CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

ALGORITMOS ESPECTRAIS DE AGRUPAMENTO EM REDES SOCIAIS DE COAUTORIA

Juliana Maria de Sousa Costa

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Tecnologia, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de Mestre

Orientadores

Leonardo Silva de Lima
Rafael Garcia Barbastefano

Rio de Janeiro

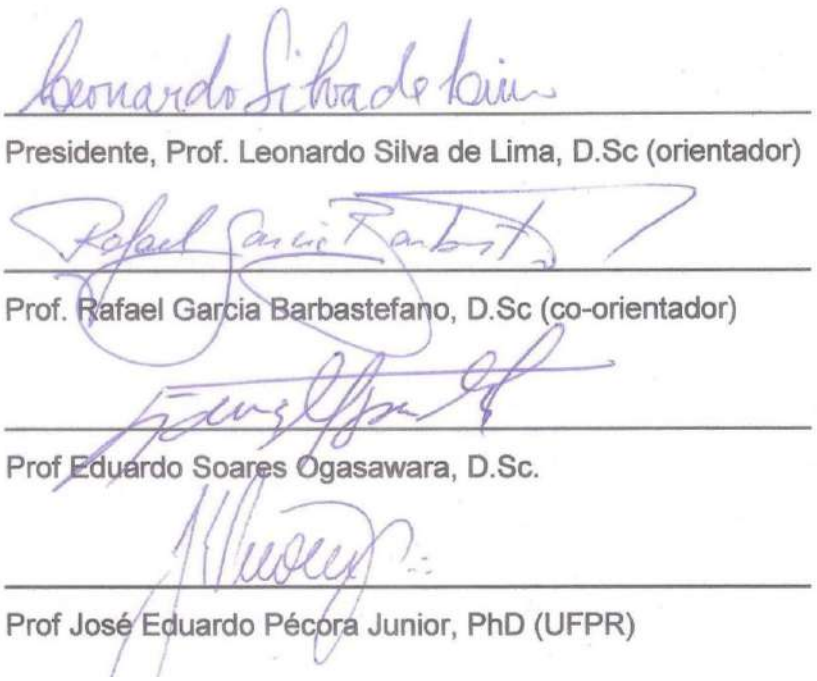
Maio de 2014

ALGORITMOS ESPECTRAIS DE AGRUPAMENTO EM REDES SOCIAIS DE COAUTORIA

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Tecnologia do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ, como parte dos requisitos necessários à obtenção do título de Mestre.

Juliana Maria de Sousa Costa

Aprovada por:



Presidente, Prof. Leonardo Silva de Lima, D.Sc (orientador)

Prof. Rafael Garcia Barbastefano, D.Sc (co-orientador)

Prof Eduardo Soares Ogasawara, D.Sc.

Prof José Eduardo Pécora Junior, PhD (UFPR)

Rio de Janeiro

Maio de 2014

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

C837 Costa, Juliana Maria de Sousa
Algoritmos espectrais de agrupamento em redes sociais de coautoria / Juliana Maria de Sousa Costa.—2014.
xiv, 108f. : il.color. , grafs. , tabs. ; enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2014.
Bibliografia : f. 104-108
Orientadores : Leonardo Silva de Lima
Rafael Garcia Barbastefano

1. Teoria dos Grafos. 2. Algoritmos. 3. Redes Sociais. 4. Análise espectral. I. Lima, Leonardo Silva de (Orient.). II. Barbastefano, Rafael Garcia (Orient.). III. Título.

CDD 511.5

DEDICATÓRIA

A Deus, o autor da vida, que me fortalece a cada dia e é o principal Mestre para mim.

A meus pais, que sempre permaneceram incansáveis em seu apoio, dedicação e amor em toda a minha vida. Vocês que me ensinaram tanto sobre a vida e o valor de tudo, agradeço imensamente por serem parte de quem sou.

“Eu o fortalecerei e o ajudarei; eu o segurarei com a minha mão direita vitoriosa.”

Isaías 41:10b

AGRADECIMENTOS

Gostaria de agradecer a todos aqueles que colaboraram de alguma forma para a realização dessa dissertação, seja através de conhecimentos, conselhos, apoio, e até mesmo pela compreensão nas minhas ausências.

Ao meu orientador Leonardo Lima pela amizade, orientação, apoio prestado e paciência durante esses dois anos. Agradeço pela confiança e por acreditar que teríamos um bom resultado final, mesmo quando tudo ainda parecia sem sentido ou os resultados não surgiam conforme esperado. Obrigada pela dedicação prestada, pelas reuniões, correções e ajudas que foram pouco a pouco permitindo que esse trabalho fosse formado.

Ao meu co-orientador Rafael Barbastefano, que sempre se mostrou disponível, pronto a me atender, ajudar nos tantos problemas que eu levava, inclusive nos momentos finais. Obrigada pelo cuidado, carinho, paciência e atenção, além de todas as ajudas e ideias inestimáveis que contribuíram para o desenvolvimento dessa Dissertação, a cada estalar de dedos, as soluções apareciam.

A todo o corpo docente do Programa de Pós-Graduação em Tecnologia do CEFET/RJ, pois contribuíram para minha formação.

À minha irmã pelo seu amor, cuidado e que muitas vezes, apesar das reclamações, compreendeu a minha ausência e falta de tempo. Obrigada pela sua amizade e cuidado em todos os momentos.

Aos meus amigos que estiveram presentes, nos momentos bons e ruins, oferecendo apoio, cuidado, paciência, atenção, alegrias e que muito contribuíram para tornar essa caminhada mais alegre e tranqüila. À Esquerda Festiva que, sem dúvidas, foi a melhor turma de Mestrado, obrigada pela amizade, carinho, conversas e todos os momentos que passamos juntos, aprendi com cada um e sou grata por fazerem parte da minha história. Agradeço à Patrícia Mattos, pela sua amizade há tantos anos, por todo apoio, pelos conselhos motivadores, por ouvir minhas reclamações, pelo incentivo e por fazer parte de mais essa etapa tão importante.

Ao CNPq pelo financiamento desse trabalho.

RESUMO

ALGORITMOS ESPECTRAIS DE AGRUPAMENTO EM REDES SOCIAIS DE COAUTORIA

Juliana Maria de Sousa Costa

Orientadores:

Leonardo Silva de Lima

Rafael Garcia Barbastefano

Resumo da Dissertação de Mestrado submetida ao Programa de Pós-graduação Tecnologia do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de Mestre

O uso de autovalores e autovetores de matrizes associadas a grafos tem se mostrado útil para extrair informações topológicas de redes. O presente estudo visa a aplicação de algoritmos baseados nos autovetores das matrizes laplaciana e laplaciana sem sinal de grafos para identificar agrupamentos em redes de coautoria. Os experimentos computacionais foram realizados com as redes de coautoria produzidas a partir do sítio do DataBase systems and Logic Programming (DBLP), base de dados que contém diversas publicações da área da Ciência da Computação nos principais periódicos da área desde 1932. Diversas redes de coautoria foram geradas a partir do DBLP em diferentes períodos de tempo com a finalidade de testar os agrupamentos espectrais e identificar as formações de grupos de pesquisa. A qualidade dos resultados obtidos foi avaliada de acordo com a medida de modularidade e, nos casos testados, os métodos espectrais tiveram melhor desempenho que um algoritmo clássico de agrupamento da literatura.

Palavras-Chave: Teoria dos Grafos; Análise de Redes Sociais; Teoria Espectral

Rio de Janeiro

Maio de 2014

ABSTRACT

SPECTRAL CLUSTERING ALGORITHMS IN COAUTHORSHIP SOCIAL NETWORKS

Juliana Maria de Sousa Costa

Advisor(s):

Leonardo Silva de Lima

Rafael Garcia Barbastefano

Abstract of dissertation submitted to Programa de Pós-graduação in Technology - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of Master

The use of eigenvalues and eigenvectors of matrices associated with graphs are useful for extracting topological information networks. The present study aims to apply algorithms based on the eigenvectors of the Laplacian and signless Laplacian matrices of graphs to identify clusters of co-authorship networks. The computational experiments were carried out with the co-authorship networks produced from the site of the DataBase Systems and Logic Programming (DBLP), a database that contains several publications from the field of Computer Science in the major journals since 1932. Several co-authorship networks were generated from the DBLP at different periods of time in order to test the spectral clusters and identify the structure of research groups. The quality of the results was evaluated according to the measure of modularity and, in the cases tested, the spectral methods performed better than a classical clustering algorithm in the literature.

Keywords: Graph Theory; Social Networks Analysis; Spectral Theory

Rio de Janeiro

May 2014

Sumário

Capítulo I - Introdução 1
I.1 Trabalhos relacionados 3
I.2 Objetivos 4
I.3 Metodologia 4
Capítulo II - Conceitos básicos 9
II.1 Teoria dos grafos 9
Capítulo III - Redes Sociais 15
III.1 Análise de Redes Sociais 16
III.1.1 Medidas de centralidade 17
III.1.2 Medidas de coesão 21
III.1.3 Hipótese do mundo pequeno 23
III.1.4 Redes aleatórias 26
III.2 Redes de coautoria 28
Capítulo IV - Agrupamento 31
IV.1 Algoritmo de partição *k-means* 32
IV.2 Algoritmo espectral 33
IV.2.1 Autovalores de matrizes 34
IV.2.2 Matrizes de Adjacência e de Grau 35
IV.2.3 Matriz Laplaciana 35
IV.2.4 Matriz Laplaciana sem Sinal 36
IV.2.5 O Algoritmo 36
IV.3 Medidas para análise dos resultados 39
IV.3.1 Modularidade 40
IV.3.2 Modelo de Programação Linear Inteira 42
Capítulo V - Descrição da Base de dados DBLP 43
Capítulo VI - Experimentos computacionais 52
VI.1 Caso 1: redes seguindo a Lei de Potência 52
VI.2 Caso 2: 03 redes de coautoria extraídas da literatura 54

VI.3 Caso 3: redes extraídas da base DBLP 69

Capítulo VII - Conclusões 102

VII.1 Melhorias e Trabalhos Futuros 103

Referências Bibliográficas 104

Lista de Figuras

Figura I.1 Fases de desenvolvimento do estudo 5

Figura II.1 (a) Grafo simples; (b) grafo trivial; (c) grafo com aresta múltipla, (d) grafo com laço; (e) grafo orientado 10

Figura II.2 (a) Exemplo de grafo em que o máximo conjunto independente possui cardinalidade 2 (u,v); (b) Exemplo de grafo em que o maior clique é igual a 3 (u, v, w) 11

Figura II.3 (a) Grafo completo com 4 vértices; (b) Grafo completo com 5 vértices 11

Figura II.4 (a) Exemplo de grafo bipartido com $|X|=3$ e $|Y|=2$; (b) Grafo bipartido completo com $|X|=3$ e $|Y|=2$; (c) Estrela com $|X|=1$ e $|Y|=5$ 12

Figura II.5 Exemplo de caminho com 4 vértices 12

Figura II.6 (a) Ciclo de comprimento 3 – Triângulo; Ciclo de comprimento 5 13

Figura II.7 (a) Grafo conexo; (b) Grafo desconexo com 2 componentes 13

Figura II.8 (a) Árvore; (b) Floresta com 2 árvores 13

Figura II.9 (a) Grafo 2-regular com 4 vértices; (b) Grafo 3-regular com 6 vértices 14

Figura II.10 (a) Grafo G com 5 vértices; e (b) seu complementar $\bar{G}$ 14

Figura III.1 Exemplo de centralidade 18

Figura III.2 Grafo desconexo com 3 componentes 22

Figura III.3 Exemplos de coesão sob a ótica de conectividade das redes (a) pouco coesa, (c) muito coesa 23

Figura III.4 Exemplo de rede em que diferentes comunidades estão ligadas por poucas ligações 25

Figura IV.1 Exemplo de agrupamento de um conjunto de dados usando *k-means* 33

Figura IV.2 Exemplo do agrupamento com algoritmo espectral 39

Figura IV.3 Exemplos de redes com Modularidade (a) positiva, $Q=0,4896$ e (b) negativa, ... 41

Figura V.1 Publicações indexadas por ano 44

Figura V.2 Tipos de publicações indexadas na base DBLP 45

Figura V.3 Total de publicações de 1965 a 2013 por autor 47

Figura V.4 Número de publicações por ano 48

Figura V.5 Número de autores que publicaram em cada ano de 1965 a 2013 48

Figura V.6 Porcentagem de publicações em coautoria e individual 49

Figura V.7 Quantidade de artigos com diferentes números de autores 49

Figura V.8 Porcentagem de publicações individuais e em coautoria de cada autor 50

Figura V.9 Número de periódicos em que os autores publicaram de 1965 a 2013 51

Figura V.10 Os 31 periódicos em que os autores mais publicaram 51

Figura VI.1 Resultados dos algoritmos espectrais e Modelo de Programação Linear Inteira 53

Figura VI.2 Rede de pesquisadores do CNPq 55

Figura VI.3 Componente gigante da rede de coautoria de publicações sobre Análise do ciclo de vida 55

Figura VI.4 Componente gigante da rede de coautoria da revista Química Nova 56

Figura VI.5 Gráfico log-log dos graus dos vértices versus a frequência dos graus da (a) Rede do CNPq; (b) Rede de Análise do Ciclo de Vida; e (c) Rede Química Nova 58

Figura VI.6 Valores da medida modularidade para diferentes quantidades de *clusteres* em cada algoritmo na Rede CNPq 63

Figura VI.7 Valores da medida modularidade para diferentes quantidades de *clusteres* em cada algoritmo na Rede ACV 64

Figura VI.8 Valores da medida modularidade para diferentes quantidades de *clusteres* em cada algoritmo na Rede Química Nova 66

Figura VI.9 Partições geradas pelo algoritmo espectral com matriz Laplaciana para a Rede CNPq 67

Figura VI.10 Resultado da maior partição encontrada para a Rede CNPq 68

Figura VI.11 Resultado da segunda maior partição encontrada para a Rede CNPq 69

Figura VI.12 Rede I – Publicações de 1965 a 1994 70

Figura VI.13 Rede II – Publicações de 1995 a 2001 71

Figura VI.14 Rede III – Publicações de 2002 a 2007 71

Figura VI.15 Rede IV – Publicações de 1995 a 2007 72

Figura VI.16 Rede V – Publicações de 2008 a 2013 72

Figura VI.17 Rede VI – Publicações de 1965 a 2013 73

Figura VI.18 Distribuição dos graus da (a) Rede I, (b) Rede II, (c) Rede III 79

Figura VI.19 Distribuição dos graus da (a) Rede IV, (b) Rede V, (c) Rede VI 80

Figura VI.20 Valores da medida modularidade para diferentes quantidades de *clusteres* em cada algoritmo na Rede I (1965-1994) 82

Figura VI.21 Valores da medida modularidade para diferentes quantidades de *clusteres* em cada algoritmo na Rede II (1995-2001) 83

Figura VI.22 Valores da medida modularidade para diferentes quantidades de *clusteres* em cada algoritmo na Rede III (2002-2007) 85

Figura VI.23 Valores da medida modularidade para diferentes quantidades de *clusteres* em cada algoritmo na Rede IV (1995-2007) 87

Figura VI.24 Valores da medida modularidade para diferentes quantidades de *clusteres* em cada algoritmo na Rede V (2008-2013) 89

Figura VI.25 Valores da medida modularidade para diferentes quantidades de *clusteres* em cada algoritmo na Rede VI (1965-2013) 91

Figura VI.26 Partições geradas pelo algoritmo *k-means* para a Rede I 92

Figura VI.27 Resultado da maior partição encontrada para a Rede I 93

Figura VI.28 Partições geradas pelo algoritmo espectral com matriz Laplaciana sem sinal para a Rede II 94

Figura VI.29 Resultado da maior partição encontrada para a Rede II 94

Figura VI.30 Partições geradas pelo algoritmo espectral com matriz Laplaciana para a Rede III 95

Figura VI.31 Resultado da maior partição encontrada para a Rede III 96

Figura VI.32 Partições geradas pelo algoritmo espectral com matriz Laplaciana para a Rede IV 97

Figura VI.33 Resultado da maior partição encontrada para a Rede IV 98

Figura VI.34 Partições geradas pelo algoritmo espectral com matriz Laplaciana para a Rede V 98

Figura VI.35 Resultado da maior partição encontrada para a Rede V 99

Figura VI.36 Partições geradas pelo algoritmo espectral com matriz Laplaciana para a Rede VI 100

Figura VI.37 Resultado da maior partição encontrada para a Rede VI 100

Lista de Tabelas

Tabela I.1 Redes de coautoria utilizadas 7

Tabela VI.1 Redes geradas segundo Lei de Potências 53

Tabela VI.2 Propriedades das redes CNPq, ACV e Química Nova 57

Tabela VI.3 Resultados para os 25 autores com maiores valores de grau normalizado de cada rede 59

Tabela VI.4 Resultados para os 25 autores com maiores valores de proximidade normalizada de cada rede 60

Tabela VI.5 Resultados para os 25 autores com maiores valores de intermediação normalizada de cada rede 61

Tabela VI.6 Resultados da modularidade para diferentes valores de *k* na rede CNPq 62

Tabela VI.7 Resultados da modularidade para diferentes valores de *k* na rede ACV 64

Tabela VI.8 Resultados da modularidade para diferentes valores de *k* na rede Química Nova 65

Tabela VI.9 Propriedades das redes extraídas da base de dados DBLP 73

Tabela VI.10 Valores de Grau, Proximidade e Intermediação normalizados dos 25 autores nas Redes I e II 74

Tabela VI.11 Valores de Grau, Proximidade e Intermediação normalizados dos 25 autores nas Redes III e IV 75

Tabela VI.12 Valores de Grau, Proximidade e Intermediação normalizados dos 25 autores nas Redes V e VI 75

Tabela VI.13 Propriedades das componentes gigantes das redes da base de dados DBLP 76

Tabela VI.14 Resultados dos 10 autores com maiores graus das componentes gigantes das Redes da DBLP 77

Tabela VI.15 Resultados dos 10 autores com maiores proximidades das componentes gigantes das Redes da DBLP 77

Tabela VI.16 Resultados dos 10 autores com maiores proximidades das componentes gigantes das Redes da DBLP 78

Tabela VI.17 Resultados da modularidade para diferentes valores de *k* na Rede I (1965-1994) 81

Tabela VI.18 Resultados da modularidade para diferentes valores de *k* na Rede II (1995-2001) 83

Tabela VI.19 Resultados da modularidade para diferentes valores de *k* na Rede III (2002-2007) 84

Tabela VI.20 Resultados da modularidade para diferentes valores de *k* na Rede IV (1995-2007) 86

Tabela VI.21 Resultados da modularidade para diferentes valores de *k* na Rede V (2008-2013) 88

Tabela VI.22 Resultados da modularidade para diferentes valores de *k* na Rede VI (1965-2013) 90

Capítulo I - Introdução

O estudo da formação de comunidades em diferentes redes é importante, pois permite a identificação de grupos com comportamentos semelhantes e relações de proximidade entre alguns dos vértices, o que sugere relações de parceria entre pesquisadores, empresas, amizades; e também ajuda na detecção das principais áreas de pesquisa em uma rede científica ou comportamentos de compras de clientes em sites, por exemplo. Para essas análises são utilizadas medidas de proximidade e algoritmos que identificam os nós que apresentam características mais similares, agrupando-os em um mesmo *cluster*, e busca separar aqueles com dissimilaridades entre si.

