

Rumo a Duas Décadas de Publicações em Banco de Dados: Uma Análise dos Artigos do JIDM*

Elena Blanco¹, Josué Santos¹, Maristela Holanda², Vanessa Braganholo¹,
Marcos Bedo¹, Daniel de Oliveira¹

¹Instituto de Computação – Universidade Federal do Fluminense (UFF)

²Universidade de Brasília (UnB)

{elenablanca, josuels}@id.uff.br, mholanda@unb.br

{vanessa, marcosbedo, danielcmo}@ic.uff.br

Resumo. O *Journal of Information and Data Management (JIDM)*, criado em 2010, é o principal periódico brasileiro dedicado à área de Banco de Dados (BD). Este artigo analisa a comunidade científica associada ao JIDM no período de 2010 a 2025, considerando três eixos: produtividade, redes de coautoria e tópicos de pesquisa. A metodologia combina análise bibliométrica e análise de redes de colaboração entre pesquisadores e instituições. Os resultados indicam um núcleo estável de autores recorrentes, com concentração moderada da produção e presença crescente de Mineração de Dados e de Aprendizado de Máquina entre os tópicos publicados. Esses achados contribuem para a compreensão da evolução da área de BD no país e podem subsidiar decisões editoriais futuras do periódico.

Abstract. The *Journal of Information and Data Management (JIDM)*, established in 2010, is the leading Brazilian journal dedicated to the field of Databases. This paper analyzes the scientific community associated with JIDM from 2010 to 2025, considering three dimensions: productivity, co-authorship networks, and research topics. The methodology combines bibliometric analysis with collaboration network analysis among researchers and institutions. The results indicate a stable core of recurring authors, with a moderate concentration of scientific output and a growing presence of Data Mining and Machine Learning among the published topics. These findings contribute to understanding the evolution of the Database field in Brazil and may support future editorial decisions of the journal.

1. Introdução

A produção científica apoia-se fortemente em comunidades organizadas em torno de conferências e periódicos [Newman 2004]. Esses fóruns têm um papel importante não só na disseminação de resultados, mas também na formação e no aperfeiçoamento de pessoas, ao favorecer o estabelecimento de colaborações e a consolidação da pesquisa. No Brasil, diversos periódicos contam com apoio institucional da Sociedade Brasileira de Computação (SBC) e são disponibilizados na plataforma *SBC OpenLib (SOL)*¹. Entre eles, o *Journal of Information and Data Management (JIDM)*², criado em 2010, é o principal periódico nacional dedicado à área de Banco de Dados (BD).

*Os autores agradecem o apoio financeiro da CAPES, do CNPq e da FAPERJ. Ferramentas de IA generativa foram utilizadas exclusivamente para revisão gramatical.

¹<https://sol.sbc.org.br/>

²<https://journals-sol.sbc.org.br/index.php/jidm>

A análise de coautoria tem sido amplamente empregada na literatura para caracterizar comunidades científicas e compreender sua dinâmica de produção [Noorden 2014, Singh et al. 2024]. No contexto brasileiro, algumas iniciativas já foram conduzidas, mas não no JIDM [Lima et al. 2015, Silva et al. 2017, Carvalho et al. 2024a, Carvalho et al. 2023, Carvalho et al. 2024b]. Esse tipo de análise é particularmente oportuno no momento atual, em que a área de BD vem sendo fortemente impactada pelos avanços em Inteligência Artificial (IA), em especial pela IA generativa e pelos Modelos de Linguagem de Larga Escala (LLMs) [Chang et al. 2024]. Examinar os 16 anos do JIDM permite observar essas transformações e compreender como a comunidade científica associada ao periódico tem evoluído ao longo do tempo.

Entretanto, um desafio anterior à análise de redes de colaboração científica está relacionado à qualidade dos dados utilizados nessas análises. Fontes públicas como a SOL podem apresentar inconsistências, *e.g.*, abreviações variadas para um mesmo nome de autor e diferentes grafias para uma mesma instituição, introduzindo ruídos que podem fragmentar artificialmente as redes e impactar os indicadores obtidos. Dessa forma, além da análise da comunidade científica, é necessário garantir um processo adequado de preparação dos dados para que os resultados representem de forma mais fiel a estrutura de colaboração existente. Com esse objetivo, este artigo analisa as publicações do JIDM no período de 2010 a 2025, combinando análise bibliométrica, análise de redes de colaboração entre pesquisadores e instituições e análise da distribuição de tópicos de pesquisa. Para realizar tal análise, foram utilizados o refinamento progressivo dos dados, seguindo a arquitetura *Medallion*, e a aplicação de técnicas de resolução de entidades textuais para tratar inconsistências, como variações nos nomes de autores e de instituições. Espera-se, assim, contribuir para a compreensão da evolução da área de BD no país, evidenciando seus principais atores, padrões colaborativos e tendências emergentes.

