



Detecção de comunidades sobrepostas em Redes Sociais Científicas

Vitor Araújo Cautiero Horta

JUIZ DE FORA
NOVEMBRO, 2017

Detecção de comunidades sobrepostas em Redes Sociais Científicas

VITOR ARAÚJO CAUTIERO HORTA

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciências da Computação

Bacharelado em Ciência da Computação

Orientador: Victor Ströele De Andrade Menezes

JUIZ DE FORA

NOVEMBRO, 2017

DETECÇÃO DE COMUNIDADES SOBREPOSTAS EM REDES SOCIAIS CIENTÍFICAS

Vitor Araújo Cautiero Horta

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS
EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTE-
GRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE
BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Victor Ströele De Andrade Menezes
Doutor

Heder Soares Bernardino
Doutor

José Maria Nazar David
Doutor

JUIZ DE FORA
22 DE NOVEMBRO, 2017

Agradeço primeiramente aos meus pais Pedro Paulo e Márcia, e meus irmãos André, Bruno e Pedro, pela confiança, incentivo, e por fornecerem todas as condições que foram necessárias para esta realização. Aos meus amigos do clube trezentos por estarem do meu lado desde o início desta jornada.

A Léia, minha namorada, pelo apoio incondicional em todos os momentos.

As empresas em que trabalhei e que sempre apoiaram os meus esforços neste curso.

Por fim, não posso deixar de agradecer a todos os professores e colaboradores da UFJF que fizeram parte desta conquista. Os ensinamentos e o apoio de cada um foram essenciais.

Resumo

Comunidades em redes sociais são compostas por pessoas com interesses comuns, que influenciam ou são influenciadas por elas mesmas. Neste trabalho são aplicados conceitos de análise de redes complexas para verificar o nível de influência entre os pesquisadores, analisando a estrutura da rede social científica e suas comunidades. São propostos um modelo baseado em grafo bidirecional para analisar a influência entre os pesquisadores e algoritmos para analisar a estrutura da rede, identificar comunidades científicas e localizar pesquisadores multidisciplinares. Para avaliação do modelo e dos algoritmos é utilizada uma base de dados científicos de grande porte em um caso de uso. Os resultados apontam para a viabilidade e eficácia da solução.

Conteúdo

1	Introdução	4
2	Agrupamento em Redes Sociais	7
3	Rede Social Científica	10
3.1	Modelagem da Rede	10
3.2	Análise Topológica	11
3.2.1	Distribuição de Grau	12
3.2.2	Distribuição de Influência	13
4	NetSCAN: algoritmo de agrupamento	16
5	Experimentação	19
5.1	Experimentos em R^2	19
5.2	Experimentos em Rede Social Sintética	21
5.3	Experimentos em Redes Sociais Reais	22
6	Rede científica real	24
6.1	Análise dos Resultados	27
6.2	Análise das sobreposições	28
6.3	Comunidades sobrepostas ao longo do tempo	33
6.3.1	Mudança de área de atuação	33
6.3.2	Mudança de grupo de pesquisa na mesma área de atuação	35
6.3.3	Atuação simultânea em múltiplos grupos de pesquisa	36
7	Considerações Finais	39
	Bibliografia	41

1 Introdução

Uma rede complexa representa um conjunto de objetos que se conectam de maneira não-trivial. Para se entender como os objetos se relacionam e se agrupam é necessária uma análise cuidadosa sobre a estrutura e características dessas redes (WASSERMAN; FAUST, 1994). Os objetos, vértices ou nós, podem ser utilizados para representar diversos cenários da vida real como, por exemplo, pessoas através de uma relação social, cadeias alimentares, interações moleculares, dentre outros. Dessa forma, ao conhecer a estrutura da rede em que estes elementos estão inseridos é possível identificar quais são as principais características que influenciam na formação de suas interações.

Como as redes complexas de domínios distintos muitas vezes compartilham de comportamentos semelhantes (BARABÁSI; BONABEAU, 2003), ao classificar as redes baseando-se apenas em sua topologia pode-se antecipar futuras ligações entre os objetos ou até mesmo influenciar e motivar a formação de novas interações. Estes motivos e a grande interdisciplinaridade do tema têm motivado cada vez mais os estudos e desenvolvimento de algoritmos e técnicas para analisar a topologia das redes, definir agrupamentos para identificar comunidades, localizar elementos influenciadores, conectores e difusores de informações, dentre outros.

Segundo Cross e Parker (2004) existem quatro tipos de pessoas em uma rede social: *central connectors*, que possuem um elevado número de relacionamentos; *boundary spanners*, que conectam diferentes grupos de pessoas; *information brokers*, que são pessoas influentes que comunicam com subgrupos mantendo um grande grupo conectado, ou conectando dois grupos; e *peripheral people*, que estão nas extremidades da rede e precisam de ajuda para aprimorar suas conexões. É importante analisar todos estes tipos de pessoas para caracterizar a rede social.

Com relação a pessoas importantes em redes sociais, os estudos focam nas suas conexões diretas (CROSS; PARKER, 2004) (GUILLE, 2013), com o objetivo de identificar *central connectors*. No entanto, os relacionamentos indiretos também são importantes e devem também ser considerados em análises de redes sociais (GRABOWICZ et al.,

2012)(YANG; LESKOVEC, 2012). Deste modo, *information brokers* podem influenciar pessoas mesmo quando não existe conexão direta entre eles. Isto ajuda a disseminar certos tipos de informação e a conectividade ao longo da rede (CROSS; PARKER, 2004).

Redes sociais científicas são caracterizadas como tipos específicos de redes complexas que representam as relações sociais estabelecidas pelos pesquisadores. Neste contexto, caracterizar os pesquisadores pode ajudar a entender como eles colaboram entre si e como as comunidades de pesquisa se relacionam. O problema é como identificar pesquisadores que, de alguma forma, ajudam a manter a conectividade da rede, disseminando informação e conectando grupos e subgrupos de pesquisa. Dessa forma, o objetivo deste trabalho é identificar *central connectors* e *research-information brokers* visando identificar pesquisadores influentes em uma rede social científica. Os conceitos e técnicas de redes complexas serão utilizados para analisar a interação entre pesquisadores em uma rede social científica, identificando (i) pesquisadores influentes que participam de duas ou mais comunidades simultaneamente, i.e, *research-information brokers* conectando dois ou mais grupos; (ii) e potenciais influenciadores em uma comunidade científica específica, i.e, *central connectors* de um grupo que conecta subgrupos.

Para isso, foi utilizada neste trabalho a base de dados DBLP (LEY, 2002), por ser uma das bases de dados mais utilizadas no estudo de redes sociais científicas de grande porte e por conter contribuições científicas que ultrapassam um milhão de pesquisadores e mais de nove milhões de interações entre eles (LEY, 2002). A rede social científica é representada por um grafo bidirecional cujos vértices representam os pesquisadores e as arestas correspondem às relações científicas entre eles. Um grafo direcionado é utilizado para que seja possível definir o grau de influência entre os pesquisadores, permitindo a identificação de pesquisadores influenciadores e influenciados em suas áreas de pesquisa.

Além da modelagem de uma rede social científica bidirecional que permite analisar os níveis de influência entre os pesquisadores, outra contribuição deste trabalho é a proposta e desenvolvimento de um algoritmo de agrupamento que considera as características de pesquisas multidisciplinares de alguns pesquisadores, permitindo que eles pertençam a mais de um grupo. Com essa abordagem é possível identificar pessoas que possuem mais de uma área de atuação, e que participam de duas ou mais comunidades

científicas. Outra característica importante do algoritmo proposto é o sub-agrupamento, cujo objetivo é identificar subgrupos de comunidades de pesquisa.

Assim, como contribuições deste trabalho, destacam-se: (i) a aplicação de técnicas de agrupamento em uma base de dados de grande volume; (ii) a definição de um modelo para análise da influência entre pesquisadores baseado em grafo bidirecional; (iii) detecção de comunidades científicas e seus subgrupos de pesquisa; (iv) e identificação de pesquisadores multidisciplinares e seus diferentes níveis de influência.

Este trabalho avança as pesquisas do NEnC (Grupo de Pesquisa em Engenharia de Conhecimento) em análise de redes sociais científicas (STRÖELE; ZIMBRÃO; SOUZA, 2013) (STRÖELE et al., 2016)(STRÖELE et al., 2015)(ALMEIDA et al., 2016).

Além da Introdução, este trabalho está organizado da seguinte forma: o Capítulo 2 apresenta os trabalhos relacionados a agrupamento em redes sociais, o Capítulo 3 descreve o modelo proposto para a realização deste estudo e analisa a topologia deste modelo, no Capítulo 4 é descrito o algoritmo de agrupamento desenvolvido, no Capítulo 5 são feitos experimentos para avaliar o desempenho do algoritmo, um estudo em uma rede real de grande porte é realizado no Capítulo 6 e, finalmente, no Capítulo 7 são apresentadas as considerações finais.

2 Agrupamento em Redes Sociais

Diversos algoritmos para detecção de comunidades em redes sociais vêm sendo elaborados com o objetivo de identificar grupos nos quais seus membros possuem maior similaridade entre si e maior dissimilaridade com membros de outros grupos (HAN; KAMBER; PEI, 2012).

Para encontrar métodos de detecção de comunidades sobrepostas em redes sociais mais utilizados foi conduzido um estudo terciário. O objetivo de um estudo terciário é encontrar estudos secundários que, por sua vez, ajudam a levantar o estado da arte de uma determinada área.

Através da busca por artigos que levantam métodos de detecção de comunidades sobrepostas, foram obtidos cinco resultados, sendo eles (FORTUNATO, 2010)(Puig-Centelles, Anna and Ripolles, Oscar and Chover, Miguel, 2008)(WANG WENZHONG TANG; WANG, 2015)(XIE; KELLEY; SZYMANSKI, 2011)(PAPADOPOULOS et al., 2012). Nestes artigos são apresentados métodos com diferentes abordagens. Alguns exemplos são métodos de particionamento (HLAOUI; WANG, 2004), hierárquicos (NEWMAN; GIRVAN, 2004) (LUO et al., 2011) e baseados em densidade (FALKOWSKI; BARTH; SPILIOPOULOU, 2007)(BHAT; ABULAISH, 2012).

Os métodos de particionamento como (HLAOUI; WANG, 2004) tem como objetivo separar os dados em k grupos minimizando alguma função de similaridade entre os pontos e os centros destes grupos. Apesar da popularidade destes métodos, as suas principais limitações são o alto custo computacional e a necessidade de determinação prévia do número k de grupos.

Newman em (NEWMAN; GIRVAN, 2004) e Luo em (LUO et al., 2011) utilizam métodos hierárquicos divisivos para detecção de comunidades. Tais métodos consistem essencialmente em retirar as arestas com maior índice de centralidade (*edge betweenness*), e identificam comunidades em um dendograma. Essa abordagem é frequentemente utilizada, porém a alta complexidade computacional do método não favorece o seu uso em redes sociais de grande porte.

O DENGGRAPH (FALKOWSKI; BARTH; SPILIOPOULOU, 2007) foi desenvolvido utilizando a estratégia de agrupamento por densidade, através de uma implementação do DBSCAN (ESTER et al., 1996) adaptada para grafos não-direcionados. Como resultado, este algoritmo alcança um melhor desempenho em performance e maior capacidade de detectar ruídos em redes com maior volume de dados.

Uma abordagem para previsão de casos de dengue foi desenvolvida também utilizando o algoritmo DBSCAN com intuito de estimar a incidência de casos de dengue em cidades brasileiras a partir das mensagens relacionadas ao assunto *dengue* coletadas em redes sociais (OLIVEIRA et al., 2015).

Como o DBSCAN em sua implementação original não prevê que o conjunto de dados pode possuir múltiplas granularidades na densidade, (GIALAMPOUKIDIS et al., 2016) desenvolveram o DBSCAN*-Martingale. Este algoritmo adapta os parâmetros de densidade de forma iterativa com a finalidade de descobrir grupos com diferentes níveis de similaridade de seus respectivos membros. Entretanto, um processo iterativo se faz muito custoso quando aplicado em um conjunto com grande volume de dados. (LI et al., 2016) observaram a necessidade de se considerar a direção nas relações dos membros em redes sociais e propuseram um método que utiliza a estratégia de agrupamento por densidade para detecção de comunidades em grafos direcionados. Entretanto, não consideram a possibilidade de um elemento pertencer a dois ou mais grupos distintos.

