Metodologia Orientada a Ciência de Dados em Grafos para Avaliação de PPGs

Breno Santana Santos [*,**] **Ivanovitch Silva** [*] **Elvis Melo** [*]

[*] *Programa de Pós-Graduação em Engenharia Elétrica e de Computação, Universidade Federal do Rio Grande do Norte, RN, Natal, (e-mails: {breno.santos.038, elvis.melo.016}@ufrn.edu.br; ivanovitch.silva@ufrn.br).*
[**] *Universidade Federal de Sergipe, SE, Itabaiana, (e-mail: breno1005@hotmail.com)*

Abstract: The mapping and analysis of scientific knowledge makes it possible to identify the dynamics and/or growth of a particular research field or to support strategic decisions related to different research entities, based on bibliometric and/or scientometric indicators. However, with the exponential growth of scientific production, a systematic and data-oriented approach to the analysis of this large set of scientific data becomes increasingly essential. Thus, in this work, a data-oriented methodology was proposed, combining techniques of Machine Learning and Complex Network Analysis, for the extraction of implicit knowledge in scientific production bases, in addition to its validation through a study of case in the Brazilian Engineering IV. The results suggest the feasibility of the proposal, indicating the main researchers, prominent areas and partnership networks. Therefore, the proposed methodology has the potential to implement and expand strategic and proactive decisions of post-graduate programs aiming for a growing impact on society.

Resumo: O mapeamento e análise do conhecimento científico permite identificar a dinâmica e/ou crescimento de um determinado campo de pesquisa ou apoiar decisões estratégicas relacionadas às diversas entidades de pesquisa, a partir de indicadores bibliométricos e/ou cientométricos. Contudo, com o crescimento exponencial da produção científica, torna-se cada vez mais essencial uma abordagem sistemática e orientada a dados para análise desse conjunto volumoso de produções. Desse modo, neste trabalho, foi proposta uma metodologia orientada a dados, combinando técnicas de *Machine Learning* e de Análise de Redes Complexas, para extração de conhecimento implícito em bases de produções científicas, além de sua validação por meio de um estudo de caso nas Engenharias IV brasileiras. Os resultados sugerem a viabilidade da proposta, indicando os principais atores, áreas de destaque e redes de parcerias. Portanto, a metodologia proposta tem o potencial de instrumentar e expandir decisões estratégicas e proativas dos programas de pós-graduação visando um impacto, como consequência, cada vez maior na sociedade.

Keywords: Data Science; Complex Network Analysis; Machine Learning; Post-Graduate Programs; Scientometrics; Bibliometrics.

Palavras-chaves: Ciência de Dados; Análise de Redes Complexas; Aprendizado de Máquina; Programas de Pós-Graduação; Cientometria; Bibliometria.

1. INTRODUÇÃO

A ciência pode ser entendida como o conjunto de teorias e métodos, comprovados rigorosa e sistematicamente, contidos no estado da arte. Ademais, na maioria dos casos, o conhecimento é gerado pelos pesquisadores, de forma isolada ou combinada, a partir da adição de novos conceitos a um campo científico (Araújo et al., 2020; Sugimoto and Larivière, 2018). Para Sugimoto and Larivière (2018), a pesquisa científica é uma atividade social complexa, visto que é realizada em diversos ambientes por inúmeros atores, além de contar com diversas etapas e tarefas. Desse modo, tais atividades são inerentemente difíceis de mensurar diretamente e sem apoio de ferramentas computacionais. Logo, o processo de mensurar a pesquisa científica é feito por meio da tradução de suas atividades em unidades mensuráveis, os quais são comumente conhecidos na Bibliometria e Cientometria por indicadores (Camargo and Barbosa, 2018; Sugimoto and Larivière, 2018).

Bibliometria pode ser entendida como o estudo dos aspectos quantitativos da produção, disseminação e uso de informações registradas, bem como desenvolve padrões e modelos matemáticos para mensurar tais processos, podendo apoiar tomadas de decisões e elaborar previsões a partir de seus resultados (Araújo et al., 2020). Por outro lado, Cientometria é definida como o estudo dos aspectos quantitativos da ciência, seja como disciplina, seja como atividade econômica. Outrossim, ela utiliza indica-

dores quantitativos para mensurar uma disciplina específica (Guirado et al., 2020).

Com o crescimento exponencial de produções científicas, estudos bibliométricos e cientométricos passaram a necessitar de um processo sistemático o qual permite analisar esse volumoso conjunto de dados e mapear a dinâmica de uma disciplina ou campo de pesquisa (Camargo and Barbosa, 2018). Desse modo, a Ciência de Dados torna-se uma potencial candidata para esse fim, visto que é comumente definida como uma metodologia que possibilita a extração de conhecimento a partir de dados, principalmente na era do *Big Data* (Igual and Seguí, 2017).

Ademais, é bastante comum, em estudos da Bibliometria e Cientometria, o uso dos mecanismos convencionais da Análise de Redes Complexas (ARC) para investigar os relacionamentos entre as entidades científicas, a exemplo de redes de coautoria e de citação para analisar a colaboração entre instituições de pesquisas (Camargo and Barbosa, 2018). ARC é uma disciplina em ascensão que investiga como reconhecer, descrever, analisar e visualizar redes complexas (Igual and Seguí, 2017; Menczer et al., 2020). Entende-se por rede complexa como um conjunto de entidades relacionadas (conectadas), onde a estrutura dessa rede é considerada não trivial (Menczer et al., 2020).

