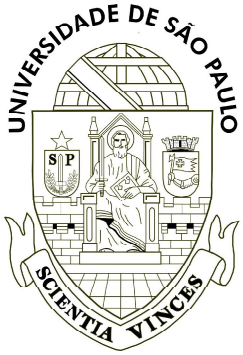


UNIVERSIDADE DE SÃO PAULO
Instituto de Física de São Carlos

Interação Entre Áreas Científicas Sob a Perspectiva de Externalidade de Redes Complexas

Autor: Matheus da Silva Fonseca
Orientador: Prof. Dr. Gonzalo Travieso
(Grupo de Computação Interdisciplinar - GCI)

07 de Agosto de 2023
São Carlos - SP

UNIVERSIDADE DE SÃO PAULO
Instituto de Física de São Carlos
Trabalho de Conclusão de Curso - Monografia

Interação Entre Áreas Científicas Sob a Perspectiva de Externalidade de Redes Complexas

Trabalho de Conclusão de Curso apresentado ao Instituto de Física de São Carlos da Universidade de São Paulo, para obtenção do título de Bacharel em Física Computacional.

Autor: Matheus da Silva Fonseca
Orientador: Prof. Dr. Gonzalo Travieso

São Carlos
2023

Resumo

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, permitindo melhoras em sua compreensão e estrutura. Redes complexas tem sido frequentemente utilizadas como um subsídio para esse tipo de investigação. Ao mesmo tempo, a identificação e obtenção de propriedades de bordas traz frequentemente informações relevantes sobre estruturas matemáticas e sistemas físicos. Juntando estes dois tópicos, construímos redes complexas que representam subáreas da *Física* e da *Ciência da Computação* derivadas de artigos da Wikipédia, seguidas da classificação dos nós das redes em "nós de borda" e "nós de núcleo", a fim de verificar se esta classificação pode fornecer informações relevantes sobre as subáreas. Uma vez que não existe consenso na literatura para identificar a borda de redes complexas, utilizamos um método baseado na Acessibilidade dos nós, uma medida que quantifica a capacidade de acessar a partir de um nó específico os demais nós da rede de forma uniforme, dada uma dinâmica de caminhada na rede (e.g, caminhada aleatória auto excludente). Assim, classificamos os nós da borda como aqueles que têm uma acessibilidade de nó inferior à um limiar e como núcleo os demais. A fim de estudar a influência da fronteira nas propriedades da rede, propomos uma nova medida chamada externalidade, definida como a razão entre a quantidade de nós de borda e o total de nós da rede. Esta nova medida é comparada com medidas frequentemente utilizadas na literatura, como o grau médio e o menor caminho médio, através de métodos estatísticos como coeficientes de correlação de Pearson e Spearman e a análise de componentes principais (PCA). Também criamos redes reduzidas em que as subáreas são mapeadas em nós que mantém sua divisão em borda e núcleo. Estes nós são ligados se as subáreas que representam compartilharem artigos comuns e cada nó pode ter mais do que um tipo de aresta que podem representar os artigos compartilhados que pertencem à borda de ambas as subáreas, ao núcleo de ambas ou à borda de uma e ao núcleo da outra. Os resultados obtidos providenciaram interessantes informações sobre a representação do conhecimento nesses dois campos de estudos. Primeiro, nós pudemos verificar das medidas topológicas que as redes obtidas possuem baixo menor caminho médio, enquanto o grau médio, o número de nós e o coeficiente de aglomeração apresentam variâncias notáveis. Uma correlação positiva foi observada entre a medida externalidade e o menor caminho médio. Outros resultados incluem identificação de algumas subáreas que tendem a ser mais interdisciplinares. Em adição, nós pudemos observar diferentes conexões entre as subáreas considerando bordas e os nós centrais (núcleo).

Palavras-chave: Modelagem, Bordas, Conhecimento, Redes-complexas

1 Introdução

No mundo natural existem diversos entes e fenômenos que mudam constantemente no tempo e no espaço. A ciência não é exceção. Conforme a humanidade se desenvolveu, a ciência emergiu como uma forma de melhor entender o ambiente que nos cerca e fazer previsões valiosas. Enquanto nos seus primórdios, pesquisadores se aventuravam em diversos temas de seu interesse, o grande aumento de resultados e conhecimentos culminou em um aumento progressivo de complexidade, eventualmente levando à especializações de grandes campos em áreas e subáreas de pesquisa. Porém, ao mesmo tempo que surge essa divisão, a ciência também demanda uma integração efetiva entre os saberes produzidos, gerando interações entre eles.

Assim, estudos sistemáticos da estrutura da ciência, caracterizada pela conciliação entre divisão e integração, podem nos levar a um melhor conhecimento de como ela é organizada, além de torná-la mais integrada e eficiente, contribuindo na maneira que o conhecimento científico é compartilhado e ensinado [4, 9]. Esse potencial para estudar a si mesma produziu áreas específicas de pesquisa, como *science of science*, cienciometria e meta-conhecimento [12, 13].

Entre as diversas ferramentas utilizadas no estudo da própria ciência, ressalta-se a importância da teoria de grafos, que surge com Euler em 1736 [1], e mais recentemente das redes complexas [22, 23]. De forma geral, grafos são um conjunto de objetos chamados nós ligados entre si por meio de arestas. O mapeamento de elementos de um sistema de interesse e suas interações em nós e arestas tornam esses objetos extremamente versáteis, podendo ser utilizados para estudar uma grande gama de problemas, chamados de "complexos", em diversas áreas, como sistemas biológicos, relações sociais, a ciência etc [22]. Uma descrição mais detalhada sobre redes e sistemas complexos será apresentada na seção 2.1.

Existem diversos tipos de redes utilizadas para o estudo da ciência, que levam em consideração seus próprios elementos básicos e interações. Entre eles, três bem comuns são redes de co-autoria [2], co-citação [32] e citação [18]. O primeira utiliza relações entre trabalhos conjuntos de cientistas. O segundo analisa artigos considerando pares citados em conjunto. O último relaciona artigos por meio de citações entre si. Por meio desses tipos de redes busca-se responder questões relevantes que tangem à estrutura e ao funcionamento da ciência, como: Quais tópicos de pesquisa são mais relevantes? Como são as relações entre diferentes grupos em um pais ou no mundo? Quais áreas de pesquisa se relacionam mais? Quais estudos apresentam resultados de maior interesse tecnológico? Quais as fronteira da ciência atualmente?

Ressalta-se que esses estudos têm se tornado cada vez mais efetivos graças ao surgimento e capacidade de processamento de grandes bancos de dados acessíveis via a *World Wide Web* (WWW), permitindo pesquisas que consideram uma quantidade inimaginável até algumas décadas atrás de artigos, parcerias, patentes e projetos.

Sobre os diversos temas que podem ser abordados em pesquisas sobre a estrutura e funcionamento da ciência, dois conceitos de extrema importância são interdisciplinaridade e inovação. O primeiro se refere à união de conceitos de diferentes áreas de pesquisa para abordar problemas mais amplos e diversos [25]. O segundo se refere ao potencial da ciência básica e aplicada em produzir avanços científicos e tecnológicos de interesse para a academia e indústria [14]. Dessa maneira, existem diversas abordagens que buscam interdisciplinaridade e conhecimentos inovadores [5, 14, 25].

Uma possível abordagem se refere à detecção das bordas de uma rede. O estudo das bordas de sistemas são relevantes em problemas de diversas áreas como definição das condições de contorno em equações diferenciais, na análise de imagens, delimitação de domínios em materiais e dispersão de informações e doenças. A detecção das bordas de uma rede complexa é um

problema em aberto e existem várias abordagens na literatura [28, 29, 36]. No que se refere aos estudos de cienciometria, questões que surgem são sobre quais propriedades ou características estão presentes nos elementos das bordas de uma rede que representa determinado sistema.

Neste trabalho estudamos redes de citação com textos relacionados à Física e Ciência da Computação, dividindo essas áreas em subáreas menores e realizando a detecção e análise de suas bordas, objetivando conseguir maior compreensão sobre as relações, estruturas e organizações dos diferentes conhecimentos particulares presentes em cada subárea. A base de dados utilizada foi a versão em inglês da Wikipédia [1] e as redes foram criadas mapeando páginas em nós e *hyperlinks* em arestas. Essa escolha foi feita pois entendemos essas subáreas como uma estrutura de conhecimentos em que cada um se constrói ou se acessa a partir de outro. As bordas foram detectadas utilizando a acessibilidade [34] dos nós, discutida na Seção 2.2.