O *clique percolation method* (DERÉNYI; PALLA; VICSEK, 2005) (CPM) foi um dos primeiros métodos de detecção de comunidades com sobreposições. A estratégia do CPM é detectar comunidades através da busca por k-cliques adjacentes no grafo. Um k-clique é um subgrafo totalmente conectado que contém k vértices. Dois k-cliques são considerados adjacentes se eles compartilham k-1 nós. Dessa forma as cliques adjacentes representam comunidades e os nós que pertencem a múltiplos cliques adjacentes representam as sobreposições entre estas comunidades.

O Cminer (BHAT; ABULAISH, 2012) é o algoritmo baseado em densidade que mais se assemelha com o método proposto neste trabalho. O algoritmo proposto por (BHAT; ABULAISH, 2012) é baseado no DBSCAN e suporta grafos direcionados e ponderados. No entanto, o Cminer utiliza as arestas direcionadas para calcular arestas não-

direcionadas entre dois vértices. Neste caso, a ligação entre dois vértices A e B passa a ser não-direcionada e $dist(a,b) = dist(b,a)$. Esta é a maior diferença entre este trabalho e o Cminer, já que o algoritmo proposto neste trabalho trata cada aresta direcionada separadamente.

Como diferencial deste trabalho pode-se destacar o desenvolvimento de um algoritmo baseado em densidade para detecção de comunidades em redes bidirecionais de grande porte. Também será proposta uma estratégia de subagrupamento com o objetivo de aumentar a coesão dos grupos previamente encontrados, principalmente em redes suscetíveis a grandes variações de densidade. O algoritmo proposto considera diferentes níveis de influência entre pares de pesquisadores (grafo bidirecional), e fornece também um parâmetro que pode ser ajustado para aprimorar a decisão de se incluir membros com menor influência. Além disso, o algoritmo permite que haja sobreposição entre as comunidades, localizando pesquisadores multidisciplinares.

3 Rede Social Científica

Uma rede social é um conjunto de objetos, onde cada um deles está conectado a outros objetos e pode ser representada por um grafo no qual os nós ou vértices estão ligados por arestas ou arcos. Uma rede social reflete uma estrutura social que pode ser representada por indivíduos ou organizações e suas relações. Em geral, as relações representam um ou mais tipos de interdependência (como idéia e religião) ou relacionamentos mais específicos (como troca de conhecimento, informação e amizade). Assim, através dessa estrutura social, a troca de dados e informações entre os indivíduos ou organizações pode ser estudada e analisada em diferentes níveis de detalhes (FREEMAN, 1978)(WASSERMAN; FAUST, 1994).

As Redes Sociais Científicas são tipos específicos de redes sociais que representam as interações sociais oriundas do meio acadêmico. No desenvolvimento deste trabalho a rede social científica foi construída com os dados dos pesquisadores da DBLP, previamente armazenados em um banco de dados relacional (GUÉRCIO; STRÖELE, 2016). Porém, em função do grande volume de dados e para um melhor entendimento da formação de sua estrutura, foi feita a conversão do esquema relacional para um modelo orientado à grafo.

3.1 Modelagem da Rede

No modelo da rede científica proposto neste trabalho os nós do grafo representam os pesquisadores e as arestas possuem pesos que representam o nível de influência entre eles. Neste modelo, a influência que um pesquisador exerce sobre o outro não necessariamente é igual à influência que ele recebe desse mesmo pesquisador. Por isso, o grafo social que representa esse modelo é um grafo bidirecional $G = (V, E)$, onde $V = \{V_0, V_1, \dots, V_n\}$ é o conjunto de vértices (pesquisadores) e E é o conjunto de arestas $e_{ij} = \{V_i, V_j\}$ entre os pesquisadores V_i e V_j , com $0 \leq i \leq n$ e $0 \leq j \leq n$. Os pesos (ida e volta) da relação entre dois pesquisadores são analisados de maneira diferenciada, de forma que $e_{ij} \neq e_{ji}$.

A Figura 3.1 ilustra uma abstração para o modelo proposto neste trabalho, onde dois vértices V_i e V_j estão sempre conectados por duas arestas direcionadas, com pesos independentes. O peso das arestas é definido conforme a equação 3.1, onde $IP(e_{ij})$ é o grau de influência que o pesquisador V_j exerce sobre o pesquisador V_i , $P(V_x)$ é o conjunto de todas as relações do pesquisador V_x , e $\| \cdot \|$ representa o número de elementos neste conjunto. É importante notar que o peso das arestas irá variar no intervalo de 0 a 1, valores próximos de 0 indicam uma fraca influência, enquanto valores próximos de 1 indicam uma forte influência.

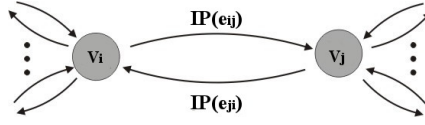


Figura 3.1: Relações entre dois pesquisadores

$$IP(e_{ij}) = \frac{\| P(V_i) \cap P(V_j) \|}{\| P(V_i) \|} \quad (3.1)$$

A análise da estrutura de uma rede social de larga escala permite o entendimento de como ocorrem os relacionamentos entre os elementos desta rede. Assim, baseando-se neste modelo, a próxima seção analisa a topologia da rede social de co-autoria (DBLP) com o objetivo de caracterizar sua estrutura e verificar como os relacionamentos entre os pesquisadores estão distribuídos.

3.2 Análise Topológica

Após a construção da rede social científica DBLP, seguindo o modelo proposto, é possível analisar a sua topologia buscando por características não triviais que facilitem o entendimento de sua estrutura e o estudo de como são construídas as ligações entre os pesquisadores. Um grande desafio que possui impacto na manipulação e análise desta rede é o volume de dados envolvido, já que, no total, existem 1.306.546 vértices e 9.915.146 arestas.

3.2.1 Distribuição de Grau

Inicialmente, para analisar a estrutura da rede foi calculada a sua distribuição de graus, propriedade que caracteriza a topologia de uma rede complexa, e é obtida calculando-se quantos nós possuem determinado grau para todos os valores de graus existentes na rede. A distribuição de grau pode ser utilizada para identificar o tipo da rede que, em geral, podem ser classificadas como redes aleatórias, livres de escala, modulares, mundo pequeno, dentre outras (WASSERMAN; FAUST, 1994). A equação 3.2 foi adotada para calcular essa distribuição, onde $f(k)$ é a fração de vértices que possuem grau igual a k .

$$f(k) = \frac{\text{Número de vértices com grau } k}{\text{Número total de vértices}} \quad (3.2)$$

Com intuito de identificar o tipo da rede social científica deste trabalho, foi calculada a distribuição de graus dos seus elementos. Os gráficos da Figura 3.2 apresentam os resultados obtidos, onde é possível observar que há um grande número de pesquisadores com poucas relações (grau baixo) e poucos pesquisadores com muitas relações (grau alto). Esta característica é típica das redes livres de escala, nas quais poucos elementos centralizam a maioria das relações da rede. No entanto, para que uma rede seja classificada como livre de escala, é preciso que a sua distribuição de grau siga uma lei de potência. A partir do gráfico da Figura 3.3 é possível comparar a distribuição de grau da rede da DBLP com a sua curva ajustada para uma lei de potência.

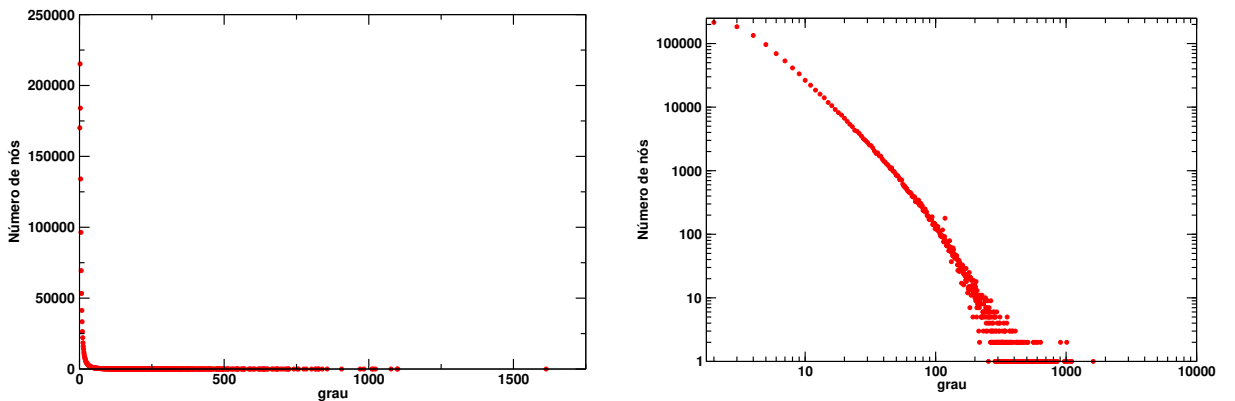


Figura 3.2: Distribuição de grau (esquerda) e gráfico log-log correspondente (direita)

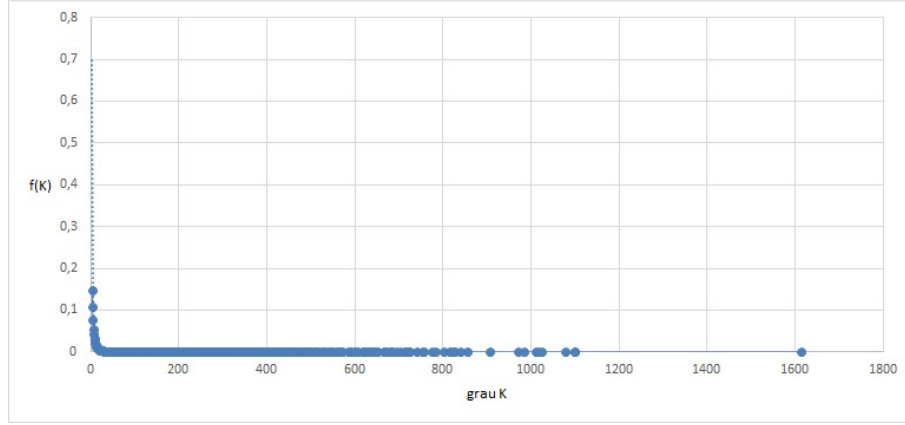


Figura 3.3: Distribuição de grau e curva ajustada da lei de potência

Nota-se que apesar de possuírem um comportamento semelhante, existe uma distância grande entre a distribuição de grau da DBLP e a curva da lei de potência para alguns pontos no gráfico. Dessa forma, não é possível concluir que a distribuição de grau da rede DBLP segue de fato uma lei de potência. Apesar disso pode-se dizer que, devido ao comportamento de cauda pesada da sua distribuição de grau, esta rede tende a receber novas ligações nos vértices que possuem maior grau. Logo, quanto maior o número de colaborações de um pesquisador maior a chance deste pesquisador se conectar a novos pesquisadores.

3.2.2 Distribuição de Influência

A partir do comportamento da distribuição de grau da rede, é possível verificar a coerência da métrica de influência proposta para a definição do peso das arestas do grafo (equação 3.1). Para isso, a influência de cada vértice foi calculada e a distribuição de grau foi realizada considerando os pesos das arestas ao invés do grau de cada nó. Por fim, um gráfico da distribuição de influência dos pesquisadores foi obtido. Como o peso de cada aresta varia de 0 a 1, a distribuição da influência também possui a mesma variação.

Na análise de distribuição de grau foi detectado que existem poucos nós com grau elevado. Por outro lado, o gráfico da Figura 3.4 nos mostra que uma grande quantidade de nós possui a influência máxima (indicado pela seta). Para que essas duas análises sejam verdadeiras de forma simultânea, é necessário que a rede da DBLP não seja totalmente conectada, sendo formada por componentes conexas independentes que contêm poucos

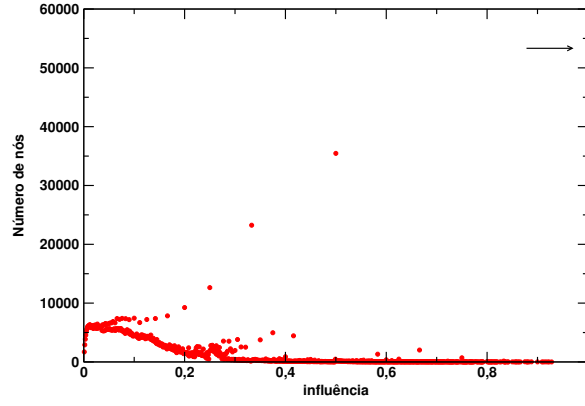


Figura 3.4: Influência dos pesquisadores.

nós de muita influência. Por exemplo, em componentes independentes com apenas dois vértices, ambos terão influência máxima um sobre o outro e ainda assim terão grau 1. Isso significa que existem vários grupos independentes de pesquisadores que trabalham de forma “isolada”, já que não se conectam com a maior parte da rede.

