UNIVERSIDADE FEDERAL DO OESTE DO PARÁ
INSTITUTO DE ENGENHARIA E GEOCIÊNCIAS
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

MATHEUS LIMA DO COUTO

CARACTERIZAÇÃO DE REDES DE COAUTORIA DE BOLSISTAS DE PRODUTIVIDADE DO CNPQ: UMA ABORDAGEM BASEADA NA CIÊNCIA DE DADOS

Santarém - Pará
2025

MATHEUS LIMA DO COUTO

CARACTERIZAÇÃO DE REDES DE COAUTORIA DE BOLSISTAS DE PRODUTIVIDADE DO CNPQ: UMA ABORDAGEM BASEADA NA CIÊNCIA DE DADOS

Trabalho de Conclusão de Curso apresentado como requisito para obtenção do grau de Bacharel em Ciência da Computação, pela Universidade Federal do Oeste do Pará, no Instituto de Engenharia e Geociências.
Orientador: Prof. Dr. Fábio Manoel França Lobato

Santarém - Pará
2025

Dados Internacionais de Catalogação-na-Publicação (CIP)
Sistema Integrado de Bibliotecas – SIBI/UFOPA

C871c Couto, Matheus Lima do

Caracterização de redes de coautoria de bolsistas de produtividade do CNPQ: uma abordagem baseada na ciência de dados. / Matheus Lima do Couto. - Santarém, 2025.

32 p. : il.

Inclui bibliografias.

Orientador: Fábio Manoel França Lobato.

Trabalho de Conclusão de Curso (Graduação) – Universidade Federal do Oeste do Pará, Instituto de Engenharia e Geociências, Bacharelado em Ciência da Computação.

1. Rede de coautoria. 2. Ciência de dados. 3. Lattes. 4. Cientometria. I. Lobato, Fábio Manoel França, *orient*. II. Título.

CDD: 23 ed. 005.1

Bibliotecária - Documentalista: Cátia Alvarez – CRB/2 843

MATHEUS LIMA DO COUTO

CARACTERIZAÇÃO DE REDES DE COAUTORIA DE BOLSISTAS DE PRODUTIVIDADE DO CNPQ: UMA ABORDAGEM BASEADA NA CIÊNCIA DE DADOS

Trabalho de Conclusão de Curso apresentado como requisito para obtenção do grau de Bacharel em Ciência da Computação, pela Universidade Federal do Oeste do Pará, no Instituto de Engenharia e Geociências.

Conceito: 10
Data de aprovação: 21/02/2025

Banca Examinadora:

Documento assinado digitalmente
FABIO MANOEL FRANCA LOBATO
Data: 18/03/2025 14:44:55-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Fábio Manoel França Lobato - Orientador
Universidade Federal do Oeste do Pará

Documento assinado digitalmente
CLAUDIR OLIVEIRA
Data: 18/03/2025 15:41:51-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Claudir Oliveira
Universidade Federal do Oeste do Pará

Documento assinado digitalmente
MARCELINO SILVA DA SILVA
Data: 18/03/2025 15:38:08-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Marcelino Silva da Silva
Universidade Federal do Oeste do Pará

AGRADECIMENTOS

Em primeiro lugar, gostaria de expressar minha gratidão aos meus familiares pelo apoio e incentivo durante este percurso. Em especial, à minha mãe Elisangela, ao meu pai Heverton e à minha irmã Marcele, que sempre acreditaram em meu potencial e estiveram ao meu lado em todos os momentos.

Agradeço ao meu orientador, Prof. Dr. Fábio Lobato, por suas valiosas orientações e sábios conselhos, que foram fundamentais para me guiar ao longo desta trajetória acadêmica.

Aos meus amigos do LACA, meu mais sincero agradecimento por estarem sempre presentes, compartilhando desafios, conquistas e muitos momentos de aprendizado. Mais do que colegas, vocês se tornaram uma parte essencial dessa jornada, tornando cada dia mais leve e especial. As risadas, as conversas e o apoio mútuo se transformaram em boas memórias e fizeram dessa caminhada uma experiência única. Sou muito grato por ter vivido tudo isso ao lado de vocês.

A todos, meu muito obrigado por fazerem parte dessa caminhada e por contribuírem, de maneira única, para que este trabalho se tornasse realidade.

"Erra aquele que não principia a aprender por supor que já é tarde." Sêneca

RESUMO

As redes de coautoria são amplamente utilizadas para avaliar a qualidade da produção de literatura científica e das colaborações entre investigadores e instituições. Identificamos uma lacuna na literatura em relação à análise das interações entre bolsistas de produtividade do CNPq. Para preencher essa lacuna, utilizamos a ciência de dados para caraterizar as redes de coautoria para obter uma visão complementar dessa política pública crítica para a promoção da pesquisa de excelência no Brasil. Para isso, coletamos 12.345 CVs Lattes de pesquisadores e analisamos aproximadamente 400.000 publicações. Os resultados mostraram maior colaboração nos estratos mais altos de bolsas (e.g., 1A e 1B). Outros achados de interesse da pesquisa estão relacionados às discrepâncias regionais e à equidade de gênero. O estudo contribui para uma melhor compreensão da dinâmica social dos beneficiários de bolsas de produtividade, apoiando a avaliação desta importante política de promoção da investigação.

Palavras-chave: Rede de coautoria, Ciência de dados, Lattes, Cientometria.

ABSTRACT

Co-authorship networks are widely used to evaluate the quality of scientific literature productions and collaborations between researchers and institutions. We identified a gap in the literature regarding the analysis of interactions between CNPq productivity fellows. To fill this gap, we used data science to characterize co-authorship networks to obtain a complementary overview of this critical public policy for promoting excellent research in Brazil. To this end, we collected 12,345 researchers' Lattes CVs and analyzed approximately 400,000 publications. The results showed greater collaboration in the higher strata of grants (e.g., 1A and 1B). Other research findings of interest are related to regional discrepancies and gender equity. The study contributes to a better understanding of the social dynamics of productivity grant recipients, supporting the evaluation of this relevant research promotion policy.

Keywords: Co-authorship network, Data Science, Lattes, Scientometric.

SUMÁRIO

1 INTRODUÇÃO 8

2 TEXTO DO ARTIGO 10

2.1 Introdução 10

2.2 Trabalhos Relacionados 11

2.3 Metodologia 12

2.4 Resultados 14

2.5 Considerações Finais 18

3 CONSIDERAÇÕES FINAIS 19

REFERÊNCIAS 21

APÊNDICE A – ARTIGO ORIGINAL 23

ANEXO A – REGISTRO DE PARTICIPAÇÃO DO AUTOR NO EVENTO 31

1 INTRODUÇÃO

Este trabalho foi desenvolvido durante o período que o autor atuou como bolsista de Iniciação Científica no Laboratório de Computação Aplicada. Neste período o autor cursou a disciplina de Tópicos Avançados em Inteligência Artificial sendo ministrada pelo Prof. Dr. Fábio Manoel França Lobato na qual abordou tópicos relacionados a ciência de dados, inteligência artificial, extração do conhecimento em bases de dados, ciclo de vida da ciência de dados. Sendo este período de intenso aprendizado para o autor e sob a orientação do docente da disciplina, foi proposto, como trabalho final o envio de um artigo que correlacionava os novos conhecimentos dos assuntos ministrados na disciplina com os adquiridos no período de iniciação científica, onde se observou a oportunidade de trabalhar conceitos voltados a cientometria, grafos, dentre outros, para analisar o grande *corpus* de dados do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) acerca dos bolsistas de produtividade, tendo em vista a relevância do tema e a facilidade no acesso aos dados, pois os mesmos são de caráter aberto.

O trabalho foi inicialmente enviado ao *Brazilian Workshop on Social Network Analysis and Mining* (BraSNAM), entretanto foi rejeitado. Posteriormente foi reestruturado com base no *feedback* enviado pelos revisores do BrasNAM, ademais a pesquisa teve o escopo expandido sendo submetida ao *Symposium on Knowledge Discovery, Mining and Learning* (KDMiLe), onde obteve o aceite e posteriormente a apresentação feita na 12° edição do evento que ocorreu no ano de 2024, em Belém. Apresentação no qual recebeu elogios, sendo bem avaliado no evento, e por este motivo foi escolhido como trabalho de conclusão de curso.

O trabalho (COUTO et al., 2024) foi apresentado pelo autor no dia 18 de novembro no ano de 2024 na 12° edição do evento Symposium on Knowledge Discovery, Mining and Learning (KDMiLe), que ocorreu na cidade de Belém, no estado do Pará e foi sediado na Universidade Federal do Pará (UFPA), como consta no na Figura 1.1, onde mostra o autor apresentando seu artigo no primeiro dia do evento. A experiência no evento foi muito enriquecedora por proporcionar tanto a apresentação do trabalho quanto a troca de ideias com especialistas e profissionais da área. Durante o evento, o autor pôde assistir a diversas apresentações de trabalhos de pesquisadores da área, o que favoreceu a troca de conhecimento e trouxe novas visões sobre as aplicações da inteligência artificial e a extração de conhecimento usando ciência de dados. Dessa forma, foi possível observar os estudos de fronteira nesses domínios, especialmente em aprendizado de máquina e inteligência artificial.

Figura 1.1 – Apresentação do Autor no Evento KDMile

O trabalho foi recebido de maneira bastante positiva, com interações construtivas por parte da plateia, que questionou algumas abordagens apresentadas e destacou como poderia aplicá-las em seus próprios projetos, evidenciando a relevância do tema abordado. Ademais, o trabalho segue as linhas de pesquisas desenvolvidas pelo grupo de pesquisa do Laboratório de Computação Aplicada (LACA), que visa explorar a aplicação de Inteligência Artificial e Ciência de Dados em diversos campos, como a análise de dados textuais e a extração de conhecimento a partir de grandes quantidades de dados, sendo expandido para novos conhecimentos como para area de grafos e cientometria. Desta forma, este trabalho vem sendo continuado por um membro do grupo de pesquisa, onde está sendo ampliado as análises feitas, para futuras publicações.

