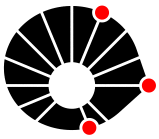

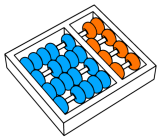

graphology: Análise de Redes de Colaboração Científica na UNICAMP

P. S. Azevedo *R. Interian*

Relatório Técnico - IC-PFG-25-04
Projeto Final de Graduação
2025 - Julho

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE COMPUTAÇÃO

graphology: Análise de Redes de Colaboração Científica na UNICAMP

Pedro Sader Azevedo* Ruben Interian*

Resumo

Este trabalho investiga as dinâmicas de colaboração científica de autores da Universidade Estadual de Campinas (UNICAMP) no período de 2000 a 2025, com ênfase em colaborações externas. Por meio da modelagem usando grafos, foram analisadas relações entre autores, instituições e publicações, e foram identificadas comunidades orgânicas de colaboração. Foi atribuída uma instituição predominante a cada comunidade, o que permitiu o estudo de interações entre comunidades da UNICAMP e comunidades externas. O efeito da primeira colaboração externa no crescimento do impacto (número de citações acumuladas) dos autores foi avaliado. Contrariando a hipótese inicial, foi observada uma desaceleração média desse crescimento após tal colaboração. Os resultados indicam que o impacto científico não responde de forma imediata ou linear à expansão da rede colaborativa, apontando para a necessidade de investigações mais aprofundadas sobre os fatores que influenciam o impacto dos pesquisadores.

1 Introdução

A colaboração científica tem papel central no avanço do conhecimento, permitindo a construção coletiva de soluções para problemas complexos. Assim, a análise dessas relações de colaboração faz-se fundamental para compreender o ecossistema científico, podendo revelar dinâmicas institucionais relevantes e orientar políticas públicas de incentivo à pesquisa.

Este trabalho tem como foco o estudo das colaborações científicas envolvendo autores da Universidade Estadual de Campinas (UNICAMP) entre os anos de 2000 e 2025. Modelamos essas interações usando grafos, já que essa estrutura de dados permite representar relações complexas entre entidades. O grafo foi construído partir de dados de publicações científicas do *Scopus*, então armazenado e analisado utilizando o Neo4j, um banco de dados orientado a grafos.

Nosso objetivo central é investigar a formação de comunidades de colaboração e, em especial, o papel das colaborações externas à UNICAMP no impacto dos pesquisadores da instituição. A hipótese inicial era que essas colaborações externas impulsionariam o crescimento do impacto científico individual, medido por citações. No entanto, os resultados observados desafiam essa suposição e sugerem uma dinâmica mais complexa, que deve ser investigada mais a fundo em trabalhos futuros.

*Instituto de Computação, Universidade Estadual de Campinas, 13081-970 Campinas, SP.

2 Revisão Bibliográfica

A análise de redes de colaboração científica tem se consolidado como uma abordagem robusta na área de cientometria e bibliometria, especialmente por meio da modelagem de dados usando grafos. No cenário brasileiro, destacam-se os trabalhos do Grupo de Análise de Redes Sociais e Cientometria (GARSC), da EACH-USP, que aplicam algoritmos de detecção de comunidades e métricas de centralidade para investigar relações de coautoria na pós-graduação e no currículo Lattes, contribuindo significativamente para a compreensão das dinâmicas de colaboração no país [1] [2].

Diversos estudos internacionais complementam esse panorama ao propor abordagens computacionais avançadas. Lopez-Rodriguez e Ceballos exploram a modelagem de indicadores cientométricos por meio de uma ontologia estatística implementada no banco de grafos Neo4j [3], tecnologia também utilizada neste projeto. No contexto da detecção de comunidades, Šubelj, van Eck e Waltman comparam algoritmos como Louvain e Infomap em redes de citação, discutindo seus impactos na qualidade dos agrupamentos encontrados [4]. Uma análise metodológica similar, porém em menor escala, foi realizada neste trabalho.

Finalmente, o algoritmo node2vec [5] representa um marco na aplicação de aprendizado de representação em grafos, permitindo extrair *embeddings* vetoriais a partir da estrutura topológica das redes – uma linha promissora para estudos futuros envolvendo aprendizado de máquina sobre grafos bibliométricos.

3 Justificativa

A análise de redes de colaboração científica fornece subsídios importantes para compreender como o conhecimento é produzido, difundido e consolidado dentro de comunidades acadêmicas. Em especial, o estudo das colaborações externas – isto é, entre pesquisadores de diferentes instituições – permite avaliar o grau de conectividade institucional e a diversidade dos vínculos científicos estabelecidos.

A UNICAMP, como uma das principais instituições de pesquisa do Brasil, apresenta um ecossistema científico complexo e articulado. Investigar suas redes de co-autoria e, particularmente, o impacto das colaborações externas sobre o desempenho científico de seus pesquisadores, pode revelar padrões de atuação relevantes e apontar caminhos para novas estratégias institucionais.

Do ponto de vista metodológico, o uso de dados da base *Scopus*, estruturados em um grafo no Neo4j e analisados por meio de algoritmos de detecção de comunidades, representa uma abordagem moderna e eficaz para captar e estudar essas dinâmicas.

4 Objetivos

O objetivo geral deste trabalho é investigar a rede de colaborações científicas envolvendo autores afiliados à UNICAMP entre 2000 e 2025, com ênfase na análise do impacto das colaborações externas sobre o crescimento das citações dos pesquisadores da instituição.

4.1 Objetivos Específicos

- Extrair e processar dados bibliográficos da base Scopus referentes a publicações com participação de autores da UNICAMP.
- Modelar os dados como um grafo de colaboração científica, contendo vértices do tipo Autor, Documento e Instituição.
- Identificar comunidades de colaboração por meio de algoritmos de detecção baseados em modularidade.
- Atribuir instituições predominantes a autores e comunidades, a fim de classificar colaborações como internas ou externas à UNICAMP.
- Avaliar, por meio de análise quantitativa, a evolução do impacto científico (medido por citações) antes e depois da primeira colaboração externa dos autores.
- Comparar os padrões de crescimento do impacto entre autores que colaboraram externamente e um grupo de controle com apenas colaborações internas.

5 Desenvolvimento do Trabalho

Este projeto foi desenvolvido em múltiplas etapas integradas. Ele se iniciou com a obtenção dos dados de colaboração científica, que foram processados e inseridos em dois tipos de bancos de dados: relacional (PostgreSQL) e orientado a grafos (Neo4j). Em seguida, foram executados e comparados três algoritmos de detecção de comunidades em grafos. A melhor atribuição de comunidades foi usada na etapa seguinte, que consistiu em definir uma instituição predominante a cada autor e comunidade — essencial para identificação de colaborações externas. Com esse grafo enriquecido, foi possível avaliar a influência da primeira colaboração externa de autores da UNICAMP em seus números cumulativos de citações. Para isso, foi comparada a taxa de variação dessa métrica antes e depois da colaboração.

