

Algoritmos de Detecção de Comunidades: análises quantitativa e qualitativa

Vitor Alves Iglesias¹, Denis dos Alves², Milena Eugênio da Silva², Evandro Coggo Cristofolletti³, Thais Dibbern¹, Milena Pavan Serafim¹, Priscila C. Berbert Rampazzo¹

¹ Faculdade de Ciências Aplicadas, Universidade Estadual de Campinas.

² Departamento de Política Científica e Tecnológica, Universidade Estadual de Campinas.

³ Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas.

v184027@dac.unicamp.br, denisalvesvgp@gmail.com,
miledasilva1@gmail.com, evcoggo@unicamp.br, tdibbern@unicamp.br,
milenaps@unicamp.br, priscila.rampazzo@fca.unicamp.br

RESUMO

Este trabalho explora a aplicação de seis algoritmos de detecção de comunidades – Leiden, Louvain, Girvan-Newman, *Label Propagation*, *Walktrap* e Infomap – em três bases de dados extraídas da plataforma Overton.io. O objetivo principal é comparar o desempenho desses métodos por meio de análises quantitativas, através da métrica de modularidade, e qualitativas, considerando a coerência temática dos agrupamentos gerados. Os resultados demonstram variações entre os algoritmos, tanto nos valores de modularidade quanto na organização semântica das comunidades. Dentre os métodos avaliados, os algoritmos de Leiden e Louvain se destacaram por apresentar os maiores valores de modularidade para as três bases, além de produzir partições mais coerentes do ponto de vista temático.

PALAVRAS CHAVE. Agrupamento. Heurísticas. Detecção de Comunidades.

Tópicos: AS&DS

ABSTRACT

This work explores the application of six community detection algorithms – Leiden, Louvain, Girvan-Newman, *Label Propagation*, *Walktrap*, and Infomap – in three databases extracted from the Overton.io platform. The main objective is to compare the performance of these methods through quantitative analyses, using the modularity metric, and qualitative analyses, considering the thematic coherence of the generated clusters. The results demonstrate variations between the algorithms, both in modularity values and in the semantic organization of the communities. Among the evaluated methods, the Leiden and Louvain algorithms yielded the highest modularity values for the three databases and produced more coherent partitions from a thematic perspective.

KEYWORDS. Clustering. Heuristics. Community Detection.

Paper topics: AS&DS

1. Introdução

Os avanços computacionais e tecnológicos fizeram crescer a obtenção de informações. Isso resultou em um crescimento exponencial da quantidade de dados gerados e armazenados. Consequentemente, surge um grande desafio: como extrair e processar informações relevantes deste vasto conjunto de dados.

Em Aprendizado de Máquina, as técnicas de Agrupamento (*Clustering*) têm como objetivo particionar um conjunto de dados em subconjuntos, chamados de grupos ou *clusters*, de forma que os elementos dentro de cada subconjunto sejam mais semelhantes entre si do que em relação aos elementos de outros subconjuntos. Essa divisão é realizada a partir de critérios de agrupamento, que verificam quão similar ou dissimilar um dado é quando comparado a outro. Dependendo da escolha desses critérios, encontramos diferentes formas de agrupar o conjunto de dados. Em relação à similaridade, quanto maior o valor observado, mais parecido são os dados (exemplo: coeficiente de correlação). Em relação à dissimilaridade, quanto maior o valor observado, menos parecidos os dados são (exemplo: distâncias). Os métodos clássicos de agrupamento dividem em quatro grupos: agrupamento hierárquico, agrupamento baseado em centróide, agrupamento baseado em grafo e agrupamento baseado em densidade [Xu e Tian, 2015]. Cada algoritmo tem seu desempenho dependente no tipo de base de dados, tipicamente dependente da dimensão do espaço vetorial no qual os pontos estão distribuídos, número de amostras e formato dos grupos.

As técnicas de agrupamento de dados baseadas em rede [da Silva et al., 2019] têm se tornado cada vez mais populares. Redes são estruturas relacionais definidas por entidades que se integram e interagem dinamicamente entre si [Teixeira et al., 2008]. Nas redes, as entidades são representadas por vértices e seus relacionamentos por arestas. Em algoritmos de agrupamento baseados em rede, o primeiro passo é transformar o conjunto de dados em uma rede. Em seguida, um algoritmo de detecção de comunidade é usado para particionar os vértices da rede em comunidades. Finalmente, as comunidades são interpretadas como grupos do conjunto de dados original. Um dos benefícios da utilização de algoritmos de detecção de comunidade é que eles podem examinar a estrutura topológica que surge da transformação dos dados em uma rede, além dos atributos dos objetos no conjunto de dados [Querobim, 2021].