O restante deste trabalho está organizado como segue: a Seção 2 apresenta o texto original do artigo, ipsis litteris, preservando sua estrutura e subdivisões. A Seção 3 é destinada às considerações finais e perspectivas futuras.

2 TEXTO DO ARTIGO

As subseções seguintes apresentam o artigo de forma fiel, mantendo a estrutura e os conceitos originais, com o objetivo de fornecer uma compreensão mais aprofundada dos objetivos, métodos, resultados e conclusões apresentados.

2.1 INTRODUÇÃO

A ciência de dados está intrinsecamente relacionada com a tomada de decisões estratégicas (PROVOST; FAWCETT, 2013). Este campo da ciência da computação fornece um conjunto de recursos para se extrair conhecimento a partir dos dados, em particular, análise baseadas em grafos têm ganhado atenção principalmente pela explosão das redes sociais (VATTER et al., 2023). Neste contexto, as redes sociais representam essencialmente as associações entre as entidades que partilham cada tipo significativo de interligação e podem ser empregadas para modelar os fenômenos que ocorrem no mundo real, a análise da estrutura de colaboração em pesquisa está entre as funções mais destacadas destas redes (Reza HabibAgahi et al., 2022). Tais análises são importantes para identificar aspectos como a qualidade das produções bibliográficas, colaborações entre áreas e instituições, além de direcionar entidades avaliadoras e de fomento à pesquisa.

No Brasil, um dos principais órgãos de financiamento e fomento à pesquisa científica e desenvolvimento tecnológico é o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), que agrega diferentes tipos de programas, prêmios, parcerias, bolsas de estudo e de pesquisa para alocação e distribuição de recursos (ALVES; SILVA, 2022). Dentre estes, destacam-se a Bolsa de Produtividade em Pesquisa (PQ) e a Bolsa de Produtividade em Desenvolvimento Tecnológico e Extensão Inovadora (DT). Conforme descrito na Chamada CNPq Nº 09/2023[1], essas bolsas visam valorizar pesquisadores com produção científica, tecnológica e de inovação de destaque em suas respectivas áreas do conhecimento, bem como incentivar o aumento da produção nessas esferas. O CNPq também, tem gerência da plataforma Lattes que agrega as informações de cada pesquisador acerca das produções, colaborações, filiações, dentre outras informações; sendo esta, uma fonte de informação relevante para análise de coautoria no cenário brasileiro (JUSTINO et al., 2023).

As bolsas de produtividade PQ e DT são organizadas hierarquicamente em três níveis estruturados em ordem crescente, da seguinte maneira: Pesquisador 2 (PQ/DT-2); Pesquisador 1 (PQ/DT-1) e Pesquisador Sênior (PQ/DT-Sr). Ademais, a categoria

[1] http://resultado.cnpq.br/2426435389813022

PQ/DT-1 é classificada em quatro categorias diferentes (ordem crescente): D, C, B e A, resultando nas subcategorias 1D, 1C, 1B e 1A, respectivamente (OLIVEIRA et al., 2018). Além dos critérios relativos ao pesquisador, são avaliados também (a) o mérito científico do projeto, (b) o desempenho do pesquisador quanto à relevância, originalidade e repercussão da produção científica, (c) a atuação contínua na formação de recursos humanos em nível de Pós-Graduação, (d) a inserção internacional, (e) a participação do proponente em edições científicas e em atividades de gestão acadêmica e científica, entre outros (NEIVA et al., 2022). O CNPq aplica cerca de 25% da verba total na manutenção das bolsas destes pesquisadores (NEIVA et al., 2022).

Neste ensejo, o presente trabalho tem como objetivo identificar e caracterizar os bolsistas de produtividade do CNPq e sua dinâmica de coautoria. Foram coletados 12.345 *Curriculum Vitae* (CV) dos bolsistas de produtividade em pesquisa e desenvolvimento tecnológico da plataforma Lattes, com aproximadamente 400 mil publicações. Também foram analisados aspectos ligados à equidade de gênero, regionalidade, áreas de atuação e coautoria. O estudo seguiu boas práticas de ciência de dados, seguindo a metodologia *Data Science Trajectory* (DST) (MARTíNEZ-PLUMED et al., 2021). Os resultados obtidos possibilitam a avaliação desta importante política de fomento à pesquisa.

2.2 TRABALHOS RELACIONADOS

(ORNELLAS et al., 2020) realizaram análise de redes de colaboração com foco nas áreas da medicina veterinária e administração. Os resultados obtidos apontam que as redes de colaboração tem um papel importante no crescimento da produção científica no país, bem como impulsionam na formação de mestres e doutores. No entanto, vale destacar a limitação do estudo quanto as áreas de abrangência. (WELTER et al., 2021) analisaram redes de coautoria nos Programas de Pós-Graduação (PPGs) de contabilidade do quadriênio de 2013-2016. Os autores evidenciaram um aumento na colaboração entre pesquisadores no Brasil ao comparar a estudos anteriores, embora a rede de coautoria entre os docentes possua baixa densidade, indicando pouca cooperação entre os mesmos. (WERNECK et al., 2023) analisaram a coautoria em artigos publicados no período de 2000 a 2020, indexados na *Web of Science, SciELO* e *Scopus*. O principal grupo amostral da pesquisa são docentes vinculados a PPGs de Artes, em Minas Gerais. No entanto, vale destacar as limitações destes estudos quanto a quantidade diminuta de artigos avaliados, influenciando diretamente na qualidade e confiança dos resultados.

(DIAS et al., 2019) apresentam um estudo sobre as redes de colaboração científica dos doutores brasileiros. Ao todo foram coletados mais de 200.000 CVs englobando dados do período de 1960 a 2014. Os resultados demostraram uma visão abrangente da evolução na colaboração entre doutores na ciência nacional. Pequenos grupos de pesquisa,

frequentemente da mesma área e instituições, podem ser observados como contribuintes ativos, ao passo que é evidenciado um aumento na colaboração entre grupos massivos nas grandes áreas do conhecimento. O trabalho é de grande relevância para uma análise temporal de colaboração de pesquisadores no cenário nacional, bem como a utilização de grafos para identificação de comunidades de pesquisa, influenciando diretamente na escolha metodológica do presente estudo. O presente trabalho faz um estudo dos atuais bolsistas de produtividade cadastrados na plataforma *Lattes*, fazendo análises baseadas em distribuição de gênero, geográfica e visualização de redes de coautoria presente nestes níveis de pesquisa sendo este um diferencial em relação ao trabalho apresentado dado que o mesmo se limita a análise até 2014.

(PERES et al., 2019) analisaram as produções científicas no Brasil na última década com foco na área da ciência da computação. Os autores utilizaram técnicas de mineração de texto aplicadas às informações das produções dos pesquisadores da área registradas na plataforma Lattes e apresentaram a evolução dos temas de interesse na pesquisa do campo da ciência da computação no Brasil, demonstrando que os temas acompanham a evolução tecnológica. No presente estudo, buscou-se expandir as áreas para cobrir todo o cenário de PQ/DT como visto na Tabela 2.1, analisando as características dos bolsistas de produtividade do CNPq utilizando outras técnicas como análise das redes de coautoria, possibilitando uma visão mais geral da ciência em todo país.

Tabela 2.1 – Comparação dos correlatos

Trabalhos	**Gênero**	**Geográfica**	**Coautoria**	**Escopo**
(ORNELLAS et al., 2020)	✗	✗	✓	Colaboração nas áreas de medicina veterinária e administração
(WELTER et al., 2021)	✗	✗	✓	Colaboração entre programas de pós-graduação de contabilidade
(WERNECK et al., 2023)	✗	✗	✓	Coautoria entre Docentes do programa de pós-graduação em Artes
(DIAS et al., 2019)	✗	✗	✓	Colaboração de doutores brasileiros entre 1960 e 2004
Presente Estudo	✓	✓	✓	Caracterização de bolsistas de produtividade do Brasil

2.3 METODOLOGIA

A metodologia adotada para este trabalho foi o modelo *Data Science Trajectory* (MARTíNEZ-PLUMED et al., 2021). Esta abordagem representa uma evolução do bem-estabelecido processo *CRoss-Industry Standard Process for Data Mining* (CRISP-DM). O DST visa expandir o escopo do CRISP-DM para englobar projetos com uma natureza mais exploratória, refletindo a evolução contínua da área de ciência de dados ao longo

das últimas décadas (MARTíNEZ-PLUMED et al., 2021). Desta maneira, a seleção desta abordagem para este estudo se deu em virtude da necessidade exploratória para construção de redes de coautoria, e da inclusão de diversas etapas de manipulação e preparação dos dados para viabilizar a construção da rede. A instanciação do DST para o presente trabalho envolveu as etapas descritas a seguir.

