Modelagem e análise de redes sociais através de hipergrafos

Matheus H. B. dos Santos
matheusrickbatista@aluno.ufsj.edu.br
Universidade Federal de São João del-Rei

Vinícius da F. Vieira
vinicius@ufsj.edu.br
Universidade Federal de São João del-Rei

Carolina R. Xavier
carolinaxavier@ufsj.edu.br
Universidade Federal de São João del-Rei

Jussara M. de Almeida
jussara@dcc.ufmg.br
Universidade Federal de Minas Gerais

ABSTRACT

Complex networks are a powerful tool for understanding phenomena in the most diverse contexts. However, modeling networks as graphs, as it is centered on pairwise relationships, offers limitations in modeling many-to-many interactions, as is the case with collaboration in scientific articles. This work provides an overview of central concepts for the use of hypernetworks as models for representing social relations, discussing advantages and disadvantages, challenges and opportunities.The comparison of network and hypernetwork models built on CSBCSet, a database of scientific articles published in CSBC, allows exploring the impact of using hypernetworks to study the phenomenon of coauthorship of scientific articles.

KEYWORDS

hiper-redes, redes sociais, co-autoria, hipergrafos, CSBCset

1 INTRODUÇÃO

A representação de sistemas complexos como redes, onde elementos são modelados como vértices e suas interações são modeladas como arestas, tem se mostrado como uma poderosa ferramenta para a compreensão de fenômenos nos mais diversos contextos [3, 5, 12]. Particularmente, a modelagem da relação entre pessoas como redes sociais fornece informação valiosa sobre a forma como sociedades se organizam, ideias são propagadas e doenças são espalhadas, apenas para citar alguns exemplos [3, 4].

Em uma forma mais simples, as relações em uma rede social podem ser modeladas por grafos nos quais as arestas conectam pares de indivíduos, representados por vértices. Entretanto, esse tipo de representação pode falhar em capturar aspectos importantes dos fenômenos aos quais esses sistemas estão relacionados. Em um contexto de colaboração científica, é impossível desconsiderar a natureza intrínseca das relações muitos-para-muitos sem que haja uma perda significativa de informação e a modelagem através de grafos impede que se tenha clareza, por exemplo, se uma clique de tamanho três representa a colaboração em pares dos autores em três trabalhos distintos ou uma colaboração conjunta dos mesmos autores em uma única publicação [2, 6, 10, 13].

In: IV Concurso de Trabalhos de Iniciação Científica (CTIC 2024). Anais Estendidos do XXX Simpósio Brasileiro de Sistemas Multimídia e Web (CTIC'2024). Juiz de Fora/MG, Brazil. Porto Alegre: Brazilian Computer Society, 2024.

ISSN 2596-1683

Hipergrafos são generalizações de grafos nas quais as *hiperarestas* são capazes de modelar relações poliádicas e, por isso, podem ser utilizados como um modelo mais apropriado para sistemas complexos que apresentam relações muitos-para-muitos. Assim, a generalização, possibilitada pelas hiperarestas, adiciona um grande poder de representação do ponto de vista de modelagem. Porém, essa mudança de perspectiva ao se representar as relações de alta-ordem demanda que toda a teoria de grafos, assim como algoritmos e ferramentas, tomados como base para a análise de redes complexas, sejam revisitados.

Dentro desse cenário, este trabalho tem como objetivo explorar a utilização de hipergrafos como modelos para representação de redes sociais e é guiado pelas seguintes questões de pesquisa: *QPI) Como as relações muitos-para-muitos representadas na CSBCSet podem ser caracterizadas sob o ponto de vista de hiper-redes? QPII) Qual o impacto do uso de modelos de hipergrafos para a compreensão de fenômenos relacionados à coautoria de artigos científicos em relação a modelos baseados em grafos clássicos?* Para que essas questões sejam respondidas, a representação de coautoria a partir da CSBCSet será analisada, de maneira comparativa, tomando como base redes modeladas pelas suas relações em pares (grafos) e redes modeladas pelas suas relações muitos-a-muitos (hipergrafos).