5.1 Obtenção dos Dados de Colaboração Científica

A primeira etapa na realização desta pesquisa foi a criação de um *pipeline* de *ETL* (*Extract*, *Transform*, *Load*) [6] para obter e tratar os dados dos artigos e das colaborações científicas. O *pipeline* foi implementado utilizando a linguagem de programação Python [7]. Seguem as descrições detalhadas de cada componente deste *pipeline*.

5.1.1 *Extract*

Os dados de colaboração científica foram extraídos da base de dados *Scopus*, da Elsevier [8]. Isso foi feito utilizando `pybliometrics` [9], uma biblioteca Python que expõe *wrappers* convenientes para acessar a API REST do Scopus [10]. Essa biblioteca inclui uma abstração do *endpoint* Scopus Search [11], que permite a execução de consultas complexas baseadas em uma sintaxe própria de busca [12].

Foram coletados dados de artigos com ao menos um autor associado à UNICAMP dos últimos vinte e cinco anos, isto é, de 2000 a 2025. No Scopus, cada instituição possui um identificador único. Assim, foi possível utilizar o identificador da UNICAMP ("60029570") para buscar por artigos com autores associados à ela. O processo de obtenção de dados foi feito em lotes, com uma consulta para cada ano, em ordem decrescente (primeiro os anos mais recentes). Isso diminuiu o tempo de retorno e o tamanho da resposta de cada consulta. O trecho do programa responsável por essa tarefa segue reproduzido no Código 1.

```
for year in range(END_YEAR, START_YEAR - 1, -1):
    query = f"aF-ID({UNICAMP_AFFILIATION_ID}) AND PUBYEAR = {year}"
    search = ScopusSearch(query, subscriber=True)
```

Código 1: Consulta para obtenção de dados de colaboração científica da UNICAMP.

O retorno dessas consultas são vetores chamados *namedtuples*, que possuem os seguintes campos:

Campo	**Descrição**
`eid`	Identificador único do documento no Scopus.
`doi`	Identificador DOI do documento.
`pii`	Identificador PII do documento.
`pubmed_id`	Identificador PubMed do documento.
`title`	Título do documento.
`subtype`	Subtipo abreviado do documento (ex: "ar").
`subtypeDescription`	Descrição do subtipo do documento.
`creator`	Primeiro autor do documento.
`afid`	Lista de IDs das instituições associadas.
`affilname`	Lista de nomes das instituições associadas.
`affiliation_city`	Lista de cidades das instituições.
`affiliation_country`	Lista de países das instituições.
`author_count`	Número de autores do documento.
`author_names`	Lista de nomes dos autores.
`author_ids`	Lista de IDs dos autores.
`author_afids`	Lista de IDs das instituições dos autores.
`coverDate`	Data de publicação no formato AAAA-MM-DD.
`coverDisplayDate`	Data de publicação por extenso.
`publicationName`	Nome do veículo de publicação.
`issn`	ISSN do veículo de publicação.
`source_id`	ID Scopus do veículo de publicação.
`eIssn`	ISSN eletrônico (eISSN).
`aggregationType`	Tipo de veículo de publicação.
`volume`	Volume do veículo de publicação.
`issueIdentifier`	Número da edição do veículo.

`article_number`	Número do artigo no volume.
`pageRange`	Intervalo de páginas do documento.
`description`	Resumo do documento.
`authkeywords`	Palavras-chave definidas pelos autores.
`citedby_count`	Número de citações do documento.
`openaccess`	1 se open access, 0 caso contrário.
`freetoread`	1 se leitura gratuita, 0 caso contrário.
`freetoreadLabel`	Nome do tipo de acesso gratuito.
`fund_acr`	Sigla da agência de fomento.
`fund_no`	Identificador do financiamento.
`fund_sponsor`	Nome da agência de fomento.

Tabela 1: Descrição dos campos retornados pela API

Inicialmente os objetos de resposta da API foram salvos sem qualquer tipo de alteração em arquivos do tipo *pickle* [13], o formato de serialização nativo da linguagem Python. Esses arquivos foram colocados em diretórios criados e nomeados automaticamente. No nome desses diretórios é possível identificar o intervalo de anos escolhido, bem como uma *timestamp* do momento de extração, no formato padronizado ISO 8601 [14].

5.1.2 *Transform*

Os dados brutos foram carregados a partir dos arquivos *pickle* e estruturados em *DataFrames*, utilizando a biblioteca *pandas* [15]. Nessa etapa do *pipeline* segregamos os dados das diferentes entidades, a fim de facilitar a etapa seguinte, de carregamento dos dados. Tais entidades são: Instituição (*Institution*), Documento (*Document*), Autor (*Author*), e Autoria (*Authorship*). Essa última é uma entidade associativa, que relaciona as três primeiras. Esta etapa é importante, pois Instituições, Documentos, e Autores formam uma relação ternária de cardinalidade $n_1 \cdot n_2 \cdot n_3$. Por exemplo, o mesmo autor pode estar afiliado a várias instituições na mesma publicação (1 *Author*: n *Institution*), e a mesma instituição pode ter vários autores afiliados a ela na mesma publicação (1 *Author*: n *Institution*).

Assim, os dados foram separados e salvos nos seguintes arquivos TSV: “institutions.tsv”, para dados de instituições; “documents.tsv”, para dados de documentos; “authors.tsv”, para dados de autores; e “authorships.tsv”, para dados de autorias.

Cada instituição foi vinculada a seu identificador único, nome, cidade, e país. Cada documento foi vinculado a seu identificador único, título, data de publicação, contagem de citações, agência de fomento, veículo de publicação, entre outros atributos. Cada autor foi associado a seu identificador único e ao seu nome. Por fim, o identificador único de cada autor foi associado ao identificador único dos seus documentos, ao identificador único da instituição à que estava afiliado naquele documento, e a um valor booleano que indica se este é o primeiro autor do documento (*true*) ou não (*false*). O resultado final desse processamento foi salvo num arquivo “authorships.tsv”.

5.1.3 *Load* - Banco de Dados Relacional

A fim de possibilitar a realização de consultas complexas aos dados obtidos, decidimos inseríos em um Banco de Dados Relacional com suporte à linguagem de consultas SQL. Para isso, foi escolhido o *Relational Database Management System* (RDBMS) PostgreSQL. As operações no banco de dados relacional foram realizadas utilizando a biblioteca de *Object-Relational Mapping* (ORM) SQLModel, que foi selecionada por sua concisão e excelente suporte ao sistema de anotação de tipos do Python.