Compreender a estrutura de comunidades em uma rede fornece percepções valiosas sobre a organização e o funcionamento dos sistemas complexos. Essas comunidades podem representar grupos de indivíduos com interesses comuns em redes sociais, áreas de documentos em redes de coautoria, linhas de pesquisa em grupos de pesquisadores, etc. Identificar comunidades permite entender melhor a dinâmica e os padrões de interação entre os elementos da rede, facilitando a análise de sua estabilidade, propagação de informações e até mesmo a identificação de potenciais pontos de influência. Assim, a detecção de comunidades desempenha um papel crucial na compreensão dos sistemas complexos e na formulação de estratégias eficazes em uma variedade de domínios.

Este trabalho tem o intuito de explorar algoritmos de agrupamento e de detecção de comunidades em redes com enfoque multidisciplinar. A detecção de comunidades é de grande importância em sociologia, biologia e ciência da computação [Rostami et al., 2023], disciplinas onde as relações entre sistemas são frequentemente representadas através de grafos. Este problema é muito difícil e ainda não foi resolvido de forma satisfatória, apesar do enorme esforço de uma grande comunidade interdisciplinar de cientistas que trabalharam nele nos últimos anos [Fortunato, 2010]. O objetivo geral é realizar uma análise e comparações entre algoritmos de detecção de comunidades e aplicá-los em uma situação de interesse para realização de análises e obtenção de informações.

2. Algoritmos de Detecção de Comunidades

O estudo de diferentes algoritmos de detecção de comunidades é importante por várias razões [Newman e Girvan, 2004]. Cada rede pode ter características únicas, como tamanho, densidade, distribuição de conexões e padrões de comunidade; é importante identificar o algoritmo mais adequado para cada tipo de rede. Algoritmos de detecção de comunidades podem se basear em diferentes princípios e técnicas: otimização de modularidade, propriedades estruturais, métodos de aprendizado de máquina, entre outros. Explorar uma variedade de abordagens permite entender melhor as vantagens e limitações de cada uma e entender como interpretar as informações resultantes dos agrupamentos. Nesta seção apresentamos os principais algoritmos de detecção de comunidades da literatura.

2.1. Louvain

O trabalho de Blondel et al. [2008] introduziu o método de Louvain, um algoritmo Heurístico, com abordagem gulosa e baseada em otimização da modularidade para detecção de comunidades em grafos ponderados. O algoritmo de Louvain começa colocando cada vértice do grafo em uma comunidade distinta, uma por vértice. Depois, o algoritmo varre sequencialmente todos os vértices i calculando o ganho de modularidade ao colocar o vértice i na comunidade de qualquer vizinho j e escolhendo o vizinho j que produz o maior ganho de modularidade e unindo-o à comunidade correspondente. Este procedimento se repete até que não se identifique mais ganhos de modularidade, definindo as partições do primeiro nível (Etapa 1). Essas partições tornam-se "supervértices" e o algoritmo reconstrói o grafo calculando o peso das arestas entre todos os "supervértices". Dois "supervértices" são conectados se houver pelo menos uma aresta entre os vértices das partições correspondentes. Neste caso, o peso da aresta entre os dois "supervértices" é igual à soma dos pesos de todas as arestas entre suas partições correspondentes na partição inicial (Etapa 2). Essas duas etapas se repetem até que as comunidades não se alterem [Que et al., 2015].

2.2. Girvan-Newman ou *Edge Betweenness*

O método é um algoritmo exato baseado em centralidade de arestas e com alto custo computacional. O algoritmo remove iterativamente as arestas de maior centralidade de intermediação (*betweenness*) para revelar comunidades, partindo do conceito de que a centralidade de intermediação de um vértice i é definida como o número de caminhos mínimos entre pares de outros vértices que passam por i [Girvan e Newman, 2002].

Seu funcionamento segue quatro passos: calcular a intermediação de todas as arestas na rede (Etapa 1); remover a aresta de maior intermediação (Etapa 2); recalcular as intermediações para todas as arestas afetadas pela remoção (Etapa 3); repetir passo 2 até não restarem arestas (Etapa 4) [Girvan e Newman, 2002]. A cada remoção, os valores são atualizados para que novas "pontes" entre comunidades sejam expostas, o que possibilita a identificação de agrupamentos com maior coesão interna.

2.3. Infomap

É um método heurístico baseado em Teoria da Informação, proposto por Rosvall e Bergstrom [2008], que detecta comunidades ao minimizar a descrição de passeios aleatórios em redes. O algoritmo utiliza o fluxo de probabilidade desses passeios como um substituto para os fluxos de informação no sistema real, decompondo a rede em módulos por meio da compressão da descrição do fluxo, o que resulta em um mapa que simplifica e evidencia regularidades estruturais e relacionais [Rosvall e Bergstrom, 2008].