2 HIPERGRAFOS

O termo "*hiper-rede*" será utilizado para permitir a distinção entre sistemas modelados como hipergrafos e sistemas modelados como grafos, aos quais associa-se o termo "*rede*". De forma direta, *hipergrafos* são generalizações de grafos cujas arestas podem relacionar mais de dois vértices. De maneira mais formal, um hipergrafo $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ corresponde a um conjunto de vértices $\mathcal{V} = \{v_1, v_2, ..., v_n\}$ e uma família de hiperarestas $\mathcal{E} = \{e_1, e_2, ..., e_m\}$ na qual cada elemento é um conjunto $e_i \subseteq \mathcal{V}$, $i = 1, ..., m$.

A Figura 1 apresenta um pequeno exemplo que associa três artigos científicos a seis autores distintos (Fig. 1(a)). A Fig. 1(b) apresenta a representação desse exemplo através de um grafo clássico, no qual cada vértice representa um autor e cada aresta representa a coautoria entre um par de autores. Já a Fig. 1(c) apresenta a representação do mesmo exemplo através de um hipergrafo, no qual cada vértice também representa um autor, mas cada hiperaresta representa a colaboração de todos os coautores de um artigo. É possível perceber que a representação baseada no grafo clássico (Fig. 1(a)) não permite distinguir, por exemplo, se o triângulo formado pelos vértices D, E e F indica um único artigo dos autores D, E e F, ou se indica a existência de três artigos dos três autores tomados em pares. Por outro lado, essa distinção é possível pelo modelo de hipergrafo, que representa cada relação, originada pela colaboração

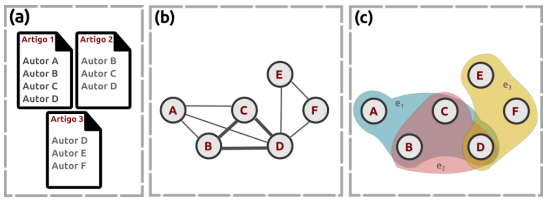


Figure 1: Exemplo da representação de autoria de artigos científicos (a) modelada como: um grafo clássico; (b) um hipergrafo (c).

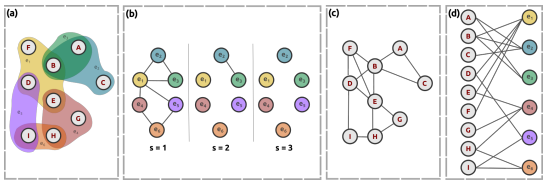


Figure 2: Projeções de um hipergrafo (a) em grafos s-linha (b), grafo clique (c), grafo bipartido (d).

em um artigo científico como uma hiperaresta, ilustrada por uma região de uma cor.

É possível estabelecer métodos para a projeção de hipergrafos em grafos sob diferentes perspectivas. A Figura 2 apresenta um exemplo de projeção de um hipergrafo (Fig. 2(a)) em três tipos de grafos: grafos s-linha, para $s = 1$, $s = 2$ e $s = 3$ (Fig. 2(b)), grafo clique (Fig. 2(c)) e grafo bipartido (Fig. 2(d)). Embora a projeção de hipergrafos em grafos seja muito útil para a aplicação de um grande conjunto de técnicas clássicas utilizadas em análise de redes, descritas e aplicadas de maneira vasta na literatura, sua utilização resulta, inevitavelmente, em modelos com uma perda de informação.

Assim, há um enorme desafio em generalizar técnicas de ciência de redes baseadas em grafos para um contexto de hiper-redes, em que as relações de alta-ordem sejam respeitadas. Askoy *et al.* [1] fazem uma revisão de definições relacionadas a hipergrafos e propõem uma série de conceitos baseados em passeios que estão na base de métodos para a definição de coeficientes e centralidades aplicados a hiper-redes que sejam análogos àqueles tradicionalmente aplicados na análise de redes.

