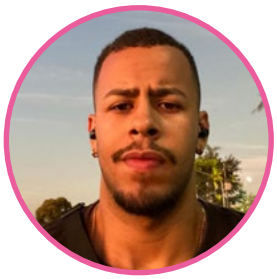
CLUSTERIZAÇÃO EM REDES MONOPARTIDAS DE AUTORES E INSTITUIÇÕES A PARTIR DOS METADADOS DOS ARTIGOS DOS ANAIS DO WIDAT DE 2017 A 2022

Clusterization in Unipartite Networks of Authors and Institutions from the Metadata of Articles of the Proceedings of WIDaT from 2017 to 2022

Agrupamiento en Redes Unipartidistas de Autores e Instituciones a partir de los Metadatos de Artículos de Annals of WDaT de 2017 a 2022

Henrique Monteiro Cristovão
Doutor em Ciência da Informação, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil
Professor, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil
Lattes: http://lattes.cnpq.br/5035919384923489
ORCID: https://orcid.org/0000-0003-2011-7022



Lucas dos Santos do Vale
Mestrando em Ciência da Informação, Universidade Federal do Espírito Santo (UFES), Vitória, ES, Brasil
Professor, Secretaria Estadual de Educação (SEDU), Vitória, ES, Brasil
Lattes: https://lattes.cnpq.br/0502914406473197
ORCID: https://orcid.org/0009-0001-5510-3520

Resumo

Introdução: o Workshop de Informação, Dados e Tecnologia (WIDaT) está em sua sexta edição em 2023. As cinco edições anteriores somam 151 artigos publicados nos anais, com a participação de 267 autores vinculados a 49 instituições em 5 países. O objetivo da pesquisa é revelar *clusters* entre autores e entre instituições. **Metodologia**: pesquisa qualitativa de natureza aplicada que se utiliza, sobretudo, de técnicas de visualização de informação com o apoio de métodos de análise de redes sociais. **Resultados**: redes de informação monopartidas entre autores e entre instituições com a indicação de *clusters* por meio de palavras-chave dos artigos. **Conclusão**: as redes de autores e de instituições por meio de palavras-chave ampliou a densidade de conexões comparativamente com as mesmas redes por meio da coautoria. A formação de *clusters* de autores e de instituições sinalizou possibilidades potenciais de aproximação entre autores com intuito de parcerias conforme suas áreas de atuação representadas por palavras-chaves de suas publicações.

Palavras-chave: visualização de informação; análise de redes; projeção bipartida, clusterização; WIDaT.

Abstract

***Introduction**: the Workshop on Information, Data and Technology (WIDaT) is in its sixth edition in 2023. The five previous editions add up to 151 articles published in the annals, with the participation of 267 authors linked to 49 institutions in 5 countries. The objective of the research is to reveal clusters between authors and between institutions. **Methodology**: qualitative research of an applied nature that uses, above all, information visualization techniques with the support of social network analysis methods. **Results**: monopartite information networks between authors and between institutions with the indication of clusters through keywords of the articles. **Conclusion**: the networks of authors and institutions through keywords increased the density of connections compared to the same networks through co-authorship. The formation of clusters of authors and institutions signaled potential possibilities of approximation between authors with the intention of partnerships according to their areas of activity represented by keywords in their publications.*

***Keywords**: information visualization; network analysis; bipartite projection, clustering; WIDaT.*

Resumen

***Introducción**: el Taller de Información, Datos y Tecnología (WIDaT) está en su sexta edición en 2023. Las cinco ediciones anteriores suman 151 artículos publicados en los anales, con la participación de 267 autores vinculados a 49 instituciones en 5 países. El objetivo de la investigación es revelar clusters entre autores y entre instituciones. **Metodología**: investigación cualitativa de carácter aplicado que utiliza, sobre todo, técnicas de visualización de información con el apoyo de métodos de análisis de redes sociales. **Resultados**: redes de información monopartitas entre autores y entre instituciones con indicación de clusters a través de palabras clave de los artículos. **Conclusión**: las redes de autores e instituciones a través de palabras clave aumentaron la densidad de conexiones en comparación con las mismas redes a través de coautoría. La formación de clusters de autores e instituciones señaló potenciales posibilidades de aproximación entre autores con la intención de alianzas según sus áreas de actividad representadas por palabras clave en sus publicaciones.*