Existem diversos algoritmos apresentados na literatura que são utilizados com o objetivo de agrupar dados de forma tal que represente bem sua estrutura e características, como o *k-means*, laplaciano espectral, *DBSCAN – Density-based spatial clustering* (Agrupamentos baseado em densidade de regiões conexas com alta densidade), métodos baseados em *grid*, cadeias de Markov, *PageRank*, SVM (*Support Vector Machine*) entre outros (DING *et al.*, 2001; FORTUNATO, 2010; GIRVAN; NEWMAN, 2002; HAN; KAMBER; PEI, 2012; KERNIGHAN; LIN, 1970; LIU; STEWART, 2011; NEWMAN, 2004b; SCHAEFFER, 2007; WU *et al.*, 2008).

O agrupamento de dados tem sido aplicado em diferentes áreas como biologia (CHEN; YUAN, 2006; JIANG; SINGH, 2010), mercado de ações (BOGINSKI; BUTENKO; PARDALOS, 2006), segmentação de imagens (SHI; MALIK, 2000) e redes sociais (DUAN *et al.*, 2012; KRINGS *et al.*, 2009; NEWMAN, 2001; RICHARDSON; MUCHA; PORTER, 2009).

Com os avanços computacionais ocorridos nas últimas décadas, os quais permitiram o aumento da capacidade de memória dos computadores e melhorias no desempenho dos *softwares*, tornou-se possível o armazenamento de uma quantidade maior de informações. E, com isso, houve o aparecimento de diversas bases de dados, criando uma necessidade para desenvolvimento e uso de técnicas e ferramentas que facilitem o seu manuseio e permitam a geração de algum conhecimento útil para estudos das áreas das quais são originários.

Existem bases com dados diversos, como informações de compras pela Internet, as quais auxiliam os *sites* a conhecerem os gostos dos seus clientes e seus padrões de compras; dados das bolsas de valores, que permitem uma análise do comportamento das ações e se há correlações entre ações de diferentes empresas; dados de buscas na Internet, através dos quais os motores de buscas podem indicar assuntos para os usuários de acordo com seu histórico. Além desses, dados de redes sociais, como *Facebook*, *MySpace*, *sites* de compartilhamento de informações, são importantes fontes de informação sobre seus usuários. Outras bases importantes são aquelas relacionadas a publicações científicas que trazem dados

sobre os pesquisadores, temas de pesquisas, entre outros, o que contribui para avanços e maior rapidez na busca de artigos, livros e outras publicações.

Com isso, muitas técnicas têm evoluído ao longo dos anos, possibilitando a análise dos dados, identificação de possíveis padrões, classificação e detecção de agrupamentos, entre outras. Além de permitirem a retirada de conhecimento com valor adicionado, mesmo de uma base que, inicialmente, aparentava ter apenas dados soltos e com pouco sentido.

Uma maneira de representar essa grande quantidade de dados é através do uso da Teoria de Grafos e Análise de Redes Sociais. Muitos problemas reais podem ser modelados através desses diagramas, chamados de grafos ou redes, os quais utilizam pontos (ou vértices) conectados por linhas (ou arestas). Alguns exemplos de situações reais são grupos de contatos, em que os vértices representam pessoas e as arestas são as relações estabelecidas entre elas, que podem ser de emprego, amizade, interesses em comum, grupos de pesquisa, etc.

A Análise de Redes Sociais (ARS) apresenta grande importância em estudos de redes diversas em que se busca detectar e analisar as diferentes características desses pontos e as relações estabelecidas entre eles. E, para isso, a ARS utiliza a Teoria dos Grafos que permite a quantificação de certas medidas verificadas na rede, como o grau dos vértices, densidade da rede, entre outras. Além dessas medidas, outras análises podem ser feitas relacionadas a modelos probabilísticos, como o estudo de redes livres de escala, a identificação da hipótese de mundo pequeno e a detecção de agrupamentos ou formação de comunidades.

A Teoria Espectral de Grafos utiliza os autovalores e autovetores de matrizes associadas a grafos para obtenção de propriedades topológicas das redes. Optou-se, nessa dissertação, por usar o algoritmo espectral empregando-se as matrizes Laplaciana e Laplaciana sem sinal. Como no problema de agrupamento dos dados busca-se agrupar aqueles com maior similaridade, em uma rede de autores, dentro dos grupos formados deve existir o maior número possível de ligações, e o mínimo possível de arestas entre grupos. Assim, para o uso da matriz Laplaciana, utilizou-se o vetor de Fiedler, autovetor associado ao segundo menor autovalor da matriz, para ser o vetor de corte dos dados, pois este vetor divide o conjunto de dados em dois grupos de forma a minimizar o número de arestas entre os grupos, ou seja, minimizar o tamanho do corte. E para a matriz Laplaciana sem sinal foi utilizado o autovetor associado ao seu segundo maior autovalor, pois o uso dessa matriz busca maximizar a coesão dentro dos grupos, de forma que haja o maior número possível de ligações entre os vértices do mesmo *cluster*.

Um exemplo de uma grande base de dados que conta com mais de dois milhões de publicações na área de ciências da computação, e que foi utilizada neste estudo, é a *Digital Bibliography & Library Project* (DBLP, 2014). Essa base de dados contém informações bibliográficas sobre os principais periódicos, publicações em conferências, teses e livros da

área de Ciência da Computação, desde o ano de 1936 até os dias de hoje. Além das redes extraídas dessa base de dados, serão utilizadas outras três redes de coautoria retiradas da literatura: redes de pesquisadores do CNPq, rede de autores que publicaram sobre Análise do Ciclo de Vida e rede de pesquisadores que publicaram na Revista Química Nova.

I.1 Trabalhos relacionados

Diversos trabalhos têm sido realizados sobre redes de coautoria, como elas se comportam, análises relacionadas a autores individualmente, os papéis desempenhados pelos cientistas nas redes, a comunicação entre diferentes grupos, e esses estudos utilizam redes de diferentes áreas, como Medicina, Matemática, Física, Conferência, entre outros. Alguns desses trabalhos buscam analisar as relações de colaboração entre autores de uma determinada rede, como eles se relacionam entre si e suas medidas de centralidade, como o estudo feito por CHEONG e CORBITT (2009).

BARABÁSI *et al.* (2002) utiliza algumas métricas, os conceitos de mundo pequeno e leis de potência para avaliar a evolução de uma rede em um período de tempo determinado, como se comporta e os resultados que variam de acordo com o tempo. TOMASSINI e LUTHI (2007) também estudaram a evolução de uma rede, eles utilizaram dados desde a formação de uma rede de pesquisadores de programação genética até um período de 20 anos de sua concepção, analisaram a formação da componente gigante, a distribuição de graus, coeficiente de clusterização e distâncias, os quais possuem valores que são dependentes do tempo. GROSSMAN (2002) publicou um estudo sobre a evolução de uma rede de colaboração em pesquisa de matemática, foram feitas análises estatísticas relacionadas a distribuição de graus dos vértices e avaliada a mudança na colaboração entre os pesquisadores com o tempo, verificando-se o número de artigos, autores, média de autores por artigo, número de artigos feitos em coautoria, etc. Esses trabalhos demonstram resultados interessantes, visto que poucos estudos fazem essa comparação da rede durante um intervalo de tempo.

No presente estudo, entretanto, como o principal objetivo é o emprego de algoritmos de agrupamento, foram realizadas apenas algumas análises nas redes escolhidas. Dentre elas estão medidas de centralidade, coesão, verificação de lei de potência e identificação dos autores mais centrais nos maiores *clusteres*, porém de não foi feita uma investigação mais aprofundada no comportamento das redes com o tempo e investigações relacionadas à prospecção tecnológica, estudos que poderão ser feitos posteriormente.

Outros trabalhos tem o foco voltado para agrupamento em redes, como o de GIRVAN e NEWMAN (2002) em que os autores propõe um algoritmo para a detecção de comunidades em redes, empregando-o em duas redes: de coautoria e em uma rede que representa uma teia alimentar de organismos marinhos. Esse algoritmo utiliza a intermediação de arestas para identificar as comunidades, removendo as arestas com maior valor de intermediação.

LUXBURG (2006) e HAGEN e KAHNG (1992) propuseram o uso de algoritmos espectrais com a matriz Laplaciana e utilizando o vetor de Fiedler. Porém nesses trabalhos foram utilizados outros algoritmos para o particionamento da rede juntamente com o espectral. NEWMAN (2006) propõe um algoritmo espectral com a matriz de adjacência buscando maximizar a modularidade para o agrupamento realizado. Já ALPERT, KAHNG e YAO (1998) utilizam o algoritmo espectral com múltiplos autovetores.

O presente trabalho se diferencia daqueles citados acima, pois utiliza o algoritmo de agrupamento espectral com a matriz Laplaciana sem sinal proposto teoricamente por LIU e STEWART (2011) a partir das entradas do autovetor associado ao segundo maior autovalor desta matriz e realiza uma comparação de desempenho com outros algoritmos já conhecidos na literatura.

I.2 Objetivos

O objetivo deste trabalho é utilizar o algoritmo espectral com as matrizes Laplaciana e Laplaciana sem sinal para a identificação de agrupamento em redes sociais de coautoria, comparando seus resultados com os obtidos por um algoritmo bastante utilizado na literatura, tendo sido escolhido o *k-means*. O algoritmo Laplaciano espectral já tem sido utilizado em diversos artigos de agrupamento (DING *et al.*, 2001; SPIELMAN, TENG, 1996; LUXBURG, 2006), porém o algoritmo utilizando a matriz Laplaciana sem sinal só foi visto no trabalho de LIU e STEWART (2011). Todavia o agrupamento desses autores utilizou cadeias de Markov. Com isso, vale ressaltar, que aplicações de autovetores da matriz Laplaciana sem sinal para agrupamentos em redes é um tema ainda bastante incipiente na literatura, o que justifica o desenvolvimento do tema. Neste trabalho, os algoritmos espectrais foram implementados e a qualidade das partições encontradas foi avaliada a partir da medida de modularidade. Outro objetivo da dissertação é realizar análises de prospecção tecnológica a partir dos grupos de pesquisa identificados pelos algoritmos espectrais.

I.3 Metodologia

A metodologia deste trabalho considerou inicialmente a realização de uma bateria de testes dos algoritmos espectrais em redes geradas artificialmente, com tamanhos que variam de aproximadamente 20 a 80 vértices, e os resultados obtidos foram comparados com um modelo exato de Programação Linear Inteira (PLI) proposto por NASCIMENTO; TOLEDO; DE CARVALHO (2010). Nos casos de redes maiores, onde o modelo de PLI não consegue gerar a solução ótima, fez-se uso da medida de qualidade denominada modularidade (NEWMAN, 2006), que mede a coesão dentro dos grupos formados, para avaliar os resultados dos algoritmos espectrais e *k-means*. Nestes casos, foram geradas 09 redes reais, 03 extraídas da

literatura e 06 a partir da base DBLP, com número de vértices variando de 184 a 5065. Os algoritmos espectrais foram implementados utilizando os softwares *Matlab* e *OCTAVE* e os parâmetros de entrada dos algoritmos foram a matriz de adjacência da rede e o número de agrupamentos desejados.

O desenvolvimento do presente estudo pode ser entendido através das fases descritas pela Figura I.1.

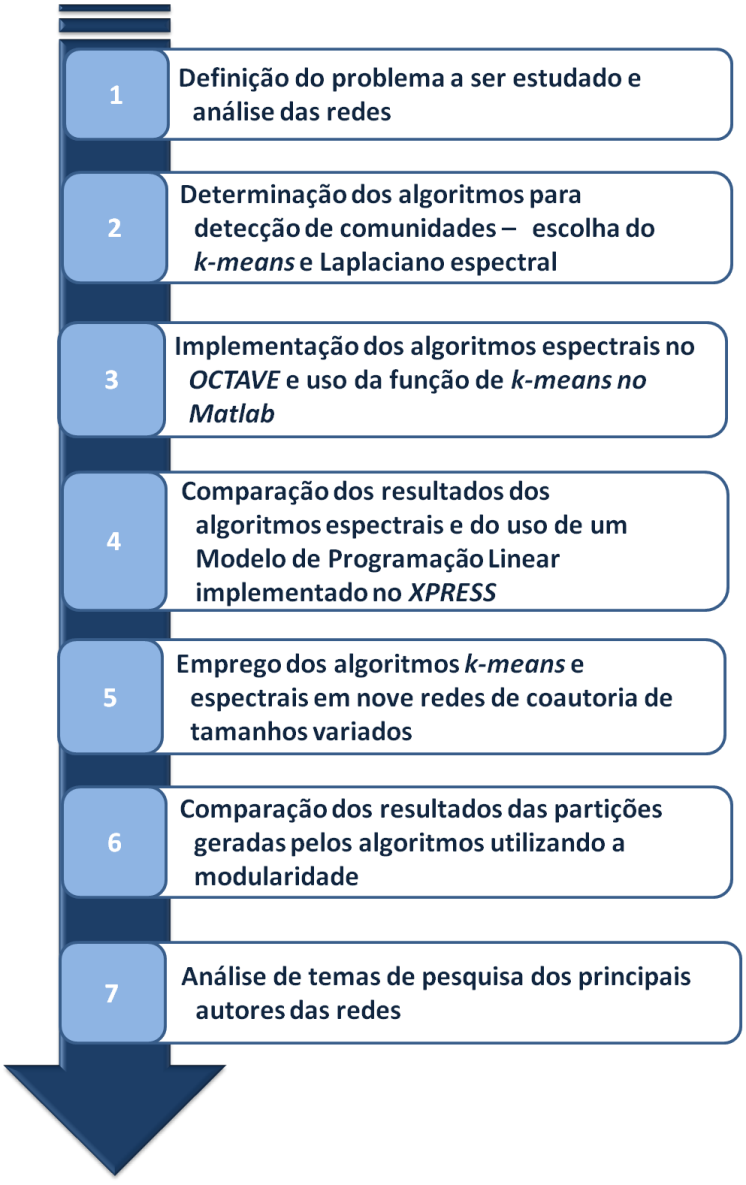


Figura I.1 Fases de desenvolvimento do estudo

Fonte: Adaptado de TAHA (2007)

A primeira fase do trabalho foi a definição do problema a ser estudado. Dentre as diferentes opções de redes sociais existentes, escolheu-se trabalhar com as redes de coautoria, buscando assim detectar grupos de autores que costumam publicar juntos e analisar algumas características presentes nessas redes. Foram utilizadas apenas redes conexas,

dessa forma dentre as redes escolhidas retirou-se a componente gigante de cada uma, garantindo o estudo de uma boa parte da rede de forma que esta não possuísse mais do que uma componente, ou seja, subgrupos não conectados. Após a geração de todas as redes foram feitas algumas análises relacionadas às suas características, centralidade, coesão, verificação da hipótese de mundo pequeno e modelos probabilísticos. Para a realização dessas análises foram utilizados os *softwares Pajek* e *Microsoft Excel*, este último principalmente para a geração de gráficos de leis de potência, e aquele para as demais medidas analisadas nas redes, como graus, densidade, coeficiente de clusterização, e outras.