2. Metodologia

A análise apresentada neste artigo segue a metodologia descrita na Figura 1, que possui cinco etapas: (i) Definição do Escopo, (ii) Extração Automática, (iii) Coleta de Metadados, (iv) Pré-processamento e Limpeza e (v) Construção da Base de Indicadores. O processo de tratamento e organização dos dados baseou-se nos princípios de refinamento progressivo da arquitetura *Medallion*. Na etapa de *Definição do Escopo*, estabelecemos como recorte temporal o período de 2010 a 2025, abrangendo desde a criação do periódico até o final do último ano. Na etapa de *Extração Automática* foram empregadas as bibliotecas *Requests* (v2.31+) e *BeautifulSoup 4* (v4.12+), adequadas a cenários em que o conteúdo está diretamente disponível em HTML. O *crawler* implementado usa o *Requests* como mecanismo de acesso às páginas *web* do JIDM na SOL, enquanto o *BeautifulSoup* é usado para navegar e interpretar a estrutura HTML retornada, permitindo localizar e extrair os metadados relevantes.

Na etapa de *Coleta de Metadados*, o *crawler* percorre automaticamente a hierarquia da SOL a partir da página principal do JIDM, identificando os volumes e números do periódico e os *links* individuais de cada artigo. Para cada artigo encontrado, foram coletados: título, lista de autores, afiliações institucionais, ano de publicação, volume e número, resumo, palavras-chave, DOI, e URL. Os dados foram inicialmente armazenados em formato bruto, preservando as informações originais da fonte antes das etapas de transformação.

Na etapa de *Pré-processamento e Limpeza*, os dados foram estruturados e tratados para reduzir inconsistências que poderiam afetar a análise da comunidade científica. Fo-

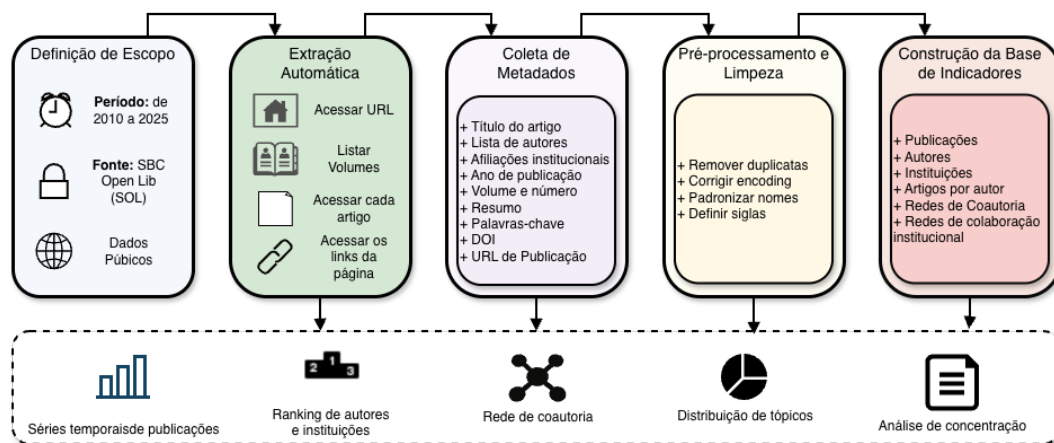


Figura 1. Visão geral da metodologia de coleta e análise de dados do JIDM.

ram realizados procedimentos de remoção de registros duplicados, correção de problemas de codificação de caracteres, normalização de nomes de autores e de instituições e resolução de variações textuais. Para isso, foram aplicadas técnicas de resolução de entidades textuais (*Entity Resolution*), com base na similaridade de *strings* e em regras heurísticas, para identificar possíveis correspondências entre diferentes representações de uma mesma entidade. Por exemplo, ocorrências como *Caetano Traina Jr.* e *Caetano Traina Junior* foram unificadas. Essa etapa é fundamental, pois diferenças textuais podem fragmentar artificialmente autores ou instituições e impactar a construção das redes de colaboração.