***Palabras clave**: visualización de información; análisis de red; proyección bipartita, agrupamiento; WIDaT.*

DOI 10.22477/vi.widat.46

1. Introdução

A área da visualização de informação tem como objetivos a revelação de padrões invisíveis a partir de dados abstratos e a possibilidade de obtenção de novas percepções, e não apenas imagens bonitas (CHEN, 2013). A representação desses dados abstratos pode ser feita em formato de gráficos e imagens diversas que favorecem a leitura de seus significados. Normalmente, aplicam-se processos de mineração de dados para potencializar revelações de relações não esperadas (HAND *et al.*, 2001). O uso de técnicas inerentes à análise de redes de informação destacam-se como importantes processos para a mineração de dados e a visualização de informação.

Na análise de redes de informação, o uso de inspeção visual possui grande capacidade para identificação de características topológicas da rede e a revelação de relacionamentos que seriam difíceis de enxergar diretamente pela observação ou por meio cálculos e inferências realizadas diretamente sobre os dados da base que a originou (NEWMAN, 2010; NOOY; MRVAR; BATAGELJ, 2018). Algoritmos de clusterização, ou agrupamento, conseguem revelar associações entre os nós de uma rede por meio da busca de padrões e cálculos sobre o quanto eles tem de potencial para se juntarem conforme suas disposições topológicas. A topologia da rede, e não os atributos de seus nós, fornece os elementos essenciais para obtenção de bons resultados advindos da análise de redes (WASSERMAN; FAUST, 1994).

O Workshop de Informação, Dados e Tecnologia (WIDaT) é um evento que, em 2023, está em sua sexta edição e, desde 2017,

> [...] foi idealizado com o intuito de reunir as comunidades acadêmicas e industriais que trabalham com dados no Brasil, por meio da oferta de um espaço de discussão e interação entre profissionais da indústria e pesquisadores das áreas de Ciência da Informação, Ciência da Computação, Engenharias e áreas afins (WIDAT, 2023, on-line).

Ao longo das suas cinco edições[1], o WIDaT publicou 151 artigos em seus anais, com a participação de 267 autores vinculados a 49 instituições em 5 países: Brasil, Portugal, Canadá, Cuba e Espanha. Sendo a grande maioria do Brasil, cobrindo 18 estados da federação.

A *clusterização* de autores e instituições em torno de interesses comuns nas cinco edições do evento pode revelar associações interessantes com potencial capacidade de fomentar futuras coparticipações em trabalhos de pesquisa. Dessa forma, o objetivo da presente pesquisa é revelar *clusters* entre autores e entre instituições por meio de visualização de informação.

1 Em 2017, Florianópolis-SC, UFSC, com os anais disponíveis em: https://repositorio.ufsc.br/bitstream/handle/123456789/180265/Anais.do.WIDAT2017.pdf.
Em 2018, João Pessoa-PB, UFPB, com anais disponíveis em: https://dadosabertos.info/enhanced_publications/idt/.
Em 2019, Brasília-DF, UnB, com os anais disponíveis em: http://widat2019.fci.unb.br/index.php/anais-widat-2019.
Em 2021, Belo Horizonte-MG, CEFET-MG, com os anais disponíveis em: https://pub.colnes.org/index.php/anis/issue/view/14.
Em 2022, Vitória-ES, UFES, com os anais disponíveis em: https://widat2022.ufes.br/wp-content/uploads/2023/04/widat-2022-anais.pdf.

2. Procedimentos Metodológicos

A pesquisa é qualitativa, de natureza aplicada e se utiliza, sobretudo, de técnicas de visualização de informação com apoio de métodos de análise de redes sociais (ARS) (MATHEUS; SILVA, 2006), pois a ARS é uma ferramenta metodológica que permite enxergar o que outras abordagens não permitem (HIGGINS; RIBEIRO, 2018; WASSERMAN; FAUST, 1994). Processos de descoberta de conhecimento (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996) foram aplicados, principalmente baseados em projeções bipartidas[2] sobre as redes de informação bipartidas[3] para torná-las redes monopartidas[4] e, assim, facilitar a sua interpretação (GAO *et al.*, 2017).

Todos os dados coletados, processados e utilizados ao longo da pesquisa, em todas as suas fases, inclusive os arquivos fontes dos resultados, estão disponíveis no repositório de dados da pesquisa[5].