A segunda etapa determinou os critérios de similaridade e a escolha dos algoritmos para detecção de comunidades: o *k-means* e o espectral com uso das matrizes Laplaciana e Laplaciana sem sinal. Destaca-se que o algoritmo espectral baseado na matriz Laplaciana é bastante conhecido na literatura e tem diversas aplicações, enquanto o emprego da matriz Laplaciana sem sinal para agrupamentos em redes foi encontrado somente no trabalho de LIU e STEWART (2011), o que justifica o desenvolvimento do tema. Na terceira etapa foram implementados os algoritmos espectrais no *OCTAVE* e modificada a função *k-means* no *Matlab* para facilitar o uso no trabalho.

Na quarta etapa foi realizada uma comparação entre os resultados dos algoritmos espectrais, implementados no *OCTAVE*, com o Modelo de Programação Linear Inteira de agrupamento proposto por NASCIMENTO; TOLEDO; DE CARVALHO (2010) e implementado no *XPRESS*. Porém, como esse modelo apresenta um tempo muito alto para encontrar a solução ótima em redes com centenas de vértices, para esta etapa os algoritmos foram aplicados em redes menores, com número de vértices variando de 20 a 80. Foram geradas cem redes aleatoriamente pelo *software Pajek* com as seguintes características: livres de escala, não direcionadas e com grau médio igual a dois.

Para o emprego dos algoritmos espectrais e *k-means*, foram selecionadas nove redes de coautoria para análise e emprego dos algoritmos de agrupamento, três dessas são redes retiradas da literatura: a rede de pesquisadores do CNPq com publicações relacionadas a grafos (BARBOSA *et al.*, 2011); a rede de pesquisadores com publicações relacionadas à Análise do Ciclo de Vida indexadas na base de dados *ISI/Web of Knowledge* (SOUZA; BARBASTEFANO, 2011); e a rede de pesquisadores com publicações na Revista Química Nova (SOUZA; BARBASTEFANO; LIMA, 2012). As outras seis redes foram obtidas através da base de dados *Digital Bibliography & Library Project* (DBLP, 2014). A Tabela I.1 apresenta as redes utilizadas e o número de autores presente nas componentes conexas gigantes de cada uma delas.

Tabela I.1 Redes de coautoria utilizadas

	Redes	Número de autores
Literatura	CNPq	207
	ACV	955
	Química Nova	2171
DBLP	Rede I (1965 - 1994)	184
	Rede II (1995-2001)	285
	Rede III (2002-2007)	1094
	Rede IV (1995-2007)	1895
	Rede V (2008-2013)	2680
	Rede VI (1965-2013)	5065

Os experimentos computacionais consistiram de duas etapas. Na primeira etapa, foram feitas diversas análises das redes de coautoria relacionadas às suas características, centralidade, coesão e modelos probabilísticos, apresentadas no Capítulo VICapítulo VI - . A segunda etapa também foi dividida em duas partes:

a) Teste dos algoritmos espectrais em 100 redes geradas pelo *software Pajek* e comparação com os resultados de um Modelo de Programação Linear Inteira;

b) Emprego dos algoritmos *k-means* e espectral nas nove redes de coautoria e avaliação dos resultados dos grupos gerados, utilizando modularidade.

Os algoritmos foram rodados com diferentes números de *clusteres* buscando-se encontrar aquele que apresentasse a melhor solução. A rotina de cálculo da medida de qualidade modularidade foi obtida em *Matlab* no sítio da *Strategic Engineering Research Group* (SERG, 2013). Para o cálculo dessa medida foram utilizadas duas funções: uma que a partir dos arquivos de partição separa os vértices em seus respectivos grupos, e outra que recebe como entradas esses grupos e a matriz de distância da rede, e através do seu valor pode-se comparar qual algoritmo apresentou o melhor resultado.

O presente trabalho está organizado da seguinte forma: o Capítulo II aborda os conceitos básicos para o entendimento de grafos e suas principais características, também empregados na Análise de Redes Sociais. O Capítulo III apresenta a definição de Análise de Redes Sociais, seu emprego em diversas áreas e algumas das medidas mais utilizadas para avaliação de redes, além do conceito de redes de coautoria e a importância de estudos dessa área. No Capítulo IV, os principais algoritmos de agrupamento de interesse desse estudo são apresentados e a medida de qualidade de um agrupamento denominada modularidade é

definida. No Capítulo V será descrita a base de dados DBLP com algumas análises sobre as publicações indexadas. No Capítulo VI serão apresentados os experimentos computacionais realizados, contando com algumas análises sobre a topologia das redes e os resultados apresentados pelas técnicas de agrupamento. Por fim, as conclusões do estudo são apresentadas e possíveis trabalhos futuros indicados.

Capítulo II - Conceitos básicos

Neste capítulo serão abordados alguns conceitos básicos relacionados à Teoria dos Grafos, os quais contribuem para uma melhor compreensão do presente trabalho.

A Seção II.1 abordará os conceitos mais gerais sobre Teoria dos Grafos, alguns dos diferentes tipos de grafos existentes e definições que serão importantes para o desenvolvimento do estudo.

II.1 Teoria dos grafos

Os conceitos clássicos de Teoria dos Grafos são de fundamental importância para estudos de Análise de Redes Sociais. Diversos invariantes de grafos como clique e conjunto independente, podem trazer informações acerca da topologia da rede como nível de agrupamento dos vértices ou atores das redes sociais. Portanto, esses e outros invariantes podem ser usados para denotar diferentes propriedades estruturais da rede.

Os conceitos apresentados nesta seção foram baseados em ABREU *et al.* (2007) e BONDY e MURTY (2008).

Um grafo (ou rede) consiste em um par ordenado $G=(V(G), E(G))$, em que $V(G)$ é o conjunto de vértices com cardinalidade $|V(G)|=n$ e $E(G)$ é o conjunto de arestas que os conectam com cardinalidade $|E(G)|=m$. Estes parâmetros n e m são denominados de ordem e tamanho do grafo, respectivamente. Um grafo em que $m=0$ é chamado totalmente desconexo ou vazio, já um grafo em que $V(G)$ é um conjunto unitário e $E(G)=\emptyset$ é chamado grafo trivial.

Define-se $e=\{i,j\}$ (ou e_{ij}) $\epsilon\ E(G)$ como uma aresta que conecta os vértices i e $j \epsilon\ V(G)$, e diz-se que e incide em i e j. Os vértices conectados por uma aresta são chamados adjacentes.

Um grafo direcionado ou um dígrafo, denotado por GD, é um par $(V(GD),E(GD))$, onde $V(GD)$ é um conjunto finito de vértices e $E(GD)$ é o conjunto de arcos e representa uma relação binária em $V(GD)$. Nessa família de grafos, o sentido obrigatório de fluxo é importante e é determinado pelo sentido atribuído ao arco. Assim, uma aresta e_{ij} indica que o fluxo ocorre no sentido do vértice i (origem) para o vértice j, sendo o vértice j adjacente ao vértice i.

Em algumas aplicações é comum aparecerem os grafos direcionados, como no caso dos problemas de fluxo em redes (fluxo máximo, fluxo a custo mínimo, caminho mínimo e outros). Em uma rede, com um vértice que é a origem e outro que é o destino, em que as arestas existentes entre esses vértices e os intermediários possuem capacidades de fluxo, o problema de fluxo máximo busca maximizar o fluxo que sai da origem e chega ao destino. No problema de fluxo a custo mínimo, o objetivo é levar uma quantidade (menor ou igual à máxima) da origem ao destino a um custo total mínimo. Já em problemas de caminho mínimo, procura-se percorrer o menor caminho possível da origem até chegar ao destino.

Exemplos de grafos direcionados podem ser uma mensagem de *email*, onde os endereços de *email* dos usuários que trocam mensagens são os vértices da rede, e as mensagens de *email* trocadas são as arestas. DUAN *et al.* (2012), em seu trabalho, utilizou um banco de dados de *email*s da empresa ENRON com o objetivo de aplicar um algoritmo de particionamento.

Outro exemplo é uma cadeia de suprimentos em que a matéria-prima sai do fornecedor para a fábrica que a utilizará, mas não ocorre o contrário. Em algumas redes de suprimentos podem ocorrer os fluxos de *i* a *j*, como de *j* a *i*, mas o custo de transporte, por exemplo, ser diferente, de uma cidade A até outra cidade B custar X reais, mas o custo para ir de B até A pode ser de Y reais, sendo Y≠X. Um exemplo de aplicação do estudo de redes em cadeias de suprimentos é o trabalho de CHOI e HONG (2002), eles apresentaram um estudo de caso utilizando as redes de suprimentos de três empresas do ramo automobilístico: Honda, Acura e DaimlerChrysler. Os autores analisaram as redes, apresentando a estrutura e o comportamento de cada uma delas em relação a seus fornecedores e fizeram algumas proposições relacionadas a seus modos de operação, como a formalização, centralização e complexidade existentes em tais relações.

Já os grafos que não possuem orientação em suas arestas são denominados não direcionados, como os que serão utilizados para representar as redes sociais neste trabalho, em que a orientação do fluxo não importa, e sim a relação existente entre os vértices da rede.

Na Figura II.1, podem ser visualizados alguns exemplos dos tipos de grafos citados anteriormente.

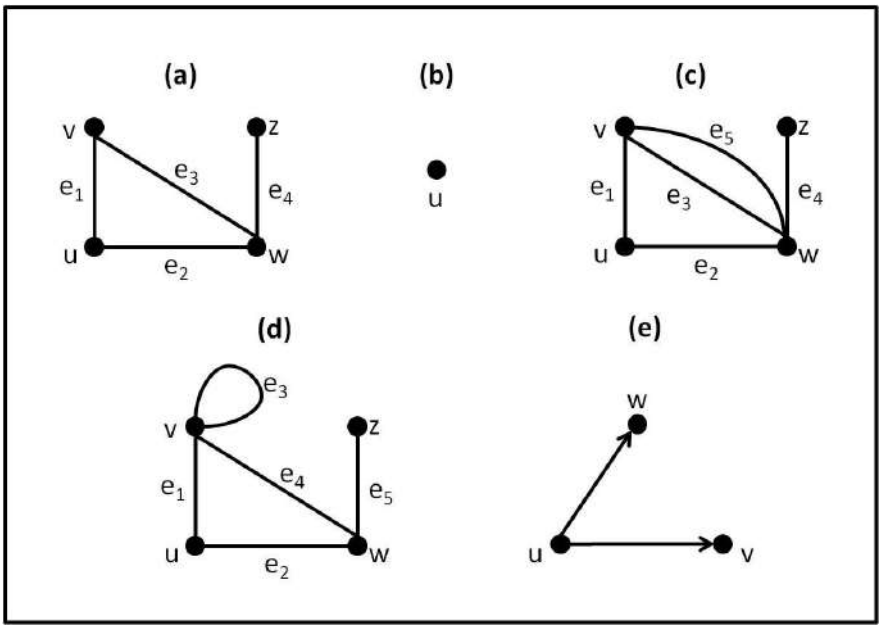


Figura II.1 (a) Grafo simples; (b) grafo trivial; (c) grafo com aresta múltipla, (d) grafo com laço; (e) grafo orientado

Neste trabalho serão considerados apenas os grafos não orientados e simples, ou seja, sem laços, sem arestas múltiplas ou paralelas, como o apresentado na Figura II.1a.

Um conjunto independente de um grafo é um subconjunto do conjunto de vértices cujos elementos não são dois a dois adjacentes. A cardinalidade do máximo conjunto independente é denominada de número de independência de G e denotada por $\alpha(G)$. Através da Figura II.2, pode-se observar que no grafo (a), a cardinalidade do máximo conjunto independente é igual a 2, sendo esse conjunto formado pelos vértices v e w, pois esse é o maior conjunto de vértices no grafo que não são adjacentes, pois não há arestas conectando-os.

Já uma clique em um grafo é um conjunto de vértices em que existem todas as ligações possíveis entre eles, todos estão ligados entre si, o tamanho da maior clique em G é denotado por $\omega(G)$. No grafo (b) nota-se que o tamanho da maior clique é igual a 3, sendo formada pelos vértices u, v e w. Uma clique forma um grafo completo, pois o grau de todos os vértices pertencentes a ela é igual a $(n-1)$.

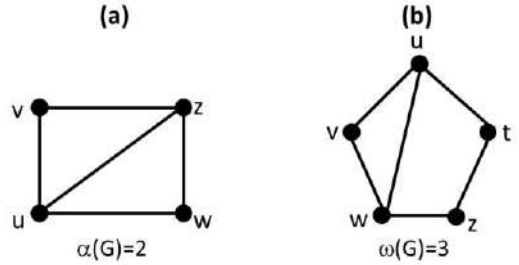


Figura II.2 (a) Exemplo de grafo em que o máximo conjunto independente possui cardinalidade 2 (u,v); (b) Exemplo de grafo em que o maior clique é igual a 3 (u, v, w)

O grau de um vértice j é o número de arestas incidentes em j, ou seja, é o número de vizinhos de j, e é denotado por d_j. Um vértice de grau zero é chamado de vértice isolado.

Algumas famílias especiais de grafos que serão utilizadas ao longo do texto são definidas aqui.

Um grafo completo com n vértices, denotado por K_n, é um grafo simples onde quaisquer pares de vértices são adjacentes. Na Figura II.3 podem ser vistos alguns exemplos.

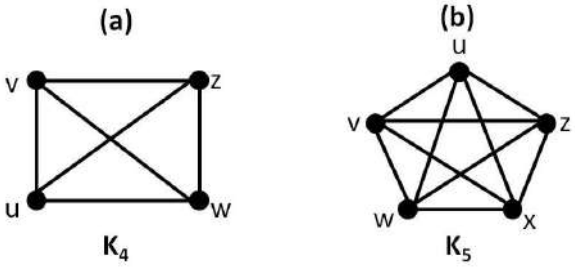


Figura II.3 (a) Grafo completo com 4 vértices; (b) Grafo completo com 5 vértices

Um grafo bipartido é um grafo que seu conjunto de vértices pode ser particionado em dois subconjuntos X e Y onde cada aresta tem um ponto inicial em X e um ponto final em Y, tal que a partição (X,Y) é denominada bipartição de G. Ou seja, os vértices pertencentes ao mesmo subconjunto não apresentam ligações entre si, apenas os vértices pertencentes a dois subconjuntos diferentes podem ter arestas conectando-os. O grafo bipartido G com partição (X,Y) é denotado por $G[X,Y]$. Se todos os vértices de X estão ligados a todos os vértices de Y, G é denominado bipartido completo. Uma estrela, denotada por S_n, é um grafo bipartido completo onde $|X|=1$ e $|Y|=n-1$. A Figura II.4 apresenta alguns exemplos de grafos bipartidos.

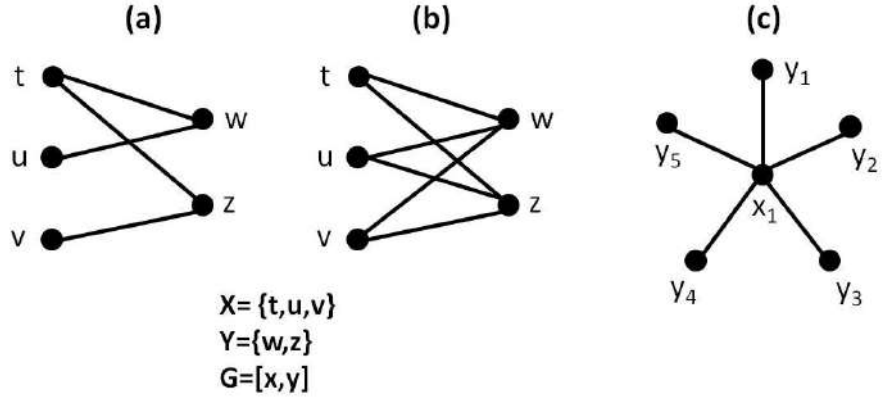


Figura II.4 (a) Exemplo de grafo bipartido com $|X|=3$ e $|Y|=2$; (b) Grafo bipartido completo com $|X|=3$ e $|Y|=2$; (c) Estrela com $|X|=1$ e $|Y|=5$

Um caminho em um grafo G é uma sequência $v_0e_1v_1e_2...v_{e-1}e_ev_e$, cujos termos são alternadamente vértices (v) e arestas (e) de G, tal que v_{i-1} e v_i são os vértices terminais de e_i, $1\leq i\leq m$. Se $v_0=x$ e $v_e=y$, diz-se que esse caminho conecta x a y. Neste grafo dois vértices são adjacentes se eles são consecutivos dentro da sequência e, caso contrário, não são adjacentes. Este grafo é denotado por P_n, um exemplo de caminho P_4 pode ser visto na Figura II.5.