O **Entendimento do Negócio** se concentrou na compreensão dos objetivos do projeto. A definição dos objetivos foi realizada em conjunto com a revisão de literatura para obtenção de informações sobre as abordagens que vem sendo utilizadas para criação e análise de redes de coautoria, conforme discutido na Seção 2.2. A etapa de **Aquisição de Dados** envolveu a coleta dos dados a partir da plataforma Lattes, o que foi feito em duas fases. A primeira, serviu para identificar currículos dos bolsistas de produtividade e o segundo para coletá-los. Para o primeiro, foi utilizado um *webCrawler* construído na linguagem *Python*. A extração (segunda coleta), foi realizada por meio da *Application Programming Interface* (API) do Lattes, utilizando a chave de acesso institucional. No total, foram coletados 12.345 currículo no período de novembro a dezembro de 2023. No **Entendimento dos Dados** conduziu-se a identificação dos atributos mais relevantes para a consolidação do conjunto de dados. Nesse sentido, foram selecionados dez atributos ao final desta etapa, incluindo: *Nome, Nome em citação, Instituição atual, Instituição de doutorado, Ano de doutorado, Áreas de atuação, Resumo do currículo, Unidade Federativa, Artigos publicados e Capítulos de livros publicados*. É importante ressaltar assim como conferências são de grande importância para computação, os capítulos de livro são influentes para áreas humanas e biológicas (SILVA et al., 2019) - por este motivo foram considerados para as análises. Posteriormente, na etapa de **Preparação dos Dados**, seguiu-se o *pipeline* convencional da mineração de textos, a saber: remoção de valores nulos, remoção de duplicatas e transformação para caixa baixa tal como em (JUSTINO et al., 2023).

A **Exploração do valor dos dados** foi feita por meio da seleção das instituições que mais publicam e inspeção suas colaborações com outras instituições, além da identificação dos autores mais ativos. Neste ensejo percebeu-se também a viabilidade de se visualizar distribuição geográfica por estados, quanto a distribuição por gênero de todos e todas os(as) bolsistas. Outro valor identificado foi o de realizar uma análise das instituições que mais cooperam, assim é possível identificar parcerias estratégicas e áreas de pesquisa em comum. Por fim, ao explorar os autores mais produtivos, foi possível identificar líderes em suas áreas de atuação. Na última etapa, de **Exploração dos resultados**, os artefatos produzidos foram melhor definidos e parametrizados - incluindo detalhamento tecnológico. Por exemplo, os gráficos para identificação de gêneros foram obtidos por meio da API alimentada por uma base do Instituto Brasileiro de Geografia e Estatística (IBGE) [1]. Foi utilizada a implemen-

[1] https://servicodados.ibge.gov.br/api/docs/nomes?versao=2#api-Ranking

tação disponível em: **<https://github.com/fabiolobato/KDMILE_CNPq.git>**. Para a criação e visualização das redes construídas foi utilizada a ferramenta *Gephi*[1]. Para este estudo foram realizadas análises usando as seguintes métricas da ferramenta: grau médio, *PageRank* e modularidade. O Grau médio permite identificar os nós com maior destaque com base no número de arestas incidentes neles, ou seja, os pesquisadores com a maior variedade de colaboração. A modularidade é uma métrica que demonstra uma relação mais forte entre um grupo de nós, destacando os *clusters*, pequenas comunidades de colaboração na grande rede. O *PageRank* foi usado como parâmetro para auxiliar na visualização dos nós centrais de cada *cluster*.

2.4 RESULTADOS

Considerando os 12.345 currículos de bolsistas de produtividade, abrangendo desde o nível 1A até o nível 2 do CNPq, foi construída uma base de dados das produções acadêmicas desses pesquisadores, totalizando 374.571 artigos completos publicados em periódicos. Para a análise geográfica, consideramos apenas os pesquisadores que informam o estado da instituição atual em seus currículos - resultando em uma amostra de 8.144 pesquisadores. A Figura 2.1 mostra a distribuição geográfica dos pesquisadores analisados, com destaque para as regiões Sul e Sudeste, onde se observa a maior concentração de bolsistas de produtividade do país.

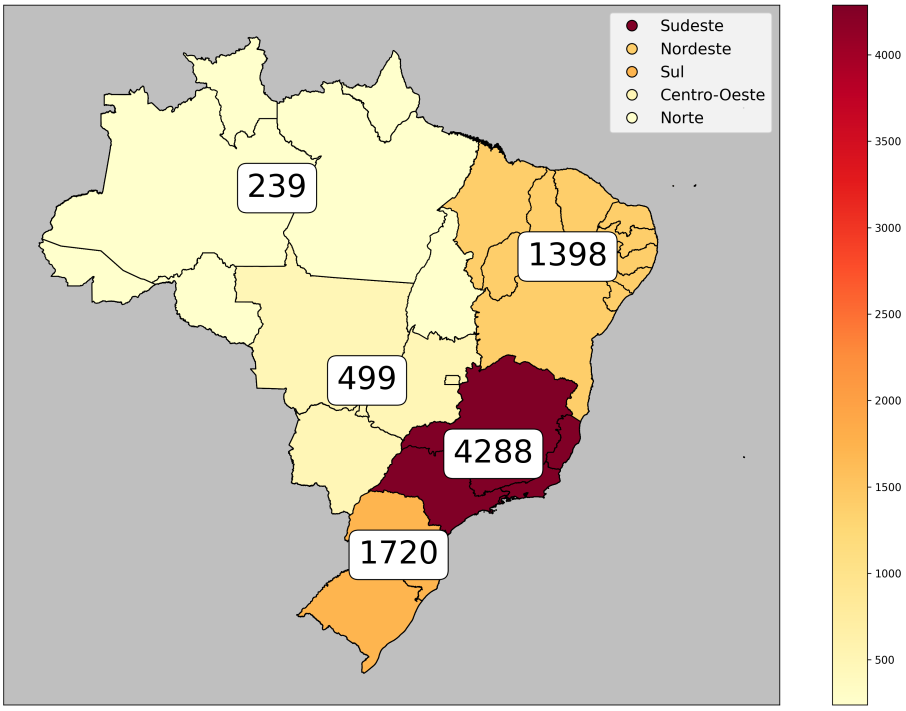

Figura 2.1 – Mapa da distribuição de pesquisadores por região.

[1]https://gephi.org/

Tabela 2.2 – Densidade de Bolsistas de cada nível por Região

Nível	Sudeste		Nordeste		Sul		Centro Oeste		Norte	
	Quantidade	(%)	Quantidade	(%)	Quantidade	(%)	Quantidade	(%)	Quantidade	(%)
1A	436	10	82	5	128	7	23	4	8	3
1B	389	9	86	6	139	8	18	3	16	6
1C	433	10	149	10	162	9	39	7	22	9
1D	648	15	195	14	269	15	70	14	30	12
2	2382	55	886	63	1022	60	349	70	163	68

No entanto, é importante considerar também as populações de pesquisadores em cada região do país, dado que um indicador importante de viés é a densidade de pesquisadores. Por exemplo, ao avaliar o nível 1A, é necessário considerar a proporção de pesquisadores desse nível em relação ao total de pesquisadores da mesma região. Deste modo, com base na distribuição geográfica dos pesquisadores ilustrada na Figura 2.1, foi possível identificar a porcentagem de pesquisadores de cada nível nas regiões, bem como mostra a Tabela 2.2.

A análise da Tabela 2.2 mostra uma significativa disparidade nos níveis de pesquisadores nas regiões do Brasil, com destaque para a região sudeste, que possui 10% dos pesquisadores no Nível 1A, em contraste com apenas 3% na região Norte. Por outro lado, as regiões norte e centro-oeste apresentam a maior predominância de pesquisadores no Nível 2 de produtividade, com 68% e 70%, respectivamente. Dessa forma, pode-se considerar a implementação de políticas eficazes de incentivo em regiões com menor concentração de pesquisadores, como o norte e centro-oeste. Além disso, ações como a criação de programas de capacitação, direcionamento de recursos e estabelecimento de parcerias entre instituições podem promover uma distribuição mais equitativa da produção científica de alto nível por todo o país (RIBEIRO et al., 2020). Foi conduzida uma análise da distribuição por gênero entre os bolsistas de produtividade, onde evidencia-se a predominância masculina, conforme ilustrado na Figura 2.2. Isso pode indicar para as instituições de fomento sobre a necessidade de políticas voltadas para exclusivamente para mulheres, conforme descrito na Chamada CNPq Nº 31/2023 [1], para alcançar uma maior equidade a distribuição gêneros dos pesquisadores (OLIVEIRA et al., 2021).

As redes de coautoria construídas são comportas por nós que representam os bolsistas de produtividade e as arestas a colaboração entre estes pesquisadores em trabalhos publicados. Para efeitos de visualização, os nomes foram ocultados e substituídos por IDs. Visando identificar as áreas com os pesquisadores e pesquisadoras mais relevantes nos níveis mais altos das bolsas PQ/DT, a rede de coautoria foi minuciosamente inspecionada nos níveis 1A e 1B. Isso permitiu identificar não apenas as áreas de destaque, mas também

[1] https://www.gov.br/cnpq/pt-br/cnpq-mcti-mmulheres-n-31-2023-meninas-nas-ciencias-exatas-engenharias-e-computacao.pdf/view

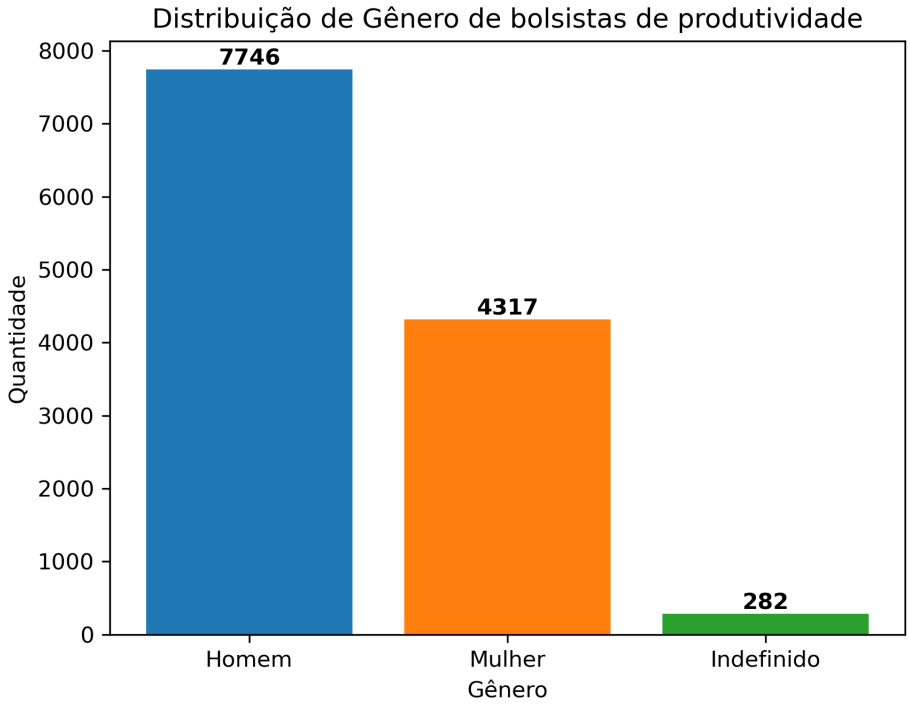

Figura 2.2 – Distribuição por gênero entre bolsistas de produtividade.

as instituições com o maior número de autores pertencentes a esses grupos de bolsistas, bem como as comunidades de pesquisa. Esse processo foi fundamental para validar as redes de colaboração identificadas, como evidenciado na Tabela 2.3.

