódicos UFSC. *Informação & Informação*, v. 20, n. 1, p. 70, 2015. ISSN 1981-8920. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/17054>. Citado na página 33.

OPSAHL, T. et al. Prominence and Control: The Weighted Rich-Club Effect. *Physical Review Letters*, v. 101, n. 16, p. 168702, oct 2008. ISSN 0031-9007. Disponível em: <http://aip.scitation.org/doi/10.1063/1.2773951https://link.aps.org/doi/10.1103/PhysRevLett.101.168702>. Citado na página 27.

PAPER, C.; CATARINA, S. Gestão Estratégica de Informações Curriculares em ICTIs. n. September 2015, p. 0–15, 2012. Citado na página 34.

PETROIANU, A. Autoria de um trabalho científico. *Revista da Associação Médica Brasileira*, v. 48, n. 1, p. 60–65, mar 2002. ISSN 0104-4230. Disponível em: <http://www.scielo.br/scielo.php?script=sci{_}arttext{&}pid=S0104-42302002000100034{&}lng=pt{&}nrm>. Citado 2 vezes nas páginas 18 e 34.

PUDIL, P.; NOVOVIČOVÁ, J.; KITTLER, J. Floating search methods in feature selection. *Pattern Recognition Letters*, v. 15, n. 11, p. 1119–1125, nov 1994. ISSN 01678655. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/0167865594901279>. Citado na página 29.

ROSA, A. Modelos formais de comunicação. *Caleidoscópio: Revista de Comunicação e Cultura*, v. 0, n. 1, 2011. Disponível em: <http://revistas.ulusofona.pt/index.php/caleidoscopio/article/view/2183>. Citado na página 21.

RUSSELL, S. J.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. New Jersey: Pearson, 2002. 1132 p. ISBN 978-0137903955. Citado na página 28.

SALZBERG, S. L. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning*, v. 16, n. 3, p. 235–240, sep 1994. ISSN 0885-6125. Disponível em: <http://www.tandfonline.com/doi/full/10.3109/02699206.2010.515375http://link.springer.com/10.1007/BF00993309>. Citado na página 68.

SAMUEL, A. L. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, v. 3, n. 3, p. 210–229, jul 1959. ISSN 0018-8646. Disponível em: <http://ieeexplore.ieee.org/document/5392560/>. Citado na página 31.

SAVIĆ, M.; IVANOVIĆ, M.; JAIN, L. C. *Complex Networks in Software, Knowledge, and Social Systems*. [s.n.], 2019. v. 148. 235–275 p. ISBN 978-3-319-91194-6. Disponível em: <http://link.springer.com/10.1007/978-3-319-91196-0>. Citado na página 32.

SHANNON, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, v. 27, n. July 1928, p. 379–423, 1948. ISSN 07246811. Disponível em: <http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf>. Citado na página 30.

SHANNON, C. E.; WEAVER, W. The Mathematical Theory of Communication. *The mathematical theory of communication*, v. 27, n. 4, p. 117, 1963. ISSN 07246811. Citado na página 30.

SOMOL, P. et al. Adaptive floating search methods in feature selection. *Pattern Recognition Letters*, v. 20, n. 11-13, p. 1157–1163, nov 1999. ISSN 01678655. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0167865599000835>. Citado 2 vezes nas páginas 29 e 30.

SPEARMAN, C. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, v. 15, n. 1, p. 72–101, 1904. Citado na página 31.

SYMEONIDIS, A. L.; MITKAS, P. A. *Agent Intelligence Through Data Mining*. New York: Springer-Verlag, 2005. v. 14. 1009–1010 p. (Multiagent Systems, Artificial Societies, and Simulated Organizations, 479). ISSN 0007-1250. ISBN 0-387-24352-6. Disponível em: <http://link.springer.com/10.1007/b136000>. Citado na página 28.

THORNTON, C. et al. Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms. *Computing Research Repository - CoRR*, aug 2012. Disponível em: <http://arxiv.org/abs/1208.3719>. Citado na página 31.