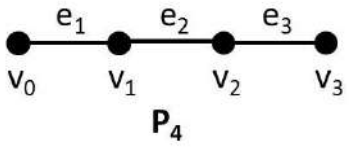


Figura II.5 Exemplo de caminho com 4 vértices

O ciclo é um caminho fechado, é um grafo simples cujos vértices podem ser arrumados numa sequência cíclica. É denotado por C_n, o ciclo C_3 é chamado triângulo. Os ciclos de comprimento ímpar não podem ser bipartidos, já os de comprimento par são grafos bipartidos. Ou seja, os vértices do grafo podem ser arrumados de modo a haver dois conjuntos de vértices

com apenas ligações entre aqueles que pertencem a conjuntos diferentes e não há ligações dentro da partição. Na Figura II.6, a seguir, tem-se dois exemplos de ciclos.

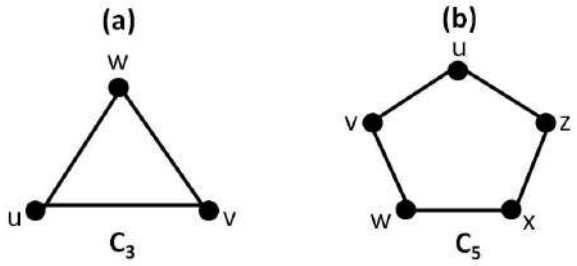


Figura II.6 (a) Ciclo de comprimento 3 – Triângulo; Ciclo de comprimento 5

Diz-se que G é conexo, se para cada par de vértices, existe um caminho ligando esses vértices. Caso contrário, o grafo G é desconexo. Se não há caminho entre dois vértices então existem subgrafos conectados, e cada subgrafo onde seus vértices não possuem ligações com vértices dos demais subgrafos são chamados componentes conexas. Em um grafo desconexo, a maior componente, ou seja, o subgrafo que apresentar maior número de vértices, é chamada de componente gigante do grafo. A Figura II.7 apresenta um exemplo de (a) grafo conexo formado pelos vértices: v, u, w e z, e um (b) grafo desconexo formado pelos vértices p, q, r, s e t, neste a componente gigante possui 3 vértices (r,s,t):

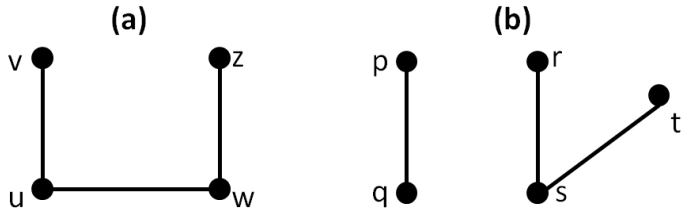


Figura II.7 (a) Grafo conexo; (b) Grafo desconexo com 2 componentes

Árvore é um grafo conexo sem ciclos, em que $m=(n-1)$ arestas. Floresta é um grafo desconexo sem ciclos. Na Figura II.8 podem ser vistas a árvore e floresta.

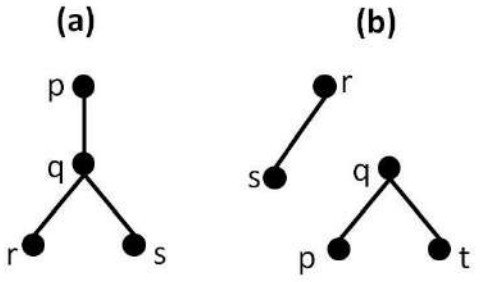


Figura II.8 (a) Árvore; (b) Floresta com 2 árvores

Grafo k-regular é um grafo em que todo vértice de G possui grau igual a k. A Figura II.9 apresenta dois exemplos de grafos regulares.

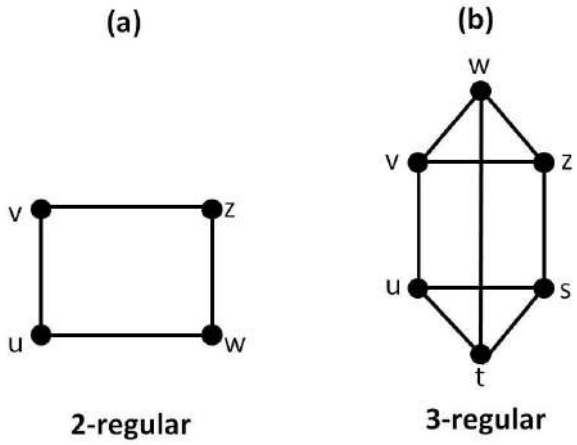


Figura II.9 (a) Grafo 2-regular com 4 vértices; (b) Grafo 3-regular com 6 vértices

Sendo G um grafo simples, o complementar $\bar{G}$ de G é um grafo simples cujo conjunto de vértices é V e cujas arestas são os pares de vértices não adjacentes em G. Na Figura II.10 podem ser observados um grafo G e seu complementar $\bar{G}$.

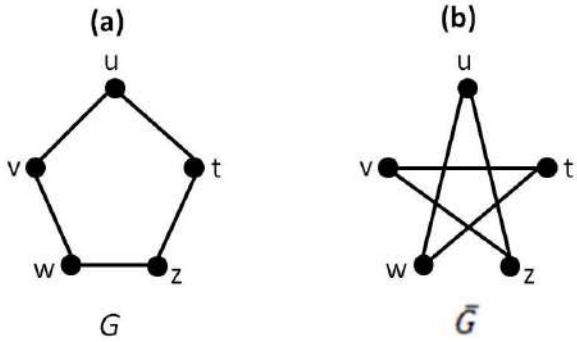


Figura II.10 (a) Grafo G com 5 vértices; e (b) seu complementar $\bar{G}$

O diâmetro de um grafo conexo, denotado por, $diam(G)$ é a distância maximal entre dois vértices, ou seja, é a máxima distância entre os comprimentos mínimos que ligam cada par de vértices. Quando G for desconexo, então $diam(G)=\infty$.O comprimento do maior ciclo em G é denominado circunferência de G, e o comprimento do menor ciclo é denominado cintura de G.

Capítulo III - Redes Sociais

Uma rede social é uma estrutura social que representa alguma relação entre um conjunto de atores sociais. Esses atores podem ser pessoas, empresas, países, Instituições de ensino; e a relação existente entre eles pode ser de amizade, proximidade geográfica, trabalho realizado em conjunto, compra e venda de produtos, fornecimento, estudos e pesquisas realizados em parceria. A Internet, *World Wide Web*, é uma das grandes redes reais atuais, a partir da qual podem ser geradas diversas redes e através da qual há uma grande quantidade de dados em tráfego.

O estudo de redes sociais começou na década de 1930, quando pesquisadores de Harvard exploravam padrões de relações interpessoais e a formação de cliques, e pesquisadores em Manchester investigavam relações de comunidades em tribos e vilas. Estudos nessa área foram reunidos entre as décadas de 1960 e 1970 quando a Análise de Redes Sociais contemporânea foi estruturada (SCOTT, 2000). Porém apenas nas décadas mais recentes a Análise de Redes Sociais passou a ter maior popularidade.

Jacob Moreno foi o primeiro a utilizar sociogramas, diagramas que representam as redes sociais, esse foi um meio encontrado por ele de facilitar a visualização das propriedades formais dessas redes (SCOTT, 2000). Esses diagramas são formados por pontos e linhas, os pontos chamados de vértices representam os atores da rede e as linhas chamadas de arestas, as relações estabelecidas entre eles.

Através do estudo das redes sociais, diversas características e informações sobre a rede podem ser detectadas, como afinidades ou gostos em comum, grupos de amigos dentro de uma rede de alunos de uma faculdade, temas de estudo de pesquisadores que desenvolvem trabalhos em conjunto e que fazem parte de uma comunidade científica mais abrangente, relações de fornecimento entre empresas. Também é possível realizar algumas análises relativas à medidas de centralidade, intermediação e proximidade, as quais estão relacionadas à importância do vértice na rede e que serão melhor explicadas na Subseção III.1.1, e também medidas de coesão, Subseção III.1.2.

Uma rede de autoria é uma rede social bipartida em que um conjunto de vértices é formado por autores e o outro conjunto é formado por artigos. Um autor está ligado aos artigos que ele publicou e não há ligações entre autores ou entre artigos. Uma rede de coautoria é formada a partir de uma rede de autoria, transforma-se a rede bipartida em uma rede unipartida, os vértices passam a representar apenas os autores e esses passam a estar ligados por uma aresta caso tenham publicado algum trabalho em conjunto. As redes de coautoria são importantes para o desenvolvimento de novas pesquisas e disseminação de informações na área acadêmica, pois em diversos projetos de pesquisa é necessária a participação de mais de um pesquisador. E, além disso, trabalhos realizados em conjunto

podem ser realizados com maior rapidez e, devido às conexões de cada autor com outros pesquisadores da rede, as informações sobre as pesquisas devem fluir de forma mais eficiente.

As próximas seções buscarão abordar o conceito de Análise de Redes Sociais e sua aplicação, e também as redes de coautoria.

III.1 Análise de Redes Sociais

A Análise de Redes Sociais (ARS) é um ramo da sociologia que analisa as relações sociais usando ferramentas da Teoria dos Grafos, que auxiliam na avaliação dos resultados através da quantificação de medidas verificadas na rede, como o grau dos vértices, densidade da rede, medidas de centralidade, entre outras. Além destas medidas, os conceitos apresentados no Capítulo II são aplicados na Análise de Redes Sociais, assim a rede pode ser conexa ou desconexa, representar uma clique, ser direcionada, etc.

O conceito de rede ressalta que cada membro da rede pode possuir ligação com outro, e cada um destes pode possuir ligações com alguns ou muitos outros. A perspectiva de redes destaca as relações existentes entre os atributos dos seus atores, não focando em avaliações individuais, mas busca observar os processos relacionais, a estrutura e características da rede estudada (WASSERMAN; FAUST, 1994).

A utilização da Teoria dos Grafos para estudo de diferentes redes de sistemas reais também vem demonstrando crescente aplicação, visto que os conceitos estudados podem ser aplicados na Análise de Redes Sociais permitindo uma avaliação mais aprofundada da rede (FORTUNATO, 2010). As redes sociais são consideradas um tipo de rede complexa, visto que elas representam redes reais.

O estudo de redes sociais possibilita a identificação de relacionamentos entre os atores da rede, como estes estão ligados entre si, os tipos de ligação entre eles, as trocas de informações que podem ocorrer, a detecção de grupos formados dentro dessas redes, e a evolução da rede ao longo do tempo (apesar de a maioria dos estudos sobre redes considerarem-na como objeto estático, sem serem feitas análises relacionadas à sua mudança com o tempo). Dessa forma, os resultados de estudos que levam em consideração essas características presentes nas redes são de grande importância em diferentes áreas, pois podem ajudar a determinar relações entre os atores, entre Instituições e países, além de ser possível analisar grupos e encontrar temas e assuntos que podem estar sendo discutidos em determinadas redes de pesquisa.

Segundo WASSERMAN e FAUST (1994), o crescente interesse que tem sido atribuído à Análise de Redes Sociais nos últimos anos é devido ao foco desse estudo nos relacionamentos entre entidades e os padrões e implicações desses relacionamentos. Os pesquisadores têm percebido que a utilização de redes para representar seus conjuntos de dados auxilia no melhor entendimento dos padrões e relacionamentos existentes, além de

conseguirem obter uma melhor visualização das informações estudadas. As relações presentes em uma rede podem ser de vários tipos, como afetiva, econômica, de autoria, de trabalho, entre outras.

Com o uso de redes para representar diferentes tipos de relações de acordo com os dados utilizados, pode-se analisar algumas características, como o fluxo de informação que passa pelos vértices e grupos formados dentro das redes. Esse fluxo de informação entre os vértices é percebido através das ligações presentes entre eles, um vértice que esteja isolado, por exemplo, não troca informação com nenhum outro da rede.

A Análise de Redes Sociais também tem sido bastante aplicada em diversas áreas de estudo, como no marketing e em Engenharia de Produção (WASSERMAN; FAUST, 1994).

Em marketing, alguns trabalhos têm buscado analisar e entender a estrutura dos relacionamentos comerciais e interpessoais, e também como se dá a comunicação nessas redes formadas e como essas relações podem influenciar o comportamento dos envolvidos (IACOBUCCI; OSTROM, 1996; WEBSTER; MORRISON, 2004). Estudos de processos de negócios é um exemplo de aplicação de redes sociais em Engenharia de Produção, como pode ser visto no trabalho de AALST e SONG (2004), em que os autores utilizaram redes sociais de mineração para melhorar a análise dos processos através do uso de técnicas e ferramentas que os auxiliaram no controle de dados e outros processos.

Na Subseção III.1.1 serão abordadas algumas medidas de centralidade bastante utilizadas na análise de redes e que auxiliam na detecção dos vértices mais influentes.

III.1.1 Medidas de centralidade

Através da Análise de Redes Sociais é possível fazer a identificação de atores que desempenham um papel importante na rede, que poderá ser como influenciador, disseminador de informações, entre outros, o que os faz ter mais visibilidade que os demais.

Centralidade foi definida por BAVELAS (1948), segundo o autor um vértice mais central é aquele que está localizado em um caminho mais curto para os outros pares de vértices da rede. A análise de centralidade dos vértices da rede permite uma melhor avaliação do papel dos atores em uma rede, de forma que atores com alta centralidade possuem mais facilidade ao acesso de informação e também têm maior potencial de disseminação das informações que chegam a eles.

FREEMAN (1978) exemplifica o conceito de centralidade utilizando a estrutura de uma estrela ou roda, o ponto no centro de uma estrela será o mais central que quaisquer outros em outra rede de tamanho similar. Porém, o centro da estrela está relacionado a alguma forma da estrutura, assim sendo, o problema é determinar uma posição dos vértices que seria única para se determinar a centralidade da rede. Três formas de analisar essa posição estão relacionadas ao grau desse vértice, pois esse terá o maior grau possível; às distâncias geodésicas entre

este vértice e o maior número possível de outros vértices (intermediação); e como este vértice central posiciona-se a uma distância mínima dos demais, estará o mais próximo deles possível (proximidade). Sendo assim, o mesmo ponto em uma estrela poderia ser considerado o mais central considerando-se as três medidas utilizadas.

Métricas para avaliar o grau de influência relativa dos vértices de uma rede são chamadas de medidas de centralidade, dentre as quais destacam-se o grau do vértice, proximidade e intermediação.

A centralidade de grau é a forma mais simples de medir a centralidade do vértice, pois seu valor é o número de arestas incidentes no vértice. Esses vértices são menos dependentes dos demais e podem conseguir informações de diferentes caminhos na rede. De acordo com SOUZA *et al.*(2012), “em uma rede de coautoria esse grau vai indicar o total de atores da rede que publicaram em parceria com um determinado ator. O grau de centralidade poderá variar, portanto, de 0 (publicação sem parceria com outros atores) até $n-1$ (publicação com todos os demais atores da rede excluindo a si próprio).”

Essa medida, porém, conduz a resultados muito focados no indivíduo sem considerar a rede como um todo e, além disso, nem sempre o ator da rede que possui o maior grau será o que terá o papel mais importante na disseminação de informações. Por exemplo, tem-se uma rede de coautoria, em que dois autores Pedro e João, que possuem graus máximos na rede, estão ligados cada um a 50 outros autores. Porém, Pedro está ligado a um grupo que não possui ou quase não possui ligações com os demais autores da rede, já João está ligado a autores que por sua vez também possuem diversas outras ligações com demais autores da rede (que não estão conectados a João), percebe-se então que João será um autor que poderá influenciar um maior número de autores que Pedro, devido às ligações que poderá percorrer através de seus adjacentes.

Como pode ser visto na Figura III.1, o grau máximo da rede é igual a 3, porém o vértice v_4, apesar de possuir grau igual a 2, apresenta uma função mais importante, pois conecta dois grupos da rede que não poderiam comunicar-se caso o v_4 não estivesse presente. Outras medidas de centralidade podem ser mais eficientes em casos como este.