2.4. Label Propagation

Label Propagation é um método heurístico baseado em propagação de rótulos. Cada vértice começa com um rótulo único e, a cada iteração, cada nó adota o rótulo que a maioria de seus vizinhos possui, com desempates resolvidos aleatoriamente [Raghavan et al., 2007]. À medida que o processo se repete, os rótulos se espalham pela rede e grupos densamente conectados rapidamente formam consenso interno sobre um único rótulo, o que define as comunidades. O algoritmo é iterado até que todos os nós tenham um rótulo que seja o mais comum entre seus vizinhos, sem necessidade de otimizar funções nem de conhecer previamente o número de comunidades.

2.5. Walktrap

Método heurístico baseado em passeios aleatórios. Usa distâncias entre vértices, medidas por passeios aleatórios, para identificar comunidades. A intuição central é que passeios aleatórios em um grafo tendem a ficar "presos" em partes densamente conectadas correspondentes às comunidades [Pons e Latapy, 2005].

O algoritmo mede a similaridade entre vértices comparando a distribuição de probabilidade de destinos após um número fixo de passos. Isso captura informação sobre a estrutura de comunidades e pode ser usado em um algoritmo aglomerativo hierárquico eficiente [Pons e Latapy, 2005]. O algoritmo inicia cada vértice em sua própria comunidade e, em cada etapa, agrupa os pares de comunidades adjacentes cuja união causa o menor aumento na variância interna.

2.6. Leiden

O algoritmo de Leiden é uma técnica relativamente recente e eficaz para detecção de comunidades em redes. Ele é uma variação do algoritmo Louvain e foi projetado para melhorar sua escalabilidade e precisão [Traag et al., 2019]. O algoritmo de Leiden aplica uma técnica de otimização visando maximizar a modularidade da rede e agrupar seus nós. O algoritmo de Leiden tornou-se popular devido à sua eficácia em detectar comunidades em redes grandes e complexas, com milhões de nós e arestas. Trata-se de um algoritmo Heurístico, baseado em modularidade e refinamento estrutural. O pseudocódigo do Algoritmo de Leiden [Nguyen, 2021] é apresentado a seguir (Algoritmo 1).

3. Métricas de Avaliação

A métrica de modularidade proposta por Newman e Girvan [2004] é uma das funções mais utilizadas e conhecidas para quantificar a estrutura de comunidades em um grafo [Que et al., 2015]. A modularidade é uma medida que varia entre -1 e 1. Valores mais altos indicam comunidades mais bem definidas.

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

onde: A_{ij} é a matriz de adjacência; k_i e k_j são os graus dos nós; m é o número total de arestas; $\delta(c_i, c_j)$ é 1, se os nós estão na mesma comunidade; 0, caso contrário. A modularidade é uma métrica que mede a qualidade da partição das comunidades comparando a densidade de arestas dentro das comunidades *versus* entre elas.

A condutância [Hasan et al., 2024] é uma métrica que indica o quão bem uma comunidade está conectada internamente em comparação às suas conexões externas; ou seja, a medida representa a proporção entre as arestas que saem da comunidade e as arestas internas, levando em conta também o tamanho da comunidade. Dado um grafo $G = (V, E)$, considere uma partição de seus vértices em dois subconjuntos disjuntos S e $T = V \setminus S$. A condutância de S é definida como:

$$\phi(S) = \frac{\text{cut}(S, T)}{\min(\text{vol}(S), \text{vol}(T))}$$

Algorithm 1 Função LEIDEN(G)

```

1: Inicialize cada nó em sua própria comunidade
2: repeat
3:   1. Otimização Local:
4:   repeat
5:     for all nó  $v$  em ordem aleatória do
6:       Encontre a comunidade vizinha de  $v$  com o maior ganho na função de modularidade
7:       Mova  $v$  para essa comunidade se o ganho for positivo
8:     end for
9:   until Nenhuma melhoria é possível
10:  2. Refinamento de Comunidades:
11:  for all comunidade  $C$  na partição atual do
12:    Identifique os componentes conexos de  $C$ :  $\{C_1, C_2, \dots, C_k\}$ 
13:    for all componente  $C_i$  em  $\{C_1, C_2, \dots, C_k\}$  do
14:      Atribua  $C_i$  como uma nova comunidade
15:    end for
16:  end for
17:  3. Agregação: Construa o grafo agregado  $G'$ 
18:  Cada nó em  $G'$  representa uma comunidade distinta de  $G$ 
19:  for all pares de comunidades  $(C_i, C_j)$  do
20:    Crie uma aresta entre os nós  $i$  e  $j$  em  $G'$  com peso igual à soma dos pesos
21:    das arestas entre membros de  $C_i$  e  $C_j$  em  $G$ 
22:  end for
23:  Substitua  $G$  por  $G'$ 
24: until Critério de convergência: número de comunidades não se alterar
25: return Partição final das comunidades:  $G = 0$ 
  
```

onde: $\text{cut}(S, T)$ é o número (ou peso) das arestas que conectam vértices em S a vértices em T ; $\text{vol}(S)$ é a soma dos graus dos vértices em S (ou seja, o volume de S).