Na etapa de *Construção da Base de Indicadores*, os dados refinados foram agregados para gerar indicadores quantitativos de produtividade científica, *e.g.*, número de artigos por autor(a), distribuição anual de publicações e medidas de concentração da produção. Além disso, os dados de coautoria foram utilizados para construir grafos de colaboração científica, nos quais os vértices representam autores e as arestas representam publicações em coautoria. Procedimento semelhante foi aplicado à análise de colaboração institucional. Também foram analisados os tópicos de pesquisa presentes nas publicações, o que permitiu identificar tendências e mudanças temáticas ao longo do período analisado. Destaca-se que todos os dados utilizados neste artigo são *públicos* e foram obtidos exclusivamente a partir de informações disponibilizadas abertamente pela plataforma SOL, sem acesso a áreas restritas ou dados sensíveis. Os *scripts* e os dados utilizados neste artigo podem ser obtidos em <https://github.com/UFFeScience/jidm-analysis>.

3. Análise dos Resultados

Os resultados obtidos a partir da aplicação da metodologia descrita na Seção 2 evidenciam aspectos relevantes da estrutura da comunidade científica brasileira de BD e áreas correlatas. A análise da Figura 2 indica a existência de um núcleo relativamente estável de pesquisadores com participação recorrente no periódico. Ao todo, identificam-se 13 autores com pelo menos 9 artigos publicados no JIDM, dos quais 3 ultrapassam a marca de 20 publicações. Esse resultado sugere a presença de grupos de pesquisa consolidados, que contribuem de forma contínua para o periódico ao longo dos anos.

O índice de Gini obtido (0,344) aponta um cenário de desigualdade moderada na produtividade científica, sem níveis extremos de concentração [da Fonseca Vieira et al. 2024]. Em particular, 10% dos autores mais produtivos respondem por 31,41% do total de

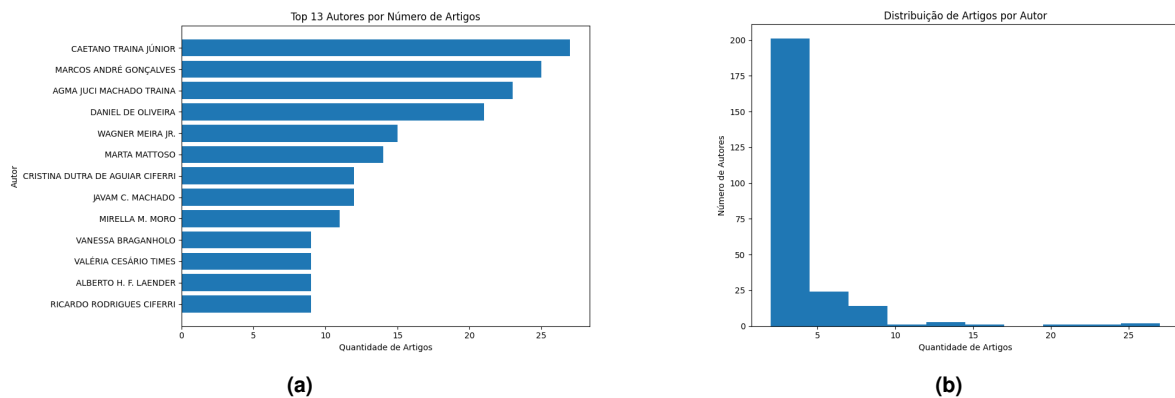


Figura 2. (a) Treze autores com maior número de publicações no JIDM (b) Distribuição da quantidade de autores em função do número total de artigos.

publicações, sendo frequente, nesse grupo, a presença de artigos desenvolvidos em coautoria com estudantes (apesar desse tipo de informação (orientadores $\times$ estudantes) não estar presente na SOL, usamos conhecimento pessoal da comunidade nessa avaliação). De forma semelhante, os 20% e 30% de autores mais produtivos concentram, respectivamente, 46,03% e 57,37% da produção científica observada. Esse padrão é compatível com o observado em periódicos consolidados, nos quais determinados autores desempenham um papel relevante na manutenção do fluxo contínuo de submissões. A Figura 3a apresenta a curva de Lorenz da distribuição de publicações por autor, evidenciando a relação acumulada entre o percentual de autores e o percentual de artigos publicados. Essa visualização permite analisar como a produção científica se concentra ao longo da base de autores e oferece uma perspectiva complementar ao índice de Gini.