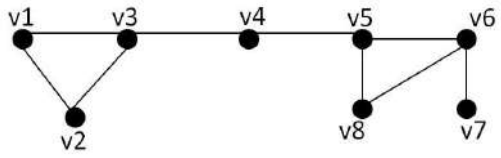


Figura III.1 Exemplo de centralidade

Quando são analisadas redes de tamanhos diferentes e são comparados os resultados apresentados por elas, não faz sentido comparar os resultados absolutos de graus dos vértices de uma rede que possui 500 vértices, por exemplo, com os de uma rede que possui 2000

vértices. Isso porque, o número máximo que os vértices dessa segunda rede podem se ligar é bem maior que os da primeira, dessa forma, em situações como essa, em que são comparadas redes de tamanhos diferentes, é necessário utilizar a medida de grau normalizada proposta por FREEMAN (1978), apresentada na equação (1), onde $a(v_i,v_j)=1$, se os vértices v_i e v_j estão ligados, e $a(v_i,v_j)=1$, caso contrário:

$$d_i' = \frac{\sum_{j=1}^{n} a(v_i,v_j)}{(n-1)}. \qquad (1)$$

Outra medida de centralidade bastante utilizada na literatura é a centralidade de intermediação (*betweenness*). A medida de intermediação de um vértice é igual ao número de caminhos mais curtos a partir de todos os vértices para todos os outros que passam por esse vértice, considera-se que a centralidade de um integrante de uma rede depende da maneira em que este integrante se faz necessário como um elo ou uma ligação entre diferentes grupos pertencentes à rede, levando em conta o menor caminho possível. Em outras palavras, quanto mais fluxo de informação passar por um integrante da rede, mais central este será. Nesta métrica, a determinação da centralidade de um ator está associada ao quociente do número de caminhos mais curtos entre todos os pares de atores em uma rede que incluem o ator em questão e o número de todos os caminhos curtos entre todos os pares da rede (FREEMAN, 1977), dada pela equação (2):

$$Int(v) = \sum_{i\neq j\neq v, i\neq v} \frac{\sigma_{ij}(v)}{\sigma_{ij}}, \qquad (2)$$

onde, $\sigma_{i,j}$ é o número de caminhos geodésicos que ligam os vértices i e j, $\sigma_{ij}(v)$ representa o número de caminhos geodésicos que ligam os vértices i e j passando por v.

Assim como o grau, quando são comparadas vértices pertencentes a redes de tamanhos diferentes, utiliza-se a medida normalizada de intermediação, equação (3):

$$Int(v)' = \sum_{i\neq j\neq v, i\neq v} \frac{\sigma_{ij}(v)}{\sigma_{ij}\times(n-1)}. \qquad (3)$$

O menor caminho, ou geodésico, entre dois vértices de uma rede, é o caminho de comprimento mínimo, sendo este comprimento a distância entre dois vértices. O diâmetro de um grafo conexo é a distância maximal entre dois vértices, ou seja, é a máxima distância entre os comprimentos mínimos que ligam cada par de vértices (BONDY; MURTY, 2008).

De acordo com a centralidade de intermediação, um ator estará numa posição privilegiada quando ele está presente em grande parte dos caminhos geodésicos entre outros pares na rede, ou seja, para que o vértice chegue a um outro deverá passar por ele, é o que ocorre na Figura III.1, em que os vértices v_5, v_6, v_7 e v_8 para chegarem aos vértices v_1, v_2 ou v_3 precisam passar pelo v_4. Assim, quanto mais caminhos geodésicos passarem por tal vértice, maior será sua importância na rede. Todavia, quando o vértice está presente em apenas alguns caminhos geodésicos, mas não em todos que conectam um certo par de vértices, então seu poder de controle do fluxo de informação diminui (FREEMAN, 1977).

Vértices com valor de intermediação alto são interessantes porque controlam o fluxo de informação na rede e podem ser necessários para a disseminação desta.

A centralidade de proximidade é inversamente proporcional à soma das distâncias geodésicas de um vértice v a todos os outros da rede. Nesse sentido, esta mede o quão próximo um vértice está dos demais; quanto menor for a distância entre o vértice e cada um dos demais, maior será a medida de proximidade. Deste modo, um indivíduo é observado como o centro da rede, se ele requisita poucos intermediários para contatar outros. O cálculo desta medida utiliza a soma de todas as distâncias entre este vértice e cada um dos demais, como pode ser visto na equação (4).

$$\boldsymbol{Prox}(\boldsymbol{v_i}) = \frac{\mathbf{1}}{\sum_{j=1}^{n} \boldsymbol{d_G}(\boldsymbol{v_i v_j})}. \qquad (4),$$

onde $d_G(v_i,v_j)$=distância entre os vértices *i* e *j*.

A medida de proximidade normalizada $Prox(v_i)'$ é definida como:

$$\boldsymbol{Prox}(\boldsymbol{v_i})' = \frac{(\boldsymbol{n-1})}{\sum_{j=1}^{n} \boldsymbol{d_G}(\boldsymbol{v_i v_j})} \qquad (5)$$

Assim, quanto mais próximos os vértices estão uns dos outros, mais facilmente as informações chegarão a cada ponto da rede. Um nó com alto valor de proximidade possui mais liberdade em relação às influências de outros nós e maior capacidade de ações independentes (WASSERMAN; FAUST, 1994).

Essas duas últimas medidas de centralidade (intermediação e proximidade) apontam diferentes meios para se calcular a centralidade de um vértice: a proximidade indica os vértices que têm acesso rápido à informação, e a intermediação aqueles que controlam o fluxo de informação. Dependendo do tipo de rede e das avaliações que são necessárias ao estudo, o uso de um ou outro valor pode ser mais relevante. Existe também o valor de grau de proximidade e intermediação totais de um grafo, que, neste caso, servem como forma de comparação com outros grafos, indicando se a informação é alcançada mais rapidamente pelos atores e/ou existe concentração de fluxos de informação.

Há diferentes formas de se medir o papel central dos atores em uma rede. Por exemplo, quando a troca realizada entre os vértices através da rede deve atingir longas distâncias, ou seja, é importante que vértices que estejam distantes recebam a informação ou o objeto da troca, a posição global deve ter um peso maior que uma centralidade local. Já em uma organização hierárquica, aquelas relações de poder que um vértice exerce sobre os demais devem ter um peso maior que as outras. Assim, é importante que se observe as características da rede avaliada para que seja escolhida a melhor forma de avaliar sua centralidade, pois não faz sentido avaliar todas as redes sob uma mesma medida.

Uma outra medida de centralidade é a de autovetor. Essa medida que foi definida por BONACICH (1987), mede o valor de um vértice de acordo com as relações que este possui com outros vértices que possuem importância na rede. Ou seja, um nó será mais central se

tiver como vizinhos nós que possuem muitas ligações e que estão ligados a outros que também possuem muitos outros vizinhos. A centralidade de autovetor, entretanto, não será objeto de estudo neste trabalho.

III.1.2 Medidas de coesão

Um dos principais pontos na Análise de Redes Sociais é averiguar quais atores estão relacionados entre si e quais não estão, e o porquê de manterem relações com determinados vértices e não com outros, além daqueles que se mantém isolados.

O conceito de coesão é definido de diferentes maneiras na literatura. As definições buscam expressar a densidade de arestas na conexão dos vértices do grafo ou de um subconjunto deles. MIZRUCHI (1993) diz que “a coesão numa rede é constituída pelas ligações existentes e as obrigações que ocorrem devido ao contato continuado.” Segundo MOODY e WHITE (2003) “a coesão estrutural é o número mínimo de atores que, se removidos, desconectam o grupo”. Para NOOY; MRVAR; BATAGELJ (2005), uma rede coesa é aquela que possui muitas arestas, o que resulta em uma estrutura mais firme.

A densidade procura mostrar a coesão de uma rede na medida em que considera o total de ligações que existem entre os componentes dessa rede dentro do total de possibilidades existentes. Ou seja, densidade é a proporção entre o número de arestas (m) existentes em uma rede e o número máximo de arestas que seria possível nessa rede $(n(n-1)/2)$. Dessa forma, quanto mais densa for uma rede, mais coesa ela será, equação (6). A densidade máxima ocorre quando todos os vértices da rede estão conectados, ou seja, a rede é completa (NOOY; MRVAR; BATAGELJ, 2005).

$$dens(G) = \frac{2m}{n(n-1)}. \qquad (6)$$

O Grau médio é uma medida de coesão da rede, assim como a densidade, e quanto maior for o grau médio, maior será a coesão da rede, equação (7). Enquanto o grau mede o número de ligações de um vértice, o grau médio é a soma dos graus dos vértices dividida pelo total de vértices da rede.

$$d(G) = \frac{\sum_{i=1}^{n} d_i}{n}. \qquad (7)$$

O conceito de coesão é importante para o estudo de agrupamentos em redes, pois redes muito coesas serão menos propensas a se dividirem em subgrupos, já redes menos coesas podem apresentar subgrupos com cliques, como ocorre em redes de coautoria onde cada grupo de autores que publicou um artigo em conjunto representa uma clique. Além disso, quando se estuda métodos para identificar agrupamento das redes, esses buscam dividir os grupos de acordo com as ligações que esses possuem entre si, com o objetivo de agrupar

aqueles mais similares e separar aqueles que possuem menos ou nenhuma relação. Com isso, percebe-se que há maior coesão dentro dos grupos do que entre eles.

As medidas de coesão baseiam-se no modo como os vértices estão interconectados. Em redes reais percebe-se que existem algumas formações de grupos dentro da rede, em que os vértices apresentam ligações com outros membros do mesmo grupo e poucas ou nenhuma ligação com outros vértices da rede. Ademais, o número de arestas presentes na rede e os graus dos vértices também demonstram resultados interessantes sobre a rede e representam medidas de coesão. Uma rede coesa é aquela que possui muitas arestas, o que resulta em uma estrutura mais firme (NOOY; MRVAR; BATAGELJ, 2005).

Quanto à conectividade, uma rede é conexa se para cada par de vértices existe um caminho ligando esses vértices. Em geral, pode haver diferentes caminhos com diferentes comprimentos ligando os vértices. A conectividade de arestas em um grafo é o número mínimo de arestas que ao serem removidas desconectam o grafo. Já a conectividade de nós em um grafo é o número mínimo de nós que ao serem removidos desconectam o grafo. Quanto mais conexo for um grafo, menor a probabilidade de ele se tornar desconexo, pois para que isso ocorra será necessária a saída de diversos vértices ou arestas presentes na rede.

Os vértices que, ao serem retirados, separam a rede em mais de uma componente são chamados vértices de corte (WASSERMAN; FAUST, 1994). A componente gigante da rede, que consiste no maior subgrafo conexo desta rede, indica o percentual de atores que possuem ligações com os demais pertencentes a essa componente, direta ou indiretamente. Em algumas redes de coautoria, a componente gigante pode abranger mais de 90% dos vértices da rede (NEWMAN, 2001). A Figura III.2 mostra um exemplo de grafo desconexo com 3 componentes, este grafo possui 2 componentes gigantes, pois existem duas componentes de mesmo tamanho que são as maiores do grafo, com 4 vértices cada.

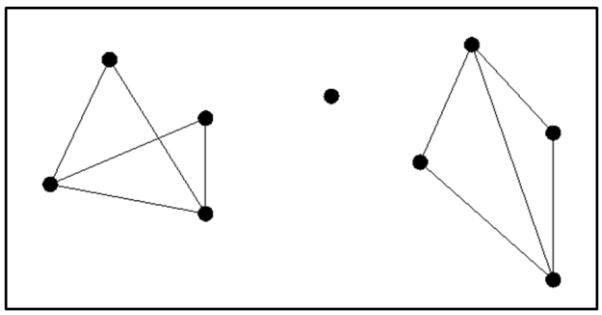


Figura III.2 Grafo desconexo com 3 componentes

Percebe-se que os conceitos de coesão e conectividade estão relacionados, visto que um grupo é coeso à medida que possui menor probabilidade de se tornar desconexo, o que está relacionado à conectividade dos vértices, ou seja, é necessário que mais vértices e/ou arestas sejam retirados para que o grafo se torne um grafo desconexo ou trivial. Essa relação é discutida por alguns autores, onde também se fala da coesão estrutural de um grupo definida

como o número mínimo de caminhos independentes que ligam cada par de vértices do grupo (MOODY; WHITE, 2003; WHITE; HARARY, 2001). Dois caminhos são independentes, internamente disjuntos, se um caminho não contém os vértices internos do outro. Dois caminhos são ditos disjuntos se não possuem nem vértices nem arestas em comum.

Neste trabalho, utiliza-se a definição de coesão encontrada em LIU e STEWART (2011): dado c subconjuntos induzidos de vértices $V1, ..., Vc$, a coesão deste agrupamento é dado pela soma dos pesos das arestas dentro de cada subgrupo de vértices de um grafo. Desta forma, diferentes definições de agrupamentos levam a diferentes valores de coesão, pois quanto mais similares forem os grupos internamente, mais forte será a coesão e, portanto, melhor é o agrupamento. Observe que esta medida é mais adequada para os propósitos desse trabalho uma vez que está diretamente relacionada ao modo de agrupamento dos vértices.

A Figura III.3 apresenta exemplos de redes com diferentes graus de coesão do ponto de vista da conectividade: a primeira rede (a) é uma rede pouco coesa, pois se o vértice $v4$ for retirado, a rede é desconectada e se formam três grupos que não se comunicam; a segunda rede (b) é um pouco mais coesa, pois nesse caso é necessário que dois vértices sejam removidos para que a rede seja desconectada, por exemplo, vértices $v4$ e $v5$; já a terceira rede (c) é um grafo completo, sendo assim uma rede muito coesa, pois todos os vértices possuem ligações com todos, seria um exemplo de uma rede com fortes relações sociais, como uma rede de amigos em que todos se conhecem.

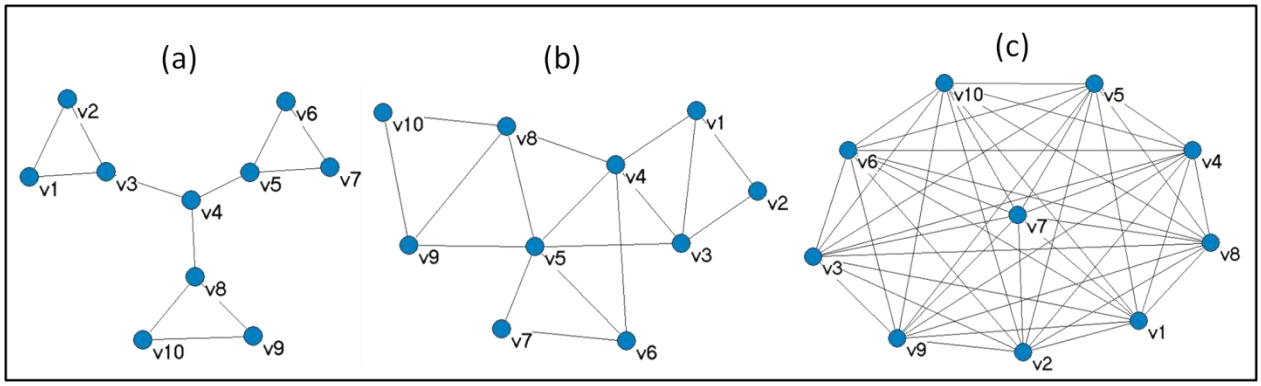


Figura III.3 Exemplos de coesão sob a ótica de conectividade das redes (a) pouco coesa, (c) muito coesa

III.1.3 Hipótese do mundo pequeno

Uma análise interessante a ser feita no estudo de redes sociais está relacionada à verificação da hipótese de mundo pequeno. De acordo com ela, uma distância média pequena entre os vértices, por volta de seis, significa que a comunidade representada pelo grafo se ajusta a tal hipótese (MILGRAM, 1967). Segundo TRAVERS e MILGRAM (1969), o modo mais simples de formular esse problema é pensar na probabilidade de duas pessoas quaisquer,

selecionadas ao acaso em uma população bem grande, como a de um país, se conhecerem, ou possuírem amigos em comum. Os resultados do estudo de MILGRAM (1967) demonstraram que, em média, pessoas escolhidas aleatoriamente possuirão cinco outras que farão parte do caminho geodésico entre elas.

O interesse principal desse estudo não era somente analisar o tamanho dessas ligações, mas também as características dos atores intermediários que fazem parte dessa rede (WASSERMAN; FAUST, 1994). Esses vértices intermediários são importantes visto que a informação que passa por eles chegará mais rápido ou não ao seu destino dependendo das distâncias e ligações desses intermediários com os demais vértices da rede, se estão conectados com o vértice final, ou, se não, o quão perto estão dele.

KARINTHY (1929) foi quem primeiro propôs a ideia dos “seis graus de separação”. Em seu trabalho chamado “*Chains*”, o autor observou que o mundo estava encolhendo, visto que as pessoas estavam mais conectadas e próximas umas das outras, e segundo o autor, qualquer pessoa poderia entrar em contato com outro indivíduo selecionado com, no máximo, cinco ligações utilizando apenas sua rede de conhecimentos pessoais. Isso porque, apesar das grandes distâncias geográficas, o aumento na densidade populacional e a facilidade de as pessoas chegarem a lugares bem distantes fazem com que os graus de separação entre duas pessoas quaisquer seja bem menor. Além disso, os meios de comunicação que fazem com que uma pessoa que está em um lado do globo terrestre saiba o que se passa do lado oposto, também ajudam a diminuir as distâncias.

Ainda que o ser humano viva em comunidades afastadas entre si, nota-se facilmente que com poucas ligações entre essas comunidades a distância média entre os vértices da rede diminui bastante, como na Figura III.4. Alguns exemplos são quando alguns vértices estão ligados a outros que possuem muitos conhecidos, tornando a distância entre eles menor, pessoas com amigos em outros países também facilitam na diminuição de distâncias mesmo entre desconhecidos (NEWMAN, 2000). WATTS e STROGATZ (1998) mostram que a inserção de arestas conectando os vértices acelera ainda mais a rapidez com que a informação se espalha.