THORNTON, C. et al. Auto-WEKA. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13*. New York, New York, USA: ACM Press, 2013. p. 847. ISBN 9781450321747. Disponível em: <http://dl.acm.org/citation.cfm?doid=2487575.2487629>. Citado na página 37.

VANZ, S. A. d. S. As redes de colaboração científica no Brasil (2004-2006). p. 204, 2009. Citado na página 18.

VIEW, M.; APRIL, C.; GOOG, N. Alphabet Announces First Quarter 2018 Results. v. 598, 2018. Citado na página 17.

WASSERMAN, S.; FAUST, K. *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press, 1994. 825 p. ISBN 9780511815478. Disponível em: <http://ebooks.cambridge.org/ref/id/CBO9780511815478>. Citado na página 26.

WATSON, C. G. *brainGraph: Graph Theory Analysis of Brain MRI Data*. [S.l.], 2019. R package version 2.7.0. Disponível em: <https://github.com/cwatson/brainGraph>. Citado na página 56.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world' networks. *Nature*, v. 393, n. 6684, p. 440–442, 1998. ISSN 00280836. Disponível em: <http://www.nature.com/nature/journal/v393/n6684/abs/393440a0.html>. Citado na página 24.

WEBB, A. R. *Statistical Pattern Recognition*. Chichester, UK: John Wiley & Sons, Ltd, 2002. v. 2. 668 p. ISSN 15524922. ISBN 9781119952954. Disponível em: <http://doi.wiley.com/10.1002/9781119952954>. Citado na página 28.

WINKLER, R.; BARR, A. *Google Reports Better-Than-Seen Revenue Growth - WSJ*. 2014. Disponível em: <http://www.wsj.com/articles/google-reports-22-revenue-growth-1405628219>. Citado na página 17.

WITTEN, I. H.; FRANK, E.; HALL, M. A. *Data Mining: Practical Machine Learning Tools and Techniques*. [S.l.]: Elsevier, 2011. 664 p. ISBN 9780123748560. Citado na página 37.

ZHANG, S.; HU, Y.; BIAN, G. Research on string similarity algorithm based on Levenshtein Distance. *Proceedings of 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference, IAEAC 2017*, n. 1, p. 2247–2251, 2017. Citado na página 32.

ZHENG, Y. et al. A novel hybrid algorithm for feature selection. *Personal and Ubiquitous Computing*, Personal and Ubiquitous Computing, v. 22, n. 5-6, p. 971–985, oct 2018. ISSN 1617-4909. Disponível em: <http://link.springer.com/10.1007/s00779-018-1156-z>. Citado na página 28.

ZHOU, S.; MONDRAGON, R. The Rich-Club Phenomenon in the Internet Topology. *IEEE Communications Letters*, v. 8, n. 3, p. 180–182, mar 2004. ISSN 1089-7798. Disponível em: <http://ieeexplore.ieee.org/document/1278314/>. Citado na página 27.

APÊNDICE A – DIAGRAMA DE ENTIDADE E RELACIONAMENTO DO BANCO DE DADOS

Neste Apêndice apresentamos o Diagrama de Entidade e Relacionamento do banco de dados gerado como resultado do processo de coleta e pré-processamento dos dados. Ao todo foram projetadas 14 tabelas sendo que as principais são: Pesquisadores, Publicações e Publicações de Pesquisadores que é a intersecção entre elas.