Assim, conforme é destacado por Camargo and Barbosa (2018), se faz necessária uma abordagem sistemática para a extração de conhecimento em fontes de produção científica, possibilitando identificar a dinâmica e/ou crescimento de uma determinada área ou apoiar decisões estratégicas relacionadas às entidades de pesquisa. Desse modo, conforme ilustrada na Figura 1, a abordagem de análise de produções científicas proposta neste estudo está alinhada com a problemática destacada por Camargo and Barbosa (2018), uma vez que, a partir de dados brutos e não estruturados, serão modeladas redes complexas contemplando o sistema encapsulado nos dados em questão, bem como, quanto mais específica for a entidade científica analisada (afiliações ou palavras-chave, por exemplo), mais *insights* e informações relevantes serão extraídas a partir da combinação dos mecanismos de ARC e Aprendizado de Máquina (AM).

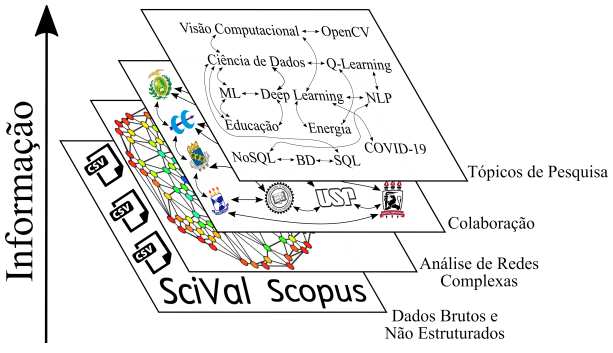


Figura 1. Abordagem Proposta de Análise de Produção Científica.

Vislumbrando aprimorar o processo de análise de produções científicas, foi desenvolvida, neste trabalho, uma metodologia para extração de padrões e conhecimento em publicações acadêmicas — indexadas na base *Scopus* e coletadas por meio da ferramenta *SciVal*[1] —, combinando técnicas de AM e de ARC. Outrossim, a abordagem proposta foi validada experimentalmente por intermédio de um estudo de caso envolvendo os Programas de Pós-graduação (PPGs) brasileiros da área de avaliação das Engenharias IV, sub-área da Engenharia Elétrica e notas 5, 6 e 7. Esses programas tem uma importância significativa para a formação de recursos humanos altamente especializados para a área de Engenharia, além de representar um subconjunto de programas de pós-graduação com produções de alto impacto, consolidados ou em processo de consolidação. Os resultados demonstraram a factibilidade da proposta indicando os principais atores, áreas de destaque, assim como redes de parcerias. A metodologia apresentada tem o potencial de instrumentar e expandir decisões estratégicas e proativas dos programas de pós-graduação visando um impacto, como consequência, cada vez maior na sociedade.

Dessa forma, as principais contribuições desse trabalho são:

- Metodologia de análise de produções acadêmicas, combinando técnicas de ARC e AM;
- Processo experimental para avaliação da metodologia proposta.

Esse artigo está organizado da seguinte forma: a seção 2 explana brevemente os principais conceitos sobre Análise de Redes Complexas. Na seção 3, são discutidos os trabalhos relacionados. A metodologia proposta é descrita na seção 4, bem como, sua avaliação empírica é detalhada na seção 5. Os resultados são apresentados na seção 6. Finalmente, é discutida, na seção 7, a conclusão e trabalhos futuros.

2. ANÁLISE DE REDES COMPLEXAS

O mundo que nos cerca possui diversos exemplos de sistemas complexos, a exemplo de interações entre proteínas, dispersão de doenças infecciosas, colaboração científica, entre outros. Segundo Zweig et al. (2016), tais sistemas complexos são extremamente complicados de serem compreendidos, pelo fato de possuírem diversos componentes, direta ou indiretamente, relacionados entre si. Ainda, conforme Menczer et al. (2020), uma vez que cada sistema complexo possui suas peculiaridades, esses sistemas podem ser representados por uma representação gráfica, denominada de rede complexa, a qual define as interações entre componentes de um sistema em questão. Desse modo, não é possível entender completamente um sistema, a menos que se compreenda totalmente sua rede.

Assim, entende-se por rede complexa como um conjunto de entidades relacionadas (conectadas), onde a estrutura dessa rede é considerada não trivial (Menczer et al., 2020; Zinoviev, 2018). Em outras palavras, de acordo com Zweig et al. (2016), uma rede é uma coleção de objetos, em que alguns pares desses objetos estão conectados por algum tipo de ligação. Esses objetos frequentemente são conhecidos por nós, bem como as ligações ou relações entre esses nós são chamadas de *links*. Contudo, com o propósito de uma análise mais precisa das propriedades de uma rede complexa, é necessário o uso de uma abordagem

[1] https://www.scival.com/

mais matemática, isto é, representar uma rede em um grafo (Menczer et al., 2020; Zinoviev, 2018).

Formalmente, um grafo pode ser definido como $G(V, E)$, onde V denota o conjunto de todos os vértices (nós) de G, e E o conjunto de todas as arestas (*links*) de G, bem como cada aresta $e = (u, v) \in E \mid u, v \in V$ representa uma ligação/aresta entre os nós u e v (Zinoviev, 2018; Zweig et al., 2016). Ademais, uma vez que exista uma aresta e interligando os nós u e v, pode-se dizer que ambos os nós são vizinhos entre si. Desse modo, neste trabalho, a rede de colaboração possui como nós os PPGs e instituições parceiras, enquanto que as arestas são representadas por uma colaboração científica, isto é, uma publicação realizada em conjunto.

Outro fator importante para a modelagem de uma rede complexa é a escolha do tipo de grafo. Existem diversas formas de representação, sendo as mais comuns aquelas relacionadas à direção ou força de uma aresta. Um grafo $G(V, E)$ é não direcionado, se toda aresta é bidirecional, ou seja, toda aresta é simétrica e recíproca, portanto, uma aresta e entre os nós u e v em G pode ser representada como $e = (u, v) = (v, u)$. Por outro lado, um grafo $G(V, E)$ é dito direcionado, se toda aresta é direcional, isto é, assimétrica e não recíproca, logo, uma aresta e entre u e v deve ser apenas representada por $e = (u, v) \mid (u, v) \neq (v, u)$ (Zinoviev, 2018; Zweig et al., 2016).