Para verificar o comportamento descrito anteriormente, foi elaborado um algoritmo, com base nos estudos desenvolvidos em [Silva et. al., 2016], cuja função é identificar as componentes conexas da rede. O algoritmo foi desenvolvido considerando o volume de dados da rede e o seu pseudocódigo está descrito no Algoritmo 1.

Algoritmo 1 Buscar Componentes

```

1: função BUSCARCOMPONENTES(RAIO)
2:    $contComponente \leftarrow 0$  ▷ Inicializa contador
3:    $no \leftarrow BuscarNoSemComponente()$  ▷ Nó inicial
4:   enquanto existir no faça
5:      $contComponente \leftarrow contComponente + 1$ 
6:      $atualizaComponente(no, contComponente)$  ▷ Atualiza componente
7:      $vizinho \leftarrow BuscarVizinho(no)$  ▷ Busca no que possui caminho
8:      $atualizaComponente(vizinho, contComponente)$ 
9:      $vizinhos \leftarrow BuscarVizinhos(vizinho, raio)$  ▷ Busca vizinhos em um determinado raio
10:    para  $i \leftarrow 1$  até tamanho(vizinhos) faça
11:       $vizinho \leftarrow vizinhos[i]$ 
12:       $atualizaComponente(vizinho, contComponente)$ 
13:    fim para
14:  fim enquanto
15: fim função

```

A parametrização do *raio* utilizado na função *BuscarVizinhos* não influencia o resultado final do algoritmo, mas possui grande impacto na sua performance e viabilidade, pois permite que o algoritmo seja executado em computadores com capacidade de processamento limitada. Um valor elevado para o *raio* irá diminuir o tempo de execução do algoritmo visto que a função irá retornar um conjunto maior de vértices, porém, o consumo de memória aumenta consideravelmente. Isso acontece porque, com o aumento do *raio*, ocorre a diminuição no número de repetições e um aumento no número de vértices

que são armazenados em memória. Assim, este algoritmo permite que mesmo computadores com baixa capacidade de processamento encontrem todas as componentes conexas de uma rede social de grande porte.

A Tabela 3.1 apresenta o resultado da execução do algoritmo descrito. Percebe-se que a rede social da DBLP possui um grande número de componentes independentes e de diversos tamanhos. Porém, existe uma componente dominante com 1.100.504 vértices. Como o foco do estudo de redes complexas está em analisar o comportamento de redes com um número elevado de vértices foi considerada, neste trabalho, apenas a componente dominante da rede e a distribuição de influência foi recalculada para verificar o seu comportamento. O gráfico da Figura 4.1 apresenta um resultado mais próximo do esperado, ou seja, um comportamento de cauda pesada onde poucos vértices possuem alto índice de influência. A distribuição de influência possui um comportamento de cauda pesada, exceto por alguns pontos que aparecem fora do comportamento padrão do gráfico. Estes pontos representam grupos de pesquisadores que se conectam à maior componente conexa através de uma ligação fraca. Porém, esses grupos possuem alta conectividade entre seus membros, o que resulta em um conjunto de vértices que não se encaixam no comportamento de lei de potência.

Tabela 3.1: Componentes encontradas através do algoritmo 1

Número de nós	Número de componentes	Número de nós	Número de componentes
1	63525	17	15
2	23830	18	11
3	10777	19	13
4	5120	20	4
5	2630	21	4
6	1482	22	6
7	786	23	4
8	483	24	2
9	317	25	2
10	207	26	3
11	129	27	1
12	84	30	1
13	52	31	1
14	35	45	1
15	27	1100504	1
16	20	Total	109573

4 NetSCAN: algoritmo de agrupamento

Após as análises da topologia da rede social científica DBLP, percebe-se que ela possui características de uma rede complexa livre de escala, que possui baixa conectividade global e alguns nós centralizadores, ou seja, alguns nós altamente conectados. Neste contexto surge o interesse em identificar quais são os grupos de pesquisa mais bem definidos, quais são os membros de cada grupo, bem como os pesquisadores com maior influência sobre seus respectivos grupos. Para responder essas questões, é proposto e desenvolvido um algoritmo de agrupamento por densidade, chamado NETSCAN, baseado no algoritmo DBSCAN (ESTER et al., 1996) e que possa ser aplicado em redes de grande porte como a DBLP.

Uma das principais diferenças com relação ao DBSCAN é que o NETSCAN considera que a distância entre dois vértices depende da direção do relacionamento que está sendo analisada. Além disso, permite que um mesmo vértice seja incluído em mais de um grupo e, dessa forma, é possível identificar pesquisadores que contribuem em diferentes áreas e grupos de pesquisa distintos, mesmo que esses pesquisadores não sejam caracterizados como centralizadores.

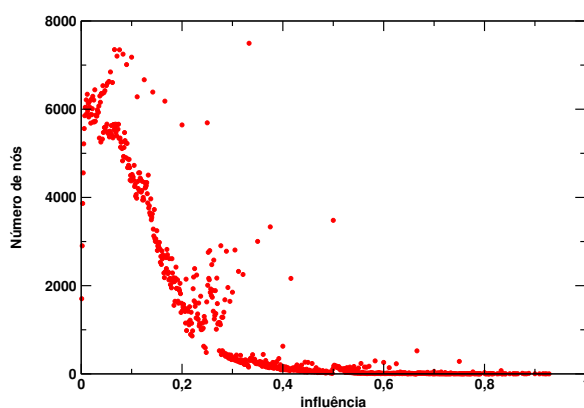


Figura 4.1: Distribuição da Influência na maior componente conexa

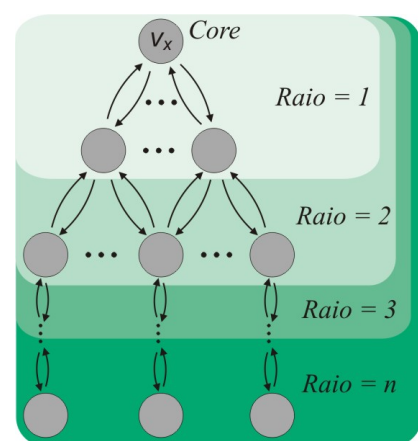


Figura 4.2: Ilustração do parâmetro "raio"

O NetSCAN possui três parâmetros, sendo dois obrigatórios (ε e $minPts$), com características semelhantes aos parâmetros do DBSCAN, e um opcional, que permite buscar por elementos influenciados pelo *core* em uma profundidade maior, ou seja, pode

ser definido quantas camadas de vizinhos a partir do *core* serão analisadas, conforme ilustrado na Figura 4.2, onde o valor máximo do parâmetro *raio* é igual ao diâmetro da rede científica dado por n . O parâmetro ε define a influência mínima para que um vértice seja considerado influenciador de seu vizinho e *minPts* indica o número mínimo de vizinhos próximos que um nó deve possuir para ser considerado um nó centralizador (*core*). O pseudocódigo do algoritmo está descrito no Algoritmo 2.

O primeiro passo do algoritmo é escolher de forma aleatória um vértice v_x que ainda não foi visitado e então é feita uma busca por região que retorna um conjunto com todos os vizinhos de v_x que são influenciados por ele com uma influência maior ou igual a ε . Se a quantidade de vértices retornados na busca for maior ou igual a *minPts*, um grupo G é formado, o elemento v_x é definido como um *core* que será posteriormente expandido. Expandir o elemento v_x significa agrupar seus vizinhos em G , e então procurar se existe algum *core* dentre seus vizinhos. Caso algum vizinho de v_x seja também um *core*, seus vizinhos também serão analisados e agrupados em G .

Nós que não possuem um número de vizinhos maior ou igual que *minPts* serão marcados como *noise* e, caso sejam vizinhos próximos e consequentemente agrupados por algum *core*, passam a ser considerados como *border points*, sendo vinculados ao grupo de seu *core*. Se o parâmetro opcional *Raio*, que possui valor padrão igual a 1, for modificado, a função *BuscaPorRegiao* irá incluir em sua busca os vértices que estejam a mais de uma aresta de distância do nó que está sendo analisado, ou seja, irá procurar além da camada de *Raio* igual a 1 (Figura 4.2). Neste caso, se v_x é um *core*, v_y é um *border point* influenciado por v_x e v_z é um *noise* influenciado por v_y , considerando um *Raio* maior ou igual a 2, a busca por região irá considerar que v_z é um vizinho influenciado por v_x e, portanto, será agrupado se tornando um *border point* vinculado ao grupo do *core* v_x . O uso desse parâmetro permite que pesquisadores influenciados por intermédio de outros pesquisadores sejam vinculados ao grupo do influenciador central, possibilitando uma análise mais detalhada dos níveis de influência dos pesquisadores.

Por questões de performance e para garantir que o processo de agrupamento possa ser interrompido e retomado posteriormente sem perda de informações, as funções *BuscaPorRegiao*, *Agrupar*, *MarcarCore* e *MarcarNoise* são implementadas através de *queries*

Algoritmo 2 Network Scan

```

1: função NETSCAN(GRAFO, EPS, MINPTS, RAO)
2:   clusterId  $\leftarrow$  0
3:   enquanto N = buscaNoNaoClassificado(grafo) faça
4:     clusterId  $\leftarrow$  clusterId + 1
5:     vizinhos  $\leftarrow$  BuscaPorRegiao(N, grafo, eps, raio)
6:     expandirNo(N, vizinhos, grafo, eps, minPts, clusterId, raio)
7:   fim enquanto
8: fim função

1: função EXPANDIRNO(N, VIZINHOS, GRAFO, EPS, MINPTS, CLUSTERID, RAO)
2:   quantidadeVizinhos  $\leftarrow$  tamanho(vizinhos)
3:   se (numVizinhos < minPts) então
4:     marcarNoise(N)
5:     return false
6:   fim se
7:   marcarCore(N)
8:   agrupaVizinhos(N, vizinhos, grafo, eps, minPts, clusterId, raio)
9: fim função

1: função AGRUPAVIZINHOS(N, VIZINHOS, GRAFO, EPS, MINPTS, CLUSTERID, RAO)
2:   agrupar(N, vizinhos, clusterId)
3:   para i  $\leftarrow$  1 até tamanho(vizinhos) faça
4:     N  $\leftarrow$  vizinhos[i]
5:     novosVizinhos  $\leftarrow$  BuscaPorRegiao(N, grafo, eps, raio)
6:     expandirNo(N, novosVizinhos, grafo, eps, minPts, clusterId, raio)
7:   fim para
8: fim função

```

executáveis no próprio banco de dados Neo4j.

5 Experimentação

Para avaliar o desempenho do NetSCAN foram realizados experimentos em três cenários distintos: (i) espaço euclidiano R^2 (ii) redes sintéticas (iii) redes do mundo real. Em cada cenário foram induzidos grafos bidirecionais cujas estruturas foram modeladas considerando os diferentes tipos de relacionamento entre os indivíduos.

Nos experimentos realizados em R^2 as arestas foram definidas a partir da distância euclidiana e todos os vértices possuem ligações de ida e de volta entre si.

Já no caso das redes foi necessário definir outra medida para o peso das arestas, já que os relacionamentos nessas redes, a princípio, não possuem uma definição de distância e também não são direcionados. Para isso a seguinte medida de peso baseada na métrica de influência foi aplicada:

$$Peso_{AB} = \frac{\| P_A \cap P_B \|}{\| P_B \|} \quad (5.1)$$

onde $\| P_A \cap P_B \|$ corresponde ao número de vértices em comum entre P_A e P_B e $\| P_B \|$ representa o número total de vértices ligados a P_B .

5.1 Experimentos em R^2

O objetivo dos experimentos em R^2 é de avaliar a capacidade do algoritmo em detectar grupos nos casos clássicos de agrupamento, ou seja, casos em que os vértices ligados são equidistantes entre si e não existe a noção de direcionamento nos relacionamentos.