As redes foram analisadas por métodos de estatística multivariada. Sobre as redes foram tomadas medidas comumente utilizadas na literatura: Coeficiente de aglomeração, número de nós, grau médio e menor caminho médio [8, 23]. Além disso, propomos uma nova medida chamada externalidade definida como a razão entre nós identificados como borda e o número total de nós. Também foram produzidas redes reduzidas para sumarizar as relações entre borda das diversas subáreas.

2 Metodologia

2.1 Grafos e Redes Complexas

Redes complexas (ou grafos em uma linguagem mais matemática) são, basicamente, um conjunto de objetos, chamados de nós, ligados entre si. Os itens que ligam os nós são chamados de arestas. Apesar de conceitualmente simples, na prática, qualquer sistema discreto pode ser representado por uma rede, mapeando elementos e interações dos sistemas em nós e arestas. Também é possível torná-las mais sofisticadas, por exemplo, criando tipos diferentes de nós e arestas, associando eles a valores numéricos ou até gerando uma direção única para a conexão entre dois nós. Algumas de suas possíveis aplicações são estudos de caminhos em superfícies de energia [10], na modelagem da World Wide Web [17] e na frequência de citações em artigos [27], além de muitos outros [22]. A Figura 1 apresenta um exemplo de grafo.

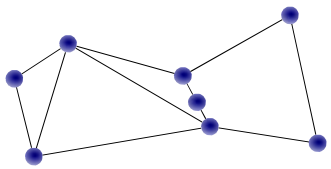


Figura 1: Exemplo de grafo.

Atualmente, graças à grande disponibilidade de bancos de dados e recursos computacionais, se torna possível o estudos de grandes sistemas que produzem redes com milhares ou até milhões de nós. Dessa forma, para estudar tais redes, torna-se necessário incorporar métodos estatísticos. Para isso, precisamos definir e obter medidas, ou seja, valores quantitativos extraídos a partir das redes e capazes de trazer informações sobre a sua estrutura e funcionamento. Entre as medidas, algumas das mais comuns e fundamentais são [22]:

[1] https://www.wikipedia.org/

- grau médio ($\langle k \rangle$): Definida como o valor médio de conexões dos nós $\langle k \rangle = \frac{2L}{N}$, onde L é o número de arestas e N o número de nós da rede, e quantifica o quão conectada é a rede.

- coeficiente de aglomeração (médio) (CA): Definido como $CA = \frac{3 \times \text{número de triângulos na rede}}{\text{número de triplas de vértice conectadas}}$ e quantifica o quão frequente dois nós conectados entre si estão conectados com um terceiro[2].

- menor caminho médio (MCM): Definido como $MCM = \frac{1}{N(N+1)} \sum_{i,j} d_{ij}$ com d_{ij} sendo o menor caminho entre os nós i e j, *i.e.*, a menor quantidade de arestas que são necessárias passar para caminhar do vértice i ao vértice j.

É possível, por exemplo, comparar redes reais com modelos teóricos a partir dessas medidas, como as redes de pequeno mundo [35] caracterizadas por valores altos do CA e baixos do MCM comparadas com redes aleatórias [11]. Além dessas, novas medidas para a obtenção de informações sobre o sistema foram surgindo na literatura ao longo das últimas décadas [8].

Também é importante saber se é possível relacionar essas medidas entre si para diferentes tipos de redes. Por exemplo, se duas medidas apresentam correlação, pode se tornar desnecessário realizar a obtenção de ambas, pois, a partir do valor de uma é possível saber de forma suficientemente boa o valor da outra.

2.2 Acessibilidade

A ideia de borda de uma rede complexa pode ser relacionada com processos dinâmicos de caminhada na rede. Em especial, caminhadas aleatórias podem permitir o desenvolvimento de uma medida capaz de identificar as bordas de uma rede. As caminhadas aleatórias em grafos consistem em processos dinâmicos no qual um agente se desloca pelos nós de uma rede utilizando suas arestas como "caminhos" seguindo alguma regra probabilística. A forma mais simples de caminhada aleatória consiste do agente se mover de um nó para qualquer um dos nós conectados a ele com igual probabilidade, porém, é possível sofisticar o método ao tornar, por exemplo, as probabilidade proporcionais ao peso das arestas ou colocando a regra de que o agente nunca retorne para um nó já visitado, que será o caso deste trabalho. Cada movimento no qual o agente se desloca de um nó para outro é chamado de passo. Habitualmente também coloca-se um número fixo de passos que o agente realiza em uma caminhada, e chamamos esse número de comprimento da caminhada.

A dinâmica se relaciona com a detecção de borda de redes por meio do conceito de diversidade de um nó. Por diversidade nos referimos, qualitativamente, à capacidade de um agente de, em uma caminhada aleatória, acessar outros nós da rede partindo de um nó inicial. Assim, quanto mais nós o agente conseguir acessar e com probabilidade mais uniforme, mais diverso é o nó inicial. Espera-se que nós mais centrais das redes possuam maior capacidade de acesso ao resto da rede e, portanto, sejam mais diversos que os nós da borda, conforme apresentado em [6, 31, 33, 34].

Uma forma de quantificar essa diversidade, ou, em outras palavras, esse grau de variabilidade entre os possíveis destinos de uma caminhada aleatória que parte de um nó específico é por meio da entropia, uma quantia muito conhecida para quantificar na física o "grau de desordem" de um sistema, ou em teoria de informação para a quantidade de informação necessária para descrever um sistema. Assim, dada uma caminhada aleatória, definimos a entropia de

[2]Uma definição alternativa para o coeficiente de aglomeração também é encontrada na literatura, porém, não abordamos ela neste trabalho. Maiores de detalhes podem ser obtidos na Seção III. B. de [22]

diversidade $H_h(\Omega, i)$ de um nó i com relação ao conjunto Ω composto por todos os demais nós da rede e a um comprimento de caminhada h como

$$H_h(\Omega, i) = -\sum_{j \in \Omega} \begin{cases} P_h(j,i)\log(P_h(j,i)), \text{ se } P_h(j,i) \neq 0 \\ 0, \text{ se } P_h(j,i) = 0 \end{cases}, \tag{1}$$

onde $P_h(j, i)$ é a probabilidade de se chegar ao nó j partindo do nó i em uma caminhada de comprimento h. É fácil verificar que nos casos limites em que um nó permite acessar apenas um outro nó $H_h(\Omega, i) = 0$ e que no caso dele permitir o acesso a todos os demais nós da rede com mesma probabilidade, $H_h(\Omega, i)$ é máximo.

Por fim, definimos ainda a acessibilidade de um nó como a exponencial da sua entropia de diversidade:

$$A_h(\Omega, i) = \exp[H_h(\Omega, i)]. \tag{2}$$

A acessibilidade pode ser útil em casos que seja necessário normalizar a entropia de diversidade [34]. No nosso caso, ambas podem ser utilizadas, porém, utilizamos a $A_h(\Omega, i)$ pela sua implementação eficiente no software *CVAccessibility* disponível em [3].

Optamos por utilizar a caminhada aleatória auto-excludente para evitar retornos aos artigos já visitados. A escolha do valor h foi tomada como o menor caminho médio arredondado para baixo, de modo a manter a caraterização da acessibilidade, pois, valores muito maiores podem fazer com que qualquer nó acesse qualquer outro nó. As probabilidades de transição serão proporcionais ao peso das arestas, computados com base na semelhança do conteúdo das páginas, conforme o método explicado na Seção 2.4 e valendo zero se as páginas não possuem *hyperlinks* entre si. Para evitar ruídos numéricos nos cálculos, removemos as arestas com peso inferior à 10^{-6}, representando menos de $0,5\%$ do total de arestas.

Podemos então identificar como borda os nós com acessibilidade abaixo de um determinado limiar. Tomamos esse limiar como 30% da acessibilidade média da respectiva rede pois esse valor produziu bons resultados. Chamamos de N_b o número de nós identificados como pertencentes à borda. Iremos nos referir aos demais nós como pertencentes ao núcleo da rede ou nós do núcleo.

A partir disso pudemos definir uma nova medida chamada externalidade (E) dada pela equação

$$E = \frac{N_b}{N}, \tag{3}$$

onde N é o número de nós total da rede. Ou seja, a externalidade representa o tamanho relativo da borda.