Com relação à força de uma ligação entre nós, um grafo pode ser ponderado ou não. Assim, um grafo não ponderado são aqueles cujas arestas não possuem nenhum peso associado. Por outro lado, um grafo ponderado possuem arestas com pesos associados às arestas, indicando o quão forte é uma ligação entre os componentes de uma rede (Zinoviev, 2018; Zweig et al., 2016). Assim, no contexto deste estudo, uma rede de colaboração entre PPGs e instituições de pesquisa pode ser representada por um grafo não direcionado e ponderado, uma vez que a colaboração científica é uma relação bidirecional entre os envolvidos, bem como quanto mais os parceiros colaboram com as PPGs, mais consolidadas e frutíferas serão as parcerias.

Uma vez que a rede complexa esteja devidamente modelada e representada por um grafo, é possível explorar todo potencial das técnicas de ARC, sendo uma delas a quantificação da importância de um nó na rede, comumente conhecida por centralidade (Menczer et al., 2020; Zweig et al., 2016). Existem diversas métricas de centralidade, sendo que as mais conhecidas e exploradas pela comunidade científica são *betweenness*, *eigenvector* e *weighted degree* (Zinoviev, 2018). A centralidade *betweenness* de um nó v pertencente a um grafo G é a fração de todos os pares de caminhos mais curtos que passam por v, calculada pela Equação 1, onde V é o conjunto de vértices do grafo G; $\sigma(s, t)$ o número de caminhos mais curtos entre os vértices s e t; e $\sigma(s, t \mid v)$ o número de caminhos mais curtos entre s e t que contém v (Zinoviev, 2018; Zweig et al., 2016). Desse modo, uma PPG com alto *betweenness* indica que a mesma é potencialmente crucial para o desenvolvimento de pesquisas nas instituições parceiras, bem como, caso haja um rompimento com essa PPG, isso impactará negativamente as atividades científicas de suas colaboradoras.

$$B(v) = \sum_{s,t \in V} \frac{\sigma(s, t \mid v)}{\sigma(s, t)} \tag{1}$$

Por outro lado, a centralidade *eigenvector* de um vértice v pertencente a um grafo G é definida como a soma das centralidades *eigenvector* dos vizinhos de v, dividido por λ (o maior autovalor da matriz de adjacência de G). Essa métrica é calculada pela Equação 2, onde V é o conjunto de vértices do grafo G; $A = (a_{v,y})$ a matriz de adjacência, isto é, $a_{v,y} = 1$ se o vértice v está ligado ao vértice y, e $a_{v,t} = 0$, caso contrário; λ uma constante. Em síntese, um alto *eigenvector* para um determinado nó determina que os vizinhos desse nó também possuem altos valores para essa métrica (Zweig et al., 2016). Desse modo, a partir dela, é possível localizar PPGs ou afiliações com alto prestígio científico.

$$E(v) = \frac{1}{\lambda} \sum_{y \in V} a_{v,y} E(y) \tag{2}$$

Por fim, a centralidade *weighted degree*, ou grau ponderado, de um vértice v pertencente a um grafo G corresponde ao somatório dos pesos das arestas que conectam o nó v aos seus vizinhos, calculada pela Equação 3, onde $N(v)$ é o conjunto de nós adjacentes a v; w_y o peso da aresta que interliga v ao seu vizinho y (Opsahl et al., 2010).

$$D^w(v) = \sum_{y \in N(v)} w_y \tag{3}$$

Com os conceitos de ARC devidamente explanados, serão discutidos, na próxima seção, os trabalhos correlatos a este estudo.

3. TRABALHOS RELACIONADOS

Diante dos diversos estudos encontrados, que realizaram análise de produções acadêmicas de programas de pós-graduação com o auxílio de técnicas de Análise de Redes Complexas, de forma direta ou indireta, destacam-se os trabalhos realizados por Lança et al. (2018), Mugnaini et al. (2019) e Silva et al. (2018).

Lança et al. (2018) caracterizou a produção científica dos pesquisadores em Ciência da Informação, por meio da análise dos indicadores bibliométricos relacionados à evolução anual das publicações, produtividade por programa, publicações por qualis, *ranking* dos periódicos com mais publicações e a colaboração e sua relação com a produtividade nas publicações dos Programas de Pós-Graduação em Ciência da Informação. A amostra analisada compreendeu 417 currículos Lattes de pesquisadores, que atuam como docentes nos 23 Programas de Pós-Graduação em Ciência da Informação no Brasil. Após as análises, observou-se crescimento nas publicações, o que indica o crescimento na pesquisa em Ciência da Informação. Também a quantidade relevante de publicações em estratos B1, podendo indicar uma busca pela internacionalização da pesquisa na área. Os resultados foram bastante reveladores para a compreensão da atividade científica em Ciência da Informação no Brasil.

Similarmente, Mugnaini et al. (2019) realizaram uma análise exploratória da dispersão (mapeamento) da produção científica de 260.663 pesquisadores doutores, registrados na Plataforma Lattes. Eles consideraram apenas os artigos completos publicados em periódicos entre os anos de 1998 e 2016 (seis ciclos avaliativos da CAPES [2]), bem como as análises contemplaram a grande área de atuação, os diferentes períodos de avaliação da CAPES, o país de publicação e a indexação em diferentes periódicos. Além dos dados dos currículos Lattes, também utilizaram informações de diversas bases de periódicos para corrigir dados errôneos constantes nos currículos. Após as análises, constatou-se uma tendência de internacionalização dos estudos, bem como a importância de periódicos nacionais para publicação de parte da produção científica de algumas áreas. Também evidenciou-se a limitação de estudos que não consideraram periódicos não indexados, bem como aqueles restritos às bases *Scopus* e/ou *Web of Sciences* (WoS). Ademais, identificou-se uma tendência de crescimento da produções em periódicos, assim como as Ciências Humanas, Sociais e Linguística tendiam a publicar em periódicos não indexados, enquanto que as demais áreas o faziam em menor proporção. Foi observado também que a indexação *SciELO* era muito utilizada, em ordem decrescente de importância, pelas Ciências Agrárias, da Saúde, Humanas e Biológicas; enquanto, para *Scopus*/WoS, destacaram-se as Ciências Biológicas, Exatas, da Saúde e Engenharias.