O banco de dados relacional foi modelado com as quatro entidades anteriormente mencionadas: *Institution*, *Document*, *Author*, e *Authorship*. O Diagrama Entidade-Relação deste banco de dados está representado na Figura 1:

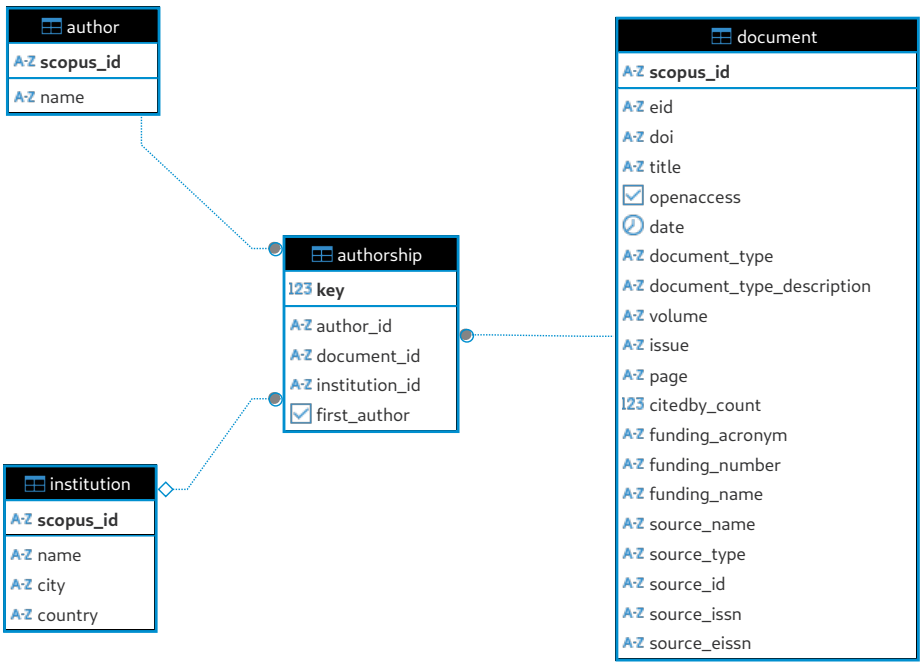

Figura 1: Diagrama Entidade-Relação do banco de dados relacional

Fica evidente o papel da entidade associativa *Authorships*: ela contém chaves estrangeiras para *Institution*, *Document*, e *Author*, modelando a relação ternária entre elas.

5.1.4 *Load* - Banco de Orientado a Grafos

Além do banco de dados relacional, os dados foram carregados no banco de dados orientado a grafos, Neo4j. Isso foi essencial para, posteriormente, executar implementações altamente otimizadas de algoritmos de grafos nos dados.

Neste banco de dados, as quatro entidades correspondem a diferentes tipos de vértices. A Figura 2 contém uma ilustração do esquema do banco de dados orientado à grafos e a Figura 3, uma ilustração de uma amostra dos registros nesse banco de dados.

Foi utilizada a função de importação de dados da interface de linha de comando (CLI) do Neo4j, que possibilita a inserção de dados em arquivos CSV. Para isso, foi necessário criar novos arquivos seguindo o formato esperado pelo Neo4j.

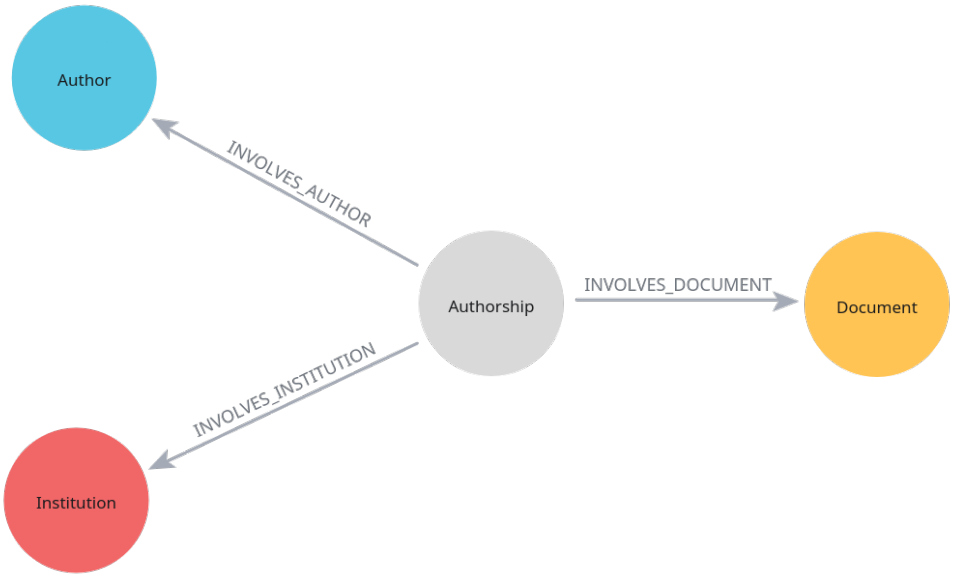

Figura 2: Diagrama do esquema banco de dados orientado a grafos.

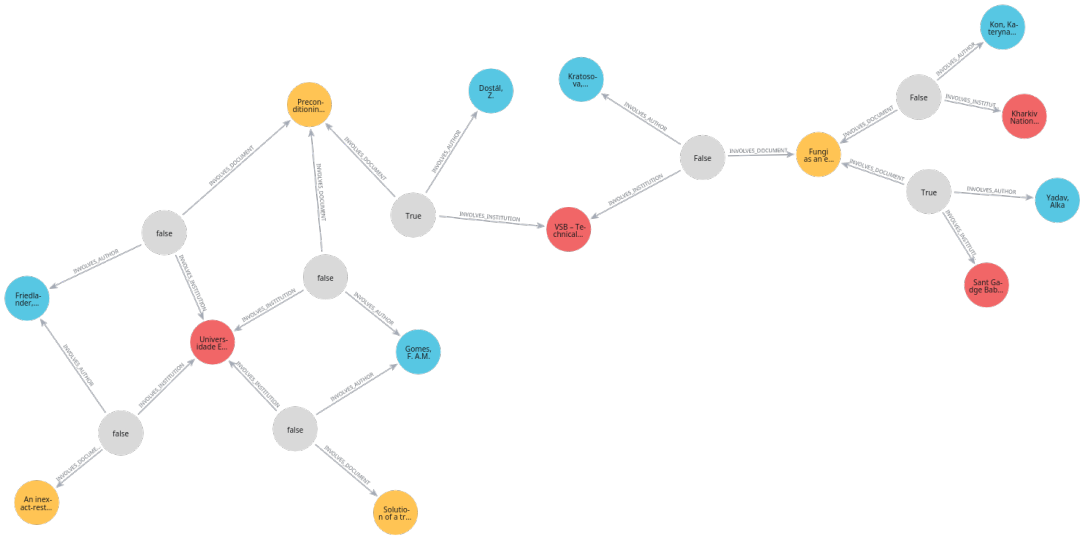

Figura 3: Amostra de registros no banco de dados orientado a grafos.

5.2 Detecção de Comunidades

Um dos requisitos para a análise de colaborações externas foi a detecção de comunidades, visto que queremos identificar colaborações com autores pertencentes à comunidades de colaboração fora da UNICAMP.

A biblioteca *Graph Data Science* (GDS) do Neo4j contém implementações de múltiplos algoritmos de detecção de comunidades. Para melhor avaliar as opções disponíveis, foi feito um *benchmark* de três destes algoritmos: *Louvain*, *Leiden*, e *Label Propagation*. No entanto, esses algoritmos foram criados para grafos com apenas um tipo de vértice e aresta, enquanto o grafo obtido até então possuía quatro tipos de vértice e três tipos de aresta.