Uma condutância baixa indica que a comunidade está bem delimitada, pois poucas arestas ligam seus vértices ao restante do grafo. Uma condutância alta sugere que a comunidade está fortemente conectada a outras partes da rede, indicando fraca separação. Assim, em problemas de detecção de comunidades, busca-se geralmente minimizar a condutância, pois valores baixos representam partições mais coesas e bem definidas.

4. Aplicações

Overton.io¹ é uma plataforma que ajuda a compreender como a pesquisa influencia as políticas. É a maior base de dados de políticas públicas e literatura cinzenta do mundo. Escolhemos como aplicação uma base de dados extraída do Overton.io com 4.662 publicações, denominadas de documentos políticos, que referenciam pesquisas ou pesquisadores da Universidade Estadual de Campinas (UNICAMP). O período da extração foi de 01 de janeiro de 1983 a 07 de julho de 2022. Considerando o conjunto de 4.662 documentos, identificamos 387 organizações distintas (fontes dos documentos políticos). Este projeto faz parte de uma pesquisa mais ampla que avalia o impacto das pesquisas universitárias em documentos políticos.

Cada documento político tem um conjunto com três palavras-chave associadas. Palavras-chave são formas de resumir os principais temas abordados por um texto. Através das palavras-chave podemos identificar ideias sob as quais as pesquisas se orientam. Escolhemos a estrutura de grafos para a análise das palavras-chave de cada conjunto de documentos políticos. Criamos uma lista de palavras-chave e avaliamos a quantidade de vezes que cada palavra apareceu em algum documento político; chamamos esta lista de *vertices*. Criamos uma lista denominada *edges*: cada

¹<https://www.overton.io/>

elemento desta segunda lista é formado por uma ligação e contabiliza a quantidade de vezes que duas palavras aparecem juntas em um mesmo documento.

Criamos uma estrutura de grafo, onde cada vértice tem peso igual ao número de ocorrências da respectiva palavra-chave no conjunto de documentos. O peso de uma aresta está associada à frequência com a qual as duas palavras-chave associadas são citadas juntas considerando todos os documentos de um conjunto.

Para a aplicação dos métodos descritos na Seção 2, selecionamos:

- 161 documentos publicados pelo Serviço de Publicações da União Europeia (POEU).
- 486 documentos publicados pela Organização Mundial da Saúde (OMS).
- 329 documentos publicados pelo Instituto de Pesquisa Econômica Aplicada (IPEA).

As 30 palavras-chave mais importantes de cada organização foram consideradas como representativas para o contexto das aplicações.

5. Resultados

As estruturas foram implementadas em linguagem de programação Python através do pacote *igraph*. Utilizamos a função *DictList* para construção do grafo com este pacote. A escolha da implementação em linguagem de programação Python e da utilização destes pacotes se justifica pela flexibilidade de adaptação e escalabilidade.

Analisando os resultados apresentados na Tabela 1, observamos que os algoritmos de Leiden e Louvain têm melhor modularidade com as bases OMS e POEU, sugerindo boa qualidade da partição nessas redes. Valores de modularidade menor indicam comunidades pouco definidas, resultado observado especialmente na base do IPEA. Ainda para a rede IPEA, Infomap e *Label Propagation* não conseguem detectar comunidades estruturadas (modularidade = 0 representada na Figura 4). Em relação à condutância (Tabela 2), quanto menor o valor, melhor é a separação das comunidades. O algoritmo Infomap apresentou excelente condutância para a base da OMS, com comunidades bem isoladas. Não foi possível calcular condutância para os algoritmos Infomap e *Label Propagation*, que apresentaram todas as palavras-chave em uma única comunidade.

Algoritmo	OMS	POEU	IPEA
Leiden	0,4336	0,4640	0,1954
Louvain	0,4336	0,4640	0,1954
Girvan-Newman	0,3956	0,4520	0,0500
Infomap	0,4256	0,4640	0,0000
Label Propagation	0,3006	0,4036	0,0000
Walktrap	0,3181	0,4226	0,1907

Tabela 1: Resultados - Modularidade, por Algoritmo e Base de Dados

Algoritmos heurísticos buscam boas soluções, mas não necessariamente a única ou a ótima solução global. Esse comportamento pode ser explicado pela ordem aleatória da etapa Otimização Local do Algoritmo 1, por exemplo, refletindo na trajetória da heurística, que pode agrupar ou separar grupos com conexões intermediárias moderadas. Exceto o algoritmo de Girvan-Newman, todos os demais algoritmos apresentados neste trabalho são heurísticos e podem apresentar soluções alternativas em diferentes rodadas.