2.3 Base de Dados

A base de dados foi produzida a partir da Wikipédia em língua inglesa acessada no dia 11/04/2022. Para conseguir acessar artigos relacionados a áreas de Física e Ciência da Computação utilizamos as chamadas páginas de "categoria" que agrupam artigos da Wikipédia com conteúdos similares [4]. Essas páginas possuem duas seções: "Subcategorias" e "Páginas na categoria". A primeira seção contém *hyperlinks* para outras páginas de categoria que são classificadas como subcategorias da página de categoria em que estão. Observa-se que esse é um

[3] https://github.com/filipinascimento/CVAccessibility
[4] https://en.wikipedia.org/wiki/Help:Category

processo recursivo no sentido de que as páginas de subcategorias possuem suas próprias subcategorias e assim sucessivamente. A segunda contém *hyperlinks* para artigos/textos relacionadas ao assunto da categoria.

Utilizamos, em princípio, as subcategorias das páginas de categorias "*Subfileds of Physics*" e "*Subfields of Computer Science*" para definir as subáreas da Física e Ciência da Computação. Por clareza, as subcategorias dessas páginas serão chamadas apenas de categorias a partir de agora. Dessa forma, os termos categorias e subárea serão utilizadas como sinônimos ao longo do resto do texto. Foram classificadas como conteúdos pertencentes a uma subárea as páginas da seção "Páginas na categoria" na própria categoria e nas suas subcategorias. Exemplificando: a categoria *Artificial Intelligence* possui sua Seção de "Páginas na categoria", então consideramos todas as páginas referenciadas por essa Seção como pertencentes a subárea de inteligência artificial, além disso, a categoria possui suas subcategorias, como, por exemplo, *Evolutionary computation*, e essas subcategorias também possuem suas Seções de "Páginas na categoria", dessa forma, também consideramos essas últimas páginas como pertencentes à subárea de inteligência artificial. Essa escolha foi feita para termos uma amostragem significante, porém, que necessitasse de uma quantidade de recursos computacionais acessível. O diagrama da Figura 2 ilustra o processo. A extração dos dados foi feita por meio da biblioteca de python Wikipedia-API [5].

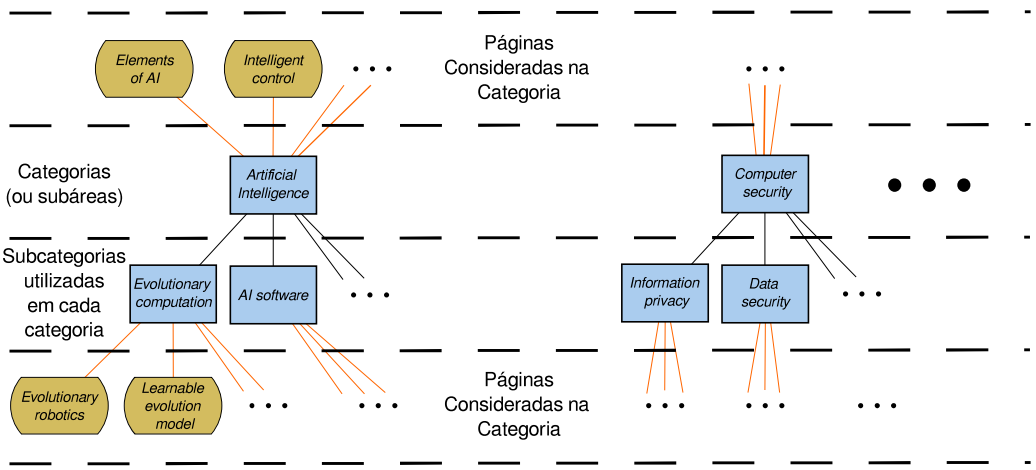


Figura 2: Diagrama ilustrando a remoção das páginas de subcategorias que também são páginas de categoria

Em seguida, foi necessário fazer uma filtragem nos dados. Isso porque existiam diversas páginas e mesmo subcategorias que não eram conteúdos científicos propriamente ditos, como páginas relacionadas a cientistas, instituições, congressos, etc.

Começamos removendo as páginas cujo título inicia com uma das seguintes sequências de caracteres: 'wikipedia:', 'portal:', 'list of:', 'help:', 'template:', 'template talk:', 'file:', 'talk:', 'user:', 'user talk:', 'wikipedia talk:', 'outline of ', 'lists of', and 'glossary of'.

Em seguida, removemos os textos da Wikipédia referenciados pelas páginas *List of physicists* e *List of computer scientists* e as páginas das subáreas que faziam uma descrição geral sobre elas, em vez de abordar algum conteúdo particular, como a página *Classical mechanics* na subárea *Classical mechanics*. Por fim, enquanto analisávamos as categorias da Wikipédia, localizamos e removemos algumas páginas que também não eram sobre conteúdos particulares

[5] https://pypi.org/project/Wikipedia-API/

de física e ciência da computação, sendo essas: *Timeline of electromagnetism and classical optics*, *User:Mdale/wikitrust.js*, *Portal:Human–computer interaction*, *Mathematical Optimization Society* e *Index of software engineering articles*.

Além das páginas individuais, algumas subcategorias não foram acessadas durante a produção das redes. Essas subcategorias estão apresentadas na Tabela 1 do Material Suplementar S1[6]. Também foram removidas as subcategorias do tipo "stub" que são um tipo de página considerada pela Wikipédia muito curtas e incompletas para providenciar cobertura sobre um assunto.

Por fim, também foram identificadas categorias que estavam como subcategorias de outras categorias, sendo essas: *Human–computer interaction* como subcategoria de *Artificial Intelligence*, *Human-based computation* como subcategoria de *Human–computer interaction*, *Formal Methods* como subcategoria de *Software engineering*, *Formal Methods* como subcategoria de *Theoretical Computer Science* and *Theory of computation* como subcategoria de *Theoretical Computer Science* para Ciência da Computação e *Quantum mechanics* como subcategoria de *Theoretical physics* no caso da Física. Optamos por manter todas essas categorias como subáreas distintas, entretanto, não incluímos como conteúdo de uma categoria o que estivesse em suas subcategorias que também fossem categorias, de forma a garantir independência entre elas. Por exemplo, na página *Subfields of Computer Science* existem as categorias *Artificial Intelligence* e *Human–computer interaction* que foram utilizadas para criar duas redes de subáreas distintas. Porém, *Human–computer interaction* também é uma subcategoria de *Artificial Intelligence*, assim, na criação da rede da subárea *Artificial Intelligence* não utilizamos as páginas de *Human–computer interaction*. O diagrama da Figura 3 ilustra esse processo.

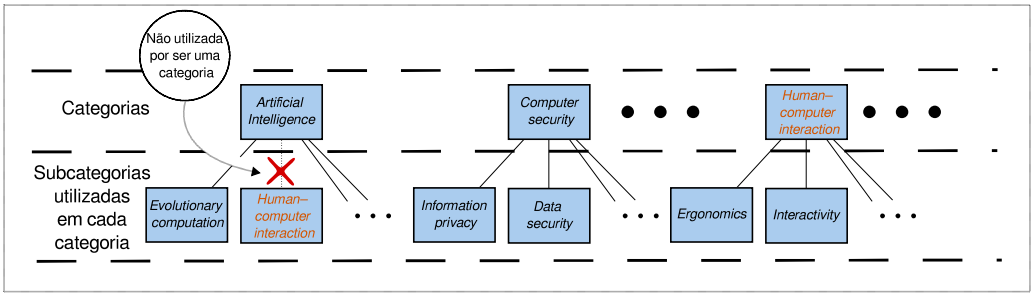


Figura 3: Diagrama ilustrando a remoção das páginas de subcategorias que também são páginas de categoria

Além disso, após um processamento gerado depois das redes serem construídas, algumas dessas categorias foram descartadas, isso será explicado melhor na Seção 2.4 onde também são feitas discussões e alterações no nome de algumas categorias com base na verificação de seus conteúdos para evitar confusões e ambiguidades.

2.4 Construção das Redes

Construímos as redes de cada subárea fazendo o mapeamento das páginas consideradas em cada uma delas em nós e dos *hyperlinks* entre essas páginas em arestas. Por simplicidade consideramos as redes como não-direcionadas. O mapeamento foi feito com a biblioteca python-igraph [7].

[6]Disponível em : https://github.com/Matheus-616/Material-Suplementar-TCC/tree/main
[7]https://python.igraph.org/en/stable/

Além disso, para uma melhor representação da relação entre os conteúdos de duas páginas foram adicionados pesos às arestas baseado na similaridade de seus textos. Para isso, antes de calcular estes pesos, fizemos um pré-processamento dos textos. Primeiro, determinamos as classes gramaticais (CL) de todas as palavras, por exemplo, verbos e substantivos [26]. Em seguida, removemos todos os números, símbolos, hifen e pontuação. As *stopwords* ("palavras de parada" ou "palavras vazias"), palavras com baixo valor semântico (e.g., "we/they/it", "on", "of", entre outras), foram também removidas. Todos os resultados foram lematizados usando o Wordnet [21, 30]. Neste passo, as CL identificadas também foram tidas em conta. A abordagem de lematização é definida como a simplificação das palavras pelo seu respectivo lema. Por exemplo, os diferentes tempos verbais de "play" (e.g., 'playing" e "played") são representados pelo mesmo lema "play". Todos estes métodos foram implementados utilizando o *Natural Language Toolkit – NLTK* [3].