Os dados (ano, nome do artigo, autores, instituições, localizações, e palavras-chave) foram coletados diretamente dos metadados nos anais dos cinco eventos, organizados em duas planilhas e, posteriormente, pré processados com o *software OpenRefine*[6].

Com métodos semi-automáticos, baseados em algoritmos de agrupamento disponíveis no *software OpenRefine*[7], as palavras-chave foram agrupadas. Grande parte desses agrupamentos foram realizados com apoio manual, orientados pela análise de conteúdo de Bardin (1977), onde a categorização dos termos se desenvolve a partir de três etapas: (i) a pré-análise, feita de modo a compreender os termos iniciais de maneira abrangente bem como suas relações; (ii) a estruturação de categorias, realizada a partir do cerne temático, resumindo os termos resultantes da primeira etapa em categorias mais abrangentes; (iii) a interpretação e a inferência das categorizações a fim de identificar e eliminar possíveis ambiguidades fazendo-se uso da intuição e da análise reflexiva e crítica, como recomendam Sousa e Santos (2020).

Os quadros com todos os agrupamentos realizados, tanto de termos equivalentes, quanto da adição de termos, encontram-se disponíveis no repositório de dados da pesquisa.

2 Projeção bipartida é uma ação que elimina um dos grupos de nós de uma rede bipartida, criando ligações entre os nós do conjunto que permaneceu na rede por meio dos nós eliminados.

3 Rede bipartida possui dois conjuntos de nós onde as ligações somente são permitidas entre nós de um conjunto com os nós do outro conjunto.

4 Rede monopartida não possui restrição de ligações entre os seus nós, isto é, um nó pode se conectar com qualquer outro nó.

5 Repositório com arquivos de dados e resultados da pesquisa, disponível em: https://github.com/henrique-cristovao/data_repository_widat.

6 *OpenRefine* é uma ferramenta de código aberto utilizada para a limpeza e transformação de dados. Disponível em https://openrefine.org/.

7 A descrição dos métodos de agrupamentos semi-automáticos, fornecidos pelo *software OpenRefine*, está disponível em: https://openrefine.org/docs/manual/cellediting#cluster-and-edit.

Percebeu-se que, após as ações de agrupamentos e adição de termos, alguns deles ficaram duplicados para o mesmo artigo, conforme mostra o Quadro 1 para o artigo 'A' com a repetição do termo 'infometria'.

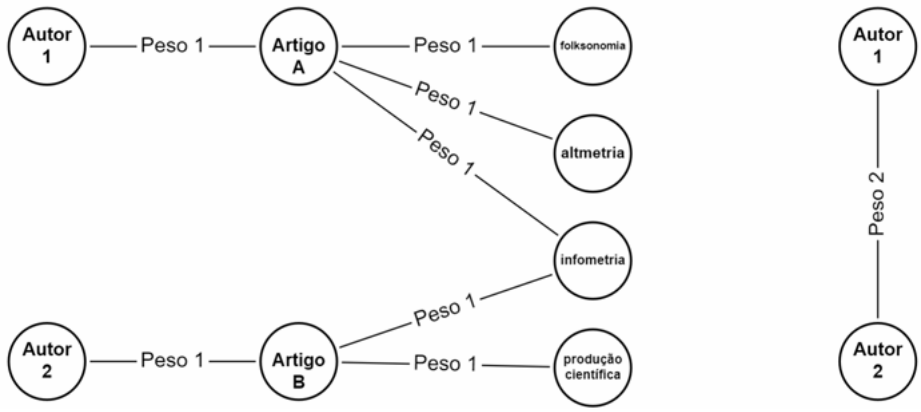
Quadro 1 - Agrupamento de termos

Artigo	Termo agrupado	Termo adicionado
A	*foksonomia*	*infometria*
A	*altmetria*	~~*infometria*~~
B	produção científica	*infometria*

Fonte: autoria própria (2023)

Esse problema poderia gerar pesos incorretos após a execução de projeções bipartidas na rede. Assim, todos os termos repetidos foram eliminados de forma manual. Por exemplo, a eliminação do termo 'infometria', indicada no Quadro 1, remete a uma correta rede monopartida de autores pela coautoria com o peso 2 na aresta que conecta 'Autor 1' com 'Autor 2', Figura 1. Caso contrário, se a eliminação de 'infometria' não tivesse ocorrido, o peso entre 'Autor 1' e 'Autor 2' seria 3, e estaria incorreto.

