Título em Português: Modelando o Conhecimento por Redes Complexas

Título em Inglês: Modeling Knowledge by Complex Networks

Autor: Matheus da Silva Fonseca

Instituição: Universidade de São Paulo

Unidade: Instituto de Física de São Carlos

Orientador: Luciano da Fontoura Costa

Área de Pesquisa / SubÁrea: Física Geral

Agência Financiadora: FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo

Identificando as Fronteiras da Física

Matheus da Silva Fonseca

Luciano da Fontoura Costa

Luciano da Fontoura Costa

Universidade de São Paulo

math.sf.2019@usp.br

Objetivos

Com a crescente disponibilidade de conteúdo científico na World Wide Web(WWW), se torna possível aplicar métodos científicos e matemáticos para investigar a própria ciência. Redes complexas tem sido frequentemente utilizadas como um subsídio para esse tipo de investigação[1]. No presente trabalho, nós desenvolvemos um estudo sobre a organização subjacente do conhecimento em Física e Ciência da Computação por meio da identificação e caracterização das bordas de redes representando subáreas desses campos, derivadas de artigos da Wikipédia. Com isso, buscamos entender como a borda das redes se relaciona com sua estrutura topológica delas, com o conteúdo da Wikipédia relativo as áreas estudadas e com temas como inovação e interdisciplinaridade.

Métodos e Procedimentos

As redes foram produzidas utilizando Wikipedia-API para mapear páginas da Wikipédia em nós e citações entre elas em arestas com peso proporcional a semelhança de texto. Para conseguir dividir os conteúdos em subáreas usamos uma estrutura organizacional da Wikipédia no qual suas páginas são classificadas em categorias, subcategorias, subsubcategorias, etc. Assim, cada rede foi criada utilizando as subcategorias das categorias Subcampos da Física e Ciência da Computação. Para cada subcategoria utilizamos as páginas pertencentes a ela e suas subsubcategorias. Para melhorar a confiabilidade dos dados, foi realizada uma filtragem para diminuir a quantidade de páginas não relacionadas à conteúdos, como páginas sobre congressos, jornais, cientistas ou páginas sobre a estrutura da Wikipédia.

As bordas foram identificadas considerando uma medida chamada acessibilidade dos nós das redes, que tende a ter valores menores para nós mais periféricos[2]. Assim, um nó é pertencente à borda se possuir valor menor que um determinado *threshold*, que foi escolhido como 30% da média da acessibilidade de uma respectiva rede, permitindo uma caracterização de borda com resultados interessantes.

Foram derivadas das redes medidas consolidadas da literatura (número de nós (N), grau médio (<k>), coeficiente de *cluster* médio (CC) e menor caminho médio (ASP)) [3], que foram comparadas com uma nova medida de rede chamada externalidade (b) definida como a razão entre o número de nós da borda Nb pelo número de nós total N (Equação 1).

$$b = \frac{N_b}{N} \qquad (1)$$

A comparação foi feita utilizando coeficientes de correlação de Pearson e Spearman e gráficos de dispersão. Nós também fizemos e a Análise de Componentes principais (Principal Component Analyses – PCA) considerando essas medidas, exceto o número de nós.

Além disso, uma análise de conteúdo foi feita utilizando um modelo chamado rede reduzida, no qual mapeamos as redes obtidas na etapa anterior em nós que são conectados se existem páginas em comuns nas duas redes, com peso proporcional á raiz do número de páginas. Esses nós são divididos em borda e núcleo e existem conexões do tipo borda-borda (bb), núcleo-núcleo (bn) e borda-núcleo (nb). Também analisamos os pesos das conexões e o strentgh dos nós considerando as diferentes conexões[3].

Resultados

We obtained 23 networks: 13 related to Physics and 10 to Computer Science.