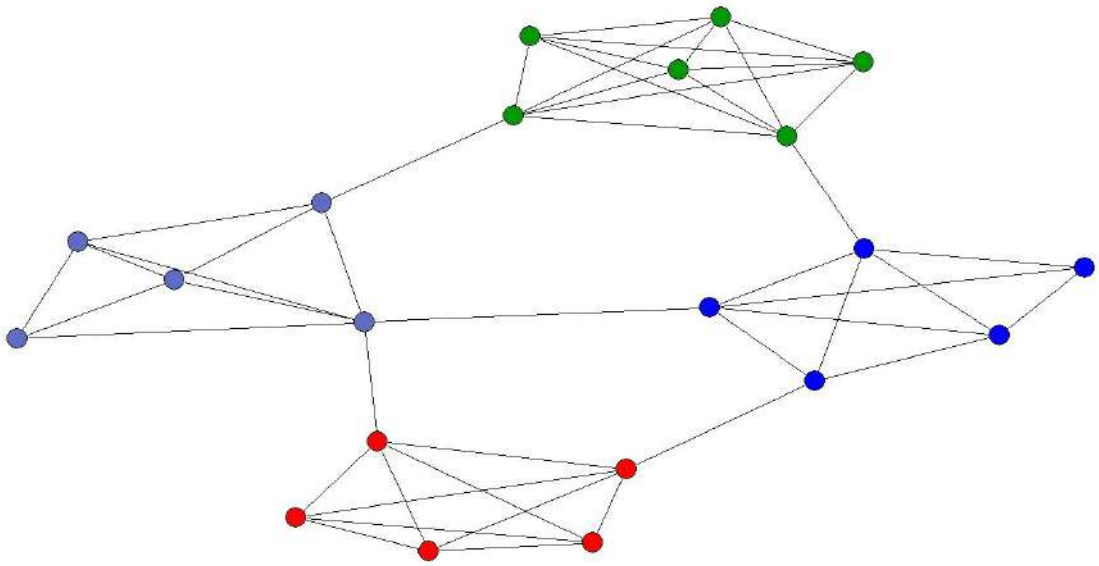


Figura III.4 Exemplo de rede em que diferentes comunidades estão ligadas por poucas ligações

No caso em que a rede é desconexa, como há um ou mais pares de vértices que não estão conectados entre si, ou seja, não existe caminho entre eles, não se pode calcular a distância geodésica, pois seria considerada infinita, então os resultados são considerados apenas para vértices conectados.

Segundo KLEINBERG (2000), o trabalho de Milgram sugere uma fonte com “pistas” subentendidas dentro da rede social, facilitando a mensagem percorrer seu caminho da fonte até o destino. O autor estudou algoritmos descentralizados, baseando-se no experimento de Milgram, em que os indivíduos da rede, possuindo apenas informações locais, tentariam transmitir a mensagem para o destino selecionado, buscando utilizar um caminho mais curto. As conclusões do trabalho de Kleinberg foram que em redes de mundo pequeno a correlação entre conexões locais e de longo alcance provê “pistas” essenciais para que se encontrem os caminhos na rede, o que permite a visualização dos caminhos que conectam os vértices e avaliação de como estes estão ligados direta ou indiretamente.

WATTS (2004) repetiu o experimento de Milgram, mas utilizando *email* em vez de cartas, os quais deveriam ser entregues para pessoas específicas, e o resultado também foi uma média de seis graus de separação.

O efeito da hipótese de mundo pequeno demonstra a velocidade média com que se pode dar a disseminação de informações nas redes. Pode-se ainda analisar os vértices que servem como ligações para que se dêem essas trocas de informações e que facilitam o acesso a elas. No caso das redes de coautoria, as informações de pesquisa podem chegar mais rapidamente a diversos grupos de cientista se as distâncias médias da rede seguirem a hipótese de mundo pequeno, o que é importante para avanços no conhecimento e produção de novas pesquisas.

III.1.4 Redes aleatórias

Em diversas situações práticas, empregam-se modelos probabilísticos para análise de certas características da rede, como no caso da distribuição dos graus dos vértices que podem seguir a distribuição de Poisson ou a lei de potências.

O estudo de grafos aleatórios começou por volta das décadas de 1950 e 1960 com Erdös e Rényi e tem contribuído para diversos resultados relacionados às propriedades de redes reais, como componente gigante, distância entre vértices, entre outras. Um exemplo de emprego de grafos aleatórios como modelo de uma rede real é em epidemiologia, onde os vértices da rede representam os indivíduos infectados e as arestas, contatos capazes de transmitir a doença, esse contatos são considerados aleatórios e não correlacionados (NEWMAN; STROGATZ; WATTS, 2001).

Os graus dos vértices de uma rede em que existe uma probabilidade p de haver ligações entre dois vértices quaisquer, que é a mesma para todos os vértices, seguem uma distribuição de Poisson e aparece uma medida central dos graus. Nessas redes, a existência de uma aresta entre dois vértices independe da existência de outra aresta qualquer. Por isso a existência de cada aresta possui uma probabilidade independente p (BOCALETTI *et al.*, 2006).

Já em redes que seguem leis de potência, não aparece medida central, e a probabilidade de haver ligações entre dois vértices quaisquer não é igual para todos os vértices da rede. Estas redes são também chamadas redes livres de escala. Dessa forma, a média dos graus não demonstra com fidelidade o que ocorre na rede. A lei de potência é semelhante a outras leis, como a Regra 80-20 de Pareto, em que o autor demonstra modelos reais, como a da renda da população em que 20% da população possui 80% de toda a renda e os outros 80% dividem os 20% restantes. E também a lei de Zipf, que segundo o autor, a tendência humana do menor esforço faz com que uma pequena quantidade de palavras seja usada com alta frequência e outras são usadas raras vezes (HOFSTAD, 2010).

Diversas redes seguem lei de potência, como a Internet, redes biológicas e sociais. Redes de colaboração também costumam ser desse tipo, o que significa que um pequeno número de vértices (nesse caso, autores) possui muitas interações com outros vértices, enquanto um grande número de autores possui poucas interações (BARABÁSI *et al.*, 2002). Esse comportamento pode ser verificado em redes de colaboração científica, visto que autores que publicam com diversos outros e possuem mais trabalhos publicados devem possuir uma probabilidade mais alta de que outros vértices se liguem a ele, do que aqueles autores que trabalham mais isoladamente.

PRICE (1965) fez um estudo sobre redes de colaboração científica, em que percebeu que a distribuição do número de citações de um artigo seguia a lei de potências ou distribuição de Pareto. Na rede estudada por ele, alguns poucos artigos possuíam várias citações, enquanto que a maioria havia sido citada poucas ou nenhuma vez. PRICE (1976) também

publicou um estudo em que propõe a Distribuição de Vantagem Cumulativa, também chamada de ligação preferencial, essa distribuição busca modelar os casos em que sucesso gera sucesso. Conforme percebido pelo autor, isso ocorre bastante em bibliometria e outros fenômenos sociais. Assim, um artigo que já foi citado diversas vezes possui maior probabilidade de ser citado novamente do que aquele que possui poucas ou nenhuma citação, assim como é mais provável que um autor que já publicou diversos trabalhos publique novamente do que aquele que possui poucas publicações. Esse modelo se adéqua a diferentes estudos em redes de colaboração científica.

A vantagem cumulativa ou ligação preferencial é um fenômeno bastante ocorrente em redes sociais, pessoas que possuem mais amigos têm mais chances de fazerem novas amizades, pessoas de sucesso também têm maior facilidade de terem ainda mais sucesso. Em redes sociais novos vértices estão mais propensos a se ligarem àqueles que possuem alto valor de grau do que àqueles com poucas ligações. Assim como as redes da Internet, em que um novo cliente estará mais propenso a comprar em um site mais conhecido em que diversos outros usuários já compraram do que em um site mais novo e desconhecido. Percebe-se que como as relações sociais não são igualitárias, a teoria de ligação preferencial faz bastante sentido visto que ao se ligarem a vértices que já possuem alto grau e assim mais caminhos de troca de informações, esses novos vértices terão acesso a um maior número de contatos, produtos, segurança na compra (no caso de compra pela Internet), compartilhamento de informações, estudos, etc (HOFSTAD, 2010).

Redes que seguem leis de potência apresentam uma distribuição de graus com uma função do tipo: $f(x)=kx^{-\tau}$, onde k e τ são constantes. A lei de potência visa descrever que o número de vértices de grau *x*, dado pela função *f(x)*, é diretamente proporcional a $x^{-\tau}$. Através do seu coeficiente de regressão pode-se avaliar o quão bem essa função explicou o que acontece com a distribuição de graus da rede, no caso de redes de coautoria pode-se avaliar a distribuição de graus relacionada ao número de autores e número de artigos.

Conforme apresentado por NEWMAN (2001) em seu estudo sobre a estrutura de redes de colaboração científica, as redes podem também seguir leis de potência com corte exponencial, como ocorreu no estudo realizado pelo autor. O corte exponencial ocorre porque existem mais artigos feitos por pequenos grupos do que seria previsto pela lei de potências pura. Tais redes seguiam lei de potências com corte exponencial, visto que a distribuição de número de artigos por autores não seguia uma reta no gráfico *log-log*, mas ajustava-se a uma expressão exponencial. Assim, segundo o autor, lei de potências com corte exponencial apresenta a seguinte expressão, em que τ e x_c são constantes: $f(x)=x^{-\tau}e^{-x/x_c}$,em que x representa o grau médio e x_c, o tamanho de corte.

Uma rede em que sua distribuição de graus se ajusta a lei de potências, provavelmente se ajustará a hipótese de mundo pequeno, porém o contrário não é verdadeiro, ou seja, uma

rede em que se verifica a hipótese de mundo pequeno não necessariamente terá distribuição de lei de potências. Isso porque redes que seguem lei de potência possuem alto coeficiente de clusterização e o diâmetro dessas redes aumentam logaritmicamente com o número de vértices (AMARAL *et al.*, 2000).

III.2 Redes de coautoria

Uma rede de coautoria é uma rede social não direcionada, em que os nós são autores e as arestas representam a relação de coautoria, ou seja, se dois autores publicaram algum trabalho juntos, existirá uma aresta conectando-os, caso contrário, não existirá essa ligação. O estudo dessas redes pode ajudar no entendimento de como as redes de pesquisa são formadas, como ocorre a disseminação de conhecimento e como redes de autores podem ajudar no desenvolvimento de conhecimentos de certa área, e também, é possível analisar os principais autores em certas áreas de estudo.

No ambiente acadêmico as redes podem demonstrar diversas relações existentes entre pesquisadores, como intercâmbios entre Universidades de diferentes países, uso comum de certos materiais, laboratórios, participação em um projeto, entre outras. A que será estudada nesta dissertação é a relação de coautoria.

Tem-se compreendido há algum tempo que a relação de coautoria de artigos em revistas científicas permite um melhor entendimento sobre os padrões de colaboração dentro da comunidade acadêmica. A estrutura destas redes auxilia também a detectar muitas características interessantes e úteis nessas comunidades (NEWMAN, 2004c). Essa colaboração científica permite uma maior difusão do conhecimento, além de facilitar as pesquisas que podem obter resultados mais rapidamente quando é feita por diversos pesquisadores, além de ser um meio através do qual esses profissionais desenvolvem novas relações, as quais são positivas para suas carreiras e estudos (CRANE, 1972).

Segundo um estudo realizado por SIN (2011), que analisou mais de 7000 artigos publicados em seis periódicos reconhecidos (ARIST, IP&M, JAMIA, JASIST, MISQ e *Scientometrics*) durante os anos de 1980 a 2008, o número de publicações feitas em coautoria, fossem os autores do mesmo país ou de diferentes países, aumentou ao longo dos anos. Percebeu-se também, em seu estudo, que artigos com maior número de autores possuem maior chance de serem citados, o que pode encorajar a formação de colaborações científicas.

Nas redes de coautoria, dependendo do horizonte de tempo estudado pode-se observar que há evolução na sua formação. Essa evolução na rede acadêmica ocorre visto que novos pesquisadores juntam-se através de publicações de seus primeiros trabalhos e aqueles mais antigos aposentam-se e então deixam de fazer parte da rede. Os autores ainda podem alterar suas relações de coautoria durante suas carreiras de pesquisa, o que altera as relações encontradas em diferentes períodos de tempo na rede estudada (DUAN *et al.*, 2012).

Alguns trabalhos que buscam estudar as relações entre os autores em redes de coautoria calculam as métricas de centralidade (grau, proximidade e intermediação) para todos os vértices envolvidos. Como pode ser observado no trabalho de SOUZA *et al.* (2012), em que foi estudada uma rede de autores que publicaram em um certo período de tempo na Revista Química Nova. Através desse trabalho conseguiram observar algumas relações entre os autores, países e Instituições, e algumas outras características, como número de autores por artigo e, dessa forma, puderam estudar as ligações existentes em tal rede. NEWMAN (2001) também realizou algumas análises em três redes de coautoria construídas a partir de bancos de dados de artigos científicos nas áreas de física, pesquisa biomédica e ciências da computação. Ele utilizou algumas medidas de centralidade para avaliar as redes, tais como grau, proximidade e intermediação, e também percebeu o fenômeno de mundo pequeno, além de ter calculado o coeficiente de clusterização de tais redes.

O coeficiente de clusterização é uma propriedade que mede a probabilidade de dois vértices que possuem um terceiro vértice em comum também estarem conectados, ou seja, o quanto a rede localmente se aproxima de um clique. Em estudos de redes sociais percebe-se que estas duas pessoas possuem uma probabilidade mais alta de se conhecerem do que duas pessoas escolhidas aleatoriamente (NEWMAN, 2004a).

Outro estudo interessante é o de KRETSCHMER (1994) que em seu trabalho sobre padrões de colaboração mostra que a freqüência de coautoria entre autores que possuem o mesmo número de publicações é maior do que entre autores com diferentes números de trabalhos publicados. KATZ e HICKS (1997) apresentaram um resultado curioso, segundo os resultados de sua pesquisa, que foi feita através de um estudo bibliométrico, publicações científicas que envolvem autores de diferentes Instituições possuem o número maior de citações que aqueles publicados por autores que pertencem à mesma Instituição.

Muitos estudos recentes de redes de vários tipos têm focado nas distâncias entre os nós da rede. Um par de indivíduos que têm um ao outro como coautor do artigo, por exemplo, possuem distância igual a um, enquanto um par que não tenha publicado junto, mas que compartilham um coautor em comum está separado por uma distância igual a dois, e assim por diante, essa é a distância geodésica entre os nós, também chamada de número de Erdös (NEWMAN, 2004c). Redes em que a maioria de seus autores estão ligados entre si, e, principalmente, quando estes possuem uma distância curta ligando um ao outro (poucos autores entre eles), são redes consideradas mais coesas, pois elas possuem um maior valor de densidade.

JIANG (2008) fez um estudo relevante onde buscou localizar atores chamados por ele de ativos em comunidades de colaboração científica baseado em análises de interação topológica. Segundo o autor, há muitos projetos de pesquisa em que um único pesquisador não pode desenvolver sozinho e precisa da ajuda de outros, os quais poderão formar um grupo

de pesquisa e assim compartilharão os estudos. Nesses grupos cada pesquisador desempenha um papel, mas sempre há um que tomará o papel de liderança, gerenciando as relações e atividades dentro do grupo. O autor chamou esses de atores ativos e, segundo ele, esses autores apresentam alto grau de ligações, possuindo assim grande influência no meio de pesquisa em que está inserido. Com isso, ele propõe a utilização de um algoritmo e apresenta estudos de caso para seu modelo. A localização destes autores ativos possui importância, pois como dito, eles possuem certa influência em sua rede e através deles pode-se avaliar novos estudos da área e até mesmo investimentos que podem estar sendo feitos em determinada área de pesquisa.

Dessa forma, percebe-se que os resultados desses estudos podem ser de grande importância em diferentes áreas de pesquisa e podem ajudar a determinar relações entre os autores, entre Instituições e países, além de ser possível analisar grupos de pesquisa e encontrar temas e assuntos que podem estar sendo estudados. No presente trabalho serão utilizadas algumas redes de coautoria, que serão analisadas e empregar-se-á alguns algoritmos de agrupamento com o objetivo de se detectar comunidades e como grupos de autores se aglomeram nessas redes, as relações entre eles, e se o algoritmo foi eficiente em seus resultados, correspondendo aos grupos mais parecidos com os reais observados.

Capítulo IV - Agrupamento

O estudo de redes emprega diversas propriedades para analisar a rede a ser estudada e, assim, detectar certas características importantes, como os graus dos vértices, densidade da rede, leis de potência, a hipótese de mundo pequeno, entre outras. Mas outra característica importante a ser considerada é a formação de estruturas de comunidade em uma rede, ou seja, vértices que possuem características similares são agrupados em um mesmo grupo. Desse modo, espera-se que haja alta concentração de arestas dentro desses grupos e poucas arestas entre os grupos, estando relacionado ao conceito de coesão, pois o objetivo é maximizar a coesão dentro dos grupos formados (FORTUNATO, 2010).