Figura 1 - Rede tripartida corrigida e o resultado da dupla projeção bipartida na rede de autores



Fonte: autoria própria (2023)

Com o objetivo de mostrar a geolocalização das participações ao longo dos cinco eventos, foi realizada uma reconciliação de dados com o *OpenRefine* para obtenção de latitude e longitude das localidades das instituições a partir da base de dados da *Wikidata*[8]. O *software Power BI Desktop*[9] foi usado para geração de elementos para colaborar com a visualização de informação, como o mapa de geolocalizações e alguns gráficos estatísticos simples sobre os eventos.

8 *Wikidata* é uma base de conhecimento livre que oferece vários serviços gratuitos, entre eles o serviço de reconciliação de dados. Disponível em: https://www.wikidata.org/.

9 *Microsoft Power BI Desktop* é um conjunto de serviços de software de uso gratuito destinados à manipulação de dados e visualização de dados e de informação. Disponível em https://powerbi.microsoft.com/.

As redes de informação foram criadas no formato GML (Graph Modelling Language)[10] por meio de mapeamento dos dados coletados e com auxílio do *software OpenRefine*. A execução de ações inerentes à análise de redes de informação com foco na inspeção visual foi realizada com o *software Gephi*[11]. Para a formatação das redes foram usados algoritmos de distribuição do próprio *software Gephi* bem como o cálculo da métrica de modularidade para a formação dos *clusters*.

10 *GML* (Graph Modelling Language) é um formato para representação de grafos de fácil leitura por humanos e com uma capacidade semântica razoável para configurar as características da rede, dos nós e das arestas. Disponível em: https://en.wikipedia.org/wiki/Graph_Modelling_Language/.

11 *GEPHI* é um *software* de código aberto utilizado para visualização, análise e manipulação de redes e grafos. Disponível em https://gephi.org/.

3. Resultados e Discussão

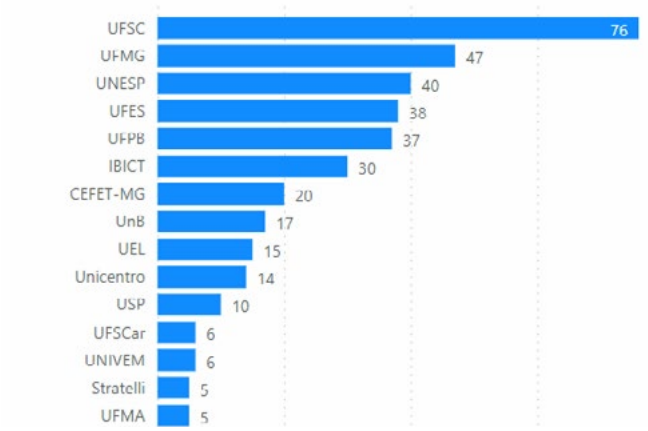
Para um entendimento mais amplo e contextualizado dos *clusters* gerados entre autores e entre instituições, conforme objetivo geral da pesquisa, foram criados alguns elementos visuais secundários. A Figura 3 mostra a geolocalização das instituições dos autores com destaque crescente para locais com maior número de autores.

Figura 3 - Geolocalização das instituições dos autores

Fonte: elaboração própria com apoio dos softwares *OpenRefine* e Power BI Desktop (2023).

A distribuição de participação de autores ao longo dos cinco anos de eventos foi: 48 autores em 2017; 108 autores em 2018; 52 em 2019; 43 em 2021 e 82 em 2022, sendo que no ano de 2020, marcado pelo início da pandemia de COVID-19, o evento não aconteceu. O Gráfico 1 apresenta as 15 instituições com maior número de participações em autorias nos anais.