Uma delas é a definição de *s-caminho*, que conecta um par de vértices v_i e v_j em um hipergrafo utilizando apenas hiperarestas que compartilham ao menos s vértices e permite definir a *s-distância* $s - l(v_i, v_j)$, que caracteriza o tamanho do menor s-caminho entre v_i e v_j em um hipergrafo. Dessa forma, as noções de centralidade de proximidade (*closeness centrality*) e centralidade de intermediação (*betweenness centrality*), tradicionalmente utilizadas em grafos clássicos, podem ser utilizadas para hipergrafos como *s-closeness centrality* e *s-betweenness centrality*, respectivamente[1]. A definição de centralidade de grau (*s-degree centrality*) de um vértice v_i em um hipergrafo também pode ser facilmente estendida daquela aplicada a grafos e pode ser calculada proporcionalmente ao número de hiperarestas que possuem ao menos s vértices às quais o vértice v_i pertence.

Uma das tarefas mais importantes na ciência de redes é a identificação de comunidades que, de maneira consensual pode ser descrita como a busca por grupos de vértices com alta densidade interna quando comparada ao volume de ligações externas [12]. Em um esforço de identificar comunidades em hiper-redes, Kumar *et al.* ([9]) definem um método baseado no tradicional método de Louvain [7], frequentemente utilizado para a caracterização de comunidades em grafos. Kumar *et al.* apontam que uma estratégia simples para generalização do problema de identificação de comunidades em hiper-redes poderia ser a projeção do hipergrafo em um grafo clássico e a aplicação de algoritmos tradicionais, mas que essa abordagem faria com que informações essenciais das hiperarestas fossem perdidas.

3 MATERIAIS E MÉTODOS

Os experimentos neste estudo são conduzidos tomando como base a CSBCSet [8], um conjunto de dados sobre publicações acadêmico-científicas nos eventos mais longevos realizados no Congresso da Sociedade Brasileira de Computação (CSBC) [2]. A CSBCSet apresenta metadados de todas as publicações disponíveis entre 2013 e 2022 de 97 edições de dez eventos. Há na CSBCSet um total de 4961 autores, já desambiguados pelos na base de dados, e 1997 publicações distintas.

A partir de um conjunto de $\mathcal{A}$ autores e um conjunto de suas $\mathcal{P}$ publicações são construídas uma rede e uma hiper-rede, representadas, respectivamente, por um grafo $\mathcal{G} = (\mathcal{V}, \mathcal{L})$ e por um hipergrafo $\mathcal{H} = (\mathcal{V}, \mathcal{E})$. Cada autor é representado por um vértice em $\mathcal{G}$ e $\mathcal{H}$. Cada publicação $p \in \mathcal{P}$ é representada por uma clique em $\mathcal{G}$, cujas arestas $l_k \in L$ relacionam os pares de vértices (v_i, v_j) correspondentes a p e por uma hiperaresta $e_k \in E$ que relaciona o conjunto de autores v_i correspondentes a p.

Considerando a definição de Aksoy *et al.* [1], são também construídas redes baseadas em grafos s-linha para $s = 1$, $s = 2$ e $s = 3$, que serão utilizados na investigação sobre os autores mais centrais, tomando como base os métodos de *s-degree centrality*, *s-closeness centrality* e *s-betweenness centrality*. Para o restante do trabalho, as projeções da hiper-rede em grafos grafos s-linha, para $s = 1$, $s = 2$ e $s = 3$, serão chamadas, respectivamente, de *Hiper-rede* $s = 1$, *Hiper-rede* $s = 2$ e *Hiper-rede* $s = 3$.