O NetSCAN foi implementado para considerar relacionamentos direcionados e, por isso, os grafos para representar as instâncias em R^2 possuem arestas de ida e de volta entre todos os vértices e com peso definido pela distância euclidiana entre eles.

Para este cenário foram realizados experimentos em três conjuntos de dados distintos que foram gerados em (SEMAAN, 2013):

- 200Data, que representa pontos em uma distribuição normal em R^2

- Numbers, representações dos algarismos de 0 a 9
- 2Face, pontos que constituem um desenho em forma de face combinado com pontos que formam o algarismo 2.

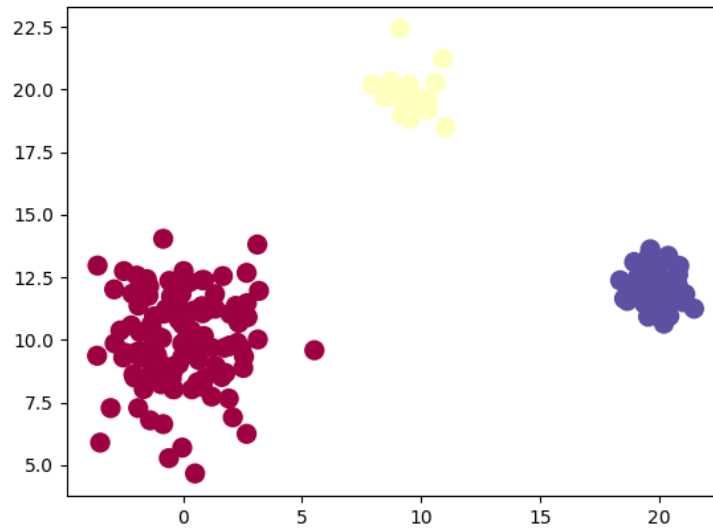


Figura 5.1: Resultado do agrupamento para a instância 200-data

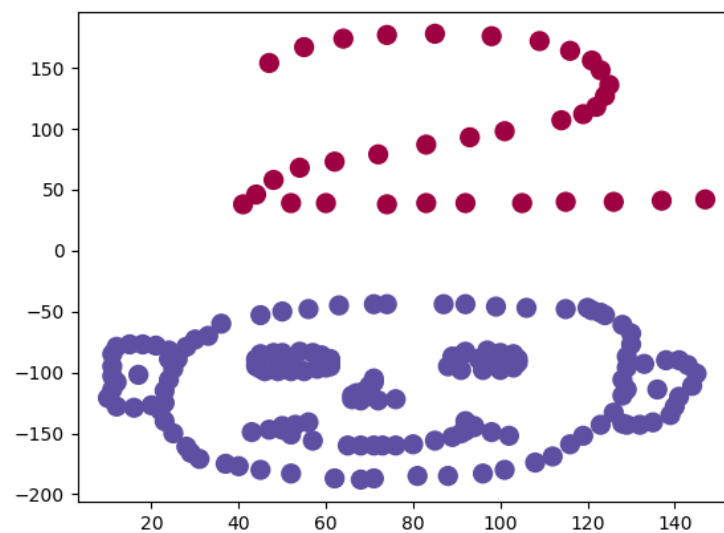


Figura 5.2: Resultado do agrupamento para a instância 2face

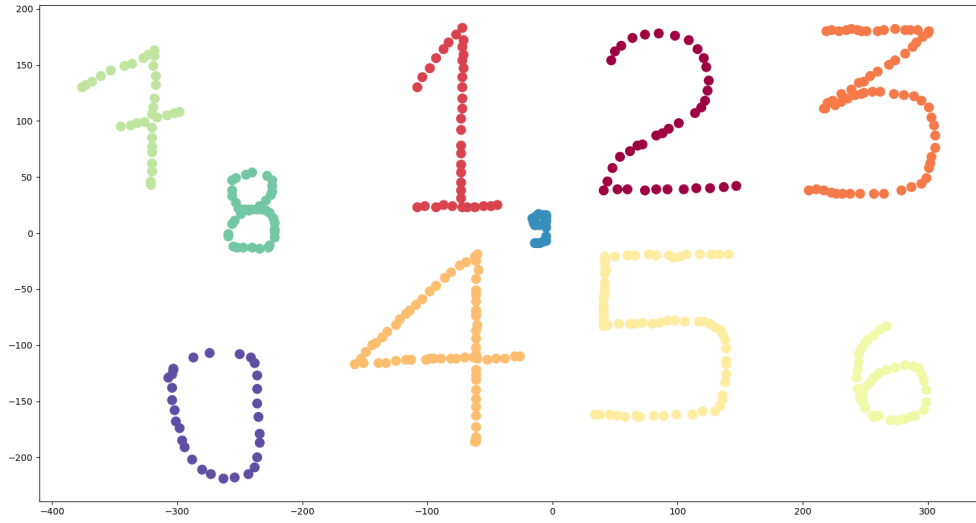


Figura 5.3: Resultado do agrupamento para a instância numbers

A figura 5.2 mostra que o algoritmo proposto reconheceu os dois grupos distintos na instância 2Face: os pontos que formam a face e o algarismo 2, enquanto que nas figuras 5.1 e 5.3 pode-se perceber a detecção eficaz dos grupos no conjunto 200Data e dos números definidos no conjunto Numbers considerando que se trata de um algoritmo baseado em densidade.

5.2 Experimentos em Rede Social Sintética

Apesar dos resultados satisfatórios do algoritmo nas instâncias em R^2 , tais instâncias não são capazes refletir todas as características estruturais de uma rede social como, por exemplo, relacionamentos assimétricos entre os indivíduos e sobreposições dos grupos.

Assim, para observar o comportamento do NetSCAN em um cenário com características de uma rede social, foram realizados experimentos em uma rede sintética e duas redes reais de pequeno porte.

A rede artificial da Figura 5.4 foi construída para possuir 3 comunidades bem definidas e que se sobrepõem através de um vértice central a rede. Nota-se a dificuldade em atribuir um grupo correto para o vértice central de número 16, já que o mesmo interage com as 3 comunidades na mesma intensidade.

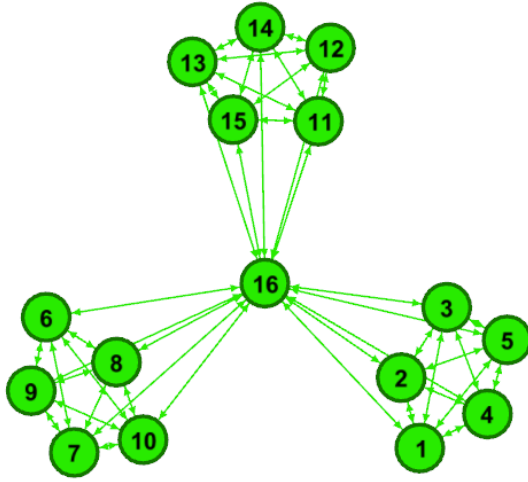


Figura 5.4: Rede social artificial para realização do experimento.

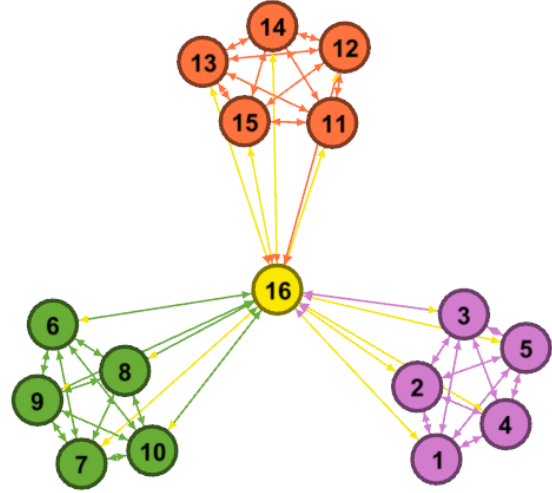


Figura 5.5: Resultado do agrupamento realizado sobre a rede artificial

Como mostrado na Figura 5.5, o NetSCAN detecta as 3 comunidades e considera que existe uma sobreposição entre estes grupos, fazendo com que o vértice 16 (amarelo) participe de todos os grupos simultaneamente, alcançando o resultado esperado para esta instância.

5.3 Experimentos em Redes Sociais Reais

Para avaliar o comportamento do algoritmo sobre redes reais foram utilizadas duas redes conhecidas da literatura. A primeira (Figura 5.6) é uma rede social conhecida como Zachary's Karate Club Network (ZACHARY, 1976). Ela mostra a relação de amizade de 34 membros do clube de karatê. Nesta instância existem 34 nós, 78 arestas e 2 comunidades reais.

O algoritmo detectou as duas comunidades existentes, considerando 3 indivíduos como outliers (azul) e 4 indivíduos que participam das duas comunidades ao mesmo tempo (amarelo), caracterizando uma sobreposição entre estes grupos.

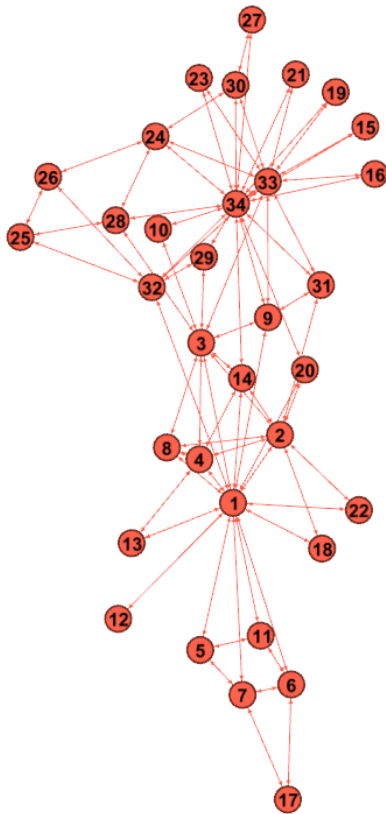


Figura 5.6: Comunidade de karatê

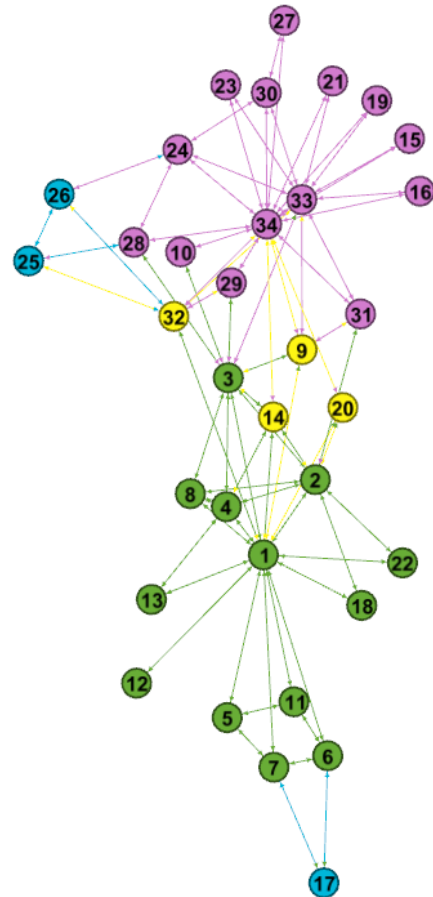


Figura 5.7: Resultado do agrupamento realizado sobre a rede karatê

A Figura 5.8 ilustra uma rede real de proteínas (MEENA; DEVI, 2015) que contém 21 vértices, 61 arestas, três grupos distintos e dois vértices sobrepostos.

O NetSCAN novamente encontrou as três comunidades e identificando corretamente os vértices 8 e 9 (amarelo) como sendo sobrepostos.



Figura 5.8: Rede de proteínas com sobreposição em 2 vértices

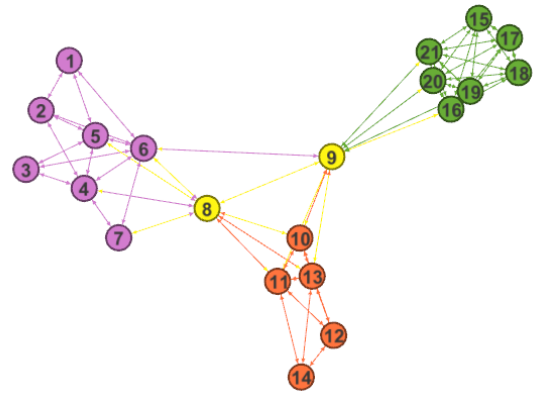


Figura 5.9: Resultado do agrupamento realizado sobre a rede de proteínas

6 Rede científica real

Após analisar o comportamento do algoritmo proposto e verificar sua capacidade de identificar comunidades e sobreposições em vários tipos de dados, foi feita uma análise de uma rede social científica real. Os objetivos desta análise são de identificar comunidades de pesquisa e pesquisadores influentes (*central connectors* and *research-information brokers*).