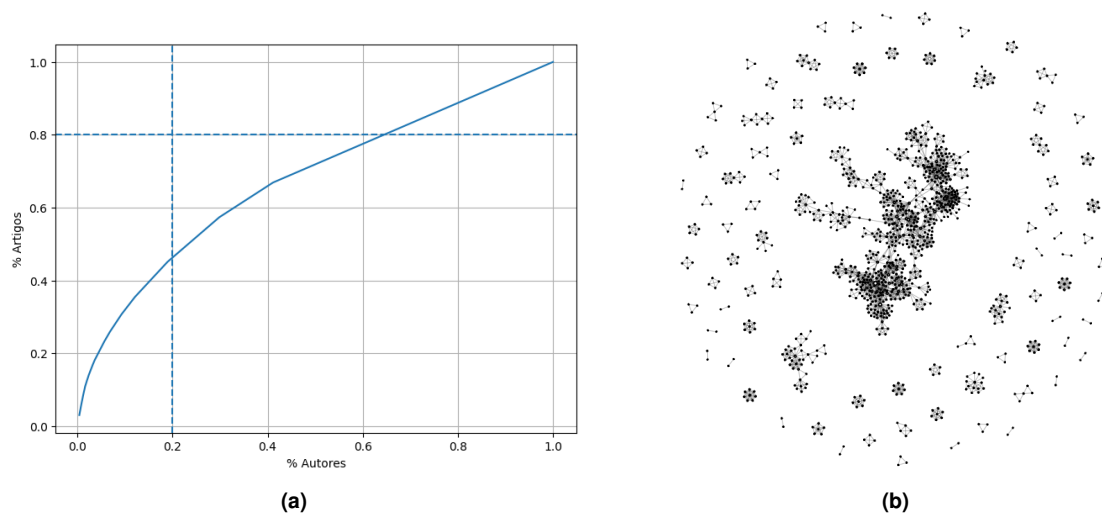


Figura 3. (a) Curva de Lorenz da concentração de publicações por autor (b) Rede de coautoria entre pesquisadores com publicações no JIDM entre 2010 e 2025.

A estrutura de colaboração entre autores (Figuras 3b e 4a) e instituições (Figura 4b) sugere uma comunidade relativamente colaborativa. Redes de coautoria dessa natureza costumam refletir a presença de grupos de pesquisa ativos, de programas de pós-graduação consolidados e de relações acadêmicas de longa duração, *e.g.*, colaborações entre orientadores e ex-orientandos. Ao mesmo tempo, a presença de autores “periféricos” e de grupos menores in-

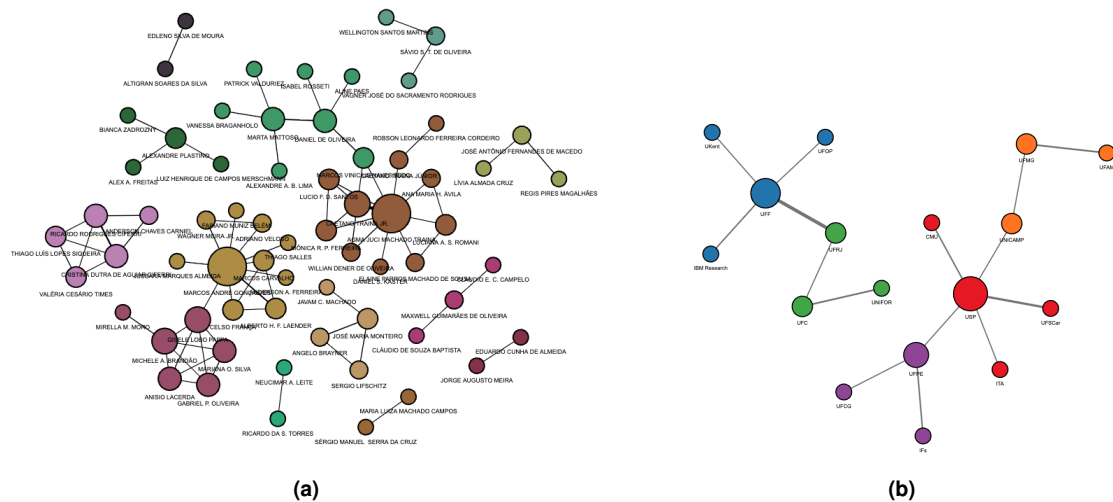


Figura 4. (a) Rede de coautoria entre pesquisadores (b) Rede de coautoria entre instituições.