Por fim, Silva et al. (2018) propôs uma metodologia de recuperação e sistematização de indicadores científicos, tecnológicos e acadêmicos para a pós-graduação da Universidade Federal do Vale do São Francisco (Univasf), a partir de ferramentas cientométricas de baixo custo. A estratégia teve como fonte principal a Plataforma Lattes do CNPq [3]. Após a realização do estudo de caso, considerou-se a estratégia metodológica viável de aplicação ao universo investigado, possibilitando conhecer diversas características da pós-graduação da Univasf.

Os trabalhos correlatos apresentados retratam o uso de técnicas de Análise de Redes Complexas para análise da produção científica de programas de pós-graduação em contextos específicos. Assim, o diferencial deste estudo é uma abordagem de análise cientométrica que combina o uso da ferramenta *SciVal* com técnicas de Análise de Redes Complexas e *Machine Learning*. É importante ressaltar que a modelagem e análise de dados de produções científicas relacionadas com tópicos/áreas de atuação, redes de cooperação entre instituições e tópicos, citações e relevância são de fundamental importância para instrumentar e municiar decisões estratégicas dos programas de pós-graduação, sendo um dos diferenciais da respectiva proposta.

4. METODOLOGIA PROPOSTA

A metodologia proposta para extrair conhecimento cientométrico em produções científicas é apresentada na Figura 2.

Primeiramente, foram criados, na ferramenta *SciVal*, os grupos de pesquisas equivalentes aos PPGs da sub-área

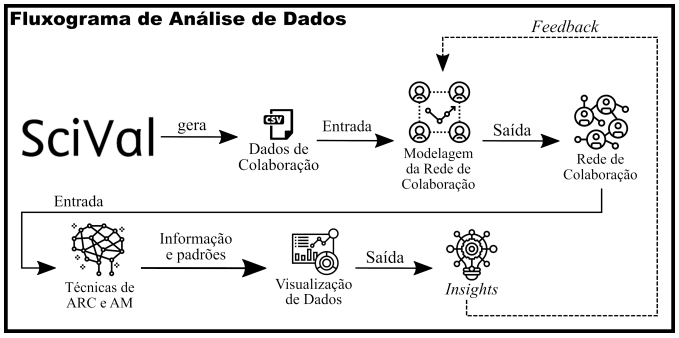


Figura 2. Fluxograma de Análise de Dados Cientométricos.

Engenharia Elétrica da área Engenharias IV, cuja nota do curso fosse superior ou igual a cinco, pelo fato desses programas possuírem as maiores notas, consolidadas ou em consolidação pela CAPES. Para a formação dos grupos de pesquisa no *SciVal*, foi feita a imputação dos 1484 identificadores *Scopus* associados aos docentes dessas PPGs. Vale destacar que, anteriormente, realizou-se um levantamento dessas PPGs e de seus respectivos docentes, utilizando a Plataforma Sucupira [4], além da realização de uma busca exaustiva e manual dos identificadores desses docentes utilizando o mecanismo de busca de autores da *Scopus* [5].

Com os grupos de pesquisa devidamente criados, foram extraídos os dados brutos de colaborações relacionados à produção científica dessas PPGs, gerando assim, o *dataset* utilizado nesse estudo. Vale destacar que, para este trabalho, esses dados contêm metadados dos trabalhos publicados no último quadriênio de avaliação da CAPES (2017-2020), além daqueles disponibilizados no ano corrente (2021). Contudo, para os anos de 2020 e 2021, os dados ainda estão incompletos pelo fato do *SciVal* não ter realizado todo o processamento necessário.

A partir dos dados brutos disponibilizados pelo *SciVal*, é factível analisar o comportamento da produção científica de qualquer entidade (pesquisador, programa de pós-graduação, grupo de pesquisa, afiliação ou país), possibilitando investigar como se deram seus esforços para ampliação do conhecimento científico. Assim, para este trabalho, esses dados foram analisados e modelados em uma rede complexa de colaboração entre instituições de pesquisa e PPGs, representando o sistema complexo implícito nesses dados.

É importante frisar que, na modelagem, por conveniência, foram considerados como pesos/atributos das colaborações (arestas) e dos PPGs (nós do tipo *program*) os principais índices cientométricos (número de autores, total de citações, número de manuscritos, total de citações por quantidade de manuscritos, entre outros) extraídos do *SciVal*. Além desses índices, os PPGs tinham o atributo *grade* atrelado ao seu conceito CAPES. Por outro lado, as instituições parceiras (nós do tipo *affil*) apenas possuíam como atributos *country_region* (país ou região de origem da afiliação) e o número de manuscritos normalizado pela quantidade de autores envolvidos.

[2] Coordenação de Aperfeiçoamento de Pessoal de Nível Superior.
[3] Conselho Nacional de Desenvolvimento Científico e Tecnológico.
[4] https://sucupira.capes.gov.br/sucupira/public/index.jsf
[5] https://www.scopus.com/freelookup/form/author.uri

Com a rede devidamente modelada, é possível identificar potenciais parceiros ou extrair conhecimento e padrões, previamente desconhecidos e não triviais, da rede de colaboração, combinando técnicas de AM e ARC. Além disso, a partir dos mecanismos de Visualização de Dados, pode-se analisar e avaliar esse conhecimento extraído e, se tais padrões e informações forem insuficientes, pode-se remodelar ou refinar a rede de colaboração até que sejam satisfatórios os *insights* obtidos, uma vez que a metodologia proposta é interativa e iterativa. Isto é, interativa no sentido de que o pesquisador participa ativamente durante todo o processo de análise, e iterativa pelo fato de retornar a etapa de modelagem da rede até que sejam obtidos resultados relevantes.