Considerando os textos pré-processados, as semelhanças textuais foram calculadas através de uma técnica de *bag-of-words* [20]. Primeiro, foi criada uma matriz em que as linhas e as colunas representam os textos e as contagens de palavras, respectivamente. Aqui, criamos esta matriz considerando todas as páginas numa rede, sem as dividir em subáreas. Para levar em conta a importância das palavras, calculamos a matriz *tf-idf* [19, 20]. Note-se que, para cada nó da rede i, existe um vetor respectivo, $\mathbf{v_i}$, na matriz *tf-idf*. Para todos os pares de nós, o peso da aresta (i, j) foi calculado como a semelhança de cosseno entre $\mathbf{v_i}$ e $\mathbf{v_j}$. Tanto a matriz *tf-idf* como a semelhança de cosseno foram calculadas utilizando a biblioteca scikit-learn [24].

Tendo produzido essas primeiras redes, foi verificado que todas possuíam uma estrutura com uma maior componente (*largest component*) contendo quase a totalidade dos seus nós, dessa forma, como nós isolados podem gerar problemas em algumas das medidas utilizadas no estudo (*e.g.* Para a medida MCM não é possível definir a menor distância d_{ij} entre dois nós em componentes não conectadas da rede), apenas a maior componente de cada rede foi considerado.

2.5 Análise da externalidade com medidas tradicionais

Nossa primeira investigação sobre as redes foi verificar se existe correlação entre a externalidade e outras medidas topológicas amplamente utilizadas na literatura. As medidas utilizadas para comparação foram: O número de nós da rede (N), grau médio dos nós ($\langle k \rangle$), o menor caminho médio (MCM) e coeficiente de aglomeração médio (CA). Para isso criamos gráficos de dispersão da externalidade pelas demais medidas analisando-os visualmente, e computamos os coeficientes de correlação de Pearson e Spearman [16] entre elas, considerando como significativas as correlações com valores absolutos dos coeficientes acima de $0,7$. Consideramos as redes da Física e da Ciência da Computação separadamente para ver se existe diferenças entre as duas áreas.

Além disso, também aplicamos a Análise de Componentes Principais (em inglês *Principal Component Analysis* - PCA) [16] para as medidas $\langle k \rangle$, MCM, CA e E considerando as redes da Física e Ciência da Computação juntas, objetivando uma compreensão geral de quais redes possuem propriedades topológicas mais semelhantes. Como o número de nós é uma medida menos relevante para a topologia da rede, optamos por não adicionar ele na PCA.

2.6 Rede reduzida

A mesma página da Wikipédia pode pertencer a mais do que uma subárea. Determinar quais páginas são essas e identificar os nós que as representam como pertencentes à borda ou ao núcleo é particularmente importante para verificar se essa classificação permite extrair

informações relevantes sobre relações entre as subáreas ou interdisciplinaridade.

Para essas análise, construímos redes reduzidas para a Física, Ciência da Computação, e ambas em conjunto. Na rede reduzida, cada subárea foi mapeada em um nó que contém as seguintes informações sobre a subárea: O seu número total de páginas, o valor da externalidade, uma lista com o nome das páginas na sua borda, uma lista com o nome das páginas no seu núcleo. Cada par de nós pode ter quatro arestas que representam quão ligadas estão as subáreas, cada uma com peso proporcional ao número de páginas compartilhadas, respetivamente, entre suas bordas (aresta *bb*), entre seus núcleos (aresta *nn*), entre borda e núcleo (aresta *bn*), e vice-versa (aresta *nb*).

Verificamos quais subáreas possuem mais páginas em comum com as demais a partir de uma generalização da medida de *strength* do nós [8]. Essa medida, habitualmente, é definida como a soma dos pesos das arestas conectadas ao nó. Porém, como cada nó na rede reduzida pode possuir quatro tipos distintos de conexão, definimos uma versão da *strength* para cada uma delas *i.e.*, *bb-strength*, *nn-strength*, *bn-strength* e *nb-strength*.

Por fim, também comparamos os pesos das arestas *bb* e *nn* para os pares de nós. Optamos por não trabalhar com as arestas *nb* e *bn* nesta parte porque, como arestas são relacionadas a pares de nós, não podemos distinguir *nb* de *bn*. Por exemplo, se considerarmos os pares (*Eletromag*, *Nuclear*), que é o mesmo que (*Nuclear*, *Eletromag*), *bn* pode referir-se a uma aresta ligando a borda de *Nuclear* ao núcleo de *Eletromag* e vice-versa. Assim, há uma ambiguidade intrínseca nesta definição, o mesmo acontecendo com *nb*.

3 Resultados

3.1 Redes obtidas

Foram produzidas com essa abordagem redes com uma grande variação no número de nós, além de redes muito pequenas para serem realizadas análises estatísticas sobre suas medidas. Dessa forma, consideramos para estudo apenas as redes com mais de 450 nós. As categorias desconsideradas foram: *Experimental physics* (205 nós), *Theory of relativity* (387 nós), *Algorithms and data structures* (158 nós), *Concurrency (computer science)* (210 nós), *Database theory* (72 nós), *Formal Methods* (390 nós), *Human-based computation* (27 nós), *Programming language theory* (329 nós) and *Soft computing* (1 nó).

Ficamos ao final com 23 redes, sendo 10 da Ciência da Computação e 13 da Física. O nome de cada uma delas está apresentada na lista abaixo, juntamente com um *label* para identificá-la, o número de nós e a externalidade de cada uma: Applied and interdisciplinary physics: *App&Inter* (2024 nós, $E = 0,25$); Astrophysics: *Astro* (1103 nós, $E = 0,15$); Atomic. molecular, and optical physics: *Ato&Mol&Opt* (460, $E = 0,16$ nós); Classical mechanics: *Classical* (974 nós, $E = 0,16$); Computational physics: *Comp* (468 nós, $E = 0,17$); Condensed matter physics: *ConMat* (1549 nós, $E = 0,20$); Electromagnetism: *Eletromag* (1367 nós, $E = 0,20$); Nuclear physics: *Nuclear* (765 nós, $E = 0,16$); Particle physics: *Particle* (1057 nós, $E = 0,12$); Quantum mechanics: *Quantum* (1685 nós, $E = 0,15$); Statistical mechanics: *StatMech* (679 nós, $E = 0,15$); Theoretical physics: *Theoretical* (1987 nós, $E = 0,15$); Thermodynamics: *Thermo* (632 nós, $E = 0,14$); Artificial intelligence: *AI* (1603 nós, $E = 0,22$); Computational science (or Scientific Computation): *SciComp* (784 nós, $E = 0,22$); Computer architecture: *Arch* (1099 nós, $E = 0,25$); Computer graphics: *Graphics* (1190 nós, $E = 0,18$); Computer security: *Security* (1365 nós, $E = 0,22$); Human-computer interaction: *HCInteraction* (1079 nós, $E = 0,22$); Mathematical optimization: *MatOpt* (1450 nós, $E = 0,16$); Software engineering:

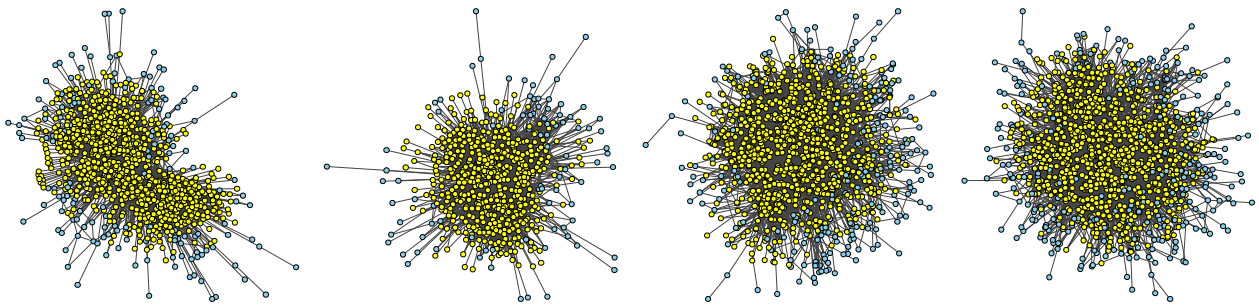
SoftEng (1098 nós, $E = 0,20$); Theoretical computer science (or Mathematical Computation): *MatComp* (1035 nós, $E = 0,21$); Theory of computation: *TheOfComp* (498 nós, $E = 0,15$).