Tabela 2.3 – Principais pesquisadores do nível 1A.

ID	Grau	Áreas de atuação	Instituição
534	24	Epidemiologia, Saúde Materno-Infantil	UFMA
430	18	Toxicologia, neurociências	UFRGS
741	18	Farmacologia Bioquímica e Molecular	UFMG
175	16	Helmintologia de Parasitos	UNESP
321	15	Clínica Médica, Educação Médica	USP
322	15	Citologia e Biologia Celular	UFRJ

A análise da Tabela 2.3 mostra a predominância das áreas relacionadas a ciência médica. Além disso, ao utilizar grafos como ferramentas de visualização destas redes, é possível notar a relação pesquisadores nesse nível, assim como os bolsistas com maior destaque (colaborações) nesta rede, estes fazendo grande contribuição na construção de comunidades de pesquisas no Brasil. A Figura 2.3 mostra o grafo construído a partir da rede de colaboração dos bolsistas do nível 1A, levando em consideração as métricas de Grau Médio e *PageRank* e modularidade. Os nós de maior destaque no grafo representam os pesquisadores com mais colaborações. É possível notar a forte relação entre os nós 741, 534 e 321. Estes pesquisadores estão entre os cinco mais importantes da rede, como mostrados na 2.3, construindo uma comunidade de coautoria em pesquisa na área médica, entre as instituições UFMA, UFMG e USP. Com relação a rede de coautoria do nível 1B, a

Tabela 2.4 – Principais pesquisadores destacados no grafo 1B.

ID	Grau	Áreas de atuação	Instituição
479	22	Nutrição e Alimentação Animal	UESB
254	16	Biologia Molecular, Genômica, Bioinformatica	USP
450	14	Fisiologia da Produção, Fitotecnia	UNESP
568	12	Microbiologia Aplicada, Sistemas de Computação, Biotecnologia	UFPE
340	11	Genética Humana e Médica, Conservação	ULBRA
461	9	Fitossanidade, Biologia Molecular	UFLA

Tabela 2.4 mostra os bolsistas de maior impacto nesse nível. É importante notar a presença de outras áreas, tais como Bioinformática, Sistemas de Computação e Biotecnologia.

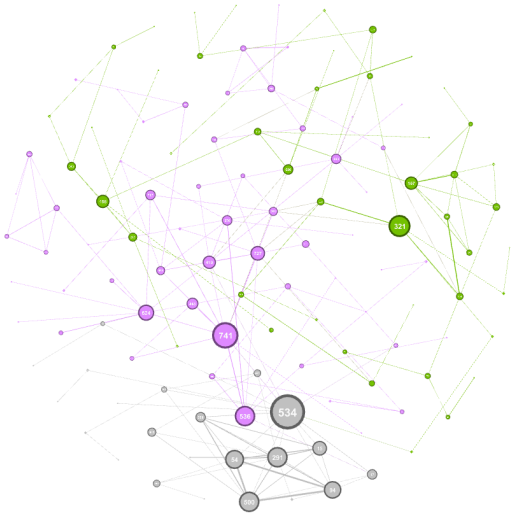

Figura 2.3 – Grafo de coautoria entre bolsistas de produtividade 1A

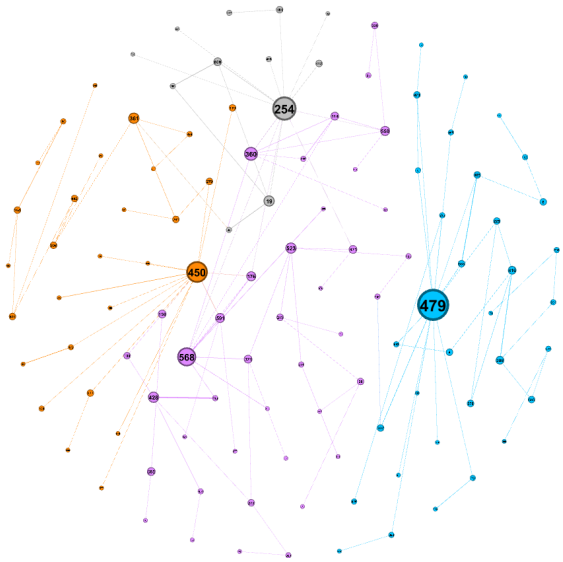

Figura 2.4 – Grafo de coautoria entre bolsistas de produtividade 1B.

Ao analisar o grafo da Figura 2.4, correlacionando com bolsistas destacados na Tabela 2.4, é possível notar a relação entre estes pesquisadores, com destaque para nós 479, 254, 450 e 568 que são os nós centrais nesta rede com maiores números de colaborações. O nó 568 faz a conexão entre as outras subredes de pesquisa destacadas no grafo. Esta análise mostra pesquisadores de áreas distintas em colaboração por meio desse conector. Ademais, também é possível perceber que estes pesquisadores são de instituições de diferentes estados brasileiros, sendo em sua maioria das regiões sul, sudeste e nordeste.

No geral, os artefatos construídos validam as análises exploratórias iniciais, destacando a predominância das universidades da Região Sul e Sudeste e Nordeste nas comunidades de pesquisa do Brasil. Essas observações destacam a interdisciplinaridade na pesquisa e uma possível necessidade de explorar mais profundamente suas atividades e colaborações. Identificar possíveis obstáculos à integração em comunidades existentes também pode ser crucial para promover uma colaboração mais diversificada e inclusiva na academia. Esses *insights* podem ser valiosos não apenas para a compreensão da dinâmica

da colaboração acadêmica, mas também para orientar agências de fomento na elaboração de políticas públicas mais adequadas e que sirvam como elementos promotores da igualdade e diversidade.

2.5 CONSIDERAÇÕES FINAIS

Neste estudo foram definidas e caracterizadas as redes de coautoria entre pesquisadores com bolsa de produtividade em pesquisa, com o objetivo de fornecer *insights* sobre a colaboração e dinâmicas de produção entre esses profissionais. A estrutura de grafos foi utilizada para representar as redes de coautoria nos níveis 1A e 1B. Além disso, foram realizadas análises exploratórias nas bases de currículos Lattes. Os resultados indicam uma predominância do gênero masculino entre os bolsistas de destaque, bem como uma concentração na grande área "Saúde". As métricas de grafo, como grau médio, classe de modularidade e *PageRank*, revelaram padrões de colaboração e sub comunidades de pesquisa. Essas descobertas têm implicações significativas para políticas de fomento à pesquisa, destacando a necessidade de promover maior equidade de gênero e regional na distribuição de recursos e oportunidades de pesquisa no país.

Apesar dos resultados promissores, é importante destacar as limitações deste estudo, bem como a restrição à análise dos níveis de bolsas 1A e 1B e as limitações de tempo e custo computacional para expansão para outras categorias e áreas do conhecimento. Além disso, investigar as porcentagens de autores com bolsas tanto PQ quanto DT, visando uma compreensão mais abrangente das dinâmicas de produção acadêmica no Brasil. Este estudo contribui não apenas para o entendimento das colaborações entre pesquisadores de alto nível, mas também fornece *insights* valiosos para o desenvolvimento de estratégias de colaboração que promovam uma comunidade de pesquisa mais equitativa e inclusiva, promovendo avanços significativos na pesquisa científica em todo o país.

AGRADECIMENTOS

Este trabalho foi apoiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)-DT-303031/2023-9, PIBIC - 161082/2023-8; e pela Fundação Amazônia de Amparo a Estudos e Pesquisas (FAPESPA) PRONEM-FAPESPA/CNPq nº 045/2021.

3 CONSIDERAÇÕES FINAIS

A importância deste estudo para a área de Ciência da Computação reside na interseção entre tecnologia, sociedade e análise de dados para embasar decisões estratégicas. Ao realizar uma caracterização dos bolsistas de produtividade do CNPq com ênfase nas redes de coautoria dos pesquisadores dos níveis 1A e 1B, utilizando a modelagem por grafos. o estudo revela padrões significativos como a predominância do gênero masculino e uma concentração geográfica nas regiões sudeste e sul do país, isso pode gerar insumos para as agências de fomento. Esses achados oferecem subsidio para agências de fomento na criação de políticas públicas.