Para definir o escopo desta avaliação foi utilizado o método GQM (KITCHENHAM; CHARTERS, 2007), que foi descrito da seguinte forma: “Analisar comunidades científicas e comunidades científicas sobrepostas geradas pelo NetSCAN com o objetivo de identificar pesquisadores multidisciplinares (*research-information brokers*) e pesquisadores com alta conectividade (*central connectors*) em relação ao potencial de influência dos pesquisadores do ponto de vista dos pesquisadores no contexto das comunidades científicas encontradas baseadas na influência dos pesquisadores em uma rede social científica.”

Baseando-se neste escopo de avaliação foram definidas uma questão de pesquisa principal e três questões de pesquisa secundárias:

- Como as comunidades científicas são organizadas considerando *research-information brokers* e *central connectors* em relação ao potencial de influência destes pesquisadores?
 - RQ1: Existem pesquisadores que participam de duas ou mais comunidades simultaneamente, caracterizando *research-information brokers*?
 - RQ2: Existem pesquisadores com um número elevado de conexões com outros pesquisadores, caracterizando *central connectors*?
 - RQ3: Através de métodos de detecção de comunidades é possível encontrar comunidades científicas reais considerando as atividades realizadas por seus pesquisadores?

Para responder a estas questões de pesquisa foi escolhido o Método de Pesquisa Histórica, já que foram avaliados os dados históricos de uma rede social científica real e que não

houve controle no comportamento dos pesquisadores nos eventos, segundo Yin (YIN, 2009). Como fontes de evidencia foram utilizados registros de arquivo e entrevistas. A primeira fonte foi usada para verificar os grupos identificados em geral. A segunda foi utilizada com a intenção de avaliar algumas comunidades e relações específicas, para alcançar uma análise mais detalhada dos resultados obtidos.

Neste estudo, o algoritmo NetSCAN foi executado na maior componente conexa da rede social científica DBLP composta por 1.100.504 pesquisadores e 9.598.808 relacionamentos. Ao todo foram identificadas 34.776 comunidades de pesquisa, utilizando os parâmetros: $\varepsilon = 1$, $minPts = 5$ e $raio = 1$.

Esses parâmetros foram definidos após alguns experimentos realizados com uma parte da base de dados. Nesses experimentos observou-se que para identificar comunidades científicas formadas por pesquisadores com influência direta, o ideal é adotar $raio = 1$ e um valor de ε menor. Dessa forma, facilita-se a localização de grandes comunidades de pesquisa nas quais os *cores* influenciam diretamente os *border points*. Por outro lado, para refinar essas comunidades e identificar subgrupos, é necessário aumentar o valor do raio, pois, em geral, os pesquisadores em seus subgrupos são influenciados por intermédio de outros pesquisadores.

Dado o grande volume de dados utilizados neste estudo, alguns agrupamentos foram selecionados com o intuito de ilustrar o comportamento do algoritmo proposto e analisar os resultados obtidos. A Figura 6.1 mostra um dos grupos definidos pelo NetSCAN onde os nós centralizadores estão representados pelos vértices verdes e os *border points* pela cor cinza. Para uma melhor visualização do grafo, as medidas de influência representadas nas arestas foram multiplicadas por 1000.

Observa-se que este grupo possui dois vértices influenciadores e que apenas dois *border points* estão ligados nestes dois vértices simultaneamente. Isso indica que este grupo possui uma grande chance de ser separado em dois subgrupos caso o parâmetro ε seja escolhido de forma mais rígida, exigindo um maior nível de influência no grupo.

O NetSCAN permite que haja sobreposição entre as comunidades de pesquisa, localizando assim pesquisadores multidisciplinares que são influenciados por comunidades de pesquisas distintas. A Figura 6.2 apresenta essa situação, na qual há um pesquisador

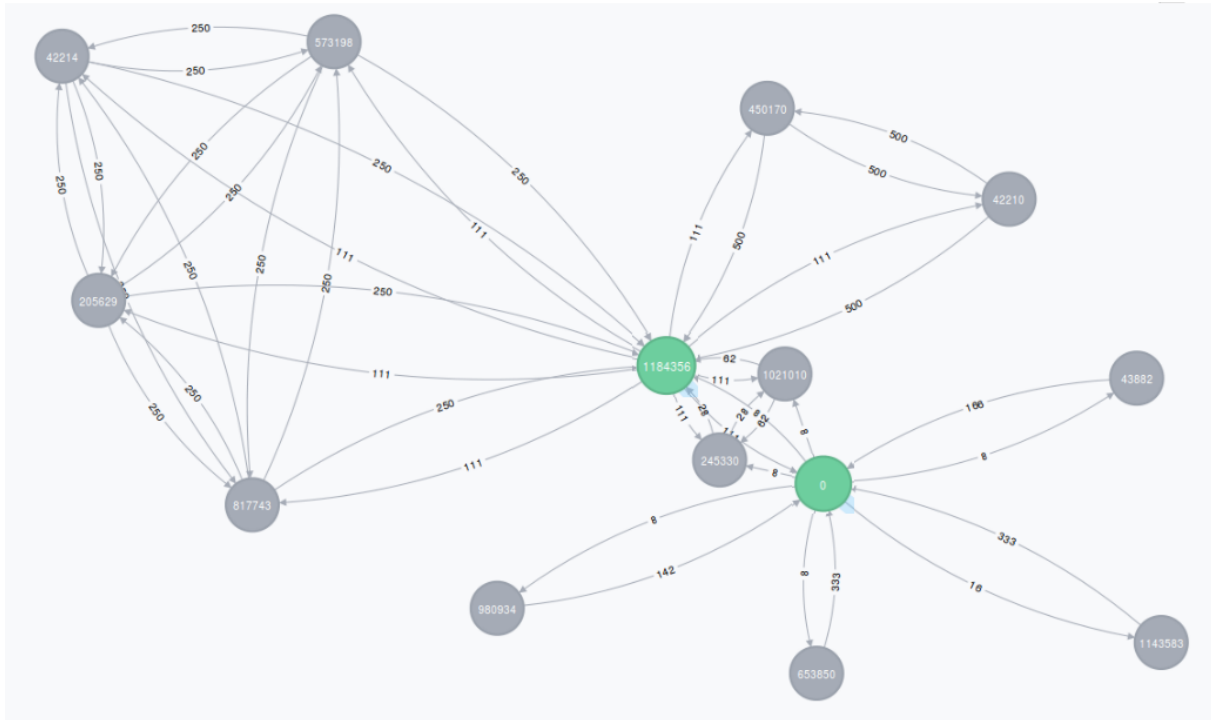


Figura 6.1: Grupo de pesquisadores identificado pelo NETSCAN

associado a dois grupos. Os nós azuis representam as duas comunidades e as relações desses vértices com os pesquisadores indicam que eles participam da comunidade. Observa-se que o nó vermelho (419600) possui ligações com membros dos dois grupos, mas só é suficientemente influenciado por membros do grupo (360998) que está mais à direita na figura e, portanto, só participa desta comunidade. Outros vértices deste conjunto também possuem propriedades peculiares como, por exemplo, o vértice mais à esquerda (338001) que exerce uma influência muito elevada a ponto de atrair um grande número de colaboradores. Além deste, o vértice central com rótulo 746362, apesar de estar ligado a poucos nós, possui uma relação de influência suficiente para participar de ambos os grupos.

Essas considerações iniciais mostram que o algoritmo é capaz de identificar comunidades científicas em uma rede de grande porte. Nas próximas seções serão realizadas análises mais detalhadas dos resultados obtidos com o intuito de responder às questões de pesquisa.

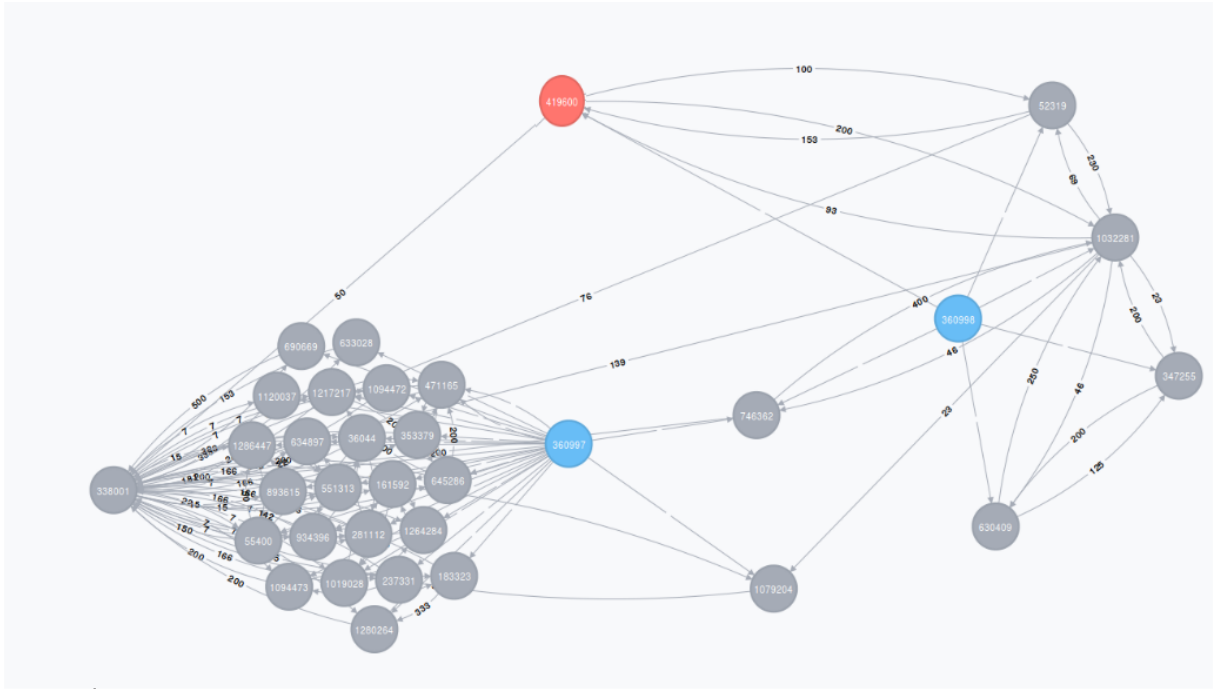


Figura 6.2: Análise de comunidades sobrepostas

6.1 Análise dos Resultados

Através do NetSCAN foram identificadas comunidades científicas, pesquisadores influenciadores e pesquisadores que participam de mais de uma comunidade. Um dos desafios encontrados, ao analisar uma rede científica com grande volume de dados, foi a variação de densidade que dificulta a parametrização das entradas do algoritmo de agrupamento. Para resolver este problema, sem a necessidade de se calcular os valores ideais de cada parâmetro repetidamente e em tempo de execução, foi proposto um método de subagrupamento que foi utilizado para refinar grupos formados por muitos pesquisadores.

Nas análises dos resultados foram encontrados vértices que possuem similaridade suficiente para participar de mais de um grupo. Neste caso, se fosse adotada uma estratégia de junção de grupos, as comunidades científicas encontradas seriam muito amplas e com baixa coesão. Para evitar este comportamento permitiu-se que as comunidades tenham sobreposição, ou seja, um vértice pode ser incluído em mais de um grupo. Este fato mostra uma característica de interdisciplinaridade do pesquisador representado.

Dada a quantidade de pesquisadores envolvidos na rede social científica analisada neste trabalho, não é possível fazer uma avaliação qualitativa com cada um deles em relação aos seus grupos de pesquisa e seus níveis de influência. Assim, através de uma

entrevista informal, com observação direta como fonte de coleta de dados, os resultados obtidos foram apresentados a seis pesquisadores para que eles os analisassem. Cinco deles reportaram que as comunidades nas quais eles estavam envolvidos estava condizente com a realidade dos seus grupos de pesquisa. Apenas um pesquisador reportou que, de fato, desenvolvia suas pesquisas com os pesquisadores que ele estava relacionado, mas considerou o grupo ao qual ele estava vinculado muito grande. Então, foi apresentado para ele os sub-agrupamentos do seu grupo original, o qual confirmou que o seu subgrupo era mais aderente à realidade de seu grupo de pesquisa.