São necessárias agora observações sobre algumas dessas subáreas. Em primeiro lugar, *Theoretical* e *App&Inter* diferem das demais no sentido que, enquanto as outras categorias parecem se referir a diferentes fenômenos ou regimes físicos, essas duas em específico parecem se referir a diferentes formas de estudo de fenômenos diversos, um teórico e o outro aplicado/experimental, mas podendo abranger qualquer fenômeno físico. Entretanto, ambas foram mantidas, porque a subárea *Theoretical*, com exceção da subcategoria *Quantum* que já foi separada pelo procedimento explicado na Seção 2.3, se refere principalmente a ferramentas matemáticas frequentemente utilizadas em física ou áreas de pesquisa da física que são predominantemente teóricas e não estão presentes nas demais categorias, como teoria das cordas e matéria escura, enquanto *Aplicação&Inter* diz respeito aos desenvolvimentos tecnológicos também não incorporados nas outras categorias.

Em segundo lugar, na Ciência da Computação, existem categorias denominadas *Theoretical Computer Science* e *Theory of Computation*. A partir de suas respectivas descrições na Wikipédia, a segunda pode ser entendida como sendo uma subárea da primeira. A primeira categoria é ampla, incluindo muitas abordagens desde teóricas a computação aplicada a áreas substancialmente diferentes, como a aprendizagem de máquina, a criptografia, a geometria computacional, etc. Por isso, passamos a designá-la por *Mathematical Computation* (MatComp). A outra categoria, *Theory of Computation*, está mais diretamente relacionada a pesquisa teórica como a identificação de problemas que podem ser resolvidos por cada tipo de modelo computacional (e.g., Máquinas de Turing, computação analógica, etc.), a eficiência dos algoritmos, entre outras possibilidades.

Em terceiro lugar, existe uma categoria denominada *Computational Science* cujo nome é particularmente semelhante a sua própria área (Ciência da Computação). No entanto, a Wikipédia classifica atualmente esta categoria como um ramo da Ciência da Computação conhecido como Computação Cientifica, relacionado com a simulação computacional, modelos e análises numéricas.

Apresentamos alguns exemplos das redes criadas na Figura 4, destacando em azul os nós da borda e em amarelo os do núcleos.



(a) Classical (b) Thermo (c) Graphics (d) Security

Figura 4: Exemplos de redes derivadas das áreas Física e Ciência da Computação. Os nós azuis são os pertencentes à borda e os amarelos os pertencentes ao núcleo das redes.

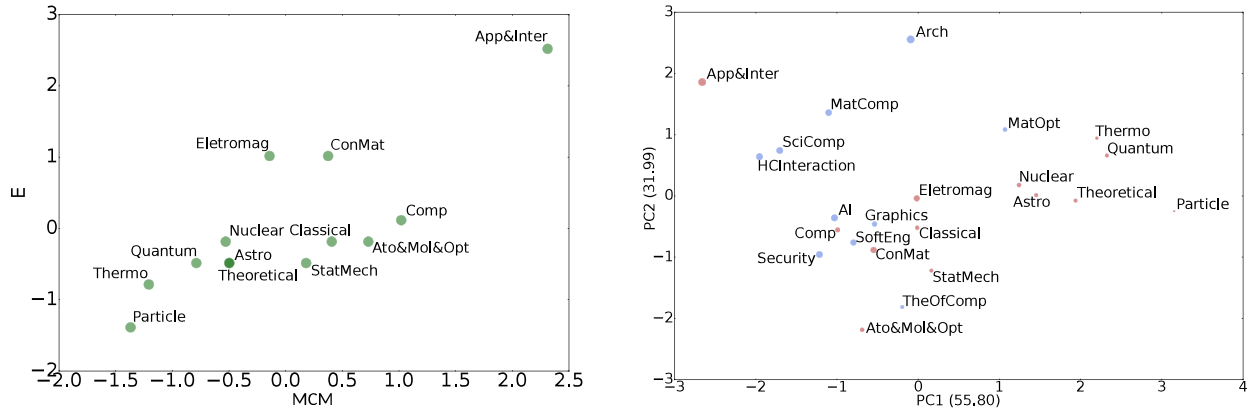
	Física		Ciências da Computação	
	Pearson	Spearman	Pearson	Spearman
N	0,47	0,21	0,65	0,53
$\langle k \rangle$	-0,50	-0,59	0,38	0,18
MCM	0,81	0,79	0,39	0,29
CA	0,25	-0,01	0,26	0,28

Tabela 1: Coeficientes de correlação de Pearson e Spearman da externalidade por: N, $\langle k \rangle$, MCM e CA para redes das subáreas da Física e da Ciência da Computação.

3.2 Análise de medidas topológicas

Iniciamos esta Seção apresentando a Tabela 1 que mostra os valores de todos os coeficientes de Pearson e Spearman obtidos para ambas as áreas.

Com base em uma análise visual e nos coeficientes obtidos, observamos baixa correlação da externalidade com o $\langle k \rangle$, o CA e N. Para o MCM as redes da Ciência de Computação também apresentarem valores baixos, porém, para as redes da Física apresentou valores consideráveis. O gráfico da externalidade pelo menor caminho médio para as redes da física estão apresentados na Figura (a) e os demais encontram-se nas Figuras do Material Suplementar S2[6]. Pela figura 5 (a) observamos que para valores menores de E os dados parecem mais concentrados em uma reta, enquanto para valores maiores parecem estar mais dispersos. Além disso, vemos que a rede *App&Inter* está mais isolada das demais em uma região de alta externalidade, o que pode influenciar principalmente no coeficiente de Pearson, desta forma optamos por fazer novamente o cálculo dos coeficientes sem essa rede e obtivemos uma queda considerável para o coeficiente de Pearson para 0,60, porém, o de Spearman, teve apenas uma pequena queda para 0,73. Assim, os resultados mostram que para as redes da física pode haver uma correlação entre as duas medidas, porém, existindo uma região que há uma dispersão dessas correlações, de forma que a externalidade ainda é capaz de trazer informações novas sobre a topologia da rede.



(a) (b)

Figura 5: (a)Gráfico de dispersão de E x MCM para a área da Física. Os valores dos coeficientes de Pearson e Spearman são, respectivamente, 0,81 e 0,79.(b) Os dois primeiros componentes principais da PCA obtidos a partir das medições $\langle k \rangle$, MCM, CA e E das redes das subáreas de Física e Ciência da Computação. A área cada ponto é proporcional ao número de nós das fronteiras como descrito no texto.

Em relação à Análise de Componentes Principais, o gráfico de dispersão apresentando as duas primeiras componentes principais (PC1 e PC2) está na Figura 5 (b). Verificamos que esses dois eixos contém grande parte da variância dos dados, aproximadamente 87%. No gráfico os pontos vermelhos são as redes da Física e em azul as da Ciência da Computação. O tamanho de cada ponto foi feito para refletir o valor da externalidade de cada nó. Para isso, inicialmente aplicamos a transformação minmax [7] na externalidade, que leva cada medida f_i de uma grandeza em em um valor $\tilde{f}_i$ dada por

$$\tilde{f}_i = \frac{f_i - f_{min}}{f_{max} - f_{min}} \tag{4}$$

onde f_{min} e f_{max} são, respectivamente, a menor e maior medida obtida da grandeza. Como esse mapeamento sempre leva a menor medida no 0, fizemos a área de cada ponto proporcional a $\tilde{f}_i$ adicionado de $0,01$.

Podemos verificar a partir do gráfico da Figura 5 (b) que as redes da Física se encontram predominantemente na região direita inferior do gráfico, enquanto as redes da Ciência da Computação se encontram mais na região esquerda superior, sendo exceções as redes *App&Inter* e *TheOfComp*. Esse fenômeno acabou gerando um pequeno acúmulo de pontos no centro da Figura e pode indicar que os conhecimentos amostrados pela Wikipédia das duas áreas se estruturam de formas distintas.