Com a abordagem proposta devidamente apresentada, será detalhado, na próxima seção, sua avaliação empírica.

5. ESTUDO DE CASO

Esta seção descreve o estudo de caso da metodologia proposta, o qual foi baseado no processo experimental apresentado por Coutelieris et al. (2018) e Santos et al. (2015). Desse modo, o principal objetivo deste estudo é analisar, descritivamente, os programas de pós-graduação em Engenharia Elétrica das Engenharias IV, com o propósito de caracterização e entendimento, com respeito aos índices cientométricos, métricas convencionais de ARC e comportamento de sua produção científica.

As próximas subseções focarão na definição e planejamento do estudo de caso. A última subseção apresentará o processo de operação desta avaliação empírica.

5.1 Questões de Pesquisa

As questões de pesquisa (QP) a serem exploradas, neste estudo, são as seguintes:

QP 1: Quais as áreas de conhecimento mais exploradas pelos PPGs?
QP 2: Quais os programas que mais firmam parcerias com instituições (inter)nacionais? E os que estabelecem menos colaborações?
QP 3: É possível determinar possíveis parcerias entre os PPGs a partir de suas instituições parceiras?
QP 4: Existe equivalência entre a classificação dos PPGs perante a CAPES com o agrupamento baseado nos índices cientométricos e métricas de ARC?

Logo, as métricas para avaliar essas questões estão descritas na Tabela 1.

5.2 Materiais e Métodos

Após a definição das questões de pesquisa e suas métricas associadas, iniciou-se o processo de seleção de objetos e/ou artefatos. Primeiramente, por conveniência, a ferramenta *SciVal* foi escolhida por possuir diversos recursos e análises básicas, a exemplo do fornecimento de dados brutos de afiliações parceiras, tópicos e áreas de pesquisas associadas a uma entidade científica (pesquisador, afiliação, grupo de pesquisa, etc). Ademais, também por conveniência, a escolha restringiu-se aos programas de pós-graduação em Engenharia Elétrica mais bem conceituados pela CAPES, cuja nota do programa fosse superior ou igual a cinco, possibilitando analisar, descritivamente, as similaridades e nuances entre esses programas bastante atuantes na Engenharia Elétrica no Brasil.

Tabela 1. Métricas/*Features* utilizadas.

Métrica	Descrição
co_authored_publications	Número de manuscritos (aresta)
scholarly_output	Número de manuscritos (nó *program*)
citations	Total de citações
co_authors	Número de autores envolvidos
citations_per_publication	Total de citações por número de manuscritos
cpp_per_coauthors	Métrica *citations_per_publication* normalizada pelo número de autores
publications_per_coauthors	Número de manuscritos por número de autores
international_collaboration_percent	Percentual de colaboração com instituições internacionais
publications_top_journal_percentiles	Percentual de publicações em periódicos extremamente relevantes
field_weighted_citation_impact	Indica como o número de citações recebidas pelas publicações de uma entidade se compara com a média mundial.
betweenness	Centralidade *betweenness*
eigenvector	Centralidade *eigenvector*
weighted_degree	Centralidade *weighted degree*

Outrossim, também por conveniência, os materiais e/ou recursos a serem utilizados são: o ecosistema Python para Ciência de Dados (pandas [6], NumPy [7], Matplotlib [8], scikit-learn [9] e outros), fornecidos pelo Google Colab [10]; NetworkX [11] e Gephi [12], biblioteca e ferramenta, respectivamente, para modelagem, visualização, análise e manipulação de redes complexas; e a metodologia de análise de produções científicas, discutida na seção 4.

5.3 Execução

Esta subseção descreve a preparação e execução desta avaliação empírica. O processo de operação foi realizado inicialmente com a configuração do ambiente para o estudo de caso, bem como o planejamento da coleta de dados.

Primeiramente, foram extraídos os dados brutos de colaborações relacionados à produção científica dos PPGs, gerando assim, o *dataset* utilizado nesse estudo. Em seguida, definiu-se a metodologia de análise, a qual foi detalhada na seção 4. Por fim, executou-se o processo de análise, previamente discutido (ver seção 4), com os artefatos necessários (ver subseção 5.2).

Após a execução, os resultados das análises foram obtidos, os quais foram baseados nas métricas previamente apresentadas (ver subsecão 5.1). Vale destacar que estes

[6] https://pandas.pydata.org
[7] https://numpy.org
[8] https://matplotlib.org
[9] https://scikit-learn.org/
[10] https://colab.research.google.com
[11] https://networkx.org/
[12] https://gephi.org/

resultados possibilitarão responder as questões de pesquisa deste estudo empírico.

Os resultados relacionados a esta análise serão apresentados na próxima seção.

6. RESULTADOS E DISCUSSÕES

Esta seção apresenta os resultados da avaliação empírica da metodologia proposta, os quais estão divididos em três partes: (i) Análise de Tópicos de Pesquisa; (ii) Análise de Colaborações entre PPGs e afiliações; e (iii) Análise de Agrupamento dos PPGs.