As análises realizadas apontam para a viabilidade tanto do modelo quanto dos algoritmos propostos. Através da avaliação realizada pode-se dizer que há indícios de que o resultado alcançado foi satisfatório e que os métodos apresentados podem ser evoluídos para alcançar uma coesão maior no agrupamento. A partir destes resultados também é possível responder parcialmente a questão de pesquisa RQ1 (Existem pesquisadores que participam de duas ou mais comunidades simultaneamente, caracterizando *research-information brokers*?), já que foram encontrados pesquisadores que participam simultaneamente de duas comunidades. Nas seções seguintes serão feitos estudos sobre as sobreposições encontradas para verificar se os pesquisadores envolvidos possuem características de *research-information brokers*.

6.2 Análise das sobreposições

Para responder, mais especificamente, a questão de pesquisa relacionada às comunidades sobrepostas RQ1 foram feitas análises quantitativas e qualitativas do resultado de agrupamento do algoritmo.

Primeiramente foi feita uma análise quantitativa e percebeu-se a existência de diferentes tipos de sobreposições, os quais serão chamados de *Border Point Overlapping Communities* (BpOC) e *Core Point Overlapping Communities* (CpOC). Para caracterizar os tipos de sobreposições considere a existência de duas comunidades C_i e C_j ; c_i e c_j como sendo vértices *cores* das comunidades C_i e C_j respectivamente; e $C_i \neq C_j$. O tipo de sobreposição BpOC é o caso no qual estas duas comunidades se sobrepõem através de um vértice *border point*. A Figura 6.3 representa esse tipo de sobreposição, onde as

comunidades de número 303 e 14 são sobrepostas pelo pesquisador 449359.

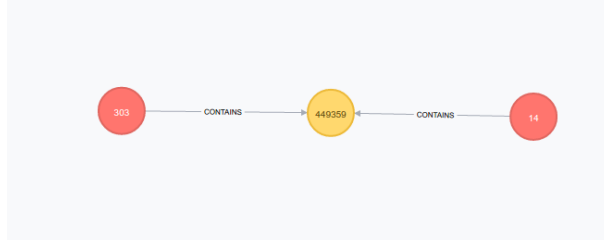


Figura 6.3: Duas comunidades (vermelho) sobrepostas por um *border point* (amarelo)

A principal característica da sobreposição BpOC é que o vértice sobreposto é um *border point* vertice (BpV) e, portanto, sua inclusão na sobreposição $C_i \cap C_j$ não expande e não altera a configuração de $C_i \cap C_j$. As sobreposições do tipo BpOC também podem ocorrer quando múltiplos *border points* são conectados aos *cores* c_i e c_j . Este tipo de sobreposição passa então a ser definida por $C_i \cap C_j = BpV^n$, onde BpV^n é o conjunto de n vértices *border points* que são influenciados por dois ou mais *cores* de duas ou mais comunidade distintas simultâneamente.

Para que esse tipo de sobreposição ocorra é necessário que um vértice *border point* (BpV) seja influenciado por *cores* pertencentes a comunidades distintas, por exemplo, pelos *cores* c_i e c_j simultâneamente. Dessa forma podemos dizer que a sobreposição BpOC acontece quando as condições definidas em 6.1 são satisfeitas.

$$\begin{aligned}
 c_i &\in C_i \\
 c_j &\in C_j \\
 c_i &\notin C_j \\
 c_j &\notin C_i
 \end{aligned} \tag{6.1}$$

$$IP(e_{BpV}, c_i) \geq eps$$

$$IP(e_{BpV}, c_j) \geq eps$$

O outro tipo de sobreposição é a CpOC, que ocorre quando duas comunidades se sobrepõem através de um vértice *core point* (CpV), como mostrado na Figura 6.4. Nesta figura o *core* sobreposto está representado pelo vértice em azul (926415), os agrupamentos são os vértices em vermelho (2046 e 53), e os vertices em cinza são *border points*

influenciados por este *core*.

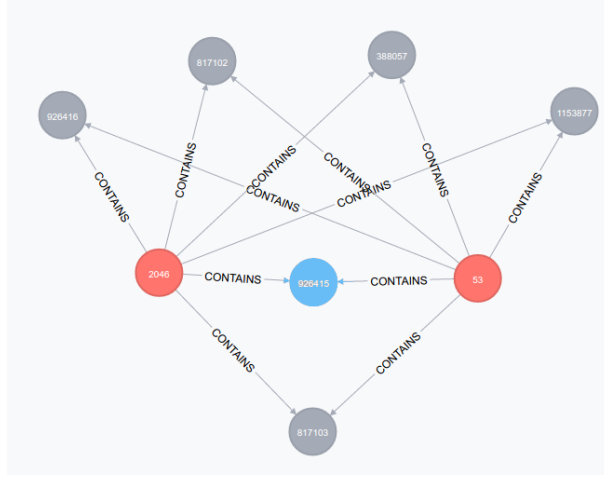


Figura 6.4: Duas comunidades (vermelho) sobrepostas por um *core point* (azul) e seus vizinhos (cinza)

Nas sobreposições CpOC, quando um vértice é incluído na sobreposição de $C_1 \cap C_2$, este vértice expande e altera a configuração de $C_i \cap C_j$. Isso ocorre porque a inclusão de um vértice caracterizado como *core point* (CpV_l) em $C_i \cap C_j$ implica que os elementos influenciados pelo CpV_l também sejam incluídos em ambas comunidades, como ilustrado na Figura 6.4. As sobreposições CpOC podem ser definidas como $C_i \cap C_j = \{CpV_l\} \cup CpV_l^n$ onde CpV_l é o *core point* vertice sobreposto e CpV_l^n é o conjunto dos n vizinhos influenciados por .

A Figura 6.5 ilustra o comportamento do NetSCAN na detecção de sobreposições do tipo CpOC. Nesta figura são representados seis passos do algoritmo para a detecção da sobreposição e a expansão causada pelo *core* sobreposto. Para facilitar a visualização apenas as arestas com influência maior que ϵ_{ps} estão representadas. Neste exemplo considere $\minPts = 3$. Os passos ilustrados na Figura 6.5 são os seguintes:

1. Vértices e ligações da rede a serem avaliados pelo NetSCAN.
2. Vértice 1 (v_1) é escolhido aleatoriamente. Como atende as propriedades do algoritmo ($\minPts$ e ϵ_{ps}), ele é definido como *core* e cria-se a comunidade C_1 com os vértices 1,2,3 e 4.
3. Como o vértice 4 (v_4) também é *core point*, ocorre sua expansão e os vértice 5 e 6 são incluídos em C_1 . Neste momento, o NetSCAN finaliza a expansão de C_1 , pois

não existem outros *cores* a serem expandidos.

4. Vértice 7 (v_7) é escolhido aleatoriamente. Como é *core* cria-se a comunidade C_2 com os vértices 4, 8 e 9. Além disso, como v_7 é um *core* que pertence a outra comunidade (C_1), caracteriza-se a sobreposição do tipo CpOC entre as comunidades C_1 e C_2 .
5. Como c_4 é *core* ocorre a sua expansão. Os vértices 5 e 6 são incluídos em C_2 causando também a expansão da sobreposição $C_1 \cap C_2$.
6. Fim do agrupamento pois não existem outros *cores* a serem avaliados.

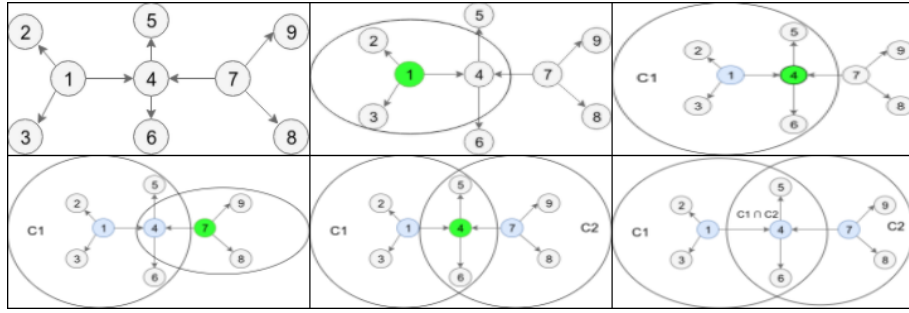


Figura 6.5: Passo a passo do tipo B de sobreposição

Com base na Figura 6.5 podemos destacar a diferença entre as sobreposições por Border Point (BpOC) e as sobreposições por Core Point (CpOC), na inclusão dos vértices 5 e 6 na interseção das comunidades. Se o vértice 4 fosse um *border point* essa sobreposição seria do tipo BpOC e os vértices 5 e 6 não iriam compor essa sobreposição.

Além do caso ilustrado na Figura 6.5, existe a possibilidade de que outros vértices do tipo *core point* ($CpV_{1...h}$) sejam influenciados por CpV_l , onde $CpV_{1...h} = \{CpV_1, CpV_2, ..., CpV_h\}$ e $CpV_{1...h} \subset CpV_l^n$, sendo h o número de *core points* influenciados direta ou indiretamente por CpV_l . Isto causa a inclusão de $CpV_{1...h}$ em $C_i \cap C_j$, a expansão destes vértices e a inclusão dos vizinhos influenciados por eles. Este processo se repete até que todos os *core points* pertencentes a $C_i \cap C_j$ sejam expandidos. Assim pode-se estender a definição de sobreposição do tipo CpOC para $C_i \cap C_j = CpV_l \cup CpV_l^n \cup \{CpV_1^n \cup ... \cup CpV_h^n\}$, onde $CpV_{1...h}^n$ são conjuntos de vértices influenciados pelos $CpV_{1...h}$.

No agrupamento realizado na rede da DBLP foram detectados e analisados os dois tipos sobreposições definidos anteriormente. Para cada um dos tipos diversas configurações foram encontradas. Existem casos envolvendo mais de duas comunidades,

múltiplos *border points* sobrepostos e até mesmo múltiplos *cores* sobrepostos. A Figura 6.6 mostra alguns destes diferentes tipos de sobreposições.

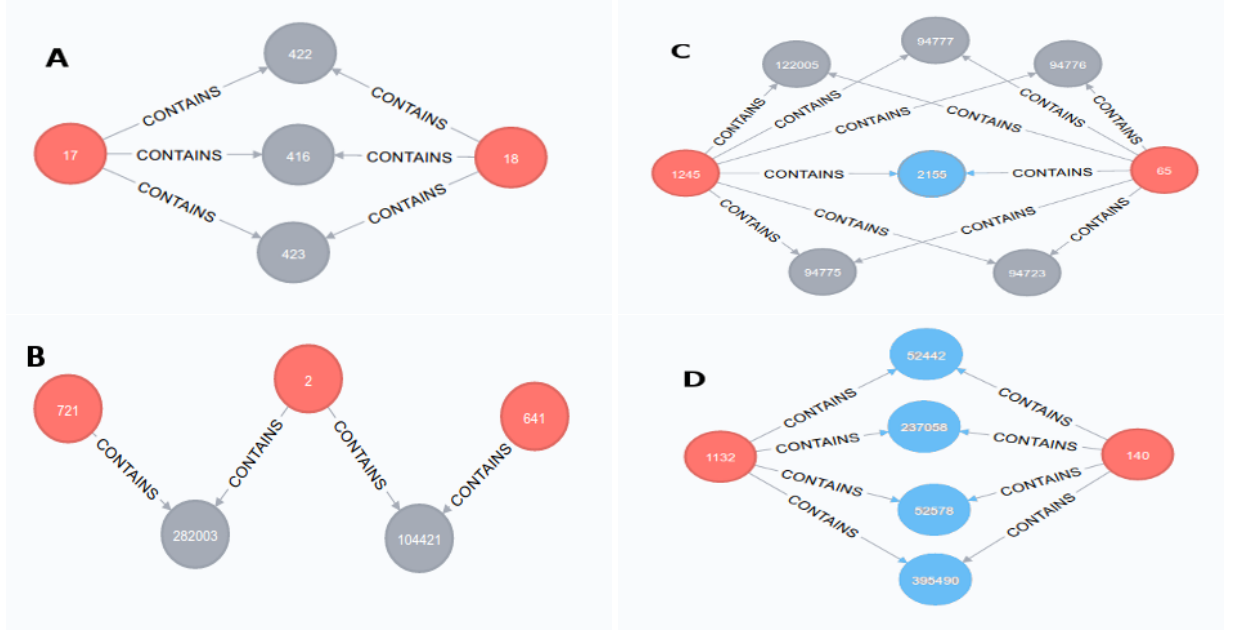


Figura 6.6: Diferentes configurações de sobreposições em comunidades (vermelho) e seus participantes (amarelo)

As configurações mostradas na figura 6.6 são: (a) três *border points* sobrepostos em duas comunidades (b) dois *border points* sobrepostos em três comunidades (c) um *core* e seus cinco vizinhos *border points* sobrepostos em duas comunidades (d) quatro *cores* sobrepostos em duas comunidades. Na figura 6.6.d apenas os *core points* sobrepostos estão representados para facilitar a visualização, mas todos os *border points* influenciados pelo *core* também pertencem à sobreposição.