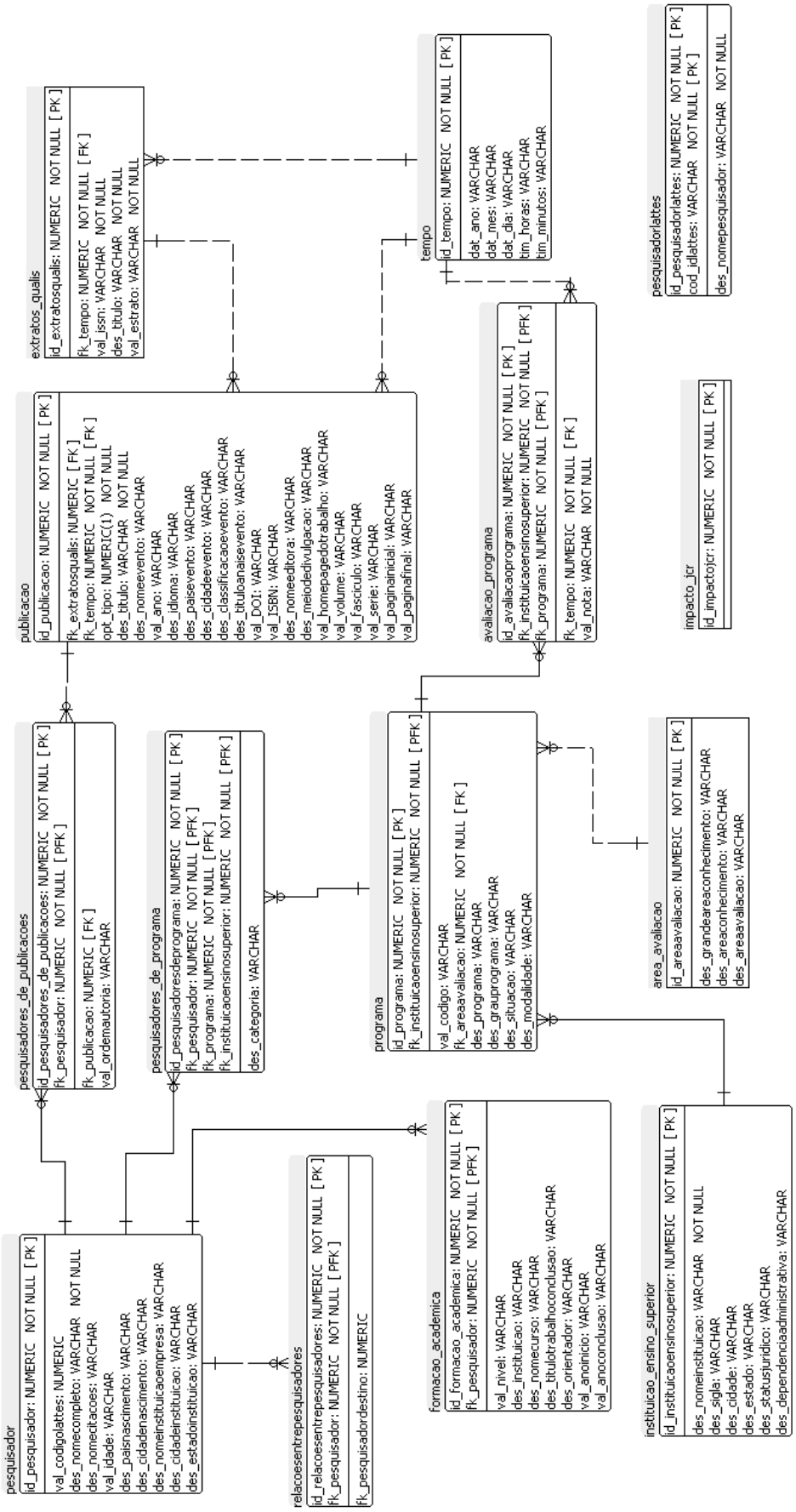

Figura 34 – Diagrama de Entidade e Relacionamento do Projeto

Fonte: Autoria própria

APÊNDICE B – DESCRIÇÃO DAS FONTES DE DADOS

Neste Apêndice são detalhadas as fontes de dados utilizados no projeto sendo que na Tabela 7 as mesmas são classificadas com um código, é informada sua origem e a descrição detalhada dela é exibida.

Tabela 7 – Detalhamento das fontes de dados utilizadas.

Código	**Origem**	**Descrição**
FD001	Sistema SUCUPIRA	Arquivo em formato CSV listando todos os docentes e seus respectivos programas, 89.255 registros no total.
FD002	Plataforma Lattes	Arquivo em formato CSV extraído da Plataforma Lattes descrito em (MENA-CHALCO et al., 2014), com todos os nomes completos dos pesquisadores cadastrados na plataforma e seus respectivos IDs Lattes, 4.55.770 registros no total.
FD003	Plataforma Lattes	Plataforma WEB onde todos os Currículos Lattes podem ser acessados e visualizados.