6.1 Análise por Tópicos

A análise exploratória de dados relacionada com os tópicos das publicações tem como objetivo elucidar e destacar em quais áreas do conhecimento os programas focam seus esforços (QP 1). As Figuras 3, 4, 5 e 6 apresentam os principais resultados, no qual o tamanho das circunferências estão relacionadas com a quantidade de produções científicas de um determinado tópico. Cada tópico é caracterizado por pertencer a uma determinada área do conhecimento e foram rotulados em todas as publicações pela ferramenta *SciVal*. A compilação de todas as áreas do conhecimento foram descritas na Figura 3. Nos resultados, percebe-se que os diferentes programas foram agrupados por região com o intuito de elucidar a regionalidade como fatores de discussão.

Figura 3. Áreas de estudo relacionadas pela ferramenta *SciVal* e indicadas como referências nas Figuras 4, 5 e 6 .

Os programas de pós-graduação Nota-5 apresentaram uma maior agregação para as áreas da Engenharia e Ciência da Computação. Entretanto, é possível também notar que as áreas da Física e Medicina se destacam em alguns programas. Em relação à regionalidade, percebe-se que os programas das regiões Norte & Nordeste apresentam uma complementaridade de áreas. Na região Sudeste, existe um destaque para a área da Engenharia, mas também é possível notar ações nas áreas da Medicina e Física. Por outro lado, na região Sul, a área da Ciência da Computação tem um maior destaque. É possível também perceber ações na área da Medicina nessa região brasileira.

A respeito dos programas de pós-graduação Nota-6, é perceptível que existe uma maior diversidade das áreas de atuação e principalmente na região Sudeste. A atuação na área

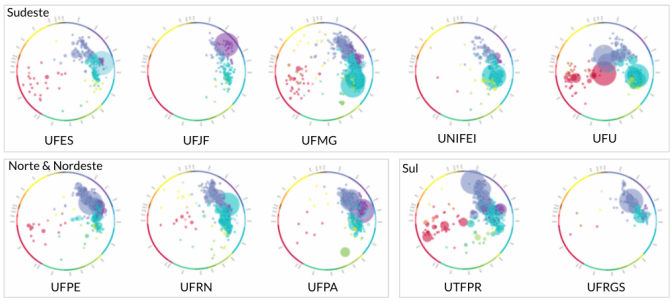


Figura 4. Análise exploratória de tópicos gerados pelos Programas de Pós-Graduação Nota-5.

da Física na região Sudeste é nítida, percebendo demais áreas como Agricultura, Medicina, Energias e Ciências Sociais também com vínculos de produções científicas nos programas de pós-graduação. As regiões Norte & Nordeste foram representadas por um programa de pós-graduação cuja área temática da Ciência da Computação apresenta um maior destaque, padrão observado também nos programas Nota-5. Os programas da região Sul apresentam uma forte atuação nas áreas de Energia, Engenharia e Ciência da Computação e, em menor quantidade, na área da Física. Reforça-se que a maior ou menor atuação refere-se ao tamanho do círculo cuja cor se relaciona com as áreas representadas na Figura 3.

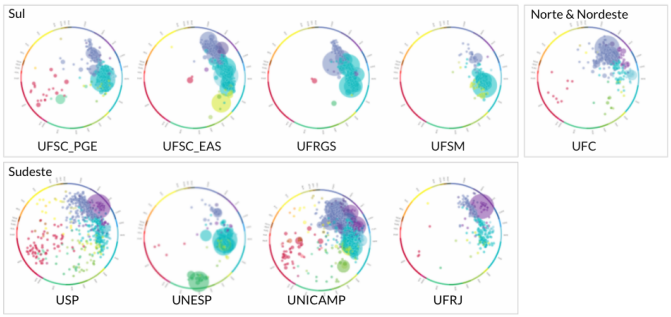


Figura 5. Análise exploratória de tópicos gerados pelos Programas de Pós-Graduação Nota-6.

Por fim, pode-se analisar os programas de pós-graduação Nota-7. Conforme classificação da CAPES, esses são programas de alta qualidade e com produções científicas relevantes. As áreas de atuação que se destacam são Engenharia, Física, Energia, Medicina e Ciência da Computação.

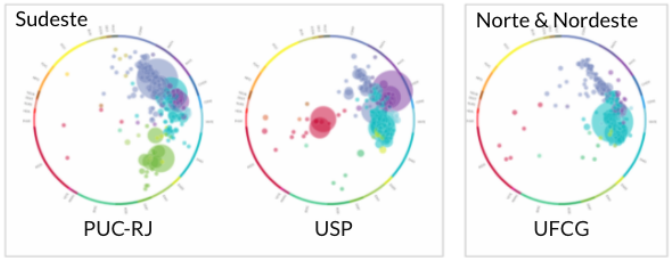


Figura 6. Análise exploratória de tópicos gerados pelos Programas de Pós-Graduação Nota-7.

6.2 Análise de Colaborações entre Programas e Afiliações

Para responder as questões de pesquisa 2 e 3, as análises descritas a seguir foram focadas nas métricas de cen-

tralidade típicas de ARC, as quais possibilitam extrair informações relevantes com base na estrutura da rede de colaboração. Primeiramente, para QP 2, a partir das métricas *betweenness*, *eigenvector* e *weighted degree*, foram determinados os programas mais importantes no aspecto de colaboração científica (ver Tabela 2).

Tabela 2. Top 10 PPgEEs mais importantes.