Após encontrar os tipos de sobreposições e configurações existentes foram feitos levantamentos de dados quantitativos sobre o resultado do agrupamento. Analisando os resultados obtidos observa-se que o número de comunidades sobrepostas representa aproximadamente 31.5% do total de comunidades encontradas. Isso indica que é comum a interação entre diferentes comunidades científicas. Percebe-se também que a maioria dos nós sobrepostos são *border points*. Isto acontece pois toda vez que um *core* pertence a duas comunidades distintos seus vizinhos também são associados a estas comunidades.

Com as análises realizadas nessa seção podemos responder à primeira e segunda questão de pesquisa (RQ2: Existem pesquisadores com um número elevado de conexões com outros pesquisadores, caracterizando *central connectors*?; RQ1: Existem pesqui-

sadores que participam de duas ou mais comunidades simultaneamente, caracterizando *research-information brokers*?). Baseando-se nas análises realizadas e com o auxílio da Figura 6.6 podemos afirmar que há elementos que representam *research-information brokers* e outros que são *central connectors*.

Na próxima seção serão estudadas as atividades dos pesquisadores sobrepostos em suas comunidades ao longo do tempo. O objetivo deste estudo é entender os motivos reais destas sobreposições e buscar evidências para responder a terceira questão de pesquisa (RQ3: Através de métodos de detecção de comunidades é possível encontrar comunidades científicas reais considerando as atividades realizadas por seus pesquisadores?).

6.3 Comunidades sobrepostas ao longo do tempo

Nesta seção é feita uma análise detalhada das comunidades sobrepostas com o objetivo de avaliar o motivo da ocorrência da sobreposição. Para tal, foi feita uma análise temporal dessas comunidades com o intuito de observar a evolução das mesmas ao longo do tempo, identificando as características das sobreposições.

Através da análise do histórico de publicações de diversos vértices sobrepostos foram encontrados três motivos reais de sobreposição, sendo estes: mudança de área de atuação, mudança de grupo de pesquisa e atuação simultânea em múltiplos grupos de pesquisa.

6.3.1 Mudança de área de atuação

A primeira causa real de sobreposição encontrada foi a mudança de área de atuação. O pesquisador analisado foi identificado como um *border point* e participa de uma sobreposição do tipo BpOC entre duas comunidades.

Analisando a evolução desse pesquisador ao longo dos anos, verificou-se que ele/ela possui publicações com membros de uma das comunidades até o ano de 2001. A partir de 2011 suas publicações foram realizadas apenas com os membros da outra comunidade. No intervalo entre 2001 e 2011 foram encontradas poucas publicações deste pesquisador, sendo que na maioria destas publicações ele foi autor único. A tabela 6.1

mostra uma síntese do histórico das publicações com estes grupos de pesquisa.

Tabela 6.1: Análise temporal das publicações deste pesquisador com suas duas comunidades C1 e C2

	First publication	Last publication with C1	First publication with C2	Last publication with C2
Year	1984	2001	2011	2015

Após analisar os anos de colaboração entre o pesquisador analisado e suas duas comunidades foi feito um levantamento das suas áreas de atuação. Para fazer este levantamento foram coletados os *abstracts* destas publicações nos períodos de 1989-2001 e 2011-2015. Os *abstracts* foram submetidos a um *word cloud* e os 10 termos mais frequentes de cada período foram coletados. Alguns termos considerados como irrelevantes (thing, give, etc.) foram retirados desta coleta. A Tabela 6.2 mostra os principais termos identificados nos dois períodos.

Tabela 6.2: Termos utilizados nas publicações nos dois intervalos de tempo

Abstract Terms (1989-2001)	Abstract Terms (2011-2015)
architecture - complexity - components development - diversity - language - product relations - software - systems	algorithm - automatic - data - information knowledge - medical - proposed - radiology semantic - system

É possível perceber uma mudança de termos utilizados nas publicações entre os dois períodos distintos. Os termos do primeiro período estão mais associados a área de arquitetura e engenharia de software. Já os termos do segundo período possuem maior semelhança com a área de ciência de dados voltada para um contexto de saúde.

Dessa forma, foi considerado que a sobreposição neste vértice ocorreu pela mudança na sua área de atuação. Essa sobreposição ocorreu em função da análise evolutiva da rede. Além disso, este vértice é o único membro em comum entre as duas comunidades. Com isso considera-se que este pesquisador possui características de um *research-information broker*, em vista do seu potencial de intermediar a colaboração entre essas comunidades.

6.3.2 Mudança de grupo de pesquisa na mesma área de atuação

Outro motivo de sobreposição encontrado foi a mudança de grupo de pesquisa em que não ocorre mudança na área de atuação. O pesquisador selecionado neste caso foi identificado como *core* e participa de uma sobreposição do tipo CpOC.

Novamente, para entender o real motivo da sobreposição, foi feita uma análise envolvendo todas as publicações deste *core*. Verificou-se que o pesquisador possui publicações com membros de uma das comunidades até o ano de 2003. A partir de 2004 suas publicações foram realizadas apenas com os membros da outra comunidade. A Tabela 6.3 mostra maiores detalhes sobre o histórico das publicações com estes grupos.

Tabela 6.3: Análise temporal das publicações do pesquisador *core* com suas duas comunidades C1 e C2

	First publication	Last publication with C1	First publication with C2	Last publication with C2
Year	1984	2003	2004	2017

Pode-se perceber através da Tabela 6.3 que o pesquisador analisado migrou da comunidade C1 para a comunidade C2 a partir do ano de 2004. Após analisar os anos de colaboração entre este pesquisador e suas duas comunidades foi feito um levantamento das suas áreas de atuação. Para fazer este levantamento foram coletados os *abstracts* das suas publicações nos períodos 1984-2003 e 2004-2017. Os *abstracts* foram novamente submetidos a um *word cloud* e os 10 termos mais frequentes de cada período foram coletados. Os termos considerados como irrelevantes foram retirados desta coleta. A Tabela 6.4 mostra os principais termos utilizados nas pesquisas deste pesquisador durante os dois períodos.

Tabela 6.4: Termos utilizados nas publicacoes deste pesquisador *core* nos dois intervalos de tempo

Abstract Terms (1984-2003)	Abstract Terms (2004-2017)
Asymptotic - bound channel - codes - error estimate - gaussian probability - signals - spl	bounds channel - codes - decoding - exponent feedback - noise observes - rate transmission

Como existe uma semelhança entre os termos dos *abstracts* apresentados na

Tabela 6.4, não é possível concluir que este pesquisador mudou sua área de atuação. Considerou-se, portanto, que este pesquisador realizou uma mudança de grupo de pesquisa sem alterar sua área de atuação.

Assim como no caso anterior, essa sobreposição caracteriza um research-information broker.

6.3.3 Atuação simultânea em múltiplos grupos de pesquisa

O terceiro motivo real de sobreposição encontrado foi a atuação simultânea em múltiplos grupos de pesquisa. Este tipo de sobreposição ocorre tanto com vértices *border point*, como com vértices *core points*. Essa sobreposição caracteriza que o pesquisador influencia (caso seja um *core point*) ou é influenciado (caso seja um *border point*) por ambas as comunidades de pesquisa. Para analisar sobreposições com essas características, investigamos o histórico de outro pesquisador representado por um vértice *border point*. Este pesquisador participa de uma sobreposição do tipo BpOC entre duas comunidades.

Primeiramente foi feita uma análise temporal das publicações deste pesquisador com cada comunidade. É possível perceber, através do gráfico da Figura 6.7, que não existe uma interrupção na atividade deste pesquisador com suas duas comunidades. Percebe-se também que entre os anos 2012 e 2016 houve colaboração deste pesquisador com as duas comunidades simultaneamente.

Além disso, não foram encontradas publicações entre outros pesquisadores de C1 e C2. Dessa forma podemos considerar que as duas comunidades são de fato distintas e se sobrepõem através do *border point* analisado.

Após caracterizar a atuação simultânea nos dois grupos de pesquisa foi feita uma análise da área de atuação do pesquisador. O objetivo dessa análise é verificar se o pesquisador trabalha com os mesmos temas em ambas as comunidades. Para isso os *abstracts* das publicações da análise anterior foram coletados e submetidos a um *word cloud*. Esta análise foi feita com os mesmos critérios dos casos anteriores.

A Tabela 6.5 mostra os principais termos encontrados para as publicações deste pesquisador em cada comunidade.

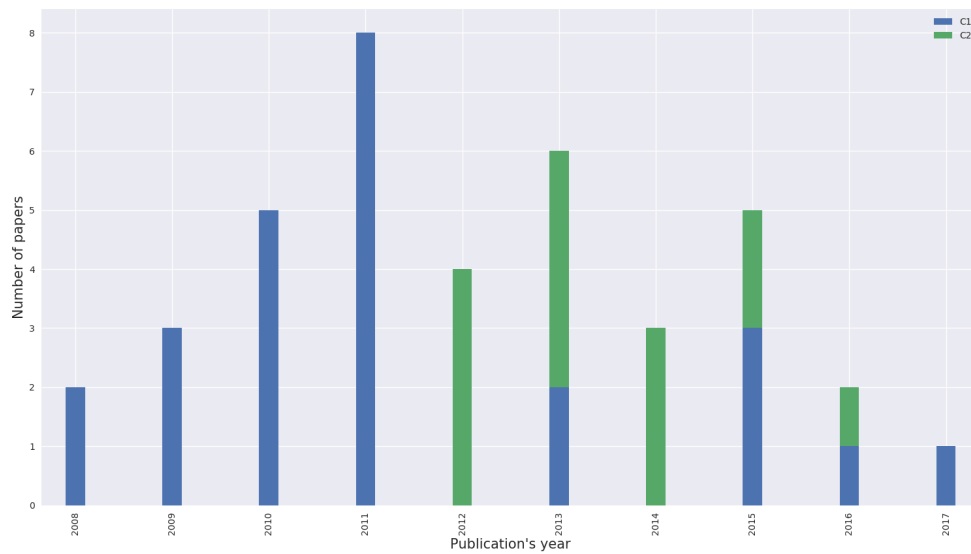


Figura 6.7: Análise temporal das publicações destes pesquisador com suas duas comunidades

Tabela 6.5: Termos utilizados nas publicações deste pesquisador nos dois intervalos de tempo

Abstract Terms C1	Abstract Terms C2
Bounds - codes - constructions - errors - kendall - matrices - modulation - rank - sampling - signals	Bits - bounds - codes - compression - constructions - encoded - errors - matrices - model - rank

Novamente os termos das publicações entre as duas comunidades são semelhantes. Por isso há indícios de que esse pesquisador trabalha com os mesmos temas de pesquisa nas duas comunidades que ele atua. Além disso, ele pode ser caracterizado como um *research-information brokers*, pois ele, de certa forma, é responsável por manter a troca de informações entre as duas comunidades científicas que ele atua.

As análises das sobreposições identificadas pelo NetSCAN nos permitem responder à primeira questão de pesquisa (RQ1: There are researchers who work in two or more communities simultaneously, characterizing research-information brokers?), pois de fato existem pesquisadores que trabalham em mais de uma comunidade de pesquisa. Mais especificamente, na análise da atuação simultânea em múltiplos grupos de pesquisa foi possível identificar pesquisadores que estão atuando em mais de uma comunidade, representando research-information brokers.

Através da detecção de core points foi possível caracterizar alguns pesquisadores

como sendo *central connectors*, respondendo assim a segunda questão de pesquisa (RQ2: Existem pesquisadores com um número elevado de conexões com outros pesquisadores, caracterizando central connectores?)