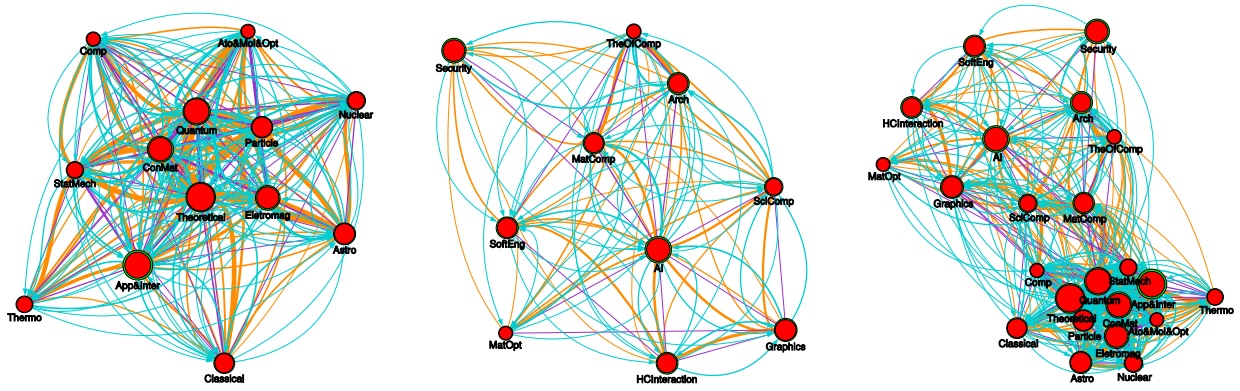
Outro fenômenos que chama atenção é um crescimento da externalidade da região direita inferior para a esquerda superior, que coincide com as regiões de acúmulo das redes da Física e Ciência da Computação, respectivamente. A partir disso, pudemos verificar que as redes da Ciência da Computação apresentam, de forma geral, maiores valores de externalidade.

3.3 Rede reduzida

A Figura 6 apresenta as redes reduzidas. Nelas, os nós na representação gráfica possuem duas regiões, uma interna com área proporcional à quantidade de nós pertencentes ao núcleo e uma externa de área proporcional ao número de nós na borda. Elas foram produzidas por um algoritmo *force-directed* de Fruchterman-Reingold [15], que distribui os nós espacialmente com base nos pesos das arestas, numa analogia com um sistema físico de cargas e molas. Assim, subáreas com mais páginas em comum ficam mais próximas. Isto permite-nos postular quais as subáreas compartilham mais conhecimentos entre si. Observe que, apesar de todos os tipos de conexões serem considerados pelo algoritmo, as conexões do tipo *nn* têm maior influência na definição da proximidade entre as subáreas no layout, pois tendem a ser mais fortes pelo fato de o núcleo ter mais artigos que as bordas.

A partir destas figuras, observa-se a existência de ligações mais fortes na rede da Física em comparação com a da Ciência da Computação, o que pode ser uma consequência das subáreas da Física estarem mais inter-relacionadas entre si, possivelmente devido à longa existência da primeira área com relação à segunda. Outra possibilidade é a área de Ciência da Computação ainda não estar tão bem representada na Wikipédia quanto as da Física a partir das páginas de categorias que utilizamos.

Na Física, existem três subáreas fortemente ligadas entre si: *Theoretical, Quantum* e *Particle*. Para além disso, *ConMat* está também fortemente ligado à *Quantum*. Estas quatro subáreas ocupam um lugar central no grafo, proporcionando pontes entre as outras subáreas. Podemos fazer algumas hipóteses sobre as suas posições centrais. Em primeiro lugar, *Theoretical* está relacionado com teorias matemáticas e formalismos que são utilizados nas subáreas



(a) (b) (c)

Figura 6: Redes reduzidas das áreas da Física (a) e Ciência da Computação (b) e ambas juntas (c). A região vermelha representa o núcleo e a verde representa a borda com ambas as áreas proporcionais ao número de artigos na parte que representam. Em roxo estão as arestas *bb*, em laranjas as *nn* e em ciano as *bn* e *nb*, com a seta apontando para a borda. A largura da aresta é proporcional à raiz quadrada do número de artigos comuns representados por ela.

analisadas com ênfase mais teórica. Isto é apoiado pelas áreas *App&Inter*, *Thermo* e *Classical* que incorporam aspectos mais aplicados, aparecendo mais perto da fronteira da rede. Em particular, *App&Inter* é uma subárea notavelmente grande, com subcategorias e páginas relacionadas com domínios interdisciplinares como *Econofísica, Geofísica* e *Oceanografia*. Assim, supõe-se que enquanto as subáreas centrais estão mais intensamente relacionadas com a interdisciplinaridade entre áreas de estudo da Física, as subáreas na fronteira da rede estão maioritariamente ligadas à interdisciplinaridade com áreas além da Física.

A centralidade resultante das áreas *Quantum*, *ConMat* e *Particle* também está possivelmente relacionada com as suas inter-relações com muitas outras áreas. Os desenvolvimentos recentes em Mecânica Quântica têm sido utilizados em muitos subdomínios da Física, enquanto que Física de Partículas, enfatizando a investigação das propriedades fundamentais da matéria, exige técnicas e tecnologias avançadas derivadas de muitas outras áreas. A Matéria Condensada tem desempenhado um papel especial no desenvolvimento de tecnologias recentes, estando próximo das áreas mais aplicadas da rede obtida.

Podemos também verificar que a ligação mais forte de *Thermo* se estabelece com *StatMech*, o que é esperado dado que grande parte da base deste último está intrinsecamente relacionada com o primeiro. Além disso, como radiação pode ser entendida como a principal fonte de informação proveniente do espaço sideral, muitas vezes gerada por processos nucleares, observam-se ligações entre *Astro* com *Eletromag* e *Nuclear*.

Para as redes da Ciência da Computação não foi possível obter tantas informações quanto no caso da Física, dada a sua estrutura mais esparsa. No entanto, ainda é possível obter conhecimentos preliminares. Para começar, *AI* é a rede mais central, possivelmente como consequência da sua particular importância atual em diversas área de Computação e suas aplicações tecnológicas. Apesar disso, não há ligações *nn* que se destacam entre *AI* e outras subáreas. No entanto, as ligações *bb* entre *AI* e *MatComp* são relativamente fortes.

TheOfComp está intensamente ligado (*bb* e *nn*) à *MatComp* e à *Arch*. A primeira destas relações seria esperável, uma vez que uma pode ser entendida como uma subcategoria da outra. No entanto, a última relação observada com *Arch* é surpreendente, porque *Theory of*

Computation concentra-se principalmente nos aspectos lógicos e matemáticos da computação, enquanto *Arch* tem relações mais próximas com hardware. Outra ligação relevante é entre *HCInteraction* e *Graphics* que pode refletir a tendência da maioria dos dispositivos de interação entre humano e máquinas envolver conceitos e recursos visuais.

Observamos que a partir da Figura 6 (c), as principais ligações entre Física e Ciência da Computação estão relacionadas com interações entre *Comp* e *SciComp*, e entre *Quantum* e *MatComp*.

Dado que a Física Computacional está relacionado com as aplicações de técnicas computacionais em problemas de física, torna-se esperado a proximidade com Computação Científica, que se centra em temas como análises numéricas e simulações. Por outro lado, os artigos partilhados entre *Quantum* e o *MatComp* tendem a estar relacionados com Computação Quântica, Informação Quântica e Estatística, como seria de esperar tendo em conta que o desenvolvimento de computadores quânticos exige a criação de novos modelos teóricos de computação, enquanto a estatística está intrinsecamente relacionada com ambas as áreas.

Agora, em relações à medida de "*strength*" generalizada, calculamos novamente os coeficientes de Pearson e Spearman considerando separadamente as redes da Física e Ciência da Computação para cada par de tipos de arestas. Aplicamos nesses dados o método de padronização [7] que consiste em subtrair de cada medida obtida de uma grandeza sua média e dividir pelo desvio padrão, para deixar ambas mesma escala. Esse valores estão presentes na Tabela 2.

Em todos os casos, os coeficiente de correlação são positivos. Entretanto, eles variam em magnitude, com alguns deles não apresentando correlações significativas. Apesar disso, os gráficos entre as subáreas são bem semelhantes e é possível ver um aumento conjunto das medidas, dessa forma, as diferenças de valores provavelmente ocorrem devido a existirem poucos pontos.

Os gráfico de dispersão para as subáreas da Física e Ciência da Computação são apresentados nas Figuras do Material Suplementar S3[6].

A tabela 3 mostra os valores de *strentgh* generalizados obtidos para cada subárea. A partir dela vemos que as três subáreas com as conexões *nn-strentgh* mais altas, que definem uma posição central na rede reduzida *Physics* (*Quantum, Theoretical* e *Particle* na Figura 6(a)), também possuem fortes conexões dos outros tipos, especialmente *Quantum*; com exceção do caso *nb-strentgh* para *Particle*, que é baixo. *ConMat*, que também está no centro (Figura 6(a)), difere das outras áreas, pois tem *bb-strentgh* normalizada maior que *nn-strentgh*.