4 EXPERIMENTOS E DISCUSSÃO

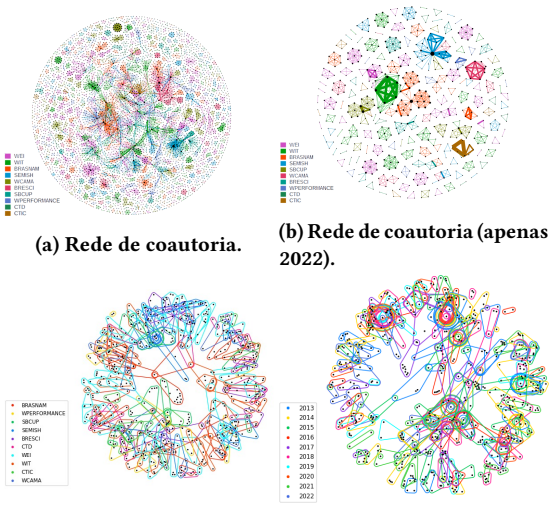
Considerando a metodologia proposta (Seção 3), esta seção apresenta os experimentos conduzidos com o objetivo de responder às questões de pesquisa (Seção 1).

A Figura 3 apresenta uma representação visual dos modelos de Rede (Figs. 3a e 3b) e Hiper-rede (Figs. 3c e 3d). É importante destacar que a Hiper-rede de coautorias resultou em uma visualização bastante poluída e pouco atrativa e, por isso, optou-se por aplicar: um filtro de todos os trabalho, mas apenas do ano de 2022 (Fig. 3c) e um filtro de todos os anos, mas apenas de trabalhos do BraSNAM (Figs. 3d).

Foi realizado um estudo sobre os autores mais centrais, que podem representar pessoas com maior importância dentro do contexto

[1]Neste trabalho, será utilizada a terminologia *s-closeness centrality* e *s-betweenness centrality* no idioma inglês, como no trabalho de Askoy *et al.* [1].

[2]https://csbc.sbc.org.br/



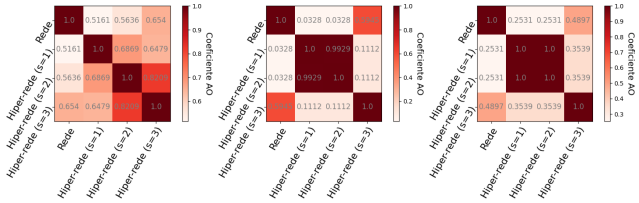
(a) Rede de coautoria.

(b) Rede de coautoria (apenas 2022).

(c) Hiper-rede de coautoria de todos os eventos (apenas 2022).

(d) Hiper-rede de coautoria de todos os anos (apenas BraSNAM).

Figure 3: Representação visual da hiper-rede e da rede de coautoria da CSBCSet.

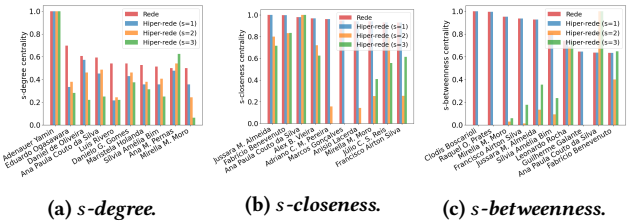


(a) *s-degree* (b) *s-closeness* (c) *s-betweenness*

Figure 4: Coeficientes AO dos *ranks* obtidos em cada modelo de (Hiper-)rede para cada centralidade estudada.

explorado. Para comparar os autores mais centrais identificados nos modelos de (Hiper-)redes estudados, foram calculados os *ranks* de autores considerando as centralidades estudadas para os modelos de Grafo, Hipergrafo ($s = 1$), Hipergrafo ($s = 2$) e Hipergrafo ($s = 3$) descritos na seção 3. Após definir um *rank* dos autores mais centrais, concentramos a análise nas *top-k* posições, com $k = 100$, de forma a minimizar o impacto de divergências que posições inferiores (e menos interessantes) pudessem trazer ao resultado. Assim, é necessário correlacionar *ranks* com elementos potencialmente diferentes e, para isso, utilizamos a métrica *Average Overlap (AO)* [14], que mede a similaridade entre *ranks* de items possivelmente diferentes. A Figura 4 apresenta o resultado dos coeficientes AO, em forma de mapa de calor, comparando os *ranks* obtidos em cada modelo de (Hiper-)rede para a *s-degree centrality* (Fig. 4a), *s-closeness centrality* (Fig. 4b) e *s-betweenness centrality* (Fig. 4c).