A partir do estudo das atividades dos pesquisadores foram encontrados indícios de que as comunidades encontradas pelo NetSCAN representam de fato comunidades científicas reais. Estes resultados apontam para a resposta da terceira questão de pesquisa (RQ3: Através de métodos de detecção de comunidades é possível encontrar comunidades científicas reais considerando as atividades realizadas por seus pesquisadores?).

7 Considerações Finais

As duas principais contribuições deste trabalho foram a análise de uma rede social científica e a proposta de um algoritmo para detecção de comunidades sobrepostas em redes bidirecionais.

Neste trabalho foi modelada uma rede social científica (DBLP), onde os elementos representam pesquisadores e suas ligações como relacionamentos de coautoria. Para essa modelagem foi adotado um grafo bidirecional para possibilitar a representação de níveis de influência diferentes entre os pesquisadores.

Um dos objetivos da análise dessa rede social científica era compreender a sua estrutura e como os pesquisadores se influenciam no meio acadêmico. Para tal, inicialmente, foi proposto e desenvolvido um algoritmo para análise da estrutura da rede considerando questões de performance e viabilidade em função do grande volume de dados dessa rede. Primeiramente, foi visto que a rede possui múltiplas componentes conexas, com uma componente dominante que contém 84.2% dos vértices. A partir do estudo da topologia da rede foi possível perceber que a rede possui características de uma rede livre de escala e que possui nós muito mais influentes do que a média do total de nós da rede.

Em um segundo momento, foi desenvolvido um algoritmo de agrupamento por densidade, apoiado pelo SGBD Neo4j e baseado no DBSCAN. Este possui adaptações para trabalhar em grafos com arestas direcionadas e com grande volume de dados. Os resultados obtidos no desenvolvimento deste trabalho apontam para a viabilidade da solução proposta.

Após testar seu comportamento básico, o algoritmo foi executado na rede DBLP. Neste estudo de caso foram encontradas diversas comunidades científicas e pesquisadores influentes. Após observar que 31.5% das comunidades encontradas possuem sobreposições com outras comunidades, foi conduzida uma análise para caracterizar estas sobreposições. Nesta análise diferentes tipos e configurações de sobreposições entre comunidades foram encontrados. Existem casos de sobreposições relacionados a mais de duas comunidades, múltiplos vértices sobrepostos e sobreposições de *border* e *core points*.

As causas reais destas sobreposições foram investigadas através do estudo individual dos pesquisadores envolvidos. Para isto os históricos de publicações destes pesquisadores foram coletados e analisados. Em muitos casos a sobreposição foi causada por um pesquisador que mudou de uma comunidade para outra. Nestes casos o pesquisador possui publicações com as duas comunidades, porém, em períodos distintos de tempo. Em alguns destes casos foi possível detectar uma mudança na área de atuação do pesquisador sobreposto. Para caracterizar essa mudança de área os *abstracts* das publicações foram coletados e submetidos a uma análise por *word cloud*. Também foram encontrados casos de mudança de comunidades sem esta alteração na área de atuação.

Outra causa real de sobreposição observada está relacionada a pesquisadores que trabalham em múltiplas comunidades simultaneamente. Para caracterizar este caso foi analisado um pesquisador que publicou trabalhos no mesmo período de tempo com duas comunidades distintas. Estas publicações ocorreram em anos consecutivos, o que dá indícios de que este pesquisador trabalhou simultaneamente nas duas comunidades.

Em trabalhos futuros uma evolução temporal das componentes conexas encontradas deve ser realizada. Esta atividade tem como objetivo verificar se as comunidades menores possuem uma grande tendência de serem incluídas na componente conexa dominante. Deseja-se também elaborar uma estratégia, baseada em métricas de grafos, para indicar a parametrização ótima no refinamento dos grupos, já que o esforço computacional nesta etapa pode aumentar consideravelmente. Por fim, avaliações adicionais devem ser realizadas, incluindo pessoas de diferentes regiões e países para que se possa explorar a solução em profundidade.

Bibliografia

ALMEIDA, R. et al. Recomendação de recursos educacionais para grupos: buscando soluções em redes sociais. In: *Anais do XXVII Simpósio Brasileiro de Informática na Educação (SBIE 2016)*. [s.n.], 2016. Disponível em: <https://doi.org/10.5753%2Fecbie.sbie.2016.996>.

BARABÁSI, A.-L.; BONABEAU, E. Scale-free networks. *Sci. Am.*, v. 288, n. 5, p. 50–59, 2003.

BHAT, S. Y.; ABULAISH, M. A density-based approach for mining overlapping communities from social network interactions. In: *Proceedings of the 2Nd International Conference on Web Intelligence, Mining and Semantics*. New York, NY, USA: ACM, 2012. (WIMS '12), p. 9:1–9:7. ISBN 978-1-4503-0915-8. Disponível em: <http://doi.acm.org/10.1145/2254129.2254143>.

CROSS, R.; PARKER, A. *The Hidden Power of Social Networks: Understanding how Work Really Gets Done in Organizations*. Harvard Business School Press, 2004. ISBN 9781591392705. Disponível em: <https://books.google.com.br/books?id=vQ3mM4Vpix8C>.

DERÉNYI, I.; PALLA, G.; VICSEK, T. Clique percolation in random networks. *Physical Review Letters*, jan. 2005. Disponível em: <http://link.aps.org/doi/10.1103/PhysRevLett.94.160202>.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: . [S.l.]: AAAI Press, 1996. p. 226–231.

FALKOWSKI, T.; BARTH, A.; SPILIOPOULOU, M. Dengraph: A density-based community detection algorithm. In: *In Proc. of the 2007 IEEE / WIC / ACM International Conference on Web Intelligence*,. [s.n.], 2007. p. 112–115. Disponível em: <http://www.witi.cs.uni-magdeburg.de/~tfalkows/publ/2007/WI-FalBarSpi07.pdf>.

FORTUNATO, S. Community detection in graphs. *Physics Reports*, v. 486, n. 3, p. 75 – 174, 2010. ISSN 0370-1573. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0370157309002841>.

FREEMAN, L. C. Centrality in social networks conceptual clarification. *Social Networks*, Elsevier BV, v. 1, n. 3, p. 215–239, jan 1978. Disponível em: <https://doi.org/10.1016%2F0378-8733%2878%2990021-7>.

GIALAMPOUKIDIS, I. et al. Community detection in complex networks based on dbscan* and a martingale process. In: *2016 11th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*. [S.l.: s.n.], 2016. p. 1–6.

GRABOWICZ, P. A. et al. Social features of online networks: The strength of intermediary ties in online social media. *PLoS ONE*, Public Library of Science (PLOS), v. 7, n. 1, p. e29358, jan 2012. Disponível em: <https://doi.org/10.1371/journal.pone.0029358>.

GUÉRCIO, H.; STRÖELE, V. Desenvolvimento de um modelo semântico para recomendação baseado em grafos. In: *BraSNAM - 5º Brazilian Workshop on Social Network Analysis and Mining, XXXVI Congresso da Sociedade Brasileira de Computação*. [S.l.: s.n.], 2016. p. 103–114.

GUILLE, A. Information diffusion in online social networks. In: *Proceedings of the 2013 Sigmod/PODS Ph.D. symposium on PhD symposium - SIGMOD13 PhD Symposium*. ACM Press, 2013. Disponível em: <https://doi.org/10.1145/2483574.2483575>.

HAN, J.; KAMBER, M.; PEI, J. *Data mining concepts and techniques, 3a edition*. Waltham, Mass.: Morgan Kaufmann Publishers, 2012. Disponível em: http://www.amazon.de/Data-Mining-Concepts-Techniques-Management/dp/0123814790/ref=tmm_hrd_title_0?ie=UTF8&qid=1366039033&sr=1-1.

HLAOUI, A.; WANG, S. A direct approach to graph clustering. In: *Neural Networks and Computational Intelligence*. IASTED/ACTA Press, 2004. p. 158–163. Disponível em: <http://dblp.uni-trier.de/db/conf/nci/nci2004.html#HlaouiW04>.

KITCHENHAM, B.; CHARTERS, S. *Guidelines for performing Systematic Literature Reviews in Software Engineering*. 2007.

LEY, M. The DBLP computer science bibliography: Evolution, research issues, perspectives. In: *String Processing and Information Retrieval*. Springer, 2002. p. 1–10. Disponível em: https://doi.org/10.1007%2F3-540-45735-6_1.

LI, X. et al. Community detection in large social networks based on relationship density. In: *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*. IEEE, 2016. Disponível em: <https://doi.org/10.1109%2Fcompsac.2016.82>.

LUO, T. et al. Detecting community structure based on edge betweenness. In: *2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. IEEE, 2011. Disponível em: <https://doi.org/10.1109%2Ffskd.2011.6019678>.

MEENA, J.; DEVI, V. S. Overlapping community detection in social network using disjoint community detection. In: *2015 IEEE Symposium Series on Computational Intelligence*. [S.l.: s.n.], 2015. p. 764–771.

NEWMAN, M. E. J.; GIRVAN, M. Finding and evaluating community structure in networks. *Physical Review E*, American Physical Society (APS), v. 69, n. 2, feb 2004. Disponível em: <https://doi.org/10.1103%2Fphysreve.69.026113>.

OLIVEIRA, D. M. de et al. Uma estratégia não supervisionada para previsão de eventos usando redes sociais. In: *SBBD*. [S.l.: s.n.], 2015.

PAPADOPOULOS, S. et al. Community detection in social media. *Data Min. Knowl. Discov.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 24, n. 3, p. 515–554, maio 2012. ISSN 1384-5810. Disponível em: <http://dx.doi.org/10.1007/s10618-011-0224-z>.

Puig-Centelles, Anna and Ripolles, Oscar and Chover, Miguel. Surveying the identification of communities. *Int. J. Web Based Communities*, Inderscience Publishers, Inderscience Publishers, Geneva, SWITZERLAND, v. 4, n. 3, p. 334–347, jul. 2008. ISSN 1477-8394. Disponível em: <http://dx.doi.org/10.1504/IJWBC.2008.019193>.

SEMAAN, G. S. *Algoritmos para o Problema de Agrupamento Automático*. Instituto de Computação, Universidade Federal Fluminense: Tese de doutorado. Tese de Doutorado, 2013.

STRÖELE, V. et al. Extração de informações para aprimorar a previsão de relacionamentos em rede social científica. In: *XXX Simpósio Brasileiro de Banco de Dados - Short Papers, Petrópolis, Rio de Janeiro, Brasil, October 13-16, 2015*. [S.l.: s.n.], 2015. p. 75–80.

STRÖELE, V. et al. Information extraction to improve link prediction in scientific social networks. In: *2016 IEEE 20th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. [S.l.: s.n.], 2016. p. 515–520.

STRÖELE, V.; ZIMBRÃO, G.; SOUZA, J. M. Group and link analysis of multi-relational scientific social networks. *Journal of Systems and Software*, Elsevier BV, v. 86, n. 7, p. 1819–1830, jul 2013. Disponível em: <https://doi.org/10.1016%2Fj.jss.2013.02.024>.

WANG WENZHONG TANG, B. S. J. F. C.; WANG, Y. Review on community detection algorithms in social networks. In: *2015 IEEE International Conference on Progress in Informatics and Computing (PIC)*. [S.l.: s.n.], 2015. p. 551–555.

WASSERMAN, S.; FAUST, K. *Social network analysis: Methods and applications*. Cambridge university press, 1994. v. 8. Disponível em: http://scholar.google.com/scholar.bib?q=info:gET6m8icitMJ:scholar.google.com/&output=citation&hl=en&as_sdt=0,5&as_vis=1&ct=citation&cd=0.

XIE, J.; KELLEY, S.; SZYMANSKI, B. K. Overlapping community detection in networks: the state of the art and comparative study. *CoRR*, abs/1110.5813, 2011. Disponível em: <http://arxiv.org/abs/1110.5813>.

YANG, J.; LESKOVEC, J. Defining and evaluating network communities based on ground-truth. In: *2012 IEEE 12th International Conference on Data Mining*. [S.l.: s.n.], 2012. p. 745–754. ISSN 1550-4786.

YIN, R. *Case Study Research: Design and Methods*. SAGE Publications, 2009. (Applied Social Research Methods). ISBN 9781412960991. Disponível em: <https://books.google.com.br/books?id=FzawIAdilHkC>.

ZACHARY, W. An information flow model for conflict and fission in small groups1. v. 33, 11 1976.