Métricas de Centralidade		
Betweenness	*Eigenvector*	*Weighted Degree*
USP (0,103)	USP-SC (1,000)	UNIFEI (675,8)
UTFPR (0,096)	UFMG (0,998)	USP-SC (422,5)
UFC (0,094)	UFSC-EE (0,961)	UFPA (391,3)
Unesp-IS (0,094)	UFPA (0,929)	Unesp-IS (380,2)
PUC-RJ (0,092)	UFRGS (0,914)	UFES (329,5)
UFRN (0,091)	UFES (0,914)	PUC-RJ (329,4)
UFES (0,090)	Unicamp (0,905)	UFRGS (308,3)
UFSC-EAS (0,089)	UTFPR (0,889)	UFRN (299,1)
UFRGS-MI (0,086)	UNIFEI (0,882)	UFMG (259,4)
UFPA (0,084)	UFRN (0,880)	UFRGS-MI (236,0)

Uma vez que as métricas de centralidade determinam o grau de importância de um nó em uma rede, logo, conforme a Tabela 2, em uma visão macro, é perceptível que os programas Nota-5 se destacam no quesito de colaboração com diversas afiliações ou que estão mais preocupados em firmar parcerias com várias instituições de pesquisa, sobressaindo-se os PPGs da UFRN, UFES e UFPA, os quais estão entre os 10 programas mais importantes mediante todas as métricas utilizadas. Por outro lado, dentre os três programas de Nota-7, para cada métrica analisada, houve apenas um programa focado em parcerias científicas, destacando-se o PPG da PUC-RJ para as métricas *betweenness* e *weighted degree*. Contudo, o PPG da USP de São Carlos (USP-SC) obteve a pontuação máxima para *eigenvector*, caracterizando-o como o programa mais influente dentre os PPGs analisados. Por fim, aparentemente, os programas Nota-6 possuem uma modesta preocupação em estabelecer colaborações com outras entidades de pesquisa, tendo entre dois a quatro ocorrências para cada métrica.

Ainda para QP 2, sob outra perspectiva, os três programas menos importantes em termos de parcerias foram: USP (0,561), UFU (0,610) e UFSM (0,653) para a métrica *eigenvector*; UFSM (0,047), UFU (0,055) e UFCG (0,069) para *betweeness*; Unicamp (8,4), UFRJ (37,4) e UFJF (52,6) para *weighted degree*. Desse modo, em geral, valores baixos para as métricas supracitadas sugerem que os programas tendem a possuir parceiros com menor possibilidades de articulação e cooperação científica com outros programas de pós-graduação.

Sejam a, b e c nós de uma rede de colaboração. Dado que b é vizinho de a, e que c é vizinho de b, logo, c pode ser considerado vizinho de a (Menczer et al., 2020). Assim, partindo-se desse pressuposto, de fato, é factível determinar possíveis parcerias entre as PPGs a partir de suas instituições parceiras, portanto, respondendo positivamente a QP 3. Para ilustrar esse pressuposto ("o vizinho do meu vizinho é também meu vizinho"), na Figura 7, são apresentados os demais programas de pós-graduação que possuem afiliações parceiras em comum com o PPG da UFRN.

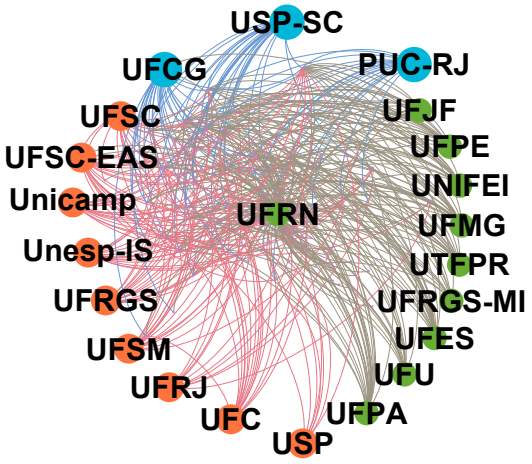
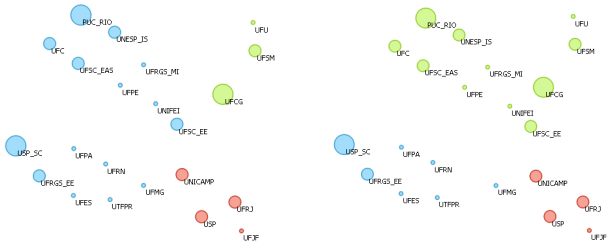


Figura 7. Possíveis programas-parceiro do PPG da UFRN.

6.3 Análise de Agrupamento dos PPGs

Para responder a questão de pesquisa 4, utilizou-se as técnicas de agrupamento *KMeans++* e *Gaussian Mixture* sob o conjunto de índices cientométricos e métricas de ARC atreladas aos PPGs analisados. Assim, por conveniência, determinou-se que a quantidade de *clusters* seria três ($k = 3$) para apreciação de QP 4, além de investigar a equivalência entre os agrupamentos detectados por esses algoritmos. Sua escolha se deu pelo fato de serem bastante utilizados pela comunidade de *Machine Learning*, bem como por possuírem apenas um único parâmetro, o número de *clusters*. Ademais, as *features* utilizadas para a clusterização foram *weighted_degree*, *betweenness*, *eigenvector*, *scholarly_output*, *citations*, *international_collaboration_percent* e *publications_top_journal_percentiles*, descritas anteriormente na Tabela 1 (ver subseção 5.1).

Inicialmente, conforme a Figura 8 (a), observa-se um grande *cluster* que comporta dois dos três PPGs Nota-7 da CAPES, com destaque a presença dos programas de Engenharias IV das instituições do Norte & Nordeste escolhidas para a abordagem neste trabalho. Além disso, vale destacar que, para visualização dos resultados da análise de agrupamento, determinou-se que o tamanho do círculo se refere à nota CAPES do PPG e a cor ao *cluster* classificado pelos algoritmos de agrupamento.



(a) *K-Means++* (b) *Gaussian Mixture*

Figura 8. Agrupamento de PPGs de Engenharias IV.

Os programas de instituições como UFU, UFSM e UFCG estão no mesmo *cluster*, levando em conta que cada um possui notas 5, 6 e 7, respectivamente. Além disso, um terceiro *cluster* é formado com instituições majoritariamente com PPGs Nota-6, sendo elas: Unicamp, USP e UFRJ, além da UFJF (Nota-5).