Padrões bastante distintos de correlação para a *s-degree centrality* podem ser observados na Figura 4, quando comparada às *s-closeness centrality* e *s-betweenness centrality*. Para a *s-degree centrality* (Fig. 4a), há uma correlação apenas moderada entre os *ranks*



(a) *s-degree.* (b) *s-closeness.* (c) *s-betweenness.*

Figure 5: Comparação de valores de centralidade das dez pessoas autoras mais centrais tomando como referência o modelo de Rede.

gerados para a Rede e a Hiper-rede ($s = 1$), o que está ligado à diferença na própria definição de grau para redes e hiper-redes. Enquanto o grau em uma rede de coautoria está relacionado à quantidade de coautores aos quais um indivíduo está relacionado, em uma hiper-rede de coautoria o grau está relacionado à quantidade de trabalhos dos quais ela participa. À medida em que o valor s aumenta para a construção do modelo de Hiper-rede, autores que participam em artigos com menos coautores são desconsiderados, e o *rank* prioriza, em suas posições mais altas, autores que colaboram em trabalhos com mais colaboradores. Dessa forma, aumenta também a correlação entre os *ranks* gerados para a Rede e as Hiper-redes ($s = 2$) e ($s = 3$). Nos *ranks* gerados para *s-closeness centrality* e *s-betweenness centrality*, observa-se que há uma baixa correlação entre os *ranks* dos autores das Redes e Hiper-redes ($s = 1$) e ($s = 2$) e apenas a Hiper-rede ($s = 3$) apresenta uma correlação significativa com a Rede. Novamente, ao desconsiderar autores com participação em trabalhos com poucos autores, o modelo de Hiper-rede ($s = 3$) coloca como mais centrais as pessoas com um maior volume de autorias, assim como o modelo de Rede. O resultado revelado pela Fig. 4 indica que o modelo de grafo clique, tradicionalmente utilizado para representar redes de coautoria, ao priorizar como autores mais centrais aqueles que possuem um volume maior de colaboração, pode estar subestimando o potencial de autores com um menor número de colaborações, mas que mantêm colaborações mais consistentes em servir como ponte intermediadora entre relações.

Com o objetivo de aprofundar o estudo sobre os autores mais centrais encontrados em cada um dos modelos de (Hiper-)redes, foi realizada uma investigação mais próxima dos nomes presentes nas primeiras posições dos *ranks*. O resultado desse experimento é apresentado na Figura 5, onde cada uma das subfiguras (Fig. 5a, 5b e 5c) apresenta a comparação para cada uma das centralidades estudadas. Para isso, o modelo de rede foi utilizado como referência e suas dez pessoas autoras mais centrais foram encontradas. Em cada uma das subfiguras da Fig. 5, o eixo-x apresenta os nomes das pessoas nas dez primeiras posições do respectivo *rank* no modelo de Rede. O eixo-y apresenta a medida de centralidade observada para essas pessoas no modelo de Rede, mas também a medida observada no modelos de Hiper-redes $s = 1$, $s = 2$ e $s = 3$, possibilitando, assim, uma avaliação da estabilidade dos *ranks* de centralidade nos diferentes modelos. Primeiramente, nota-se que há pouca concordância nas primeiras posições dos *ranks* encontrados para o modelo de Rede e os modelos de Hiper-redes, o que pode ser visto pelo fato que há um comportamento monotônico para as barras referentes ao modelo de Rede – naturalmente, já que é esse o modelo usado como