dica que o JIDM também atua como porta de entrada para novos pesquisadores e instituições emergentes, como podemos verificar nos subgrafos da Figura 3b. Para uma análise mais focada, filtramos a rede de colaboração para apresentar pesquisadores e instituições com pelo menos 4 artigos publicados em conjunto. As Figuras 4a e 4b apresentam, respectivamente, as redes de colaboração científica entre autores e entre instituições após o filtro. Em ambas as redes, os vértices representam atores (pesquisadores ou instituições), enquanto as arestas indicam a existência de publicações em coautoria. O tamanho dos nós está associado ao volume de participação, enquanto as cores representam as comunidades detectadas pelo algoritmo de agrupamento estrutural [Blondel et al. 2008].

A Figura 4a evidencia uma rede de coautoria composta por múltiplas comunidades bem definidas, o que indica que a produção no JIDM se organiza, em sua maioria, em torno de grupos de pesquisa consolidados. Observa-se a presença de componentes densamente conectados, nos quais determinados autores ocupam posições centrais e atuam como polos agregadores. Esses pesquisadores, geralmente docentes seniores, aparecem conectados a diversos coautores, provavelmente incluindo (ex-)orientandos. Também se destacam autores que atuam como pontes entre comunidades, conectando grupos que poderiam estar isolados. A presença dessas conexões entre comunidades sugere que, embora existam *clusters* bem definidos, a comunidade do JIDM não é fragmentada.

A Figura 4b complementa a análise anterior ao mostrar a colaboração entre instituições. Nota-se a existência de instituições centrais, como USP, UFMG, UFRJ, UFPE, UNICAMP e UFF, que concentram múltiplas conexões e atuam como *hubs* institucionais da rede. Todas essas universidades mantêm programas de pós-graduação nível 6 ou 7 na CAPES com grupos ativos em BD, o que ajuda a explicar suas posições de destaque. A USP aparece como principal vértice institucional, conectando-se diretamente a diversas universidades e centros de pesquisa. Esse resultado sugere uma forte capacidade de articulação nacional e de liderança na formação de recursos humanos. Observa-se, ainda, a formação de comunidades regionais bem definidas. Instituições como UFMG, UNICAMP e UFAM aparecem conectadas em um mesmo bloco, enquanto UFPE, UFCG e IFs formam outro agrupamento, com a USP atuando como ponte. De forma semelhante, a UFF e a UFRJ formam outro subgrupo de

colaboração, indicando outro eixo na comunidade. Esses padrões sugerem que a proximidade geográfica, as redes acadêmicas históricas e os projetos conjuntos continuam desempenhando um papel importante na formação de colaborações. A presença de instituições internacionais e corporativas, *e.g.*, *University of Kent* e *IBM Research*, embora periférica, é relevante, pois indica inserção externa da comunidade e colaborações além do ambiente universitário.

Entretanto, um dos resultados mais relevantes diz respeito à distribuição dos tópicos dos artigos. Conforme apresentado na Figura 5a, tópicos frequentes não estão necessariamente associados aos temas tradicionalmente centrais de BD, *e.g.*, processamento de consultas e indexação. Em seu lugar, aparecem tópicos ligados à Mineração de Dados, à Análise de Redes e, mais recentemente, ao Aprendizado de Máquina e à IA em sentido amplo. Esse movimento é consistente com transformações em curso na área de dados, conforme se observa nas tendências de busca por tópicos apresentadas na Figura 5b, nas quais as fronteiras entre Banco de Dados e Aprendizado de Máquina estão se tornando cada vez mais difusas. Parte desse padrão também se deve à publicação de *special issues* do KDMiLe, evento de Mineração de Dados. Isso contribui naturalmente para elevar a presença de tópicos não tradicionalmente considerados BD no conjunto de publicações. A análise não distingue artigos regulares e *special issues*, o que limita a atribuição direta do deslocamento observado a uma única causa. Ainda assim, os resultados sugerem que o principal desafio para os próximos anos não é escolher entre a BD tradicional e as novas áreas orientadas a dados, mas construir uma identidade editorial capaz de acomodar ambas.