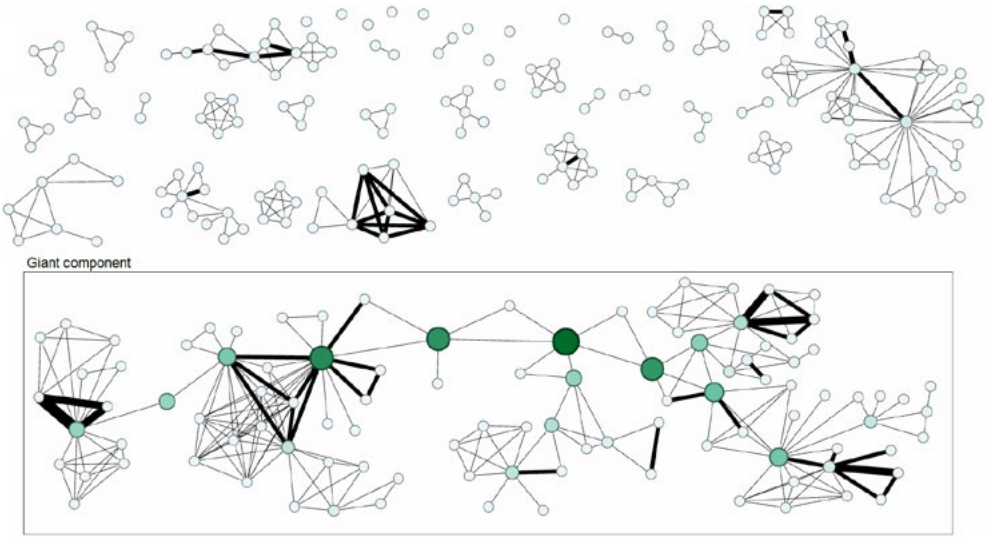
Gráfico 1 - Comparativo de instituições com cinco ou mais participações



Fonte: elaboração própria (2023)

Para a etapa da análise de redes, foi formada inicialmente uma rede de informação 4-partida contendo: artigos, autores, instituições e palavras-chave, denominada aqui de rede original. A rede monopartida da Figura 4 foi obtida com a aplicação de uma projeção bipartida sobre a rede original para conectar autores pela coautoria. Apesar dela mostrar um *giant component*[12] com quase a metade dos autores conectados, muitos deles apareceram sem conexões ou agrupados em componentes menores. Os destaques dos autores no *giant component* foi feito em função dos maiores graus de intermediação (*betweenness*)[13] entre os vários subgrupos de coautoria. Também é possível observar pequenos grupos de autores bem consolidados, com coautoria frequente, cujas arestas de suas conexões são mais espessas.

Figura 4 - Rede monopartida, autores por coautoria, destaque para nós de maior intermediação



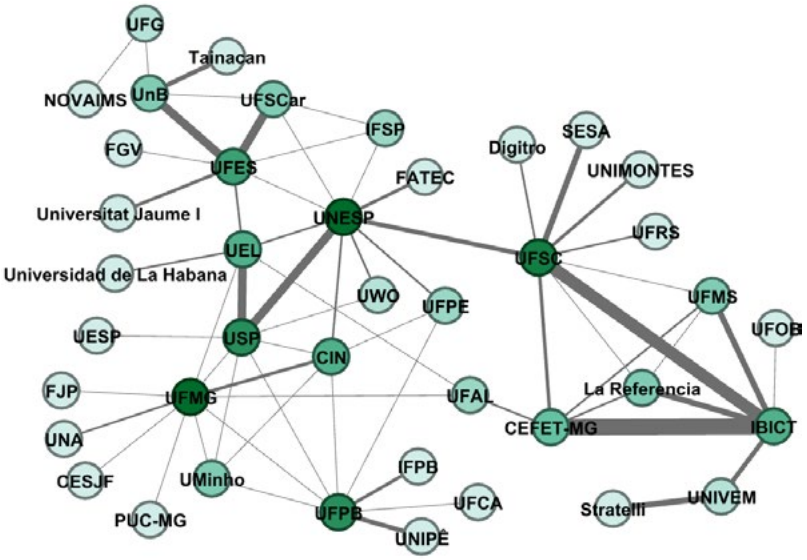
Fonte: elaboração própria com auxílio do software Gephi (2023)

12 *Giant component,* ou componente gigante, de uma rede é um componente conectado com proporções muito maiores do que os demais. Componente conectado é um conjunto de nós onde cada um possui pelo menos um caminho para os demais nós da rede.

13 *Betweenness* mede a importância de um nó quanto à sua capacidade de intermediar o fluxo com os demais nós da rede.