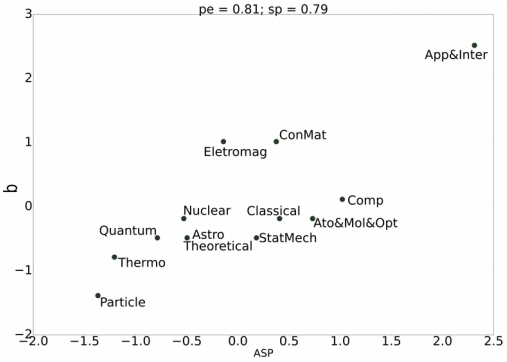

Figura 1: Gráfico de dispersão de *b* x *ASP* para a área da física. Os valores dos coeficientes de corrlação de Pearson e Spearman também estão presentes.

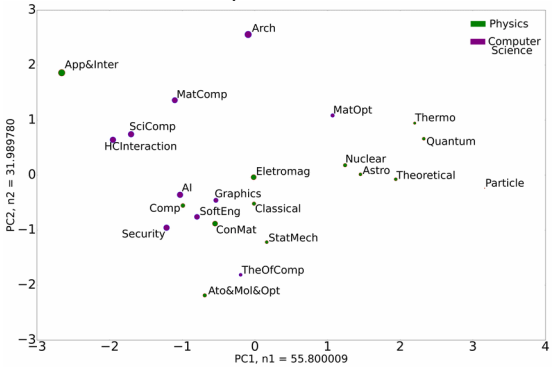

Figura 2: Dois primeiras componentes principais obtidas das medidas <*k*>, *CC*, *ASP,* e *b* das redes de subáreas da Física e Ciência da Computação.

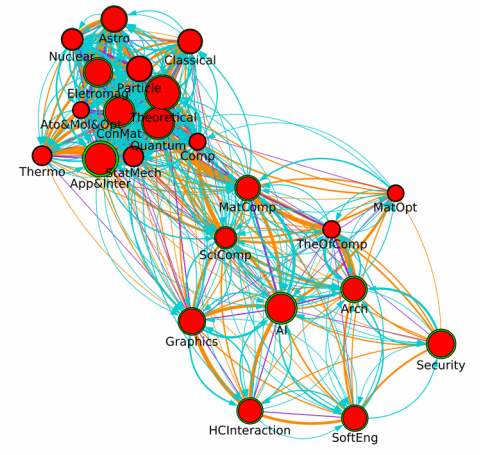

Figura 3: Rede reduzida combinada para Física e Ciência da Computação. As regiões vermelhas dos nós representam seus núcleos, enquanto as verdes representam suas bordas. As conexões em roxo correspondem às conexões bb, as laranjas às nn e as ciano as nb. As flechas apontam para a borda.

A externalidade apresentou correlação significativa apenas com a medida ASP no das redes da Física, com coeficientes de correlação de Pearson e Spearman tendo valores 0,81 e 0,79, respectivamente, conforme apresentado na Figura 1. Apesar da correlação, vemos que ainda há informações diferentes nas duas medidas, dessa forma, é possível que a externalidade seja uma medida útil para caracterização topológica da rede.

Em relação à PCA, consideramos ambas as áreas (Figura 2). No gráfico, o tamanho dos pontos é proporcional à externalidade da rede, após uma transformação MinMax.

Podemos verificar que a maior parte da variância encontra-se capturada nas duas componentes principais. Além disso, as redes da Ciência da Computação se concentram na região superior esquerda, enquanto as da física, com exceção de "Aplicada e Interdisciplinar'" na região inferior direita.

Conseguimos, a partir da rede reduzida (Figura 3), observar algumas relações como áreas mais centrais que são Física Teórica, Mecânica Quântica, Partículas, Matéria Condensada para Física e IA para Ciência da Computação. Assim como há outras mais periféricas, como Física Aplicada e Interdisciplinar. Uma possibilidade para esse resultado é que essas subáreas mais centrais se conectam (interdisciplinaridade) dentro da sua respectiva área, enquanto as periféricas podem ter relação com outras áreas que não sejam Física ou Ciência da Computação. Observou-se também que as conexões mais fortes entre Física e Ciência da Computação ocorrem pelas conexões entre Mecânica Quântica com Computação Matemática e Físcia Computacional com Computação Cientifica.