Durante a realização deste trabalho, o autor teve seu primeiro contato com a pesquisa científica e pôde vivenciar, na prática, como a computação pode ser aplicada para solucionar problemas reais. Nesse processo, desenvolveu habilidades essenciais, como leitura e escrita acadêmica, além de aprimorar sua comunicação e colaboração em equipe. A interação com os coautores, que contribuíram com suas expertises na validação dos resultados e nas revisões, foi fundamental para a elaboração do estudo.

As revisões obtidas por parte dos eventos foram de grande importância para o discente, ao trazerem novas perspectivas sobre o tema e levaram a novas análises no *corpus*. A apresentação do trabalho e a participação do discente no evento KDMile e Bracis foi de grande valia para a sua trajetória acadêmica. Propiciando ao autor o contato com pesquisadores renomados na área e com estudos de temas de fronteira na computação. Além disso, a experiência de apresentar um trabalho acadêmico foi enriquecedora, contribuindo para o desenvolvimento de habilidades de comunicação essenciais tanto para o meio acadêmico quanto para o mercado de trabalho além de enriquecer o currículo do autor.

A partir dos achados deste estudo e das colaborações feitas pela banca durante a defesa deste trabalho, tornou-se evidente a relevância de investigações futuras que aprofundem a compreensão sobre a distribuição e colaboração dos pesquisadores no Brasil. A análise das redes de coautoria não somente revelam padrões estruturais, mas também abre espaço para reflexões sobre equidade e representatividade no cenário acadêmico. Além disso, ao demonstrar o impacto da computação na análise de dados científicos, este trabalho reforça a importância de abordagens interdisciplinares que integrem tecnologia e políticas públicas, contribuindo para um ambiente acadêmico mais diverso e inclusivo.

Por fim, os trabalhos futuros desta pesquisa, já está em andamento, com o objetivo de expandir o presente estudo de modo a abranger todos os níveis de bolsas de produtividade

do CNPq, e investigar outros aspectos como internacionalização das produções cientificas, analisar Qualis dos periódicos de publicação, o fenômeno do predatismo acadêmico e a aplicação de outras técnicas de mineração de texto para a análise deste grande *corpus* da ciência brasileira.

REFERÊNCIAS

ALVES, C. A.; SILVA, L. L. da. Perfil e produção científica dos bolsistas de produtividade em pesquisa do cnpq atuantes na educação matemática. **Educação Matemática Pesquisa Revista do Programa de Estudos Pós-Graduados em Educação Matemática**, v. 24, n. 3, p. 393–429, 2022.

COUTO, M.; CARMO, F.; JR, A. J.; MARCACINI, R.; LOBATO, F. Characterization of co-authorship networks of cnpq productivity fellows: an approach based on data science. In: **Anais do XII Symposium on Knowledge Discovery, Mining and Learning**. Porto Alegre, RS, Brasil: SBC, 2024. p. 113–120. ISSN 2763-8944. Disponível em: <https://sol.sbc.org.br/index.php/kdmile/article/view/30955>.

DIAS, T. M. R.; MOITA, G. F.; DIAS, P. M. Um estudo sobre a rede de colaboração científica dos pesquisadores brasileiros com currículos cadastrados na plataforma lattes. **Em Questão**, v. 25, n. 1, p. 63–86, 2019.

JUSTINO, A.; NHOATTO, F.; SOUZA, E.; LOBATO, F. Analise orientada a dados como auxílio para tomada de decisao em gestao de pesquisa. **Revista da CGU**, v. 15, n. 27, p. 203–219, 2023.

MARTíNEZ-PLUMED, F.; CONTRERAS-OCHANDO, L.; FERRI, C.; HERNáNDEZ-ORALLO, J.; KULL, M.; LACHICHE, N.; RAMíREZ-QUINTANA, M. J.; FLACH, P. Crisp-dm twenty years later: From data mining processes to data science trajectories. **IEEE Transactions on Knowledge and Data Engineering**, v. 33, n. 8, p. 3048–3061, 2021.

NEIVA, S. C. P. C.; DUTRA, A. C.; SILVA, V. de S.; FONSECA, M. C. C. da; SILVA, C. M. Perspectivas da ciência brasileira: um estudo sobre a distribuição de bolsas de pesquisa em produtividade do cnpq ao longo do ano de 2019. **Revista Interdisciplinar Científica Aplicada**, v. 16, n. 3, p. 51–71, 2022.

OLIVEIRA, A.; MELO, M. F. d.; RODRIGUES, Q. B. d.; PEQUENO, M. Gênero e desigualdade na academia brasileira: uma análise a partir dos bolsistas de produtividade em pesquisa do cnpq. **Configurações. Revista Ciências Sociais**, Centro Interdisciplinar de Ciências Sociais, v. 27, n. 27, p. 75–93, 2021.

OLIVEIRA, C. A. d.; COSTA, B. I. R.; OLIVEIRA, M. Genealogia acadêmica: um estudo dos pesquisadores da área de ciência da informação com bolsas produtividade em pesquisa na modalidade pq-1 do cnpq. **Encontro Brasileiro de Bibliometria e Cientometria**, v. 6, p. 291–300, 2018.

ORNELLAS, A. L. M. B.; SILVA, B. G.; ARAÚJO, U. P. Sucesso na área acadêmica: evidências nas redes de coautoria de pesquisadores estrelas na área da administração e da medicina veterinária. **Estudos de Administração e Sociedade**, v. 5, n. 2, p. 41–53, 2020.

PERES, L.; CECILIO, P.; RODRIGUES, F.; SILVA, N.; ROCHA, L. An overview of brazilian researches in the computer science field in last years. In: **Anais do VII**

Symposium on Knowledge Discovery, Mining and Learning. [S.l.]: SBC, 2019. p. 9–16.

PROVOST, F.; FAWCETT, T. Data science and its relationship to big data and data-driven decision making. **Big data**, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 1, n. 1, p. 51–59, 2013.

Reza HabibAgahi, M.; Agha Mohammad Ali Kermani, M.; MAGHSOUDI, M. On the co-authorship network analysis in the process mining research community: A social network analysis perspective. **Expert Systems with Applications**, v. 206, p. 117853, 2022. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417422011083>.

RIBEIRO, D. B.; OLIVEIRA, E. F. d. A.; DENADAI, M. C. V. B.; GARCIA, M. L. T. Financiamento à ciência no brasil: distribuição entre as grandes áreas do conhecimento. **Revista Katálysis**, SciELO Brasil, v. 23, p. 548–561, 2020.

SILVA, G. C. d. O.; IRIART, J. A. B.; CHAVES, S. C. L.; ABADE, E. A. F. Characteristics of research on child labor in latin america. **Cadernos de Saude Publica**, SciELO Brasil, v. 35, p. e00031018, 2019.

VATTER, J.; MAYER, R.; JACOBSEN, H.-A. The evolution of distributed systems for graph neural networks and their origin in graph processing and deep learning: A survey. **ACM Computing Surveys**, ACM New York, NY, USA, v. 56, n. 1, p. 1–37, 2023.

WELTER, L. M.; SOUZA, Â. R. L. d.; TRAJANO, B. B.; BEHR, A. Redes de coautoria dos programas brasileiros de pós-graduação em contabilidade. **Contextus: Revista Contemporânea de Economia e Gestão**, v. 19, n. 10, p. 146–159, 2021.

WERNECK, L.; PINTO, T.; LOULA, A.; QUEIROZ, J. et al. Mapeando redes de co-autoria na comunidade acadêmica de artes em minas gerais. **SciELO Preprints**, 2023.

APÊNDICE A – ARTIGO ORIGINAL

Characterization of co-authorship networks of CNPq productivity fellows: an approach based on data science

Matheus L. do Couto[1], Fernando A. do Carmo[1], Antonio F. L. Jacob Jr[2], Ricardo M. Marcacini[3], Fábio M. F. Lobato[1]

[1] Universidade Federal de Oeste do Pará, Brazil
fabio.lobato@ufopa.edu.br
[2] Universidade Estadual do Maranhão, Brazil
antoniojunior@professor.uema.br
[3] Universidade de São Paulo
ricardo.marcacini@icmc.usp.br

Abstract.

Co-authorship networks are widely used to evaluate the quality of scientific literature productions and collaborations between researchers and institutions. We identified a gap in the literature regarding the analysis of interactions between CNPq productivity fellows. To fill this gap, we used data science to characterize co-authorship networks to obtain a complementary overview of this critical public policy for promoting excellent research in Brazil. To this end, we collected 12,345 researchers' Lattes CVs and analyzed approximately 400,000 publications. The results showed greater collaboration in the higher strata of grants (e.g., 1A and 1B). Other research findings of interest are related to regional discrepancies and gender equity. The study contributes to a better understanding of the social dynamics of productivity grant recipients, supporting the evaluation of this relevant research promotion policy.

CCS Concepts: • **Information systems** → *Data mining*; • **Computing methodologies** → *Natural language processing*.

Keywords: Co-authorship network, Data Science, Lattes, Scientometric

1. INTRODUÇÃO

A ciência de dados está intrinsecamente relacionada com a tomada de decisões estratégicas [Provost and Fawcett 2013]. Este campo da ciência da computação fornece um conjunto de recursos para se extrair conhecimento a partir dos dados, em particular, análise baseadas em grafos têm ganhado atenção principalmente pela explosão das redes sociais [Vatter et al. 2023]. Neste contexto, as redes sociais representam essencialmente as associações entre as entidades que partilham cada tipo significativo de interligação e podem ser empregadas para modelar os fenômenos que ocorrem no mundo real, a análise da estrutura de colaboração em pesquisa está entre as funções mais destacadas destas redes [Reza HabibAgahi et al. 2022]. Tais análises são importantes para identificar aspectos como a qualidade das produções bibliográficas, colaborações entre áreas e instituições, além de direcionar entidades avaliadoras e de fomento à pesquisa.