Por esse motivo, antes de executar os algoritmos de detecção de comunidades, foi necessário adicionar um novo tipo de aresta (*COLLABORATED_WITH*), cujas extremidades são sempre vértices do tipo *Author* e cujo peso é o número de vezes que estes autores colaboraram. Assim, é possível executar os algoritmos considerando apenas os vértices *Author* e as arestas *COLLABORATED_WITH*.

Para inserir as novas arestas no Neo4j, nossa abordagem foi utilizar uma consulta ao banco de dados relacional exibida no Código 2, e salvar seus resultados em um arquivo TSV no formato esperado pelo comando de importação do Neo4j. Depois disso, utilizamos a CLI do Neo4j para importar as arestas.

```
SELECT
    a1.scopus_id AS author1_id,
    a2.scopus_id AS author2_id,
    COUNT(DISTINCT au1.document_id) AS collaboration_count
FROM
    authorship au1
JOIN
    authorship au2 ON au1.document_id = au2.document_id AND
                      au1.author_id < au2.author_id
JOIN
    author a1 ON au1.author_id = a1.scopus_id
JOIN
    author a2 ON au2.author_id = a2.scopus_id
GROUP BY
    a1.scopus_id, a2.scopus_id
```

Código 2: Consulta para obtenção de arestas *COLLABORATED_WITH*.

Após inserir as novas arestas, foi possível obter um subgrafo contendo apenas vértices do tipo *Author* e arestas do tipo *COLLABORATED_WITH*, do qual uma pequena fração está representada na Figura 5.

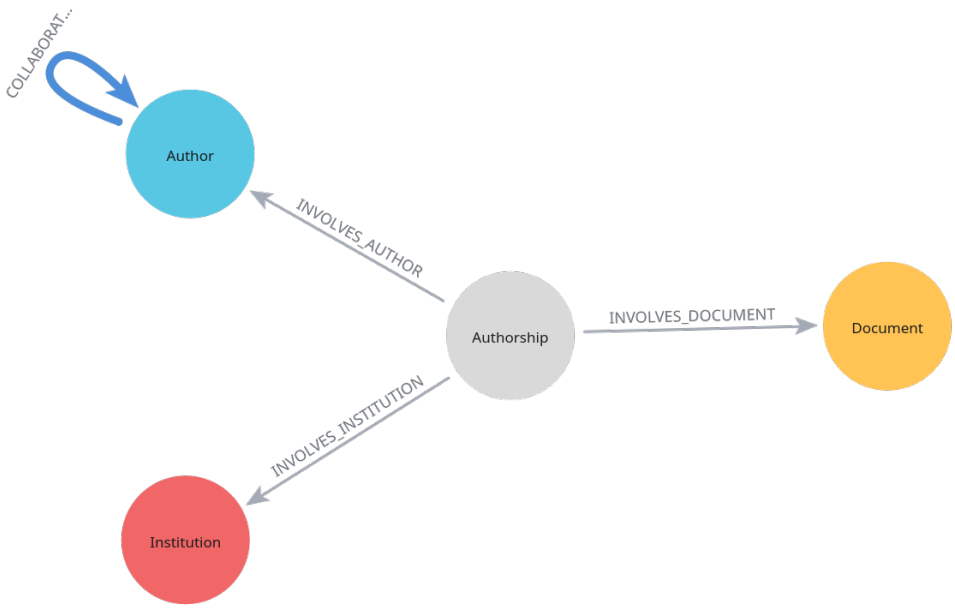

Figura 4: Diagrama do esquema atualizado do banco de dados orientado a grafos, com as arestas *COLLABORATED_WITH*

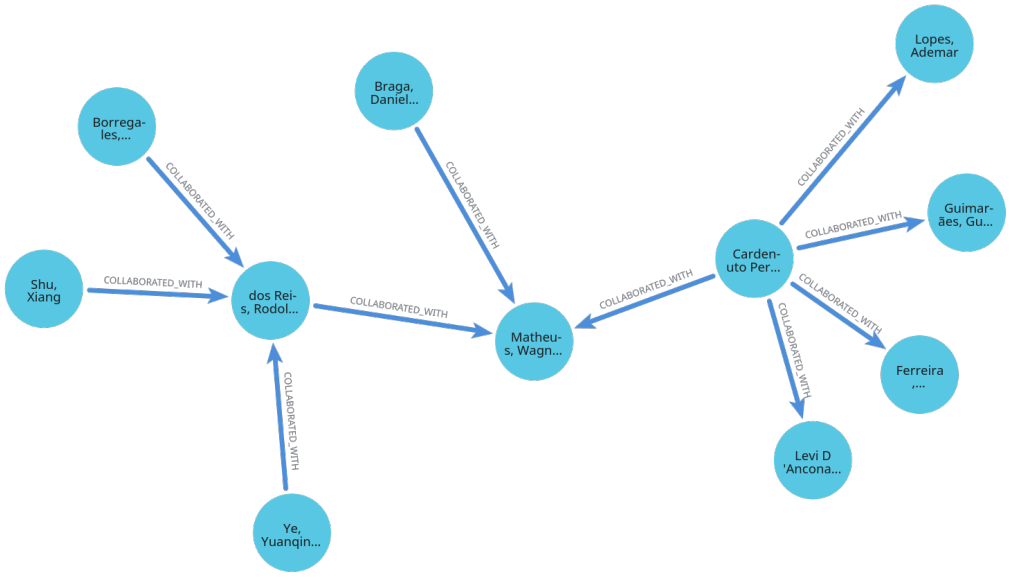

Figura 5: Amostra do subgrafo com vértices *Author* e arestas *COLLABORATED_WITH*

Os algoritmos de detecção de comunidade escolhidos (*Louvain*, *Leiden*, e *Label Propagation*) são estocásticos, ou seja, podem atribuir comunidades diferentes a cada execução.

Por isso, cada um deles foi executado dez vezes. Para avaliar a qualidade das comunidades atribuídas à cada execução, foi utilizada a métrica de modularidade [16, 17].

Inicialmente obtivemos valores de modularidade inacreditavelmente altos (em torno de 0.95). Após uma análise cuidadosa, foi possíver aferir que comunidades pequenas, de um a dois autores, estavam introduzindo significativo ruído na métrica de modularidade, distorcendo-a. A fim de mitigar esse ruído, foram incluídas na análise apenas arestas com peso maior ou igual a 2 (isto é, aquelas que representam ao menos duas colaborações entre dois autores) e apenas vértices conectados a pelo menos uma aresta. Os resultados finais do *benchmark*, com as mudanças citadas, estão reproduzido na Tabela 2:

Algoritmo	**Iteração**	**Modularidade**
Label Propagation	1	0.7356
Label Propagation	2	0.7355
Label Propagation	3	0.7449
Label Propagation	4	0.7435
Label Propagation	5	0.7355
Label Propagation	6	0.7450
Label Propagation	7	0.7435
Label Propagation	8	0.7450
Label Propagation	9	0.7447
Label Propagation	10	0.7435
Louvain	1	0.7589
Louvain	2	0.7589
Louvain	3	0.7589
Louvain	4	0.7586
Louvain	5	0.7589
Louvain	**6**	**0.7590**
Louvain	7	0.7586
Louvain	8	0.7588
Louvain	9	0.7589
Louvain	10	0.7588
Leiden	1	0.7586
Leiden	2	0.7586
Leiden	3	0.7586
Leiden	4	0.7587
Leiden	5	0.7587
Leiden	6	0.7587
Leiden	7	0.7585
Leiden	8	0.7586
Leiden	9	0.7587
Leiden	10	0.7585

Tabela 2: *Benchmark* de Algoritmos de Detecção de Comunidades

O melhor resultado foi obtido na sexta iteração do algoritmo Louvain, então essa foi a atribuição de comunidade selecionada para realização das próximas etapas.