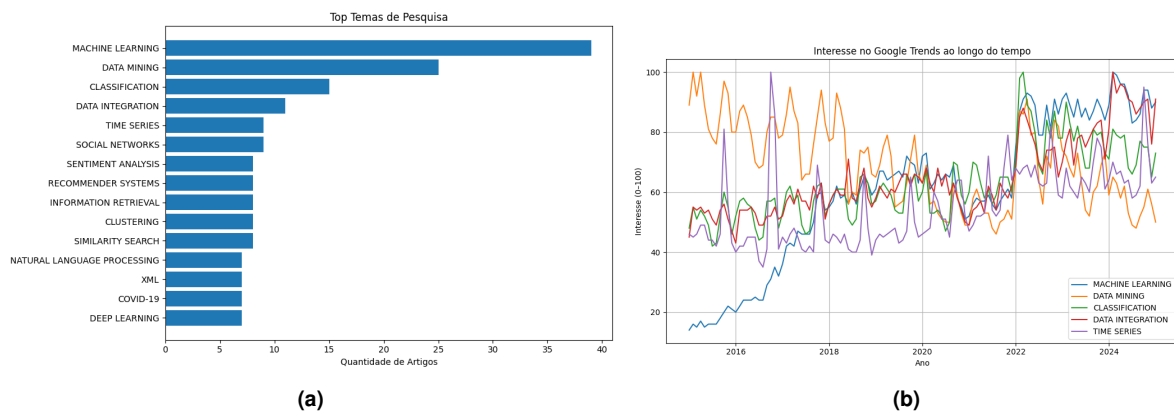


Figura 5. (a) Principais tópicos recorrentes nos artigos. (b) Evolução temporal das buscas dos temas de pesquisa.

4. Conclusões

Este artigo apresentou uma análise bibliométrica e estrutural da comunidade científica associada ao JIDM, no período de 2010 a 2025. A partir de dados públicos coletados da SOL, foi possível investigar padrões de produtividade, concentração da produção científica, redes de colaboração e evolução temática do principal periódico brasileiro da área de BD. Os resultados indicam que o JIDM possui um núcleo estável de pesquisadores recorrentes, responsável por uma parcela relevante das publicações, ao mesmo tempo em que mantém abertura a novos autores e instituições. As redes de coautoria evidenciam uma comunidade madura, organizada em torno de grupos de pesquisa consolidados e de instituições centrais. Além disso, observou-se um aumento na presença de tópicos relacionados à Mineração de Dados e ao Aprendizado de Máquina, o que reflete transformações recentes no ecossistema de dados. Em conjunto, os resultados reforçam o papel estratégico do JIDM.

Referências

- Blondel, V. D. et al. (2008). Fast unfolding of communities in large networks. *J. of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.
- Carvalho, L. P. et al. (2024a). Celebrating two decades of sbisi (2004 to 2023): A comprehensive descriptive and meta-scientific analysis. In *Anais do SBSI*, Juiz de Fora/MG. SBC.
- Carvalho, L. P., Filho, S. L., Brandão, M., Oliveira, J., Santoro, F. M., and Silva, M. (2024b). Onze anos de wei (2013 – 2023), uma análise meta-científica. In *Anais do Workshop sobre Educação em Computação*, pages 611–622, Porto Alegre, RS, Brasil. SBC.
- Carvalho, L. P., Filho, S. L., Brandão, M., Silva, M. D., Oliveira, J., and Pimentel, M. (2023). Thirteen years of webmedia (2010 to 2022) - a comprehensive descriptive and meta-scientific analysis. In *Proceedings of the 29th Brazilian Symposium on Multimedia and the Web*, page 184–192, Porto Alegre, RS, Brasil. SBC.
- Chang, Y., Wang, X., Wang, J., Wu, Y., et al. (2024). A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3):1–45.
- da Fonseca Vieira, V., Ferreira, C. H. G., et al. (2024). A network-driven study of hyperprolific authors in computer science. *Scientometrics*, 129(4):2255–2283.
- Lima, H. et al. (2015). Assessing the profile of top brazilian computer science researchers. *Scientometrics*, 103(3):879–896.
- Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(suppl_1):5200–5205.
- Noorden, R. V. (2014). Scientists and the social network. *Nature*, 512:126–129.
- Silva, T. H. P., Laender, A. H. F., et al. (2017). A profile analysis of the top brazilian computer science graduate programs. *Scientometrics*, 113(1):237–255.
- Singh, S. S., Muhuri, S., et al. (2024). Social network analysis: A survey on process, tools, and application. *ACM Comput. Surv.*, 56(8).