No Brasil, um dos principais órgãos de financiamento e fomento à pesquisa científica e desenvolvimento tecnológico é o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), que agrega diferentes tipos de programas, prêmios, parcerias, bolsas de estudo e de pesquisa para alocação e distribuição de recursos [Alves and da Silva 2022]. Dentre estes, destacam-se a Bolsa de Produtividade em Pesquisa (PQ) e a Bolsa de Produtividade em Desenvolvimento Tecnológico e Extensão

Copyright©2024 Permission to copy without fee all or part of the material printed in KDMiLe is granted provided that the copies are not made or distributed for commercial advantage, and that notice is given that copying is by permission of the Sociedade Brasileira de Computação.

Inovadora (DT). Conforme descrito na Chamada CNPq Nº 09/2023[1], essas bolsas visam valorizar pesquisadores com produção científica, tecnológica e de inovação de destaque em suas respectivas áreas do conhecimento, bem como incentivar o aumento da produção nessas esferas. O CNPq também, tem gerência da plataforma Lattes que agrega as informações de cada pesquisador acerca das produções, colaborações, filiações, dentre outras informações; sendo esta, uma fonte de informação relevante para análise de coautoria no cenário brasileiro [Justino et al. 2023].

As bolsas de produtividade PQ e DT são organizadas hierarquicamente em três níveis estruturados em ordem crescente, da seguinte maneira: Pesquisador 2 (PQ/DT-2); Pesquisador 1 (PQ/DT-1) e Pesquisador Sênior (PQ/DT-Sr). Ademais, a categoria PQ/DT-1 é classificada em quatro categorias diferentes (ordem crescente): D, C, B e A, resultando nas subcategorias 1D, 1C, 1B e 1A, respectivamente [Oliveira et al. 2018]. Além dos critérios relativos ao pesquisador, são avaliados também (a) o mérito científico do projeto, (b) o desempenho do pesquisador quanto à relevância, originalidade e repercussão da produção científica, (c) a atuação contínua na formação de recursos humanos em nível de Pós-Graduação, (d) a inserção internacional, (e) a participação do proponente em edições científicas e em atividades de gestão acadêmica e científica, entre outros [Neiva et al. 2022]. O CNPq aplica cerca de 25% da verba total na manutenção das bolsas destes pesquisadores [Neiva et al. 2022].

Neste ensejo, o presente trabalho tem como objetivo identificar e caracterizar os bolsistas de produtividade do CNPq e sua dinâmica de coautoria. Foram coletados 12.345 *Curriculum Vitae* (CV) dos bolsistas de produtividade em pesquisa e desenvolvimento tecnológico da plataforma Lattes, com aproximadamente 400 mil publicações. Também foram analisados aspectos ligados à equidade de gênero, regionalidade, áreas de atuação e coautoria. O estudo seguiu boas práticas de ciência de dados, seguindo a metodologia *Data Science Trajectory* (DST) [Martínez-Plumed et al. 2021]. Os resultados obtidos possibilitam a avaliação desta importante política de fomento à pesquisa.

2. TRABALHOS RELACIONADOS

[Ornellas et al. 2020] realizaram análise de redes de colaboração com foco nas áreas da medicina veterinária e administração. Os resultados obtidos apontam que as redes de colaboração tem um papel importante no crescimento da produção científica no país, bem como impulsionam na formação de mestres e doutores. No entanto, vale destacar a limitação do estudo quanto as áreas de abrangência. [Welter et al. 2021] analisaram redes de coautoria nos Programas de Pós-Graduação (PPGs) de contabilidade do quadriênio de 2013-2016. Os autores evidenciaram um aumento na colaboração entre pesquisadores no Brasil ao comparar a estudos anteriores, embora a rede de coautoria entre os docentes possua baixa densidade, indicando pouca cooperação entre os mesmos. [Werneck et al. 2023] analisaram a coautoria em artigos publicados no período de 2000 a 2020, indexados na *Web of Science, SciELO* e *Scopus*. O principal grupo amostral da pesquisa são docentes vinculados a PPGs de Artes, em Minas Gerais. No entanto, vale destacar as limitações destes estudos quanto a quantidade diminuta de artigos avaliados, influenciando diretamente na qualidade e confiança dos resultados.

[Dias et al. 2019] apresentam um estudo sobre as redes de colaboração científica dos doutores brasileiros. Ao todo foram coletados mais de 200.000 CVs englobando dados do período de 1960 a 2014. Os resultados demostraram uma visão abrangente da evolução na colaboração entre doutores na ciência nacional. Pequenos grupos de pesquisa, frequentemente da mesma área e instituições, podem ser observados como contribuintes ativos, ao passo que é evidenciado um aumento na colaboração entre grupos massivos nas grandes áreas do conhecimento. O trabalho é de grande relevância para uma análise temporal de colaboração de pesquisadores no cenário nacional, bem como a utilização de grafos para identificação de comunidades de pesquisa, influenciando diretamente na escolha metodológica do presente estudo. O presente trabalho faz um estudo dos atuais bolsistas de produtividade cadastrados na plataforma *Lattes*, fazendo análises baseadas em distribuição de gênero, geográfica e visualização de

[1] http://resultado.cnpq.br/2426435389813022

redes de coautoria presente nestes níveis de pesquisa sendo este um diferencial em relação ao trabalho apresentado dado que o mesmo se limita a análise até 2014.

[Peres et al. 2019] analisaram as produções científicas no Brasil na última década com foco na área da ciência da computação. Os autores utilizaram técnicas de mineração de texto aplicadas às informações das produções dos pesquisadores da área registradas na plataforma Lattes e apresentaram a evolução dos temas de interesse na pesquisa do campo da ciência da computação no Brasil, demonstrando que os temas acompanham a evolução tecnológica. No presente estudo, buscou-se expandir as áreas para cobrir todo o cenário de PQ/DT como visto na Tabela I, analisando as características dos bolsistas de produtividade do CNPq utilizando outras técnicas como análise das redes de coautoria, possibilitando uma visão mais geral da ciência em todo país.

Table I. Comparação dos correlatos

Trabalhos	**Gênero**	**Geográfica**	**Coautoria**	**Escopo**
[Ornellas et al. 2020]	✗	✗	✓	Colaboração nas áreas de medicina veterinária e administração
[Welter et al. 2021]	✗	✗	✓	Colaboração entre programas de pós-graduação de contabilidade
[Werneck et al. 2023]	✗	✗	✓	Coautoria entre Docentes do programa de pós-graduação em Artes
[Dias et al. 2019]	✗	✗	✓	Colaboração de doutores brasileiros entre 1960 e 2004
Presente estudo	✓	✓	✓	Caracterização de bolsistas de produtividade do Brasil

3. METODOLOGIA

A metodologia adotada para este trabalho foi o modelo *Data Science Trajectory* [Martínez-Plumed et al. 2021]. Esta abordagem representa uma evolução do bem-estabelecido processo *CRoss-Industry Standard Process for Data Mining* (CRISP-DM). O DST visa expandir o escopo do CRISP-DM para englobar projetos com uma natureza mais exploratória, refletindo a evolução contínua da área de ciência de dados ao longo das últimas décadas [Martínez-Plumed et al. 2021]. Desta maneira, a seleção desta abordagem para este estudo se deu em virtude da necessidade exploratória para construção de redes de coautoria, e da inclusão de diversas etapas de manipulação e preparação dos dados para viabilizar a construção da rede. A instanciação do DST para o presente trabalho envolveu as etapas descritas a seguir.

O **Entendimento do Negócio** se concentrou na compreensão dos objetivos do projeto. A definição dos objetivos foi realizada em conjunto com a revisão de literatura para obtenção de informações sobre as abordagens que vem sendo utilizadas para criação e análise de redes de coautoria, conforme discutido na Seção 2. A etapa de **Aquisição de Dados** envolveu a coleta dos dados a partir da plataforma Lattes, o que foi feito em duas fases. A primeira, serviu para identificar currículos dos bolsistas de produtividade e o segundo para coletá-los. Para o primeiro, foi utilizado um *webCrawler* construído na linguagem *Python*. A extração (segunda coleta), foi realizada por meio da *Application Programming Interface* (API) do Lattes, utilizando a chave de acesso institucional. No total, foram coletados 12.345 currículo no período de novembro a dezembro de 2023. No **Entendimento dos Dados** conduziu-se a identificação dos atributos mais relevantes para a consolidação do conjunto de dados. Nesse sentido, foram selecionados dez atributos ao final desta etapa, incluindo: *Nome, Nome em citação, Instituição atual, Instituição de doutorado, Ano de doutorado, Áreas de atuação, Resumo do currículo, Unidade Federativa, Artigos publicados e Capítulos de livros publicados.* É importante ressaltar assim como conferências são de grande importância para computação, os capítulos de livro são influentes para áreas humanas e biológicas [Silva et al. 2019] - por este motivo foram considerados para as análises. Posteriormente, na etapa de **Preparação dos Dados**, seguiu-se o *pipeline* convencional da mineração de textos, a saber: remoção de valores nulos, remoção de duplicatas e transformação para caixa baixa tal como em [Justino et al. 2023].