5.3 Atribuição de Instituições Predominantes

A cada autor foi atribuída uma instituição predominante, definida como aquela presente na maioria de suas afiliações em publicações. Isso foi guardado no grafo usando arestas de um novo tipo *HOME_INSTITUTION*. O esquema final do banco de dados baseado em grafos segue reproduzido na Figura 6.

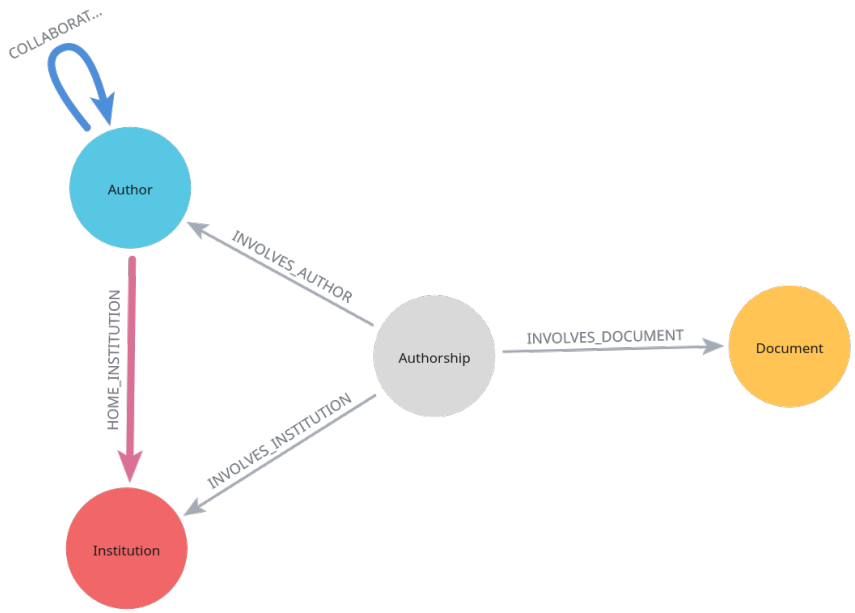

Figura 6: Diagrama do esquema final do banco de dados orientado a grafos, com as arestas *HOME_INSTITUTION*

Em seguida, a cada comunidade foi atribuída uma instituição predominante, definida como aquela mais frequente entre os autores que a compõem. Esse dado foi persistido no Neo4j como uma nova propriedade *community_institution* dos nós *Author*, que já antes armazenavam a propriedade *community* para identificar suas respectivas comunidades.

Assim, foi possível enfim definir uma colaboração externa como aquela que envolve ao menos um autor de cada um dos seguintes perfis:

Autor interno:

- Instituição predominante é a UNICAMP
- Pertence a uma comunidade da UNICAMP
- Afiliado à UNICAMP na publicação

Autor externo:

- Instituição predominante não é a UNICAMP
- Não pertence a uma comunidade da UNICAMP
- Não afiliado à UNICAMP na publicação

5.4 Análise das Ocorrências de Primeira Colaboração Externa

A influência da primeira colaboração externa sobre o impacto científico dos autores foi avaliada por meio da variação no número cumulativo de citações, definido como a soma total de citações atuais recebidas por todas as publicações de um autor. Importa destacar que os dados de citações utilizados referem-se ao total acumulado no momento da coleta, não estando disponíveis valores anuais históricos. Para cada autor, a trajetória de impacto foi representada por uma série temporal com um ponto para cada ano em que houve ao menos uma publicação.

A análise concentrou-se nos anos de 2005 a 2020, adotando uma janela de cinco anos antes e cinco anos depois de cada ano analisado. Para cada ano, os autores foram classificados em dois grupos. O grupo de controle incluiu autores que (1) nunca realizaram colaborações externas, (2) publicaram ao menos uma colaboração interna naquele ano e (3) já haviam realizado colaborações (não externas) anteriormente. O grupo experimental foi composto por autores que (1) realizaram sua primeira colaboração externa no ano analisado e (2) também possuíam histórico prévio de publicações.

Para viabilizar comparações, os valores de impacto de cada autor foram normalizados dividindo-se todos os pontos da série temporal pelo valor do impacto no ano analisado, que, portanto, passou a ter valor unitário. Em seguida, foi ajustada uma regressão linear separadamente sobre os cinco anos anteriores e os cinco anos posteriores ao ano analisado, resultando em dois coeficientes angulares por autor. Calculou-se então, para cada autor, a razão entre os coeficientes posteriores e anteriores, refletindo a mudança relativa na taxa de crescimento do impacto. Por fim, as médias dessas razões foram comparadas entre os grupos de controle e experimental, considerando todos os anos agregados, com o objetivo de identificar se a primeira colaboração externa esteve associada a uma mudança significativa na trajetória de citações.

6 Resultados

Os resultados revelaram a formação de comunidades de colaboração com grande heterogeneidade de tamanho, diversidade institucional e impacto médio. Observou-se também um crescimento contínuo no número de primeiras colaborações externas da UNICAMP ao longo dos anos. No entanto, a análise do impacto científico mostrou que essas colaborações externas não resultaram, em média, em aumento expressivo da taxa de crescimento de citações, contrariando a hipótese inicial do estudo.

6.1 Caracterização das Comunidades

Dos 183.821 autores no banco de dados, apenas 65.139 foram inseridos em comunidades devido às restrições descritas na seção 5.2. Foram formadas 837 comunidades, com uma distribuição altamente heterogênea de quantidade de membros: 759 comunidades (uma quantidade superior a 90% das comunidades) têm menos de dez membros, havendo algumas com mais de 1000 membros.

Uma característica interessante de uma comunidade é a sua "diversidade". Essa métrica é a razão entre o total de instituições dos autores na comunidade, e o total de autores membros da comunidade. A distribuição da diversidade das comunidades está representada na Figura 7.