Assim, uma questão central no agrupamento é a definição dos critérios de similaridade entre os objetos. Exemplos de critérios de similaridades são: o coeficiente de correlação entre os pares de objetos; a distância euclidiana; distância entre dois vértices no grafo, dada pelo número de arestas entre dois vértices do grafo.

As medidas de dissimilaridade satisfazem as propriedades:

- $d(x_i,x_j) \geq 0$, para todo x_i e x_j(positividade);
- $d(x_i,x_j) = d(x_j,x_i)$ (simetria);
- $d(x_i,x_j) = 0$, se e somente se $i=j$.

Se além dessas a medida também satisfizer a propriedade abaixo, essa medida será uma métrica.

- $d(x_i,x_z) \leq d(x_i,x_j) + d(x_j,x_z)$, para todo x_i, x_j e x_z (desigualdade triangular). (VICINI; SOUZA, 2005)

Em uma rede de coautoria, pode-se avaliar o processo de colaboração e o surgimento de grupos de pesquisa através das distâncias em número de ligações entre os pares de vértices e, neste caso, autores mais próximos uns dos outros podem potencialmente estar realizando pesquisa em conjunto ou têm alta chance de que isto aconteça.

No presente estudo, buscou-se focar no uso de técnicas de detecção de agrupamentos, empregando-se os métodos de clusterização: *k-means* e espectral, este último utilizando a matriz Laplaciana e a Laplaciana sem sinal. O critério escolhido para a detecção dos agrupamentos foi a distância entre os vértices no caso do algoritmo *k-means* e a relação de adjacência entre os vértices, para os algoritmos espectrais. Nas redes de coautoria avaliadas, os vértices, que representam os autores, possuem ligações entre si caso possuam alguma publicação em conjunto, assim o objetivo será agrupar aqueles que têm publicado conjuntamente e/ou aqueles que possuem o menor caminho geodésico possível entre si.

A Seção IV.1 abordará o algoritmo de partição *k-means*, como este funciona e algumas de suas vantagens e desvantagens. A Seção IV.2 apresentará os algoritmos espectrais e algumas matrizes necessárias para o entendimento do funcionamento deste. Na Seção IV.3

busca-se apresentar os métodos utilizados para a análise dos resultados gerados pelos algoritmos de agrupamento.

IV.1 Algoritmo de partição *k-means*

O algoritmo *k-means* é um método de partição para detecção de *clusteres*. De forma geral, o *k-means* utiliza como entrada o conjunto de dados contendo *n* objetos e o número de *clusteres* k.

Para o emprego do algoritmo *k-means*, neste trabalho, as entradas utilizadas para o cálculo dos *clusteres* foram as matrizes de distância das redes estudadas e o número k de *clusteres*. Esse valor de k varia de acordo com as redes, pois foram feitas diversas rodadas com diferentes valores com o objetivo de obter melhores resultados de agrupamentos.

Sendo $G=(V, E)$ um grafo conexo com n vértices tal que para todo vértice i e $j \in V$, d_{ij} é o comprimento do menor caminho entre os vértices i e j, ou seja, corresponde à distância geodésica entre eles. Para grafos não valorados, considera-se d_{ij} como o menor número de arestas percorrido entre os vértices. Para todo i, $d_{ij}=0$. Assim, a matriz distância de um grafo é uma matriz simétrica quadrada de ordem n, com todos os elementos da diagonal principal iguais a 0, e os demais valores correspondentes às distâncias entre os vértices i e j das respectivas linhas e colunas. Neste trabalho as matrizes de distâncias serão calculadas a partir das redes de coautoria, que são grafos não valorados, pois as relações entre os autores não possuem pesos.

No caso em que o grafo é desconexo, o valor da distância, d_{ij},, entre dois vértices i e j, que estão em componentes conexas diferentes, é considerado como infinito. Nos experimentos computacionais foram utilizadas apenas as componentes gigantes das redes, dessa forma, todos os grafos são conexos. Para o cálculo da matriz de distâncias utilizou-se no *Matlab* uma função chamada *graphallshortestpaths* que emprega o algoritmo de Johnson para encontrar o menor caminho para cada par de vértices da rede. A complexidade do algoritmo é de $O(n*log(n)+n*m)$, onde n é o número de vértices e m de arestas. O grafo da matriz de distâncias é um grafo completo em que as arestas são valoradas com um peso que é igual ao valor da distância entre os vértices: $w_{ij}=d_{ij}\geq 0$ e $w_{ij}=w_{ji}$.

Inicialmente, o usuário escolhe o número de *clusteres* k. O algoritmo escolherá então arbitrariamente k objetos do conjunto de dados que servirão como centros iniciais dos grupos. Calcula-se então a distância de cada objeto a cada um desses centros e eles são dispostos nos grupos em que apresentam maior similaridade, ou seja, estão mais próximos. Após a determinação de cada objeto em um *cluster*, calcula-se novamente o centro de cada um desses k grupos, e então se recalcula a distância de cada um dos n objetos a esses centros e são reagrupados nos *clusteres* de acordo com sua similaridade. Esses passos continuam até

que não haja mudanças nos grupos formados ou no caso de rede muito grande atinja-se o número de iterações possível ou determinado. Para o uso do algoritmo *k-means* foi empregada uma função já existente no *Matlab*, essa função recebia como dados de entrada a matriz de distâncias da rede e o número *k* de *clusteres* que se desejasse dividir a rede. No entanto, foi feita uma modificação nessa função, de forma que ela recebesse como entrada o arquivo da rede, e calculasse a partir dela a matriz de distância para então realizar a divisão dos *clusteres.* A Figura IV.1 apresenta os passos descritos resumidamente.

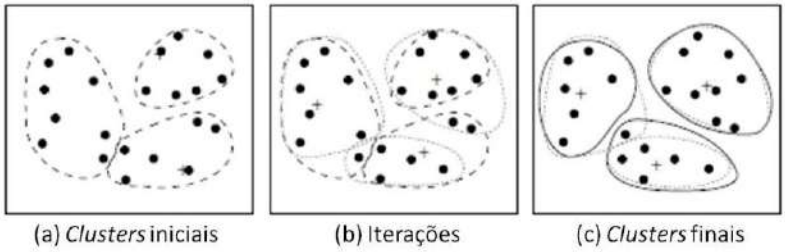


Figura IV.1 Exemplo de agrupamento de um conjunto de dados usando *k-means*
Fonte: Adaptado de HAN *et al.* (2012)

Um ponto negativo deste algoritmo é que não há garantias que ele convergirá ao ótimo global e frequentemente ele resulta em um ótimo local. Este método também não é adequado para a identificação de comunidades em conjuntos de dados com forma não convexa ou quando os vértices estão dispostos de forma tal que ou em que os grupos possuam tamanhos muito variados. Um subconjunto *X* de um espaço vetorial real ou complexo é convexo quando todo segmento de reta ligando dois pontos de *X* está contido em *X*, assim um conjunto de dados onde nem toda ligação entre dois vértices esteja contida nesse espaço vetorial, terá uma forma não convexa, e o método *k-means* não deverá gerar boas partições. Os resultados também podem depender da seleção inicial dos centros de cada grupo; na prática, é comum rodar o algoritmo diversas vezes com diferentes centros iniciais para os clusters. Além disso, o *k-means* é sensível a *outliers* e variações nos dados, isso porque uma pequena mudança nos dados pode influenciar substancialmente o valor da média (HAN; KAMBER; PEI, 2012). Por ser um algoritmo simples, de fácil entendimento e poder ser modificado para lidar com grande quantidade de dados, o *k-means* é um algoritmo muito utilizado (WU *et al.*, 2008).

IV.2 Algoritmo espectral

O agrupamento espectral inclui técnicas que particionam o conjunto de objetos em *clusteres* através do uso de autovetores de matrizes, como a matriz de similaridade e outras derivadas dela. Diferentemente do algoritmo *k-means*, que assume que o conjunto de dados possui uma forma convexa, o espectral não faz nenhuma suposição relacionada à forma do

conjunto de dados a ser particionado, o que confere uma vantagem a seu uso (LUXBURG, 2007). Essa mudança de representação induzida pelo uso de autovetores permite que as propriedades do agrupamento do conjunto de dados inicial sejam mais evidentes. Com isso, o agrupamento espectral consegue separar pontos podendo gerar bons resultados de *clusteres* (FORTUNATO, 2010).

A Teoria Espectral de Grafos é um dos principais pontos para o entendimento do agrupamento espectral. O estudo de autovalores e autovetores de grafos se propõe a fornecer informações topológicas do grafo a partir do estudo do autoespaço das matrizes associadas a tais grafos.

Há algumas características interessantes relacionadas ao estudo das matrizes dos grafos e seu espectro. Dois grafos G e H são chamados coespectrais quando possuem os mesmos autovalores com suas respectivas multiplicidades, ou seja, $spect(G)=spect(H)$.

Dois grafos G e H são isomorfos $G \approx H$ se existem duas funções bijetivas entre o conjunto de seus vértices de modo que as relações de adjacência são preservadas. As matrizes de adjacência de G e H são semelhantes, $P^T A(G)P=A(H)$, onde P é uma matriz de permutação quadrada de ordem n.

Se dois grafos são isomorfos, então eles são coespectrais, pois matrizes semelhantes possuem o mesmo polinômio característico. A recíproca, entretanto, não é verdadeira, ou seja, os grafos podem possuir o mesmo espectro e não serem isomorfos. Se todos os grafos coespectrais a G são isomorfos a G, diz-se que G é caracterizado pelo seu espectro.

Neste trabalho serão utilizadas as matrizes Laplaciana e Laplaciana sem sinal dos grafos, e através do uso de algumas propriedades relacionadas a seus autovalores e autovetores empregar-se-á o agrupamento espectral nos conjuntos de dados utilizados.

Para um melhor entendimento do agrupamento espectral, primeiramente, serão apresentadas algumas explicações sobre as matrizes utilizadas e certas propriedades e então será melhor explicado o funcionamento deste algoritmo.

Um maior aprofundamento sobre a Teoria Espectral de Grafos pode ser visto em CVETKOVIĆ e SIMIĆ (2009) e CHUNG (1996).

IV.2.1 Autovalores de matrizes

Seja B uma matriz de ordem *n* e *x* um vetor do R_n e considere o sistema de equações

$$Bx = \lambda x,\ x \neq 0,$$

onde λ é um escalar. Para λ e $x \neq 0$ que satisfazem o sistema de equações acima, diz-se que λ é um autovalor de B e x é autovetor de B associado a λ. Isso significa que a matriz B faz uma transformação linear sobre o vetor x de forma que o mesmo está sujeito somente a uma mudança na sua magnitude e/ou sentido, mas não na sua direção. O conjunto de todos os

autovalores λ de B é denominado espectro de B. Interessante ressaltar que, em geral, os autovalores podem ser números complexos. Entretanto, resultados clássicos de Álgebra Linear, garantem que o espectro de matrizes simétricas é sempre formado por autovalores reais. Neste trabalho, o interesse está focado exatamente no espectro de matrizes simétricas associadas a grafos, a saber as matrizes Laplaciana e Laplaciana sem sinal que dependem das definições de matrizes de adjacência e de graus.

O espectro de um grafo é único independente da ordenação de seus vértices, pois o polinômio característico é único. O espectro de G é representado pela sequência em ordem decrescente dos autovalores da matriz de G, $\lambda_1(G) \geq \cdots \lambda_{n-1}(G) \geq \lambda_n(G)$.

IV.2.2 Matrizes de Adjacência e de Grau

A matriz de adjacência de um grafo G, denotada por $A(G)$, é uma matriz 0-1, formada de acordo com a existência ou não de arestas entre cada par de nós. Assim, a matriz de adjacência de um grafo não direcionado G com n vértices é uma matriz de ordem n, e é representada da seguinte forma:

$$a_{ij} = \begin{cases} w_{ij}, \ se\ existe\ uma\ aresta\ conectando\ os\ vértices\ i\ e\ j \\ 0, \ caso\ contrário \end{cases}.$$

onde w_{ij} é o peso da aresta e_{ij} no grafo G. Em grafos não-valorados, estes pesos são definidos como w_{ij} =1 para todas as arestas e_{ij}. Em grafos valorados, em geral, os pesos das arestas são diferentes entre si.

Assim, para um grafo com *n* vértices, o grau de cada vértice i pode ser obtido por $d_i = \sum_{j=1}^{n} a_{ij}$. A partir destes graus, define-se a matriz diagonal de graus, denotada por D, do grafo como

$$D_{ij} = \begin{cases} d_i, se\ i = j \\ 0, se\ i \neq j \end{cases}$$

Para cada ordenação de vértices pode-se ter uma matriz de adjacência distinta.

IV.2.3 Matriz Laplaciana

A matriz Laplaciana de um grafo G é real e simétrica e pode ser definida por:

$$L(G) = D(G) - A(G),$$

onde $A(G)$ é a matriz adjacência de G e $D(G)$ é a matriz diagonal cujos elementos são os graus dos vértices de G. Prova-se que o menor autovalor associado a $L(G)$ é sempre nulo e tem como um autovetor o vetor de 1's. Assim, os autovalores reais de $L(G)$ são denotados e organizados da seguinte forma:

$$\mu_1 \geq \mu_2 \geq ... \geq \mu_{n-1} \geq \mu_n = 0.$$

Fiedler (1973) provou que um grafo G é desconexo se e somente se $\mu_{n-1}=0$. Este autovalor é amplamente conhecido como conectividade algébrica do grafo, por informar sobre a conexidade da rede. O autovetor associado a este autovalor é conhecido como vetor de Fiedler e tem sido amplamente usado para agrupamento de dados (DING *et al.*, 2001; LIU; STEWART, 2011; QIU; HANCOCK, 2006). Propriedades como esta, além da simetria e da definição de semipositividade de $L(G)$ (o que implica que todos os autovalores são maiores ou iguais a zero) constituem fatores atrativos para diversas aplicações em problemas da vida real. Maiores detalhes sobre a matriz Laplaciana e seus autovalores podem ser vistos em um *survey* sobre o tema em MOHAR (1992).

IV.2.4 Matriz Laplaciana sem Sinal

A matriz Laplaciana sem sinal, que também será utilizada no emprego do algoritmo espectral neste trabalho, é uma matriz semidefinida positiva e todos os seus autovalores são não-negativos. Ela é denotada da seguinte forma:

$$Q(G) = D(G) + A(G),$$

onde $A(G)$ é a matriz adjacência de G e $D(G)$ é a matriz diagonal cujos elementos são os graus dos vértices de G. O espectro da matriz Laplaciana sem sinal é assim representado:

$$q_1(G) \geq q_2(G) \geq ... q_{n-1}(G) \geq q_n(G) \geq 0.$$

Neste trabalho será utilizado o segundo maior autovalor da matriz Laplaciana sem sinal (q_2) no algoritmo de agrupamento. O trabalho de CVETKOVIĆ e SIMIĆ (2009) pode ser consultado para informações mais detalhadas sobre a matriz Laplaciana sem sinal e seu uso.

IV.2.5 O Algoritmo

Como citado por SCHAEFFER (2007), “no agrupamento espectral um autovetor ou uma combinação de vários autovetores são usados como uma medida de similaridade do vértice para a computação dos *clusteres*.” Neste trabalho, são desenvolvidas duas versões do algoritmo espectral: em uma delas utiliza-se o autovetor associado à conectividade algébrica do grafo (μ_{n-1}), e na outra o segundo maior autovalor da matriz Laplaciana sem sinal (q_2).

Uma das principais vantagens do agrupamento espectral reside no fato de que este não faz nenhuma suposição relacionada à forma dos *clusteres*. Ademais, para matrizes esparsas, o agrupamento espectral pode ser aplicado de forma eficaz para grandes conjuntos de dados (LUXBURG, 2007).

O algoritmo utiliza como entradas a matriz de adjacência, que será utilizada para a identificação dos agrupamentos, e o número de *clusteres* k que se pretende particionar o grafo.

A partir de $A(G)$ e $D(G)$, as matrizes Laplaciana e Laplaciana sem sinal são construídas e seus autovetores são computados. No caso da matriz $L(G)$, o vetor de corte será o vetor de Fiedler, que está associado a μ_{n-1}, para a matriz $Q(G)$ o vetor de corte utilizado será o autovetor associado ao segundo maior autovalor (q_2). Esses vetores de corte permitirão a partição dos vértices em *k* grupos. O corte corresponde a elementos com sinal positivo e negativo nas entradas do autovetor, assim os elementos com sinal positivo pertencerão a um grupo e aqueles com sinal negativo ao outro grupo. O algoritmo é aplicado recursivamente seguindo este procedimento.

A medida de qualidade utilizada para analisar a qualidade do agrupamento obtido pelos algoritmos foi a coesão. Esta medida é definida como a soma do número de arestas dentro de cada agrupamento. Em geral, nos problemas de agrupamento de redes, busca-se maximizar a coesão dentro dos grupos visto que é esperado que a maioria das arestas esteja dentro dos grupos e que haja poucas arestas entre grupos. Já o corte em uma rede é dado pela soma das arestas entre grupos. Observe que minimizar o corte é o problema dual de maximizar a coesão.