referência –, mas que não se repete para os modelos de Hiper-redes. Uma notória exceção ocorre na primeira posição observada para a *s-degree centrality* (Fig. 5a), onde há um alinhamento na primeira posição dos *ranks* para todos os modelos. Para a *s-closeness centrality* (Fig. 5b) e a *s-betweenness centrality* (Fig. 5c) nota-se que há algumas pessoas para as quais nem há valor de centralidade associado nos modelos de Hiper-redes, indicando que essas foram desconsideradas por esses modelos. Esse fato é ainda mais evidente nas primeiras posições do gráfico da Fig. 5c, o que mostra que pessoas que colaboram em artigos com poucos coautores têm papel de destaque como intermediadores de relações no modelo representado por Redes, corroborando o resultado apresentado pela Figura 4.

Os resultados das Figuras 4 e 5 evidenciam a importância da modelagem de coautorias como Hiper-redes como ferramenta para análise do fenômeno de colaboração em artigos científicos sob uma ótica alternativa àquela fornecida por modelos de Redes para a investigação local. Uma investigação da organização topológica das (Hiper-)redes foi também realizada para que as coautorias na CSBC-Set pudessem ser exploradas sob uma perspectiva global. Tomando como base o método de Louvain [7] e o método de Kumar *et al.* [9], como descrito na Seção 2, foram identificadas as comunidades considerando os modelos de Redes e Hiper-redes, repectivamente. A modularidade da partição obtida pelo método de Louvain para o modelo de Rede foi de $Q_G = 0.96$, enquanto um valor $Q_H = 0.97$ foi obtido para a partição encontrada pelo método de Kumar *et al.* Esses resultados mostram que, quando são consideradas as coautorias de trabalhos no CSBC no período investigado, há uma clara divisão topológica da (Hiper-)rede de autores, implicando em um alto valor de modularidade para ambos modelos considerados.

A extensão da sobreposição da partição obtida nos modelos de (Hiper-)redes, foi avaliada através do coeficiente *Overlapping Normalized Mutual Information (ONMI)* [11], que retorna um valor 0 quando não há informação mútua entre as partições e 1 quando há uma perfeita correlação. Quando todas as comunidades são consideradas, obtém-se um valor de ONMI = 0.91, indicando uma grande sobreposição. Com o objetivo de verificar se o coeficiente ONMI está dominado por comunidades de tamanhos muito pequenos, um filtro foi aplicado para que apenas comunidades com mais que dez vértices fossem avaliadas. Após a aplicação desse filtro, foi obtido um valor de ONMI = 0.83, indicando que há, sim, uma sobreposição concentrada nas comunidades menores, mas, mesmo sem considerá-las, a sobreposição entre as partições é bastante significativa.

5 CONCLUSÃO

Este trabalho investiga o uso de hiper-redes para a modelagem de relações sociais de alta-ordem. Nesse sentido, é importante destacar que a área ainda é incipiente e oferece um grande potencial de pesquisa, principalmente considerando a baixa quantidade de trabalhos sobre hiper-redes em língua portuguesa (nenhum trabalho em língua portuguesa foi encontrado em assunto relacionado ao aqui explorado, por exemplo).

Em relação à QPI, pode-se notar que ainda há uma certa limitação nas ferramentas computacionais para modelagem e análise de hiper-redes quando compara-se com aquelas voltadas a redes tradicionais. Porém, a análise de hiper-redes, apesar de não ser uma área de pesquisa nova, tem recebido, apenas recentemente, uma atenção mais dedicada, ao contrário da análise de redes tradicionais, que já encontra uma área bastante consolidada. Mesmo assim, há uma grande quantidade de métodos relatados na literatura para investigação de hiper-relações, tanto de maneira local, como de maneira global, o que permite que se tire proveito da riqueza na representação de relações de alta-ordem trazida pelas hiper-redes, mas fazendo análises que sejam coerentes com aquelas realizadas com ferramentas oriundas da ciência de redes. Assim, é possível avançar também na resposta à QPII, verificando que a aplicação de modelos de hipergrafos para a compreensão da coautoria de artigos científicos pode revelar aspectos importantes sobre, não apenas o número de interações realizadas por cada indivíduo, mas a maneira como se dão essas interações. Assim, pode-se compreender com mais clareza o papel de indivíduos com diferentes padrões de colaboração na intermediação de relações e, consequentemente, na difusão de ideias e conhecimento através da hiper-rede.