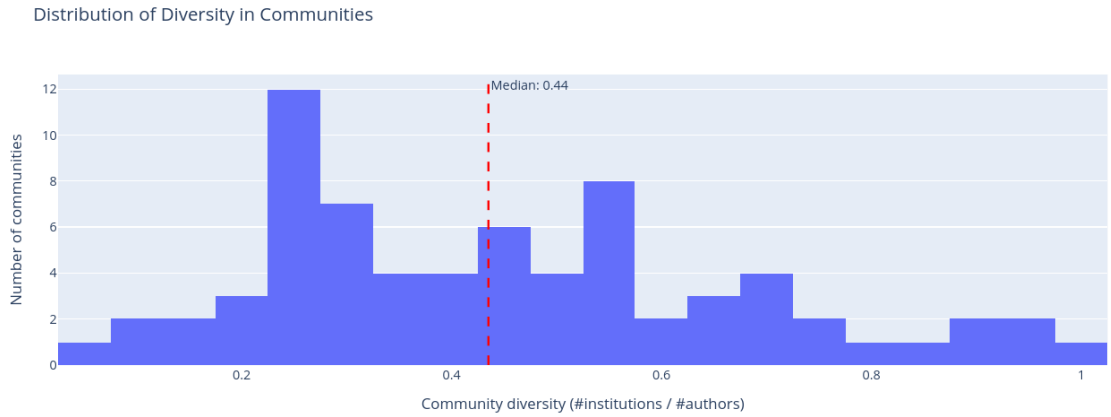

Figura 7: Histograma de distribuição de "diversidade"em comunidades

Além disso, foi determinada a distribuição de número de comunidades por instituição. Considerando o processo de extração dos dados, que selecionou apenas documentos com ao menos um co-autor da UNICAMP, é natural que ela tenha sido a instituição com mais comunidades. No gráfico da Figura 8 estão exibidas as vinte instituições com maior número de comunidades, onde se observam grandes nomes nacionais e internacionais.

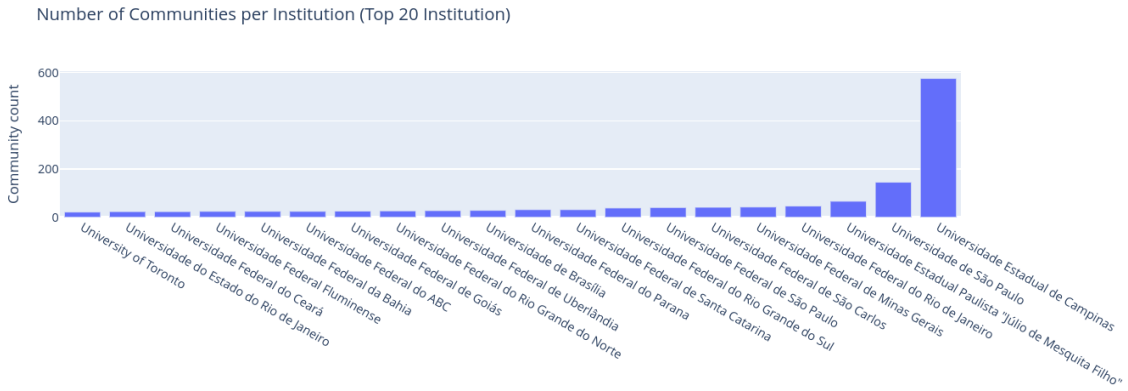

Figura 8: Gráfico de barras de número de comunidades por instituição

Por fim, investigamos também o número médio de citações por autor em cada comunidade. Novamente, a distribuição foi bastante heterogênea. Esse valor é inferior à 100 na grande maioria das comunidades (mais de 80%), e superior a 10.000 nas três comunidades com maior número médio de citações.

6.2 Caracterização da Primeira Colaboração Externa

Primeiramente, estudamos a distribuição do número de colaborações internas ocorridas antes da primeira colaboração externa. Metade dos pesquisadores faz sua primeira colaboração externa após 0 a 9 colaborações internas, como pode ser visto na Figura 9.

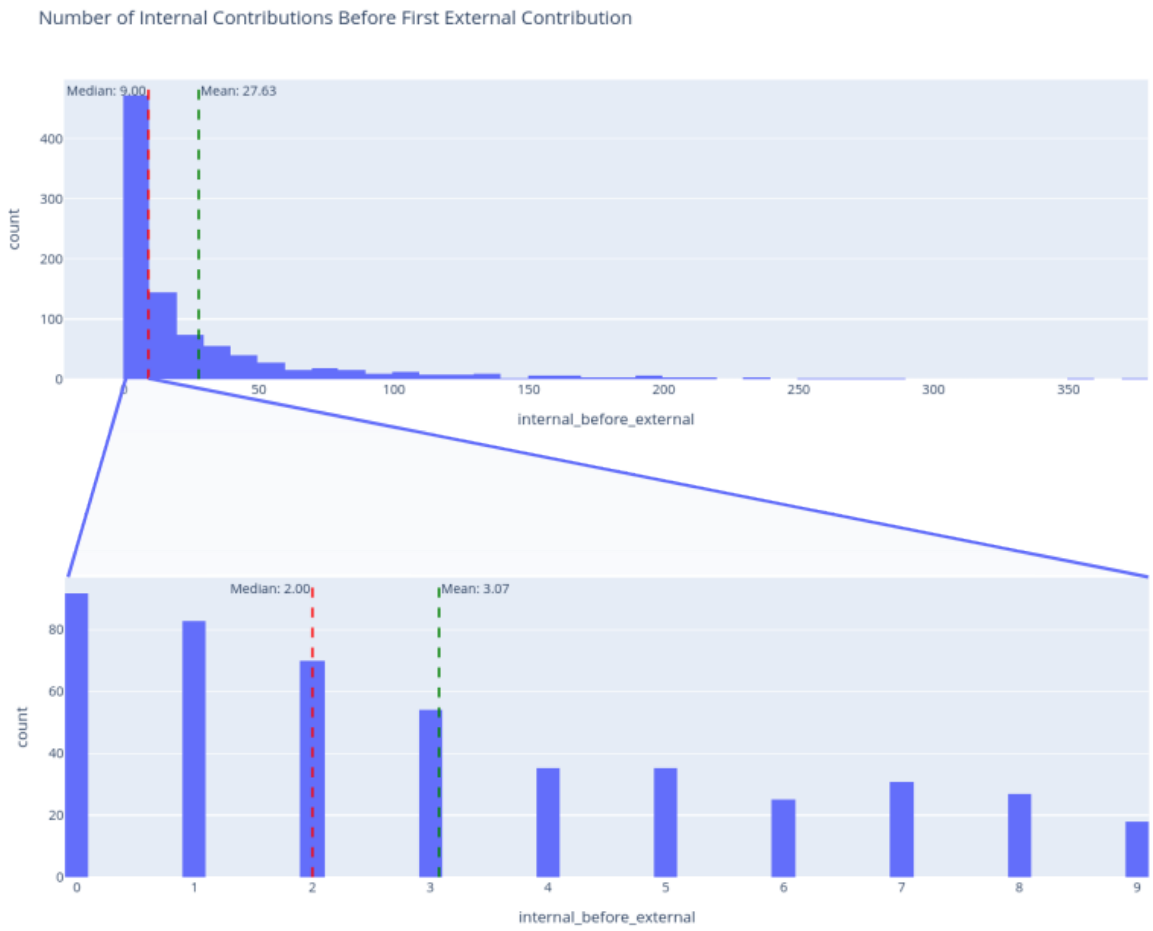

Figura 9: Número de colaborações internas antes da primeira colaboração externa

Foi estudado também o número de ocorrências de primeira colaboração externa a cada ano na UNICAMP. É possível observar no gráfico na Figura 10 uma tendência crescente nesse número, indicando que a UNICAMP está progredindo em seus vínculos com outras instituições.