Outrossim, ele reuniu instituições como PUC-RIO (Nota-7), USP-SC (Nota-7), UFRN (Nota-5) e UFC (Nota-6). Mais uma vez, todas as instituições do Norte & Nordeste estiveram em um único *cluster*, com exceção do programa da UFCG (Nota-7) que ficou juntamente com instituições do Sul/Sudeste como a UFU (Nota-5), por exemplo.

Em uma segunda análise, foi utilizado o algoritmo *Gaussian Mixture*, conforme a Figura 8 (b). Assim, uma nova reorganização foi sugerida, como, por exemplo, o PPG da UFRN (Nota-5) está junto da USP-SC (Nota-7) e UFRGS (Nota-6). Os programas da PUC-RIO (Nota-7) e UFCG (Nota-7) compõem um segundo grupo, juntamente com diversas PPGs Nota-5 e Nota-6 espalhadas no Brasil afora.

Um terceiro *cluster* continuou da mesma forma como identificado pelo *KMeans++*, sendo composto apenas por instituições do Sudeste. Esse fato pode indicar um possível comportamento similar entre os PPGs — especificamente entre seus docentes — que compõe esse *cluster*, são eles: Unicamp, USP e UFRJ (todos Nota-6) e UFJF (Nota-5).

Por fim, apesar dos agrupamentos sugeridos pelos algoritmos serem distintos da classificação dos PPGs perante a CAPES, com base na análise dos *clusters*, é perceptível que alguns PPGs podem ter perfis cientométricos similares de acordo com as *features* utilizadas. Outro fato interessante é que alguns programas de notas inferiores estavam, na maioria dos casos, no mesmo grupo de PPGs bem conceituadas, a exemplo do *cluster* contendo os PPGs da UFRN (Nota-5) e da USP-SC (Nota-7), um possível indicativo de que os PPGs de notas 5 e 6 se espelham nos programas Nota-7, visando uma melhoria de suas avaliações perante a CAPES.

7. CONCLUSÃO

Com o crescimento exponencial de produções científicas, estudos bibliométricos e cientométricos passaram a necessitar de um processo sistemático que permite analisar esse volumoso conjunto de dados. Assim, o referido trabalho propôs uma metodologia orientada a Ciência de Dados em grafos, combinando técnicas de Aprendizado de Máquina e de Análise de Redes Complexas, a fim de extrair conhecimento e padrões relevantes e implícitos em produções científicas, além da validação experimental dessa abordagem por meio de um estudo de caso.

A partir da abordagem proposta, foi possível analisar e caracterizar, descritivamente, os PPGs em Engenharia Elétrica das Engenharias IV, com base nos índices cientométricos e métricas convencionais de ARC, além de entender o comportamento de sua produção científica. Logo, os resultados demonstraram a factibilidade da proposta, indicando os principais atores, áreas de destaque e redes de parcerias. Ademais, a metodologia apresentada tem o potencial de instrumentar e expandir decisões estratégicas e proativas dos programas de pós-graduação, visando um impacto, como consequência, cada vez maior na sociedade.

Como trabalhos futuros, pretende-se ampliar o contexto de análise, aumentando-se a quantidade e diversidade de programas de pós-graduação, possibilitando uma visão geral da pesquisa *stricto sensu* no Brasil.

REFERÊNCIAS

Araújo, W.C.O., Andretta, P.I.S., and Inomata, D.O. (2020). A produção científica na universidade federal do ceará considerações bibliométricas para o período de 2009 a 2018. *Informação em Pauta*, 5(1).

Camargo, L.S.d. and Barbosa, R.R. (2018). Bibliometria, cienciometria e um possível caminho para a construção de indicadores e mapas da produção científicaca. *PontodeAcesso*, 12(3), 109–125.

Coutelieris, F.A., Kanavouras, A., Theologou, K., and Stelios, S. (2018). *Experimentation Methodology for Engineers*. Springer.

Guirado, J.R., Tavares, R.L.C., and Oliveira, M. (2020). Análise cientométrica sobre a produção científica em meditação nos periódicos da medicina. *Informação em Pauta*, 5(1), 98–121.

Igual, L. and Seguí, S. (2017). *Introduction to Data Science*. Springer.

Lança, T.A., Amaral, R.M., Rocha, E.S.S., and Maciel, R.S. (2018). Produção científica dos programas de pós-graduação em ciência da informação na plataforma lattes. In *XIX Encontro Nacional de Pesquisa em Ciência da Informação (XIX ENANCIB)*.

Menczer, F., Fortunato, S., and Davis, C.A. (2020). *A First Course in Network Science*. Cambridge University Press.

Mugnaini, R., Damaceno, R.J.P., Digiampietri, L.A., and Mena-Chalco, J.P. (2019). Panorama da produção científica do brasil além da indexação: uma análise exploratória da comunicação em periódicos. *Transinformação*, 31.

Opsahl, T., Agneessens, F., and Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social networks*, 32(3), 245–251.

Santos, B.S., Júnior, M.C., da Paixão, B.C., Santos, R.M., Nascimento, A.V.R., dos Santos, H.C., Wallace Filho, H., and de Medeiros, A.S. (2015). Comparing text mining algorithms for predicting irregularities in public accounts. In *SBSI*, 667–674.

Silva, A.P., Cunha, F., and Sobral, N. (2018). Recuperação e sistematização de indicadores científicos, tecnológicos e acadêmicos: uma proposta metodológica para a pós-graduação da univasf. In *XIX Encontro Nacional de Pesquisa em Ciência da Informação (XIX ENANCIB)*.

Sugimoto, C.R. and Larivière, V. (2018). *Measuring research: What everyone needs to know*. Oxford University Press.

Zinoviev, D. (2018). *Complex network analysis in Python: Recognize-construct-visualize-analyze-interpret*. Pragmatic Bookshelf.

Zweig, K.A. et al. (2016). *Network analysis literacy*. Springer.