Ainda no contexto de coautoria, a rede da Figura 5 mostra as conexões entre as instituições com destaque para aquelas de maior grau, isto é, com maior número de conexões de coautoria.

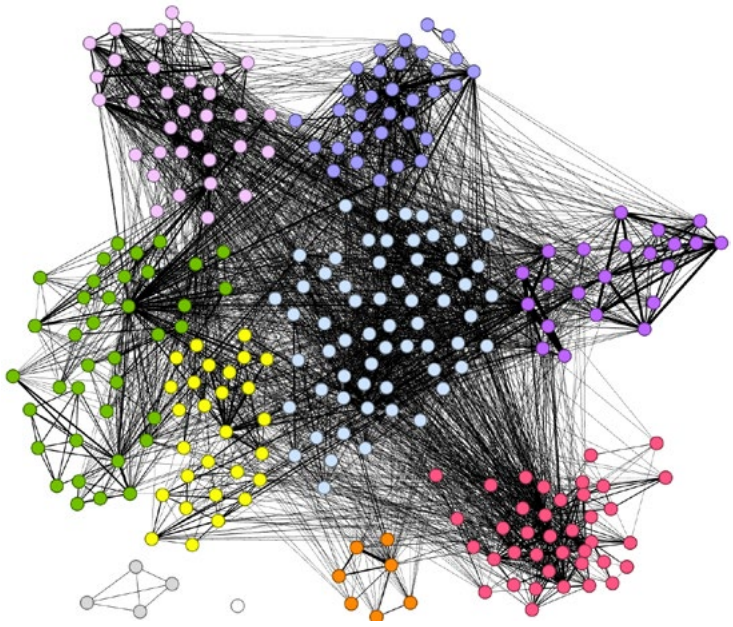
Figura 5 - Rede monopartida de instituições por coautoria com destaque de maior grau



Fonte: Captura de tela (2022)

As redes monopartidas de autores e instituições, respectivamente nas Figuras 6 e 7, foram formadas após duas projeções bipartidas sobre a rede original utilizando-se as palavras-chave. A rede de autores, Figura 6, mostra oito *clusters* bem identificados, isto é, autores com aproximação de temáticas de pesquisa representadas pelas palavras-chave de seus artigos.

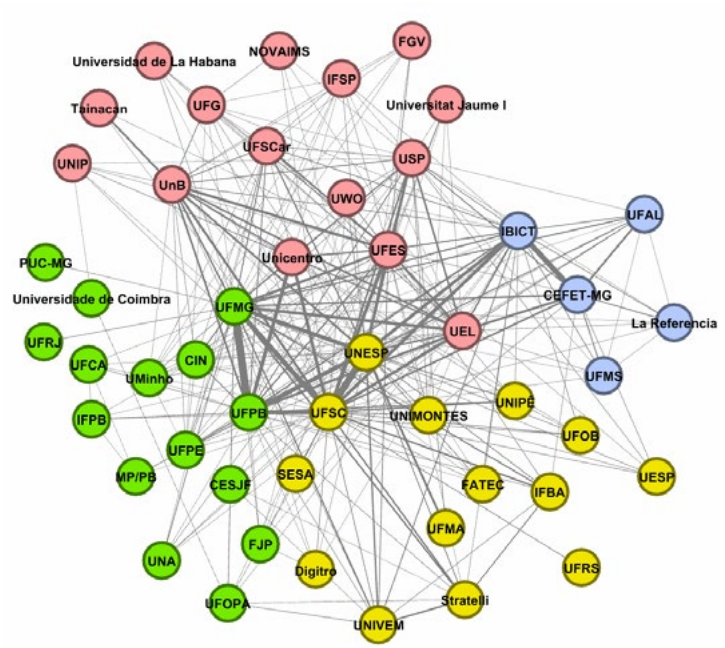
Figura 6 - Rede monopartida de autores com destaques de *clusters* formados por autores com afinidades de palavras-chave em suas publicações



Fonte: elaboração própria com apoio do software Gephi (2023)

A rede de instituições, Figura 7, mostra quatro *clusters* bem identificados onde instituições estão próximas por temáticas de pesquisa representadas por meio das palavras-chave dos artigos dos autores lotados nas instituições.