Como possíveis direções futuras, então, pode-se buscar a verificação da generalidade das observações aqui realizadas em outras bases de dados de interações sociais, inclusive de colaboração e, mais especificamente, de coautoria, mas em outros contextos. Além disso, diversos outros métodos de análise de hiper-redes podem ser empregados, como forma de confrontar os resultados aqui obtidos sob outras perspectivas.

REFERENCES

[1] Sinan Aksoy, Cliff Joslyn, Carlos Ortiz Marrero, Brenda Praggastis, and Emilie Purvine. 2020. Hypernetwork science via high-order hypergraph walks. *EPJ Data Science* 9 (12 2020). https://doi.org/10.1140/epjds/s13688-020-00231-0

[2] Alessia Antelmi. 2021. *Beyond Pairwise Relationships: Modeling Real-world Dynamics Via High-order Networks.* PhD thesis. Università degli Studi di Salerno, Salerno, Italy.

[3] Albert-László Barabási and Márton Pósfai. 2016. *Network science.* Cambridge University Press, Cambridge.

[4] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. 2004. The architecture of complex weighted networks. *PNAS* 101, 11 (2004), 3747–3752.

[5] Marc Barthélemy, Alain Barrat, Romualdo Pastor-Satorras, and Alessandro Vespignani. 2005. Characterization and modeling of weighted networks. *Physica A: Statistical Mechanics and its Applications* 346, 1 (2005), 34–43.

[6] Federico Battiston, Giulia Cencetti, Iacopo Iacopini, Vito Latora, Maxime Lucas, Alice Patania, Jean-Gabriel Young, and Giovanni Petri. 2020. Networks beyond pairwise interactions: Structure and dynamics. *Physics Reports* 874 (2020), 1–92.

[7] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (oct 2008), P10008.

[8] Silas Lima Filho, Luiz Carvalho, José Suzano, Michele Brandão, Jonice Oliveira, and Flávia Santoro. 2023. CSBCSet: Um conjunto de dados para uma década de CSBC, seus eventos e publicações. In *Anais do XII Brazilian Workshop on Social Network Analysis and Mining* (João Pessoa/PB). SBC, Porto Alegre, RS, Brasil, 240–245.

[9] Tarun Kumar, Sankaran Vaidyanathan, Harini Ananthapadmanabhan, Srinivasan Parthasarathy, and Balaraman Ravindran. 2018. Hypergraph Clustering: A Modularity Maximization Approach. arXiv:1812.10869 [cs.LG]

[10] Jürgen Lerner and Marian-Gabriel Hâncean. 2023. Micro-level network dynamics of scientific collaboration and impact: Relational hyperevent models for the analysis of coauthor networks. *Network Science* 11, 1 (2023), 5–35. https://doi.org/10.1017/nws.2022.29

[11] Aaron McDaid, Derek Greene, and Neil Hurley. 2011. Normalized Mutual Information to evaluate overlapping community finding algorithms. *CoRR* (10 2011).

[12] M. E. J. Newman. 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* 103, 23 (2006), 8577–8582. https://doi.org/10.1073/pnas.0601602103 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.0601602103

[13] Alice Patania, Giovanni Petri, and Francesco Vaccarino. 2017. The shape of collaborations. *EPJ Data Science* 6 (08 2017), 18. https://doi.org/10.1140/epjds/s13688-017-0114-8

[14] William Webber, Alistair Moffat, and Justin Zobel. 2010. A Similarity Measure for Indefinite Rankings. *ACM Trans. Inf. Syst.* 28, 4, Article 20 (nov 2010), 38 pages.