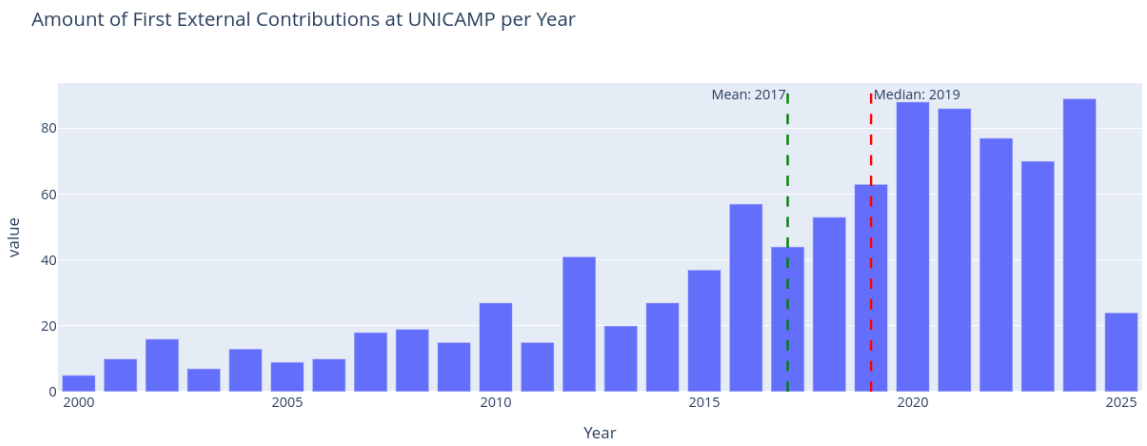

Figura 10: Número de ocorrências de primeira colaboração externa a cada ano na Universidade Estadual de Campinas (Unicamp).

6.3 Influência da Primeira Colaboração Externa no Impacto de um Autor

Seguindo os métodos descritos na Seção 5.4, foram feitas regressões lineares para obter os coeficientes lineares e angulares das linhas de melhor ajuste aos pontos anteriores e posteriores ao ano de referência para cada autor. Isso foi feito tanto do grupo experimental quanto do grupo de controle. Obtivemos p-valores estatisticamente significativos em boa parte dos casos, com 79% dos autores do grupo experimental e 66% do grupo de controle apresentando p-valores inferiores a 0,05.

Os dados referentes aos anos a partir de 2017 foram descartados por apresentarem coeficientes angulares e lineares significativamente mais baixos em comparação aos anos anteriores. Essa diferença é provavelmente devida à recência dessas publicações, que ainda não acumularam um número comparável de citações, comprometendo a consistência da análise longitudinal.

Os dados do ano de 2010 também foram removidos por se caracterizarem como *outliers*. Enquanto as razões entre os coeficientes angulares antes e depois do ano de referência permaneceram abaixo de 2,50 nos demais anos, em 2010 essa razão atingiu o valor atípico de 44,34. Uma análise mais aprofundada revelou que essa anomalia foi causada por um único autor, cujo coeficiente angular aumentou mais de 800 vezes naquele ano, distorcendo significativamente a média geral.

Inesperadamente, o aumento médio do coeficiente angular foi maior no grupo de controle do que no grupo experimental. Mais especificamente, a razão entre os coeficientes angulares anterior e posterior no grupo de controle foi 1,84 e no grupo experimental foi 1,02.

A razão obtida para o grupo experimental indica que a primeira colaboração externa pouco influenciou a taxa de crescimento do número de citações, no universo dos dados utilizados. Isso não pode, a princípio, ser explicado por uma diferença no perfil inicial dos dois grupos, visto que ambos tiveram coeficientes angulares pré-colaboração significativamente próximos (0,1938 para o grupo de controle, e 0,1815 para o grupo experimental). É possível comparar visualmente as tendências nos dois grupos usando as Figuras 11 e 12.