A Figura 1 apresenta as comunidades detectadas para a base OMS com os algoritmos de Leiden e Louvain. Neste caso, as comunidades dividem-se em: [Grupo 1] health, medicine, health

Algoritmo	OMS	POEU	IPEA
Leiden	0,3136	0,2947	0,5256
Louvain	0,3136	0,2947	0,5424
Girvan-Newman	0,6485	0,3014	0,7956
Infomap	0,1483	0,2947	-
Label Propagation	0,2776	0,4483	-
Walktrap	0,2685	0,3349	0,4786

Tabela 2: Resultados - Condutância, por Algoritmo e Base de Dados

sciences, clinical medicine, medical specialties, health care, public health, human activities, mental disorder, dengue fever, diseases and disorders; [Grupo 2] childbirth, birth control, hiv/aids, maternal death, world health organization, family planning, breastfeeding, sexually transmitted infection, abortion, caesarean section; [Grupo 3] coronavirus disease 2019, cervical cancer, human papillomavirus infection, vaccine, vaccination, covid-19 pandemic, pfizer–biontech covid-19 vaccine; [Grupo 4] randomized controlled trial, suicide. O Grupo 1 representa pesquisas amplas relacionadas à saúde, medicina clínica, saúde pública, epidemiologia de doenças e transtornos mentais, tendo como base conceitual geral as ciências da saúde e cuidados médicos. O Grupo 2 tem foco em saúde da mulher, gestação, parto, planejamento familiar, infecções sexualmente transmissíveis e mortalidade materna; representa um subcampo especializado em saúde reprodutiva e materno-infantil, articulado com políticas públicas e recomendações de órgãos como a OMS. O Grupo 3 representa pesquisas em doenças infecciosas virais (COVID-19, HPV), vacinas e imunização, com enfoque em saúde pública e prevenção de doenças, com foco específico de epidemiologia e imunologia e influenciando práticas clínicas e políticas de saúde. O Grupo 4 apresenta pesquisas que utilizam ensaios clínicos randomizados para avaliar tratamentos ou intervenções, incluindo estudos sobre comportamentos de risco à saúde (suicídio) através de métodos de pesquisa clínica, fornecendo a base científica para validação de práticas médicas aplicadas nos 3 grupos anteriores. A estrutura conceitual dos grupos pode ser organizada como: (i) nível macro: Grupo 1 (abrange todo o sistema de saúde); (ii) áreas especializadas: Grupo 2 (reprodutiva/materno-infantil) e Grupo 3 (doenças virais e vacinação); (iii) base metodológica: Grupo 4 (ensaios clínicos, pesquisa aplicada). Portanto, os algoritmos de Leiden e Louvain produziram grupos bem definidos, com temas mais específicos e coesos. O algoritmo Infomap apresentou melhor condutância (Tabela 2) incorporando o Grupo 4 ao Grupo 1.

A Figura 2 apresenta as comunidades detectadas para a base POEU com o algoritmo de Leiden, Louvain e Infomap. Neste caso, as comunidades dividem-se em: [Grupo 1] chemistry, nature, energy, chemical substances, physical sciences, biology, materials, artificial objects, applied and interdisciplinary physics, manufacturing; [Grupo 2] human activities, economy, technology, branches of science, computing, sustainability, cognition, greenhouse gas, waste; [Grupo 3] medical specialties, health, health care, medicine, health sciences, public health, clinical medicine; [Grupo 4] natural environment, earth sciences, physical geography, habitat. O Grupo 1 é o mais abrangente em termos de ciências exatas e da natureza. O núcleo é formado por química, física e biologia. A inclusão de termos como energia, materiais, objetos artificiais e manufatura indica que a comunidade representa não apenas a ciência pura, mas também suas aplicações em tecnologia e indústria. O Grupo 2 se concentra na interação entre a sociedade, a economia e o meio ambiente. Tópicos como economia, tecnologia e computação representam as atividades produtivas e a base técnica da sociedade. A presença de sustentabilidade, gás de efeito estufa e resíduos aponta para uma discussão focada no impacto ambiental dessas atividades e nas soluções para um desenvol-



Figura 1: Base OMS, Algoritmo de Leiden.

vimento mais consciente. A inclusão de cognição pode sugerir um olhar sobre a relação entre o pensamento humano e as questões sociais e tecnológicas. O Grupo 3 é o mais coeso e especializado, com foco claro na área da saúde. Os tópicos centrais são saúde, medicina, ciências da saúde e cuidados de saúde. A presença de especialidades médicas, medicina clínica e saúde pública indica que esta comunidade abrange tanto a prática médica individual quanto as estratégias de saúde populacional. O Grupo 4 é o mais voltado para o estudo do meio ambiente, ecossistemas e no estudo do nosso planeta e seus sistemas naturais. A estrutura conceitual dos grupos pode ser organizada como: (i) ciências de base: Grupo 1 e Grupo 4; (ii) ciências aplicadas e sociais.