Por fim, verificou-se que os conteúdos das bordas possuem um caráter mais aplicado e específico, enquanto os do núcleos são mais gerais.

Conclusões

Concluímos com esse projeto que a detecção e análise das bordas de uma rede complexa viabilizou uma análise mais sistemática das interconexões entre os vários tópicos envolvidos.

Agradecemos a FAPESP, especialmente processo 2021/07112-5, pelo suporte e o apoio no desenvolvimento do projeto.

Referências Bibliográficas

[1] FORTUNATO, S. et al. Science of science. Science, v. 359, n. 6379, p. eaao0185, 2018.

[2] TRAVENÇOLO, B. A. N.; VIANA, M. P.; COSTA, L. da F. Border detection in complex networks. New Journal of Physics, v. 11, n. 6, p. 063019, 2009.

[3]COSTA, L. da F. et al. Characterization of complex networks: A survey of measurements. Advances in physics, v. 56, n. 1, p. 167-242, 2007.

Identifying Physics Frontiers

Matheus da Silva Fonseca

Luciano da Fontoura Costa

Luciano da Fontoura Costa

University of São Paulo

math.sf.2019@usp.br

Objectives

With the increasing availability of scientific content in the World Wide Web (WWW), it becomes possible to apply scientific and mathematical methods to investigate science itself. Complex networks have been often used as a subsidy for this type of investigations[1]. In the present work, we aim at getting insights about the underlying organization of knowledge in Physics and Computer Science by identifying and studying the border of networks representing subareas of those fields as derived from articles in Wikipedia. So, we look for better understanding how the border is related to the topological structure of the networks, the Wikipedia content of studied areas and with innovation and interdisciplinary.

Materials and Methods

The networks were produced using Wikipedia-API to mapping the pages into nodes and citations between them into edges with weight proportional to text similarities. The content division into subareas was made using a Wikipedia classification of its pages into categories, subcategories, subsubcategories, etc. Each network was obtained using subcategoreis of Subfields of Physics and Computer Science categories. For each subcategory we used the pages belonging to it and its subsubcategories. In order to enhance data reliability, data filtering was applied to reduce the amount of pages not related to science content like congress, journals, scientists and pages about Wikipedia structure. The borders were identified using a node measurement called accessibility which tends to have small values on peripheral nodes[2].

A node was understood to belong to border whenever its accessibility value was smaller than a threshold taken as 30% of average networks accessibility, allowing an interesting border characterization.
Several measurements were obtained (number of nodes (N), mean degree (<k>), cluster coefficient (CC) and average shortest path (ASP)) [3] and compared with a new measurement called externality (b), defined as the ratio between the number of border nodes (Nb) and the number of total nodes (N) (Equation 1)

$$b = \frac{N_b}{N} \qquad (1)$$

The comparison was made using Pearson and Spearman correlation coefficients . Principal Component Analyses (PCA) was considered using these measurements, with exception to the number of nodes.
An analysis of content was implemented using a model called reduced network, in which the networks obtained in the previous process were mapped into nodes that are connected if they had common pages, with weights being proportional to the number o common articles. These nodes have been divided into border an nucleus, and the four following types of connections were considered: border-border (bb), nucleus-nucleus (nn) and nucleus-border(nb). We also analyzed the connections weights and the nodes strength considering these types of connections[3].

Results

We obtained 23 networks: 13 related to Physics and 10 to Computer Science.

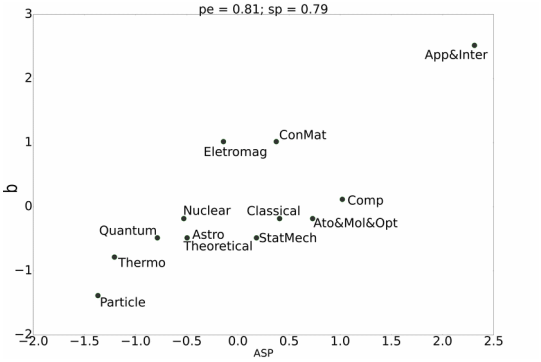

Figure 1: Scatterplot of b x ASP for the Physics areas. The values of the Pearson and Spearman correlation coefficients are also presented.