Figura 7 - Rede monopartida de instituições com destaques de *clusters* formados por instituições com afinidades de palavras-chave nas publicações de seus autores



Fonte: elaboração própria com apoio do *software Gephi* (2023)

Analisando as redes de autores e instituições das Figuras 6 e 7, formadas por meio das palavras-chave, é possível inferir que há uma proximidade potencial muito maior do que aquelas apresentadas pelas Figuras 4 e 5, formadas pela coautoria. Os resultados da *clusterização* realizada evidenciou um potencial de proximidade entre autores, mesmo considerando-se uma grande distância geográfica entre eles, conforme mostrou o mapa da Figura 3, e sinalizou possibilidades de parcerias em pesquisas com temáticas próximas.

4. Considerações Finais

A revelação de *clusters* entre autores, e entre as suas instituições, dos artigos publicados nos anais das cinco edições do WIDaT, por meio de visualização de informação, foi realizada e cumpriu o objetivo da pesquisa. A formação dos *clusters* no contexto das palavras-chave sinalizou possibilidades potenciais de aproximação entre autores com intuito de parcerias conforme suas áreas de atuação representadas por palavras-chaves de suas publicações.

Como continuidade da pesquisa indica-se o levantamento dos termos associados a cada cluster identificado nas redes, tanto de autores quanto de instituições, para tentar identificar as temáticas mais fortes que poderiam estar associadas a cada um dos *clusters* revelados para, em seguida, informar aos respectivos autores sobre possibilidades de cooperação. Considerando a possibilidade de aplicação do método dessa pesquisa em outros eventos, indica-se também a criação de rotinas automatizadas de extração de metadados de anais.

Referências

BARDIN, Laurence. **Análise de conteúdo**. Lisboa: Edições 70, 1977.

CHEN, Chaomei. **Mapping scientific frontiers**: the quest for knowledge visualization. 2. ed. London: Springer Science & Business Media, 2013.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI Magazine**, Palo Alto, v. 17, n. 3, p. 37-54, 1996. DOI: https://doi.org/10.1609/aimag.v17i3.1230.

GAO, Man; CHEN, Ling; LI, Bin; LI, Yun; LIU, Wei; XU, Yong-cheng. Projection-based link prediction in a bipartite network. **Information Sciences**, New York, v. 376, p. 158-171, Jan. 2017. Disponível em: https://doi.org/10.1016/j.ins.2016.10.015. Acesso em: 29 maio 2023.

HAND, David J.; MANNILA, Heikki; SMYTH, Padhraic. **Principles of data mining**. Cambridge, MA: MIT Press, 2001.

HIGGINS, Silvio Salej; RIBEIRO, Antonio Carlos Andrade. **Análise de redes em Ciências Sociais**. Brasília, DF: Enap, 2018. Disponível em: https://repositorio.enap.gov.br/bitstream/1/3337/1/Livro_Analise%20de%20Redes%20em%20Ci%C3%AAncias%20Sociais.pdf. Acesso em: 29 maio 2023.

MATHEUS, Renato Fabiano; SILVA, Antonio Braz de Oliveira e. Análise de redes sociais como método para a Ciência da Informação. **DataGramaZero** - Revista de Ciência da Informação, Belo Horizonte, v. 7, n. 2, p. A03. 2006. Disponível em: DataGramaZero - Revista de ... (brapci.inf.br). Acesso em: 29 maio 2023.

NEWMAN, M. E. J. **Networks**: an introduction. Oxford; New York: Oxford University Press, 2010.

NOOY, Wouter De; MRVAR, Andrej; BATAGELJ, Vladimir. **Exploratory social network analysis with Pajek**: revised and expanded edition for updated software. 3. ed. Cambridge: Cambridge University Press, 2018.

SOUSA, José Raul de; SANTOS, Simone Cabral Marinho dos. Análise de conteúdo em pesquisa qualitativa: modo de pensar e de fazer. **Pesquisa e Debate em Educação**, Juiz de Fora, MG, v. 10, n. 2, p. 1396-1416, jul./dez. 2020. Disponível em: https://periodicos.ufjf.br/index.php/RPDE/article/view/31559. Acesso em: 29 maio 2023.

WASSERMAN, Stanley; FAUST, Katherine. **Social network analysis**: methods and applications. Cambridge: Cambridge University Press, 1994.

WIDAT. Web page. 2023 [*on-line*]. Disponível em: https://widat2023.ibict.br/. Acesso em: 29 maio 2023.