As Figuras 3 e 4 apresentam as comunidades detectadas para a base IPEA com os algoritmos de Leiden/Louvain e com o algoritmo Infomap. Os resultados com os algoritmos de Leiden e Louvain foram: [Grupo 1] brasil, industry, education, constitution of brazil, sistema único de saúde, china, higher education; [Grupo 2] economics, innovation, tax, economic growth, credit, inflation, tertiary sector of the economy, investment; [Grupo 3] economic development, state (polity), sociology, poverty, municipality, são paulo (state), bolsa família, democracy, economic inequality; [Grupo 4] economy, human activities, politics, branches of science, gender, research. O Grupo 1 tem a presença de Brasil, Constituição do Brasil e Sistema Único de Saúde (SUS) estabelece um contexto nacional e institucional. Indústria e educação (incluindo ensino superior) indicam um foco em setores estratégicos para o desenvolvimento do país, enquanto China sugere uma relação de política externa ou cooperação econômica relevante. Esta comunidade representa discussões sobre o

Estado, políticas públicas e setores-chave para a nação. O Grupo 2 é altamente especializado e se concentra em conceitos e indicadores econômicos e financeiros. O núcleo do grupo é formado por economia e termos como crescimento econômico, inflação, crédito e investimento, que são fundamentais para a análise macroeconômica. A inclusão de imposto e inovação sugere uma discussão sobre políticas fiscais e fatores de crescimento. Esta comunidade aborda os motores e as medições do desempenho econômico. O Grupo 3 foca em questões sociais, políticas e de desenvolvimento regional. Tópicos como desenvolvimento econômico, pobreza, desigualdade econômica e Bolsa Família apontam para uma discussão sobre o bem-estar social e programas de combate à pobreza. A presença de democracia e Estado insere o debate em um contexto político e de governança. O Grupo 4 representa a comunidade mais genérica e conceitual. Os tópicos economia, atividades humanas, política e ramos da ciência sugerem que este grupo discute a interconexão entre as grandes áreas do conhecimento. A comunidade também aborda pesquisas e a perspectiva de gênero.

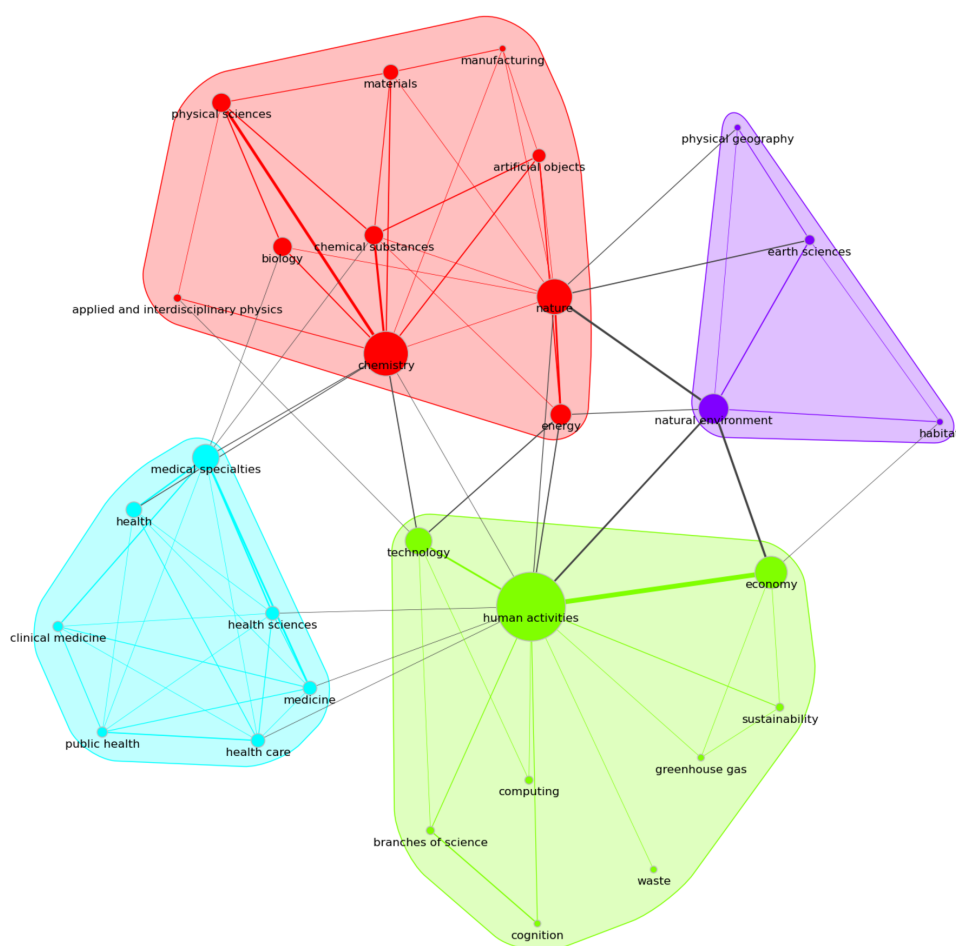


Figura 2: Base POEU, Algoritmo de Leiden.