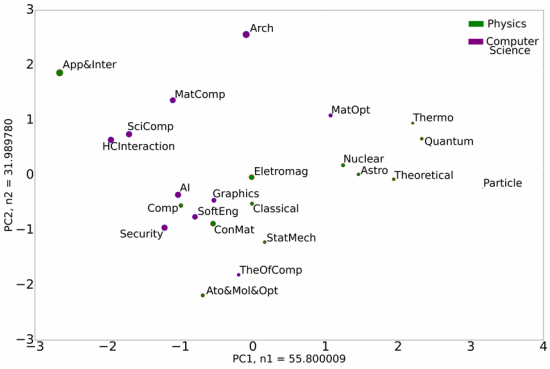

Figure 2: The first two principal components of the PCA obtained from the measurements <k>, CC, ASP and b of the subareas networks of Physics and Computer Science.

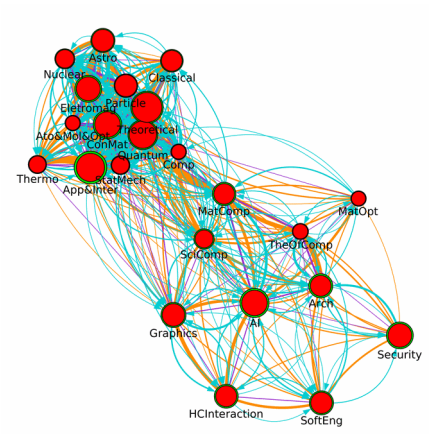

Figure 3: Combined reduced Networks for Physics and Computer Science. The red regions of the nodes represent their nucleus, while the green part represents their borders. The purple edges correspond to bb connections, the orange to nn connections and cyan represents connections nb, with the arrows always pointing to the border.

The externality presented significative correlation just with ASP for Physics cases, with Pearson and Spearman coefficients of 0.81 and 0.79, respectively, as illustrated in Figure 1. Even with this correlation, complementary information was provided by the two measurements, so that the externality can be a useful measurement to characterize the topologies of these networks.

The obtained PCA considering both areas is presented in Figure 2. In this figure, the size of the points is proportional to their externality after a MinMax transformation.

Most variance is captured by the two principal components. In addition, the networks of Computer Science are concentrated at the top right portion, while the Physics ones, with “Applied and Interdisciplinary” being an exception, appears at bottom left.

With the reduced network we could observe that some networks are more central, such as Theoretical Physics, Quantum Mechanics, Condensed Matter for Physics and AI for Computer Science. Some other networks resulted more peripheral, like Applied and Interdisciplinary. One possibe explanation for this result concerns the fact that the conenctions between these central subareas mostly take place within the respective area, while the peripheral subaraes are related to interdisciplinary relationships outside the Physics and Computer Science areas. We also observed that some of the stronger connections between Physics and Computer Science take place between Quantum Mechanics and Mathematical Computation and between Computational Physics and Scientific Computation.

It has also been verified that the contents belonging to the borders tend to be more applied and specific, while the nucleus are more general.

Conclusions

We conclude that the detection and analyses of networks borders allowed a more systematic analyses of interconnections between the diverse considered topics.

We thank FAPESP, especially process 2021/07112-5, for the support in the development of this project.

References

[1] COSTA, L. da F. et al. Characterization of complex networks: a survey of measurements. Advances in Physics, v. 56, n. 1, p. 167-242, Jan. 2007. DOI 10.1080/00018730601170527.

[2] SILVA, F. N. et al. Identifying the borders of mathematical knowledge. Journal of Physics A, v. 43, n. 32, p. 325202-1-325202-7, Ago 2010. DOI 10.1088/1751-8113/43/32/325202.

[3] TRAVENÇOLO, B. A. N.; VIANA, M. P.; COSTA, L. da F. Border detection in complex networks. New Journal of Physics, v. 11, n. 6, p. 063019-1-063019-17, June 2009. DOI 10.1088/1367-2630/11/6/063019.