A **Exploração do valor dos dados** foi feita por meio da seleção das instituições que mais publicam e inspeção suas colaborações com outras instituições, além da identificação dos autores mais ativos. Neste ensejo percebeu-se também a viabilidade de se visualizar distribuição geográfica por estados, quanto a distribuição por gênero de todos e todas os(as) bolsistas. Outro valor identificado foi o de realizar uma análise das instituições que mais cooperam, assim é possível identificar parcerias estratégicas e áreas de pesquisa em comum. Por fim, ao explorar os autores mais produtivos, foi possível identificar líderes em suas áreas de atuação. Na última etapa, de **Exploração dos resultados**, os artefatos produzidos foram melhor definidos e parametrizados - incluindo detalhamento tecnológico. Por exemplo, os gráficos para identificação de gêneros foram obtidos por meio da API alimentada por uma base do Instituto Brasileiro de Geografia e Estatística (IBGE) [2]. Foi utilizada a implementação disponível em: `https://github.com/fabiolobato/KDMILE_CNPq.git`. Para a criação e visualização das redes construídas foi utilizada a ferramenta *Gephi*[3]. Para este estudo foram realizadas análises usando as seguintes métricas da ferramenta: grau médio, *PageRank* e modularidade. O Grau médio permite identificar os nós com maior destaque com base no número de arestas incidentes neles, ou seja, os pesquisadores com a maior variedade de colaboração. A modularidade é uma métrica que demonstra uma relação mais forte entre um grupo de nós, destacando os *clusters*, pequenas comunidades de colaboração na grande rede. O *PageRank* foi usado como parâmetro para auxiliar na visualização dos nós centrais de cada *cluster*.

4. RESULTADOS

Considerando os 12.345 currículos de bolsistas de produtividade, abrangendo desde o nível 1A até o nível 2 do CNPq, foi construída uma base de dados das produções acadêmicas desses pesquisadores, totalizando 374.571 artigos completos publicados em periódicos. Para a análise geográfica, consideramos apenas os pesquisadores que informam o estado da instituição atual em seus currículos - resultando em uma amostra de 8.144 pesquisadores. A Figura 1 mostra a distribuição geográfica dos pesquisadores analisados, com destaque para as regiões Sul e Sudeste, onde se observa a maior concentração de bolsistas de produtividade do país.

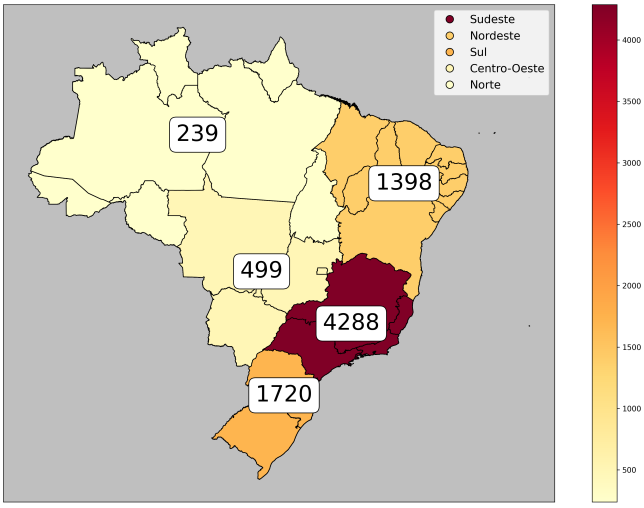

Fig. 1. Mapa da distribuição de pesquisadores por região.

No entanto, é importante considerar também as populações de pesquisadores em cada região do país, dado que um indicador importante de viés é a densidade de pesquisadores. Por exemplo, ao avaliar

[2]https://servicodados.ibge.gov.br/api/docs/nomes?versao=2#api-Ranking
[3]https://gephi.org/

o nível 1A, é necessário considerar a proporção de pesquisadores desse nível em relação ao total de pesquisadores da mesma região. Deste modo, com base na distribuição geográfica dos pesquisadores ilustrada na Figura 1, foi possível identificar a porcentagem de pesquisadores de cada nível nas regiões, bem como mostra a Tabela II.

Table II. Densidade de Bolsistas de cada nível por Região

Nível	Sudeste		Nordeste		Sul		Centro Oeste		Norte	
	Quantidade	(%)	Quantidade	(%)	Quantidade	(%)	Quantidade	(%)	Quantidade	(%)
1A	436	10	82	5	128	7	23	4	8	3
1B	389	9	86	6	139	8	18	3	16	6
1C	433	10	149	10	162	9	39	7	22	9
1D	648	15	195	14	269	15	70	14	30	12
2	2382	55	886	63	1022	60	349	70	163	68

A análise da Tabela II mostra uma significativa disparidade nos níveis de pesquisadores nas regiões do Brasil, com destaque para a região sudeste, que possui 10% dos pesquisadores no Nível 1A, em contraste com apenas 3% na região Norte. Por outro lado, as regiões norte e centro-oeste apresentam a maior predominância de pesquisadores no Nível 2 de produtividade, com 68% e 70%, respectivamente. Dessa forma, pode-se considerar a implementação de políticas eficazes de incentivo em regiões com menor concentração de pesquisadores, como o norte e centro-oeste. Além disso, ações como a criação de programas de capacitação, direcionamento de recursos e estabelecimento de parcerias entre instituições podem promover uma distribuição mais equitativa da produção científica de alto nível por todo o país [Ribeiro et al. 2020]. Foi conduzida uma análise da distribuição por gênero entre os bolsistas de produtividade, onde evidencia-se a predominância masculina, conforme ilustrado na Figura 2. Isso pode indicar para as instituições de fomento sobre a necessidade de políticas voltadas para exclusivamente para mulheres, conforme descrito na Chamada CNPq Nº 31/2023 [4], para alcançar uma maior equidade a distribuição gêneros dos pesquisadores [Oliveira et al. 2021].

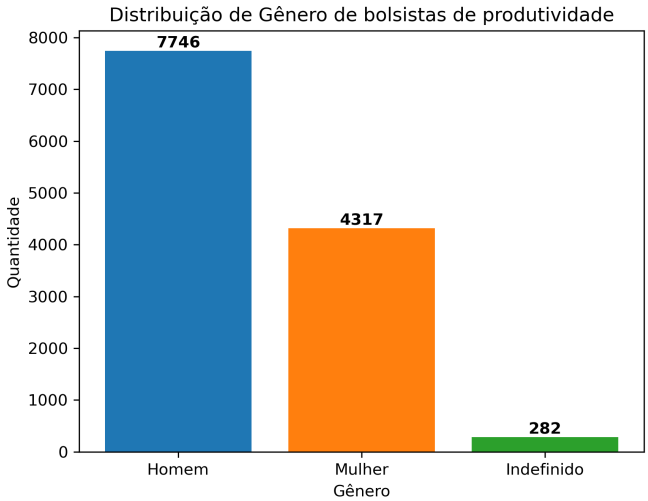

Fig. 2. Distribuição por gênero entre bolsistas de produtividade.

As redes de coautoria construídas são comportas por nós que representam os bolsistas de produtividade e as arestas a colaboração entre estes pesquisadores em trabalhos publicados. Para efeitos de visualização, os nomes foram ocultados e substituídos por IDs. Visando identificar as áreas com os pesquisadores e pesquisadoras mais relevantes nos níveis mais altos das bolsas PQ/DT, a rede de

[4]https://www.gov.br/cnpq/pt-br/cnpq-mcti-mmulheres-n-31-2023-meninas-nas-ciencias-exatas-engenharias-e-computacao.pdf/view

coautoria foi minuciosamente inspecionada nos níveis 1A e 1B. Isso permitiu identificar não apenas as áreas de destaque, mas também as instituições com o maior número de autores pertencentes a esses grupos de bolsistas, bem como as comunidades de pesquisa. Esse processo foi fundamental para validar as redes de colaboração identificadas, como evidenciado na Tabela III.

Table III. Principais pesquisadores do nível 1A.

ID	Grau	Áreas de atuação	Instituição
534	24	Epidemiologia, Saúde Materno-Infantil	UFMA
430	18	Toxicologia, neurociências	UFRGS
741	18	Farmacologia Bioquímica e Molecular	UFMG
175	16	Helmintologia de Parasitos	UNESP
321	15	Clínica Médica, Educação Médica	USP
322	15	Citologia e Biologia Celular	UFRJ

A análise da Tabela III mostra a predominância das áreas relacionadas a ciência médica. Além disso, ao utilizar grafos como ferramentas de visualização destas redes, é possível notar a relação pesquisadores nesse nível, assim como os bolsistas com maior destaque (colaborações) nesta rede, estes fazendo grande contribuição na construção de comunidades de pesquisas no Brasil. A Figura 3 mostra o grafo construído a partir da rede de colaboração dos bolsistas do nível 1A, levando em consideração as métricas de Grau Médio e *PageRank* e modularidade. Os nós de maior destaque no grafo representam os pesquisadores com mais colaborações. É possível notar a forte relação entre os nós 741, 534 e 321. Estes pesquisadores estão entre os cinco mais importantes da rede, como mostrados na III, construindo uma comunidade de coautoria em pesquisa na área médica, entre as instituições UFMA, UFMG e USP. Com relação a rede de coautoria do nível 1B, a Tabela IV mostra os bolsistas de maior impacto nesse nível. É importante notar a presença de outras áreas, tais como Bioinformática, Sistemas de Computação e Biotecnologia.