6. Conclusão

De forma geral, os algoritmos de Leiden e Louvain apresentam o melhor desempenho. Os algoritmos Infomap, *Label Propagation* e Girvan-Newman são sensíveis à estrutura da rede. As análises qualitativas dos resultados obtidos pelos algoritmos de detecção de comunidades evidenciam que, além das métricas quantitativas (como a modularidade e condutância), a interpretação

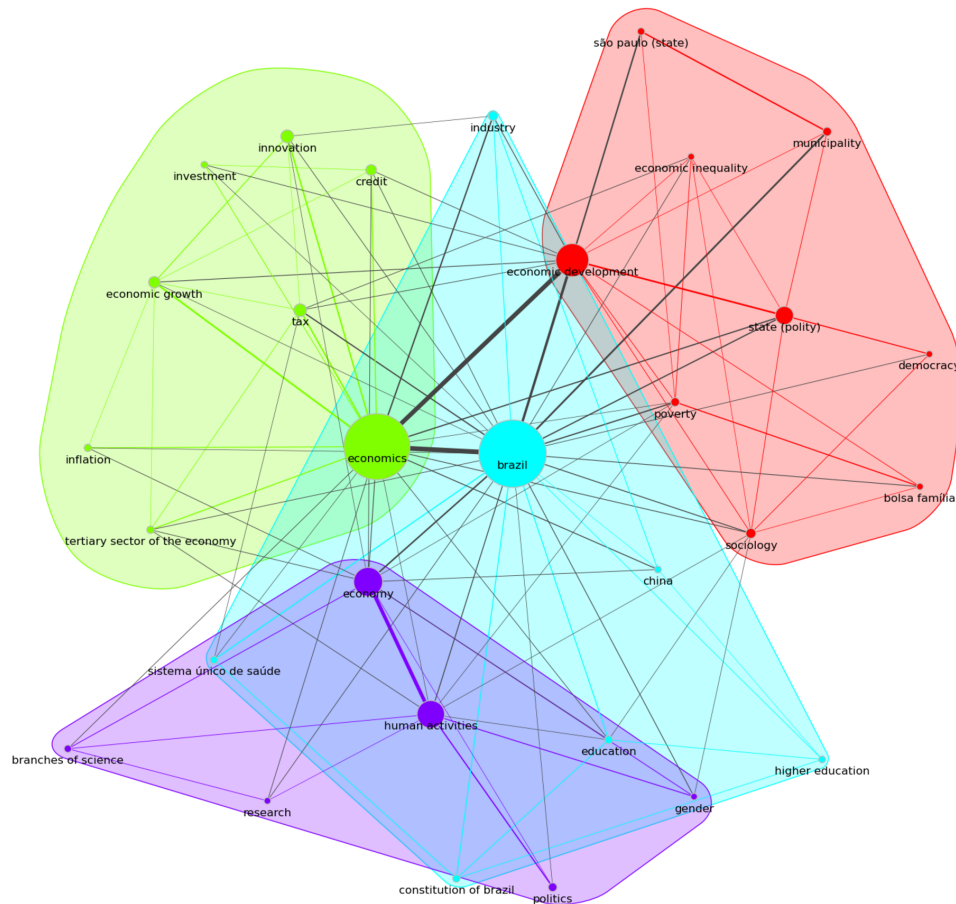


Figura 3: Base IPEA, Algoritmo de Leiden.

temática dos grupos gerados desempenha um papel fundamental na avaliação da qualidade das comunidades.

É importante destacar que esses métodos são heurísticos, ou seja, baseiam-se em estratégias aproximadas para encontrar soluções na tentativa de maximizar a conexão interna dos grupos. Como consequência, execuções distintas podem produzir partições diferentes, mesmo com os mesmos dados de entrada. Para as bases de dados apresentadas, isso foi particularmente observado com os algoritmos: Leiden, Louvain, *Label Propagation*. Isso reforça a importância de realizar análises complementares, tanto quantitativas quanto qualitativas.

Portanto, a escolha do algoritmo e a interpretação dos resultados devem considerar o contexto do problema, os objetivos da análise e o nível de detalhe desejado. A análise qualitativa, nesse sentido, não apenas complementa a análise quantitativa – aqui apresentada através das métricas de modularidade e condutância –, como permite identificar padrões temáticos e relações semânticas relevantes que podem ser decisivos para a aplicação prática da detecção de comunidades.

Referências

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., e Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10): P10008.