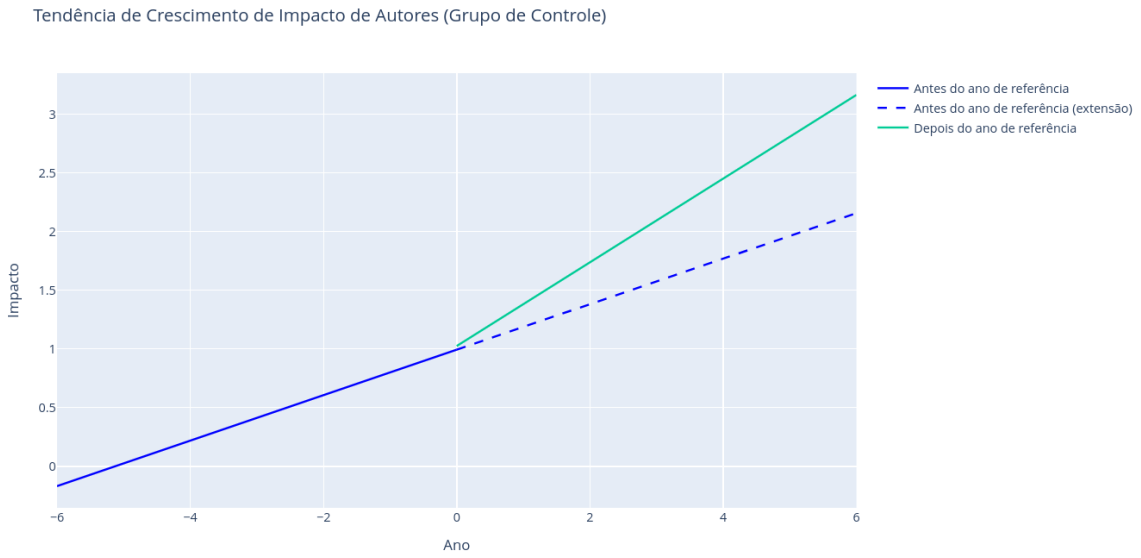

Figura 11: Linhas de melhor ajuste médias no grupo de controle

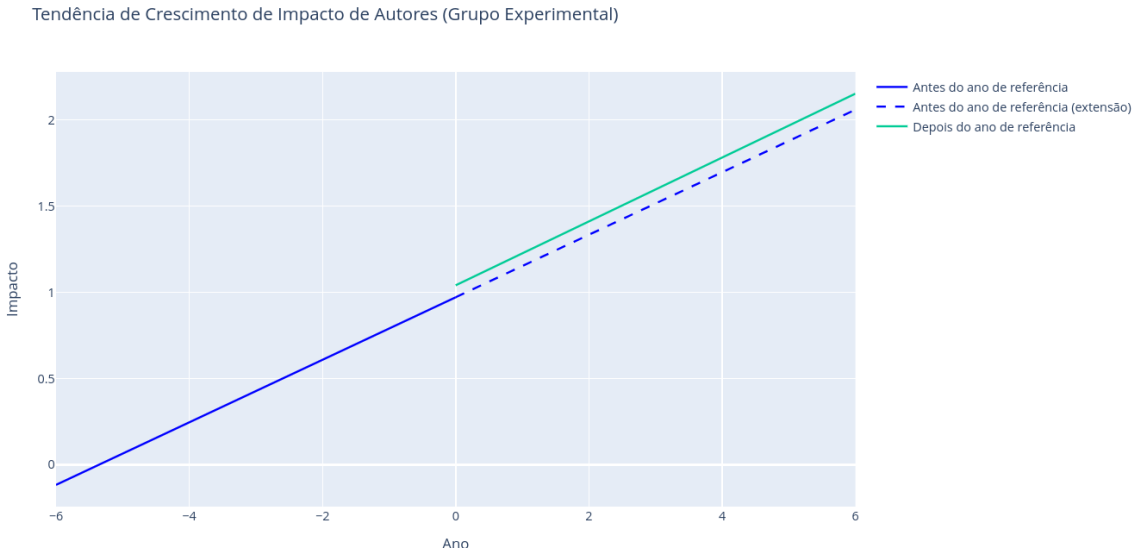

Figura 12: Linhas de melhor ajuste médias no grupo experimental

7 Discussão

Os resultados observados contrariaram a expectativa inicial de que colaborações externas estariam associadas a um aumento mais acentuado no impacto científico dos autores. Duas hipóteses são propostas para explicar essa discrepância.

A primeira, mais simples, está relacionada ao estágio de desenvolvimento da carreira dos autores. Colaborações externas tendem a ocorrer após um histórico consistente de colaborações internas, o que implica que os autores envolvidos já acumulam um número maior de citações. Como a análise foi baseada em crescimento relativo, aumentos em valores absolutos mais altos resultam em variações proporcionais menores. Por exemplo, dobrar um impacto de 1 exige apenas uma citação adicional, enquanto dobrar um impacto de vinte, exige vinte.

A segunda hipótese diz respeito a uma limitação na coleta dos dados. Como apenas publicações com ao menos um coautor afiliado à UNICAMP foram consideradas, publicações futuras de autores que migraram para outras instituições e deixaram de colaborar com a universidade podem ter sido excluídas da base. Dessa forma, parte relevante do impacto posterior desses autores pode não ter sido contabilizada, o que compromete a avaliação completa de sua trajetória científica.

8 Conclusões

Este trabalho propôs e implementou uma metodologia para analisar colaborações científicas envolvendo pesquisadores da UNICAMP, com foco na estruturação e exploração de um grafo de co-autoria extraído da base Scopus. Utilizando modelagem em grafos e algoritmos de detecção de comunidades baseados em modularidade, foi possível identificar agrupamentos orgânicos de colaboração e classificar colaborações como internas ou externas à instituição. Além disso, foi desenvolvido um método quantitativo para avaliar o impacto relativo da primeira colaboração externa sobre a trajetória de citações de cada autor.

Como desdobramentos futuros, sugere-se a validação das hipóteses levantadas por meio de um rastreamento mais abrangente das trajetórias dos autores, levando em conta afiliações institucionais posteriores. Além disso, há espaço promissor para o uso de técnicas de aprendizado de máquina em grafos, como node2vec, para detectar padrões não triviais de impacto e identificar fatores estruturais que influenciam a visibilidade científica. Essas abordagens podem complementar as análises tradicionais e fornecer novas perspectivas sobre os mecanismos de produção e disseminação do conhecimento científico.

Referências

[1] DIGIAMPIETRI, L. A. et al. Análise da evolução das relações de coautoria nos programas de pós-graduação em computação no brasil. *Revista Eletrônica de Sistemas de Informação*, Instituto Brasileiro de Estudos e Pesquisas Sociais, v. 14, n. 1, p. 1–1, Jul 2025.

[2] MENA-CHALCO, J. P. et al. Brazilian bibliometric coauthorship networks. *Journal of the Association for Information Science and Technology*, v. 65, n. 7, p. 1424–1445, Jan 2014.

[3] LOPEZ-RODRIGUEZ, V.; CEBALLOS, H. G. Modeling scientometric indicators using a statistical data ontology. *Journal of Big Data*, v. 9, n. 1, Jan 2022.

[4] ŠUBELJ, L.; ECK, N. J. van; WALTMAN, L. Clustering scientific publications based on citation relations: a systematic comparison of different methods. *PLOS ONE*, v. 11, n. 4, p. e0154404, Apr 2016.

[5] GROVER, A.; LESKOVEC, J. node2vec: Scalable feature learning for networks. *arXiv:1607.00653 [cs, stat]*, Jul 2016. Disponível em: <https://arxiv.org/abs/1607.00653>.

[6] VASSILIADIS, P.; SIMITSIS, A.; SKIADOPOULOS, S. Conceptual modeling for etl processes. *Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP - DOLAP '02*, 2002.

[7] PYTHON. *Python*. Python.org, 2025. Disponível em: <https://www.python.org/>.

[8] ELSEVIER. *About Scopus — Abstract and Citation Database — Elsevier*. 2023. Disponível em: <https://www.elsevier.com/products/scopus>.

[9] ROSE, M. E.; KITCHIN, J. R. pybliometrics: Scriptable bibliometrics using a python interface to scopus. *SoftwareX*, v. 10, p. 100263, 07 2019.

[10] ELSEVIER. *Elsevier Developer Portal — API Interface Specification*. 2025. Disponível em: <https://dev.elsevier.com/api_docs.html>.

[11] ELSEVIER. *Scopus Search API*. 2024. Disponível em: <https://dev.elsevier.com/documentation/SCOPUSSearchAPI.wadl>.

[12] ELSEVIER. *Elsevier Developer Portal — Scopus Search Guide*. 2024. Disponível em: <https://dev.elsevier.com/sc_search_tips.html>.

[13] FOUNDATION, P. S. *pickle — Python object serialization — Python 3.7.3 documentation*. 2019. Disponível em: <https://docs.python.org/3/library/pickle.html>.

[14] ISO. *ISO 8601-1:2019*. 2019. Disponível em: <https://www.iso.org/standard/70907.html>.

[15] TEAM, T. pandas development. *pandas-dev/pandas: Pandas*. Zenodo, 2024. Disponível em: <https://doi.org/10.5281/zenodo.3509134>.

[16] NEWMAN, M. E. J.; GIRVAN, M. Finding and evaluating community structure in networks. *Physical Review E*, v. 69, n. 2, Feb 2004.

[17] INTERIAN, R.; RODRIGUES, F. A. Group polarization, influence, and domination in online interaction networks: a case study of the 2022 Brazilian elections. *J. Phys. Complex.*, v. 4, n. 3, p. 035008, 2023.