	Física		Ciências da Computação	
	Pearson	Spearman	Pearson	Spearman
$bb \times nn$	0,83	0,86	0,81	0,86
$bb \times bn$	0,87	0,82	0,68	0,88
$bb \times nb$	0,76	0,76	0,73	0,60
$nn \times bn$	0,85	0,74	0,66	0,87
$nn \times nb$	0,52	0,71	0,53	0,39
$bn \times nb$	0,53	0,56	0,18	0,26

Tabela 2: Coeficientes de correlação de Pearson e Spearman entre os diferentes tipos *strength* da rede reduzida.

O mesmo ocorre para *Eletromag*, que não é tão central quanto as outras, mas tem conexões fortes e é uma rede relativamente grande. É também interessante observar que, embora

StatMech tenha pouca *strentgh* para três casos, o seu valor de *nb-strentgh* é o segundo mais elevado.

Tal como discutido na seção anterior, a subárea *App&Inter*, embora relativamente grande, tem ligações fortes apenas do tipo *bn-strentgh* e *nb-strentgh*. *Ato&Mol&Opt* têm *nb-strentgh* e *bb-strentgh* elevados, com *nn-strentgh* e *bn-strentgh* baixos.

Área	*nn*	*bb*	*nb*	*bn*	Área	*nn*	*bb*	*nb*	*bn*
AI	0,89	1,76	2,42	0,65	*App&Inter*	-0,66	-0,38	0,13	0,1
SciComp	-0,14	-0,86	0,15	-0,65	*Astro*	-0,38	-0,97	-1	-0,28
Arch	1,13	0,57	-0,67	2,28	*Ato&mol&Opt*	-0,44	0,07	0,46	-1,02
Graphics	-0,05	-0,26	0,26	-0,37	*Classical*	-0,98	-1,21	-1,5	-0,73
Security	-1,38	-0,98	-0,57	-0,85	*Comp*	-0,88	-0,67	-0,07	-0,68
Hcinteraction	0,08	-0,38	-0,87	-0,31	*ConMat*	0,51	1,75	1,62	1,16
MatOpt	1,28	-0,98	-0,87	-1,06	*EletroMag*	0,14	0,56	0,57	0,35
SoftEng	-0,91	-0,86	-0,57	-0,37	*Nuclear*	-0,78	-0,25	-0,68	-0,26
MatComp	0,08	1,05	0,05	0,85	*Particle*	1,08	0,41	-0,59	0,08
TheOfComp	1,58	0,93	0,67	-0,17	*Quantum*	2,1	1,8	0,81	2,38
					StatMech	-0,19	-0,44	1,18	-0,82
					Theoretical	1,43	0,58	0,6	0,85
					Thermo	-0,96	-1,27	-1,53	-1,13

Tabela 3: Valores das medidas de *strength* generalizadas após o processo de padronização para cada uma das subáreas.

Para além disso, no caso da Ciência da Computação, ao contrário da Física, existe apenas uma rede central (AI), sendo as suas ligações *bb-strentgh* e *nb-strentgh* superiores às restantes. O mesmo não acontece com as ligações *nn-strentgh* ou *bn-strentgh*, sendo inferiores à *TheOfComp* e *Arch* no primeiro caso e à *MatComp* e *Arch* no segundo. No entanto, as subáreas *TheOfComp* e *Arch* têm apenas uma ligação *nn-strentgh* forte entre si, e esta ligação individual pode enviesar a sua relação com as outras áreas. Estas duas subáreas apresentam também forte ligação do tipo *bb-strentgh*. Por outro lado, para além de ser a subárea mais central, AI tem uma ligação *bb-strentgh* normalizada maior do que a sua ligação *nn*, mas isto pode ser consequência do enviesamento imposto por *TheOfComp* e *Arch*. Ainda, em respeita à *Arch*, embora seu valor de *bn-strentgh* seja substancialmente elevada, o seu *nb-strentgh* é significativamente baixo, possivelmente devido a sua maior externalidade de $0,25$.

Para discutir as relações entre os pesos das conexões do tipo *bb* com *nn* é necessário apresentar a Figura 7. Ela mostra o gráfico de dispersão comparando os dois pesos analisados, em que cada ponto representa um par de subáreas e a posição no eixo horizontal representa o peso da conexão *bb* entre as duas subáreas e o eixo vertical representa os pesos das conexões *nn*. Não colocamos pontos representando pares de subáreas que não possuem nenhuma das duas conexões entre si. Novamente foi realizado a transformação de padronização das variáveis.

A Figura 7 revela uma região de maior densidade de pontos para valores de peso baixo, tanto *bb* quanto *nn*. Colorimos os pontos dessa região de azul e os demais de laranja, além disso, como os segundos representam pares de subáreas em que há uma quantidade mais significativa de páginas compartilhadas por pelo menos um dos dois tipos de regiões, marcamos o nome de cada um dos pares para fazer uma análise mais detalhada dessas conexões. Dessa forma,

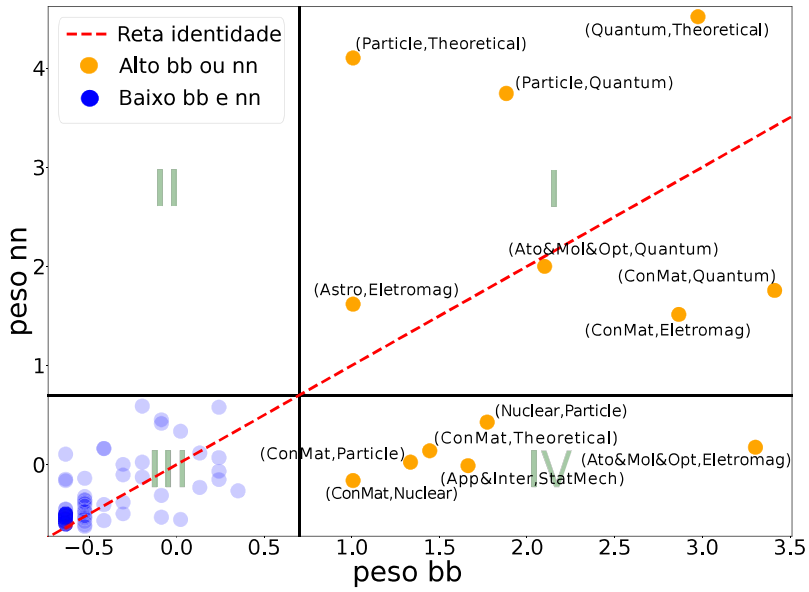


Figura 7: Gráficos de dispersão do peso das arestas *nn* pelas *bb* para cada par de subáreas com pelo menos uma destas ligações. O gráfico de dispersão foi dividido em quatro quadrantes conforme explicado no texto.

dividimos a figura em quatro quadrantes, demarcados pelas retas $x = 0,7$ e $y = 0,7$ e os analisamos individualmente: (I) Valores elevados de *bb* e *nn*; (II) Valores baixos *bb* e elevados de *nn*; (III) Valores baixos *bb* e *nn*; (IV) Valores elevados *bb* e baixos *nn*.

Pelo III quadrante ser o mais populado, optamos por calcular os coeficientes de correlação de Pearson e Spearman para os pontos presentes neles, cujos valores foram, respectivamente, $0,60$ e $0,59$. Dessa forma, verificamos que não há uma correlação significativa entre os dois tipos de peso, pelo menos, para os casos de valores mais baixos, ou seja, onde existe menor interrelação entre as áreas.

Agora, indo para o quadrante I, percebe-se, mais uma vez, que as três áreas centrais da rede da Física (*Quantum, Theoretical* e *Particle*, mostradas na Figura 6(a)), possuem fortes conexões *bb* e *nn* entre si, especialmente o segundo caso. No entanto, neste mesmo quadrante, as únicas ligações que as áreas *Particle* e *Theoretical* apresentam são entre si e com *Quantum*, enquanto *Quantum* tem ligações *bb* e *nn* mais fortes com *Ato&Mol&Opt* e *ConMat*. Outras ligações *bb* e *nn* interessantes podem ser observadas. Em primeiro lugar, a relação *Eletromag* e *Nuclear* ocorre tanto entre as bordas quanto pelo núcleo. Uma situação semelhante é válida para *ConMat* e *Eletromag*.