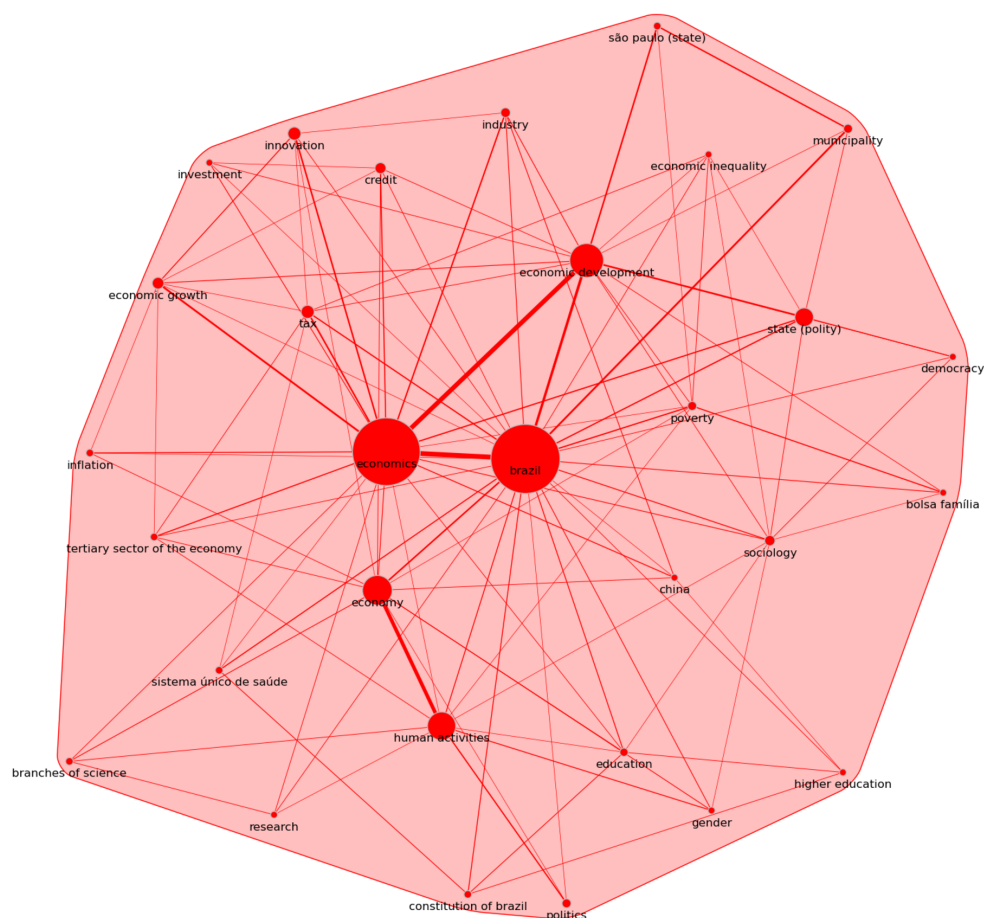


Figura 4: Base IPEA, Infomap.

da Silva, D. M., de Moura Brito, J. A., e Oliveira, C. S. (2019). Um estudo computacional comparativo entre algoritmos de agrupamento e de detecção de comunidades. *Simpósio de Pesquisa Operacional e Logística da Marinha*.

Fortunato, S. (2010). Community detection in graphs. *Physics reports*, 486(3-5):75–174.

Girvan, M. e Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826.

Hasan, M. A. U., Bakar, A. A., e Yaakub, M. R. (2024). Detecting community through user similarity analysis on twitter. In *2024 18th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, p. 1–6.

Newman, M. E. e Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113.

Nguyen, F. (2021). *Leiden-based parallel community detection*. PhD thesis, Karlsruher Institut für Technologie.

Pons, P. e Latapy, M. (2005). Computing communities in large networks using random walks.

- In *Computer and Information Sciences-ISCIS 2005: 20th International Symposium, Istanbul, Turkey, October 26-28, 2005. Proceedings 20*, p. 284–293. Springer.
- Que, X., Checconi, F., Petrini, F., e Gunnels, J. A. (2015). Scalable community detection with the louvain algorithm. In *2015 IEEE international parallel and distributed processing symposium*, p. 28–37. IEEE.
- Querobim, J. N. C. (2021). *Análise comparativa entre algoritmos de agrupamento e de detecção de comunidades em redes*. PhD thesis, Universidade Federal de São Carlos.
- Raghavan, U. N., Albert, R., e Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 76(3):036106.
- Rostami, M., Oussalah, M., Berahmand, K., e Farrahi, V. (2023). Community detection algorithms in healthcare applications: A systematic review. *IEEE Access*, 11:30247–30272.
- Rosvall, M. e Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the national academy of sciences*, 105(4):1118–1123.
- Teixeira, L., Lima, L., e Abreu, N. (2008). Grafos que modelam redes confiáveis. *Mestrado em Engenharia de Produção, COPPE, Universidade do Rio de Janeiro, Rio de Janeiro*.
- Traag, V. A., Waltman, L., e Van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):5233.
- Xu, D. e Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193.