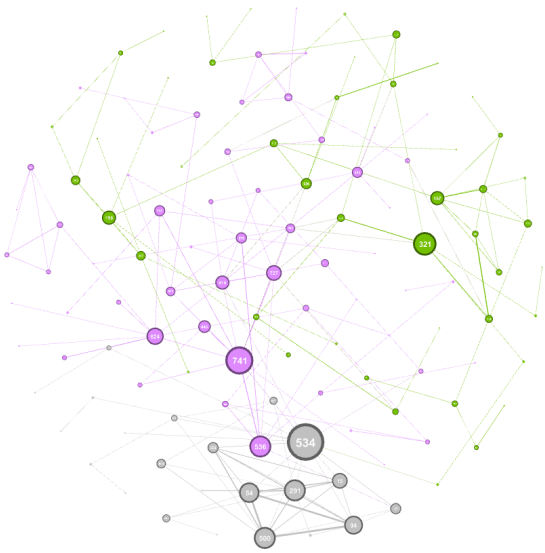

Fig. 3. Grafo de coautoria entre bolsistas de produtividade 1A

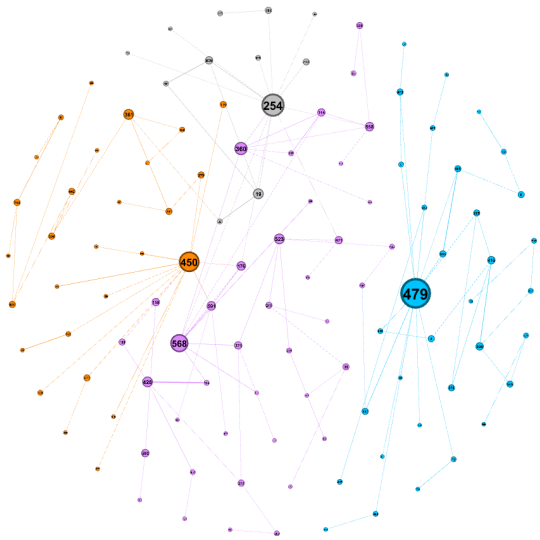

Fig. 4. Grafo de coautoria entre bolsistas de produtividade 1B.

Ao analisar o grafo da Figura 4, correlacionando com bolsistas destacados na Tabela IV, é possível notar a relação entre estes pesquisadores, com destaque para nós 479, 254, 450 e 568 que são os nós centrais nesta rede com maiores números de colaborações. O nó 568 faz a conexão entre as outras subredes de pesquisa destacadas no grafo. Esta análise mostra pesquisadores de áreas distintas em colaboração por meio desse conector. Ademais, também é possível perceber que estes pesquisadores

Table IV. Principais pesquisadores destacados no grafo 1B.

ID	Grau	Áreas de atuação	Instituição
479	22	Nutrição e Alimentação Animal	UESB
254	16	Biologia Molecular, Genômica, Bioinformatica	USP
450	14	Fisiologia da Produção, Fitotecnia	UNESP
568	12	Microbiologia Aplicada, Sistemas de Computação, Biotecnologia	UFPE
340	11	Genética Humana e Médica, Conservação	ULBRA
461	9	Fitossanidade, Biologia Molecular	UFLA

são de instituições de diferentes estados brasileiros, sendo em sua maioria das regiões sul, sudeste e nordeste.

No geral, os artefatos construídos validam as análises exploratórias iniciais, destacando a predominância das universidades da Região Sul e Sudeste e Nordeste nas comunidades de pesquisa do Brasil. Essas observações destacam a interdisciplinaridade na pesquisa e uma possível necessidade de explorar mais profundamente suas atividades e colaborações. Identificar possíveis obstáculos à integração em comunidades existentes também pode ser crucial para promover uma colaboração mais diversificada e inclusiva na academia. Esses *insights* podem ser valiosos não apenas para a compreensão da dinâmica da colaboração acadêmica, mas também para orientar agências de fomento na elaboração de políticas públicas mais adequadas e que sirvam como elementos promotores da igualdade e diversidade.

5. CONSIDERAÇÕES FINAIS

Neste estudo foram definidas e caracterizadas as redes de coautoria entre pesquisadores com bolsa de produtividade em pesquisa, com o objetivo de fornecer *insights* sobre a colaboração e dinâmicas de produção entre esses profissionais. A estrutura de grafos foi utilizada para representar as redes de coautoria nos níveis 1A e 1B. Além disso, foram realizadas análises exploratórias nas bases de currículos Lattes. Os resultados indicam uma predominância do gênero masculino entre os bolsistas de destaque, bem como uma concentração na grande área "Saúde". As métricas de grafo, como grau médio, classe de modularidade e *PageRank*, revelaram padrões de colaboração e sub comunidades de pesquisa. Essas descobertas têm implicações significativas para políticas de fomento à pesquisa, destacando a necessidade de promover maior equidade de gênero e regional na distribuição de recursos e oportunidades de pesquisa no país.

Apesar dos resultados promissores, é importante destacar as limitações deste estudo, bem como a restrição à análise dos níveis de bolsas 1A e 1B e as limitações de tempo e custo computacional para expansão para outras categorias e áreas do conhecimento. Além disso, investigar as porcentagens de autores com bolsas tanto PQ quanto DT, visando uma compreensão mais abrangente das dinâmicas de produção acadêmica no Brasil. Este estudo contribui não apenas para o entendimento das colaborações entre pesquisadores de alto nível, mas também fornece *insights* valiosos para o desenvolvimento de estratégias de colaboração que promovam uma comunidade de pesquisa mais equitativa e inclusiva, promovendo avanços significativos na pesquisa científica em todo o país.

Agradecimentos

Este trabalho foi apoiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)-DT-303031/2023-9, PIBIC - 161082/2023-8; e pela Fundação Amazônia de Amparo a Estudos e Pesquisas (FAPESPA) PRONEM-FAPESPA/CNPq nº 045/2021.

REFERENCES

Alves, C. A. and da Silva, L. L. Perfil e produção científica dos bolsistas de produtividade em pesquisa do cnpq atuantes na educação matemática. *Educação Matemática Pesquisa Revista do Programa de Estudos Pós-Graduados em Educação Matemática* 24 (3): 393–429, 2022.

Dias, T. M. R., Moita, G. F., and Dias, P. M. Um estudo sobre a rede de colaboração científica dos pesquisadores brasileiros com currículos cadastrados na plataforma lattes. *Em Questão* 25 (1): 63–86, 2019.

Justino, A., Nhoatto, F., Souza, E., and Lobato, F. Analise orientada a dados como auxílio para tomada de decisao em gestao de pesquisa. *Revista da CGU* 15 (27): 203–219, 2023.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., and Flach, P. Crisp-dm twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering* 33 (8): 3048–3061, 2021.

Neiva, S. C. P. C., Dutra, A. C., de Souza Silva, V., da Fonseca, M. C. C., and Silva, C. M. Perspectivas da ciência brasileira: um estudo sobre a distribuição de bolsas de pesquisa em produtividade do cnpq ao longo do ano de 2019. *Revista Interdisciplinar Científica Aplicada* 16 (3): 51–71, 2022.

Oliveira, A., Melo, M. F. d., Rodrigues, Q. B. d., and Pequeno, M. Gênero e desigualdade na academia brasileira: uma análise a partir dos bolsistas de produtividade em pesquisa do cnpq. *Configurações. Revista Ciências Sociais* 27 (27): 75–93, 2021.

Oliveira, C. A. d., Costa, B. I. R., and Oliveira, M. Genealogia acadêmica: um estudo dos pesquisadores da área de ciência da informação com bolsas produtividade em pesquisa na modalidade pq-1 do cnpq. *Encontro Brasileiro de Bibliometria e Cientometria* vol. 6, pp. 291–300, 2018.

Ornellas, A. L. M. B., Silva, B. G., and Araújo, U. P. Sucesso na área acadêmica: evidências nas redes de coautoria de pesquisadores estrelas na área da administração e da medicina veterinária. *Estudos de Administração e Sociedade* 5 (2): 41–53, 2020.

Peres, L., Cecilio, P., Rodrigues, F., Silva, N., and Rocha, L. An overview of brazilian researches in the computer science field in last years. In *Anais do VII Symposium on Knowledge Discovery, Mining and Learning*. SBC, pp. 9–16, 2019.

Provost, F. and Fawcett, T. Data science and its relationship to big data and data-driven decision making. *Big data* 1 (1): 51–59, 2013.

Reza HabibAgahi, M., Agha Mohammad Ali Kermani, M., and Maghsoudi, M. On the co-authorship network analysis in the process mining research community: A social network analysis perspective. *Expert Systems with Applications* vol. 206, pp. 117853, 2022.

Ribeiro, D. B., Oliveira, E. F. d. A., Denadai, M. C. V. B., and Garcia, M. L. T. Financiamento à ciência no brasil: distribuição entre as grandes áreas do conhecimento. *Revista Katálysis* vol. 23, pp. 548–561, 2020.

Silva, G. C. d. O., Iriart, J. A. B., Chaves, S. C. L., and Abade, E. A. F. Characteristics of research on child labor in latin america. *Cadernos de Saude Publica* vol. 35, pp. e00031018, 2019.

Vatter, J., Mayer, R., and Jacobsen, H.-A. The evolution of distributed systems for graph neural networks and their origin in graph processing and deep learning: A survey. *ACM Computing Surveys* 56 (1): 1–37, 2023.

Welter, L. M., Souza, Â. R. L. d., Trajano, B. B., and Behr, A. Redes de coautoria dos programas brasileiros de pós-graduação em contabilidade. *Contextus: Revista Contemporânea de Economia e Gestão* 19 (10): 146–159, 2021.

Werneck, L., Pinto, T., Loula, A., Queiroz, J., et al. Mapeando redes de co-autoria na comunidade acadêmica de artes em minas gerais. *SciELO Preprints*, 2023.

ANEXO A – REGISTRO DE PARTICIPAÇÃO DO AUTOR NO EVENTO

CERTIFICADO

BRACIS
Belém/PA - Brazil

A Sociedade Brasileira de Computação atesta a participação de

Matheus Lima do Couto

no(a) XXXIV Brazilian Conference on Intelligent Systems (BRACIS 2024),

realizado(a) entre 17 e 21 de Novembro em Belém - PA, com carga horária de 40 horas.

Profa. Thais Vasconcelos Batista
Presidente da Sociedade Brasileira de Computação

SBC
Sociedade Brasileira de Computação

Denis do Rosário
Prof. Denis Lima do Rosário
Diretor de Eventos e Comissões Especiais