Como mostra a Figura 7, três ligações da subárea *ConMat* situam-se no quarto quadrante. Observa-se que suas ligações do quadrante I possuem peso normalizado do tipo *bb* superior à *nn*. Este resultado indica que os nós da borda podem ter uma relevância elevada para esta subárea. Outro resultado interessante é o fato da subárea *App&Inter* ser a maior rede e ter a maior externalidade. No entanto, esta subárea tem apenas uma ligação forte, no quadrante IV, com a subárea *StatMech*. Isto pode refletir as várias relações tipicamente observadas entre a Mecânica Estatística e a Física Aplicada.

Verifica-se que não existem pontos no quadrante II, o que pode indicar que conexões maiores entre núcleos só existem se houver também conexões entre as bordas.

4 Conclusões e considerações finais

Com a crescente disponibilidade de conteúdos científicos na World Wide Web (WWW), torna-se possível aplicar métodos científicos e matemáticos para investigar a própria ciência. As redes complexas têm sido frequentemente utilizadas para neste tipo de investigação. No presente trabalho, construímos redes que representam subáreas da Física e Ciências da Computação a partir de páginas da Wikipédia e identificamos os nós pertencentes às suas bordas por meio da medida de acessibilidade. Definimos uma nova medida denominada externalidade da rede como a razão entre o número de nós da borda da rede pelo número total de nós da rede.

Comparamos a externalidade com o $\langle k \rangle$, MCM, CA e N, através de gráficos de dispersão e coeficientes de correlação e conseguimos identificar que a externalidade possui uma correlação positiva e significativa com o MCM para o caso de redes de Física. Porém, ela ainda pode ser capaz de trazer novas informações sobre a topologia da rede.

Outros resultados incluem a identificação de que algumas subáreas tendem a ser mais interdisciplinares, no sentido de possuírem diversas páginas em comum com outras subáreas. Além disso, pudemos verificar diferenças na quantidade de conexões entre um mesmo par de subáreas, tanto em valores absolutos quanto padronizados, ao separar borda de núcleo.

Dessa forma, concluí-se que a identificação de bordas de uma rede de conhecimento real a partir da medida de acessibilidade tem potencial para estudos e análises sobre a estrutura e organização de suas respectivas áreas. Além disso, a medida derivada dela, a externalidade, pode trazer novas informações sobre propriedades topológicas de redes.

É evidente que todas estas hipóteses precisam ser melhor testadas em trabalhos futuros. Por exemplo, é possível utilizar diferentes métodos para criar as redes, como a coautoria e a co-citação, bem como outros conjuntos de dados, como a Web of Science ou o ArXiv. Outra possibilidade é considerar outras medidas para comparar com a externalidade, testar outros limiares para definir a borda, possivelmente com base na distribuição da acessibilidade dos nós. O estudo de diferentes domínios científicos ou grandes áreas, como a matemática, a biologia, a química, ou mesmo a sociologia e a história, também pode ser considerado.

5 Referências

[1] A.-L. Barabási and M. Pósfai, *Network science*, Cambridge University Press, Cambridge, United Kingdom, 2016. OCLC: ocn910772793.

[2] L. M. Bettencourt, D. I. Kaiser, and J. Kaur, *Scientific discovery and topological transitions in collaboration networks*, Journal of Informetrics, 3 (2009), pp. 210–221.

[3] S. Bird, E. Loper, and E. Klein, *Natural language processing with python o'reilly media inc*, (2009).

[4] C. Ceribeli, H. Ferraz de Arruda, and L. da Fontoura Costa, *How coupled are capillary electrophoresis and mass spectrometry?*, Scientometrics, 126 (2021), pp. 3841–3851.

[5] C. Chen and C. Chen, *Mapping scientific frontiers*, Springer, 2003.

[6] L. d. F. Costa, *Inward and outward node accessibility in complex networks as revealed by non-linear dynamics*, arXiv preprint arXiv:0801.1982, (2008).

[7] ——, *Features transformation and normalization: A visual approach (cdt-24)*, (2020).

[8] L. d. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas, *Characterization of complex networks: A survey of measurements*, Advances in physics, 56 (2007), pp. 167–242.

[9] H. F. de Arruda, C. H. Comin, and L. d. F. Costa, *How integrated are theoretical and applied physics?*, Scientometrics, 116 (2018), pp. 1113–1121.

[10] J. P. Doye, *Network topology of a potential energy landscape: A static scale-free network*, Physical review letters, 88 (2002), p. 238701.

[11] P. Erdös and A. Rényi, *On random graphs i*, Publicationes Mathematicae Debrecen, 6 (1959), p. 290.

[12] J. A. Evans and J. G. Foster, *Metaknowledge*, Science, 331 (2011), pp. 721–725.

[13] S. Fortunato, C. T. Bergstrom, K. Börner, et al., *Science of science*, Science, 359 (2018), p. eaao0185.

[14] J. G. Foster, A. Rzhetsky, and J. A. Evans, *Tradition and innovation in scientists' research strategies*, American Sociological Review, 80 (2015), pp. 875–908.

[15] T. M. Fruchterman and E. M. Reingold, *Graph drawing by force-directed placement*, Software: Practice and experience, 21 (1991), pp. 1129–1164.

[16] F. L. Gewers, G. R. Ferreira, H. F. D. Arruda, et al., *Principal component analysis: A natural approach to data exploration*, ACM Computing Surveys (CSUR), 54 (2021), pp. 1–34.

[17] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins, *The web as a graph: Measurements, models, and methods*, in Computing and Combinatorics: 5th Annual International Conference, COCOON'99 Tokyo, Japan, July 26–28, 1999 Proceedings 5, Springer, 1999, pp. 1–17.

[18] S. Lehmann, B. Lautrup, and A. D. Jackson, *Citation networks in high energy physics*, Physical Review E, 68 (2003), p. 026113.

[19] A. G. Maguitman, F. Menczer, H. Roinestad, and A. Vespignani, *Algorithmic detection of semantic similarity*, in Proceedings of the 14th international conference on World Wide Web, 2005, pp. 107–116.

[20] C. Manning and H. Schutze, *Foundations of statistical natural language processing*, MIT press, 1999.

[21] G. A. Miller, R. Beckwith, C. Fellbaum, et al., *Introduction to wordnet: An on-line lexical database*, International journal of lexicography, 3 (1990), pp. 235–244.

[22] M. E. Newman, *The structure and function of complex networks*, SIAM review, 45 (2003), pp. 167–256.

[23] M. E. Newman, A.-L. E. Barabási, and D. J. Watts, *The structure and dynamics of networks.*, Princeton university press, 2006.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research, 12 (2011), pp. 2825–2830.

[25] A. Porter and I. Rafols, *Is science becoming more interdisciplinary? measuring and mapping six research fields over time*, Scientometrics, 81 (2009), pp. 719–745.

[26] A. Ratnaparkhi, *A maximum entropy model for part-of-speech tagging*, in Conference on empirical methods in natural language processing, 1996.

[27] S. Redner, *How popular is your paper? an empirical study of the citation distribution*, The European Physical Journal B-Condensed Matter and Complex Systems, 4 (1998), pp. 131–134.

[28] M. Semenov and K. Lelushkina, *Study of the materials microstructure using topological properties of complex networks*, in IOP Conference Series: Materials Science and Engineering, vol. 135, IOP Publishing, 2016, p. 012040.

[29] J. Shao, S. V. Buldyrev, R. Cohen, et al., *Fractal boundaries of complex networks*, EPL (Europhysics Letters), 84 (2008), p. 48004.

[30] M. Sigman and G. A. Cecchi, *Global organization of the wordnet lexicon*, Proceedings of the National Academy of Sciences, 99 (2002), pp. 1742–1747.

[31] F. N. Silva, B. A. Travencolo, M. P. Viana, and L. d. F. Costa, *Identifying the borders of mathematical knowledge*, Journal of Physics A: Mathematical and Theoretical, 43 (2010), p. 325202.

[32] H. Small, *Co-citation in the scientific literature: A new measure of the relationship between two documents*, Journal of the American Society for information Science, 24 (1973), pp. 265–269.

[33] B. A. Travençolo, M. P. Viana, and L. da F. Costa, *Border detection in complex networks*, New Journal of Physics, 11 (2009), p. 063019.

[34] B. A. N. Travençolo and L. d. F. Costa, *Accessibility in complex networks*, Physics Letters A, 373 (2008), pp. 89–95.

[35] D. J. Watts and S. H. Strogatz, *Collective dynamics of 'small-world'networks*, nature, 393 (1998), pp. 440–442.

[36] Z. Wu, X. Lu, and Y. Deng, *Image edge detection based on local dimension: A complex networks approach*, Physica A: Statistical Mechanics and its Applications, 440 (2015), pp. 9–18.