Os algoritmos espectrais estão relacionados a esses parâmetros pois estão relacionados a maximizar a coesão dentro dos *clusteres* e minimizar o valor das arestas que estão entre *clusteres*, ou seja, aqueles vértices que estão mais distantes devem estar em *clusteres* separados e por isso o somatório dessas arestas deve ser o mínimo possível para que se tenha a melhor solução. Dessa forma, pode-se observar que há relação entre os autovetores da matriz Laplaciana e a medida de corte de arestas de um grafo. Pelo quociente de Rayleigh,

$$\mu_{n-1}(G) = Min_{\substack{x^T e=0 \\ x\neq 0}} \frac{x^T L x}{x^T x} = \sum_{\{i,j\in E\}} w_{ij}(x_i - x_j)^2. \quad (8)$$

Dada uma partição V em V_1 e V_2 tal que $V_1 \cup V_2 = V$, pode-se definir um vetor de partição p como:

$p_i = \begin{cases} +1, & se\ o\ vértice\ i\ \in V_1 \\ -1, & se\ o\ vértice\ i\ \in V_2 \end{cases}$, de forma que $p^T p = n$.

Tem-se então que:

$$\frac{p^T L p}{p^T p} = \frac{1}{n}\sum_{\{i,j\in E\}} w_{ij}(p_i - p_j)^2 = \frac{4}{n}\, corte(V_1, V_2), \quad (9)$$

onde $corte(V_1, V_2) = \sum_{\{i\in V_1, j\in V_2\}} w_{ij}$. Observe que o problema de minimização (8) pode ser estimado por (9), que está diretamente relacionado ao problema de corte. Assim, o problema de agrupamento Laplaciano espectral é um problema de minimização do corte.

LIU e STEWART (2011) observam que ao utilizar-se a matriz Laplaciana sem sinal em (8), o problema é o dual do problema de minimização do corte na matriz Laplaciana. Nesse caso é um problema que originalmente tem a função objetivo de maximizar as arestas dentro dos *clusteres*. Assim, a função objetivo para o problema utilizando a matriz Laplaciana sem sinal é de maximizar a coesão e será escrita como:

$$Max\ \frac{p^T Q p}{p^T p} = \frac{1}{n}\left(4\sum_{\{i,j\in V_1\}} w_{ij} + 4\sum_{\{i,j\in V_2\}} w_{ij}\right) = \frac{4}{n}\ Max\ coesao(V_1, V_2). \quad (10)$$

Como observado na equação (10), a função objetivo para o algoritmo espectral com uso da matriz Q é um problema de maximização da coesão, dessa forma, deve-se escolher como vetor de corte para divisão da rede em grupos o autovetor relacionado ao maior autovalor, porém como todas as entradas de q_1 são positivas, optou-se pelo uso de q_2.

A mudança de representação dos pontos dos dados da rede para pontos no espaço vetorial é útil devido às propriedades da matriz Laplaciana. Ocorre uma transformação do conjunto de dados inicial em um conjunto de pontos no espaço, sendo as coordenadas desse ponto elementos dos autovetores. Essa transformação do conjunto de dados faz com que as propriedades do *cluster* se tornem mais evidentes, o que pode permitir que os grupos sejam detectados com maior facilidade (FORTUNATO, 2010).

Os algoritmos utilizados para geração dos resultados dessa dissertação foram implementados no *OCTAVE*. A divisão dos *clusteres* foi feita em forma de árvore, assim, inicialmente, o conjunto de objetos foram divididos em dois grupos, e, de acordo com o número de *clusteres* requeridos, os ramos dessas árvores eram divididos em mais dois grupos, e assim sucessivamente até que o conjunto de dados estivesse dividido em k *clusteres*.

Os dados de entrada dos algoritmos são a matriz de adjacência do grafo e o tamanho mínimo de um *cluster* para ser subdividido. O tamanho mínimo refere-se ao número de vértices presente nos *clusteres*, ou seja, se o *cluster* possui um número de vértices igual ou maior do que esse tamanho mínimo, o mesmo será subdividido. Esse procedimento é aplicado recursivamente e dividem os vértices em grupos. As sucessivas subdivisões realizadas podem ser representadas por meio de uma árvore binária. Mais detalhadamente, pode-se dizer que o algoritmo funciona do seguinte modo: inicialmente, a rede possui todos os vértices em um só *cluster*. Quando se coloca como entrada o número total de vértices dessa rede como tamanho mínimo de um *cluster* para ser subdividido, então o algoritmo calculará o vetor de corte, dado pelo autovetor adequado, e dividirá a rede em dois grupos, que representam dois ramos da árvore. Para que os ramos sejam subdivididos com diferentes quantidades de *k*, é necessário que se faça diversas rodadas modificando o tamanho mínimo de um *cluster* para ser subdividido. Assim, quando os ramos possuem um número de vértices maior do que o mínimo pedido calcula-se o novo vetor de corte e esse *cluster* é dividido em mais dois, e assim sucessivamente, até que não haja ramos finais com número de vértices maior que o número mínimo que foi dado de entrada. Ao final, é gerado um arquivo de partição com os *clusteres* respectivos a cada vértice.

Através da Figura IV.2, pode-se entender mais facilmente como ocorre esse agrupamento. Inicialmente, tem-se uma rede com *n*=1000 e o tamanho mínimo para um *cluster* é igual a 400. É calculado o vetor de corte e os vértices em que as entradas no vetor são positivas passarão a pertencer a um *cluster* e aqueles com entradas negativas a outro. São

formados dois *clusteres*, um com *n*=250 e outro com *n*=750. Como o tamanho mínimo é igual a 400, o *cluster* com *n*=750 sofre nova divisão, ou seja, um novo vetor de corte é calculado, e novamente os vértices são divididos em grupos de acordo com as entradas que apresentam no vetor de corte. Nesse exemplo, a rede é particionada em três grupos com números de vértices iguais a 350, 400 e 250, os quais são os ramos finais da árvore.

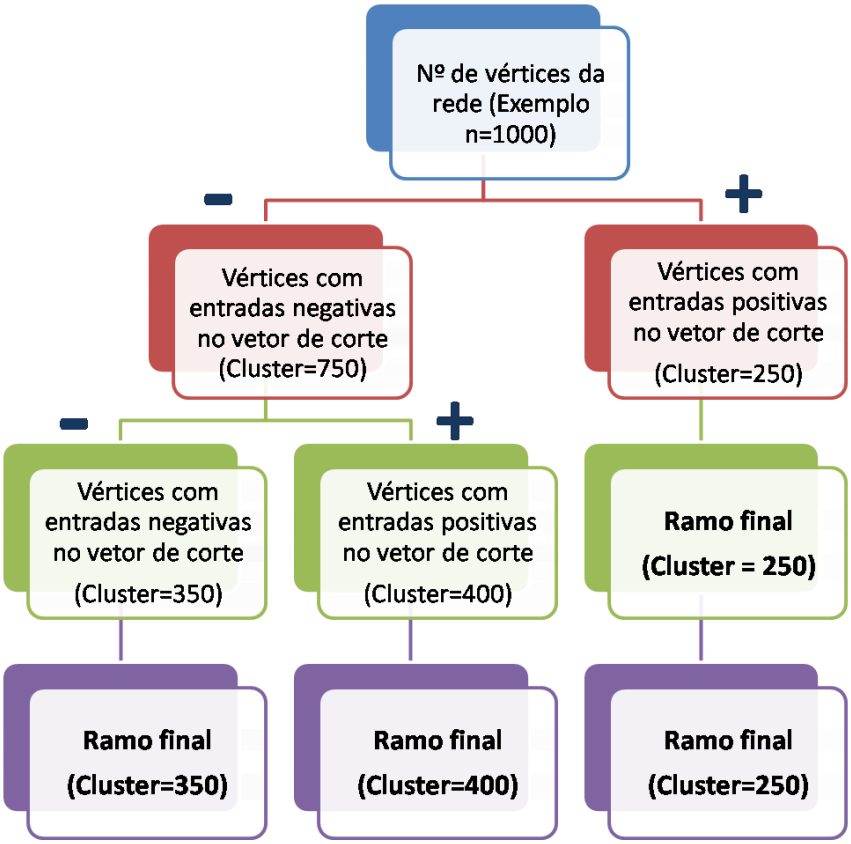


Figura IV.2 Exemplo do agrupamento com algoritmo espectral

IV.3 Medidas para análise dos resultados

O emprego de algoritmos para a detecção de comunidades é importante para a análise dos dados das redes, pois auxilia o estudo de diferentes características presentes nos atores da rede que passam a ser agrupados de acordo com a similaridade percebida entre eles. Porém os algoritmos por não serem métodos que conduzem à solução ótima possuem também algumas desvantagens que influenciam seus resultados. Com isso, algumas medidas podem ser utilizadas para avaliar os resultados obtidos pelas técnicas de agrupamento empregadas, essas medidas de qualidade ajudam a avaliar quão bons esses resultados são, ou seja, o quanto os grupos formados conseguiram ajustar o conjunto de dados de modo que esses *clusteres* apresentassem grande similaridade interna (entre vértices do mesmo *cluster*) e baixa similaridade externa (entre vértices de *clusteres* diferentes). Existem métricas de qualidade que

também auxiliam na comparação de métodos diferentes empregados na identificação de agrupamentos em um mesmo conjunto de dados.

Serão empregadas duas formas de comparação dos resultados obtidos pelos algoritmos. A primeira será baseada em um Modelo de Programação Linear Inteira e será utilizada nos testes em redes pequenas com tamanhos que variam de aproximadamente 20 a 80 vértices. A solução resultante desse modelo que buscará a solução ótima para os agrupamentos das redes será comparada às soluções dos algoritmos espectrais para que se possa avaliar os melhores resultados. Já para redes maiores, como o Modelo de PLI utilizado não funciona bem para tais redes, será utilizada a medida de qualidade chamada modularidade, através do seu uso será possível fazer a comparação dos resultados obtidos para os agrupamentos de tais redes.

A Subseção IV.3.1 apresentará a medida Modularidade, e o Modelo de Programação Linear Inteira será exposto na Subseção IV.3.2.

IV.3.1 Modularidade

De acordo com NEWMAN (2006), a modularidade baseia-se na ideia de que um grafo aleatório não deve possuir uma estrutura de comunidade. Dessa forma, é feita uma comparação entre a densidade real das arestas em um subgrafo e a densidade que se esperaria neste subgrafo se os vértices do grafo tivessem sido conectados aleatoriamente. Assim, redes com valores altos de modularidade possuem alta coesão dentro dos grupos formados e entre os diferentes grupos as ligações são esparsas. Ou seja, a modularidade compara o número de arestas dentro de um grupo com o número esperado de arestas que se podem encontrar no cluster se a rede fosse aleatória, com o mesmo número de vértices e onde cada vértice mantém o seu grau.

Esta medida é dada por uma função que é maximizada para encontrar uma partição que divide a rede em dois grupos. Considerando um grafo G não direcionado com n vértices e m arestas que pode ser dividido em 2 comunidades, se o vértice i pertence a comunidade 1 então $s_i=1$, se j pertence a comunidade 2, $s_j=-1$, e A_{ij} é a matriz de adjacência do grafo, a função de modularidade pode ser calculada pela equação (11):

$$Q = \frac{1}{4m}\sum_{ij}\left(A_{ij} - \frac{d_i d_j}{2m}\right)\left(s_i s_j + 1\right) = \frac{1}{4m}\sum_{ij}\left(A_{ij} - \frac{d_i d_j}{2m}\right)s_i s_j, \qquad (11)$$

onde $\frac{1}{2}\left(s_i s_j + 1\right) = 1$ se os vértices i e j pertencem ao mesmo grupo e 0, caso contrário. Os graus dos vértices são representados por d_i e d_j, e o número de arestas esperado entre os vértices i e j se as arestas foram colocadas aleatoriamente é igual a $\frac{d_i d_j}{2m}$.

A equação (11) é utilizada para calcular a modularidade apenas para redes divididas em 2 comunidades, o particionamento hierárquico é uma aproximação que pode ser utilizada para calcular a medida para redes com diferentes números de *clusteres*, pois através do seu uso pode-se particionar cada um das 2 comunidades em 2 outras subcomunidades e assim por diante de acordo com o número de comunidades da rede.

A equação de modularidade pode também ser escrita como $Q = \frac{1}{4m} s^T B s$, onde s é o vetor coluna cujos elementos são s_i e B é uma matriz simétrica tal que: $B_{ij} = A_{ij} - \frac{d_i d_j}{2m}$.

O valor da modularidade pode ser positivo ou negativo, sendo que valores positivos indicam a possível presença de estrutura de comunidade, pois esse valor é alcançado quando o número de arestas dentro de cada *cluster* excede àquele esperado em um grafo aleatório. Além disso, valores altos de modularidade indicam boas partições, porém o valor dessa medida não deve ser usado para comparar grafos de tamanhos diferentes (FORTUNATO, 2010). Assim, quanto maior o valor da Modularidade, melhor a divisão feita pelo algoritmo de agrupamento. Na Figura IV.3 são apresentados dois exemplos de redes de mesmo tamanho, porém com valores de modularidade bem diferentes. A rede (a) apresenta uma estrutura de comunidade e a partição encontrada possui valor positivo para a modularidade, já a rede (b) é uma rede gerada de forma aleatória, não possuindo estrutura de comunidade e, como apresentado na figura, a partição gerada possui valor negativo para a modularidade.

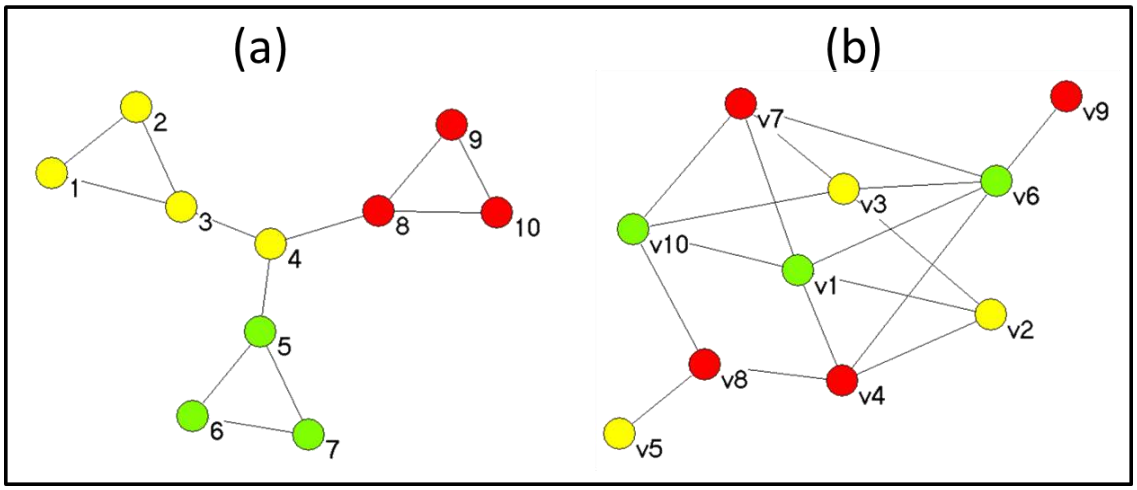


Figura IV.3 Exemplos de redes com Modularidade (a) positiva, $Q=0,4896$ e (b) negativa, $Q = -0,1142$

FORTUNATO e BARTHÉLEMY (2007), SARKAR e DONG (2011), em seus trabalhos sinalizaram alguns pontos negativos para a modularidade. Um deles é que essa medida é válida somente para grafos unipartidos, e, além disso, grupos sobrepostos não são revelados, e essa medida não consegue detectar subestruturas pequenas escondidas dentro de comunidades maiores.

IV.3.2 Modelo de Programação Linear Inteira

O Modelo de Programação Linear Inteira utilizado foi o mesmo estudado por NASCIMENTO *et al.* (2010), que por sua vez baseou seu trabalho em RAO (1971) cuja proposta era apresentar modelos para minimizar o somatório dos quadrados das distâncias dentro dos grupos; e uma alternativa era minimizar a máxima distância dentro do grupo, ou seja, minimizar o diâmetro. O autor apresentou diversos modelos que pudessem analisar os diferentes objetivos requeridos para se obter a solução ótima para o agrupamento. A partir desses modelos, NASCIMENTO *et al.* (2010) sugeriram o modelo a seguir:

$$Min \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij} y_{ij}$$

Sujeito a:

$$\sum_{k=1}^{M} x_{ik} = 1,\ i = 1, \dots, N \quad (1)$$

$$\sum_{i=1}^{N} x_{ik} \geq 1,\ k = 1, \dots, M \quad (2)$$

$$x_{ik} \in \{0,1\},\ i = 1, \dots, N,\ k = 1, \dots, M \quad (3)$$

$$y_{ij} \geq x_{ik} + x_{jk} - 1,\ i = 1, \dots, N,\ j = i+1, \dots, N,\ k = 1, \dots, M \quad (4)$$

$$y_{ij} \geq 0,\ \ i = 1, \dots, N,\ j = i+1, \dots, N \quad (5)$$

onde d_{ij} é a distância entre os vértices *i* e *j*; *N* é o número de vértices; *M* é o número de *clusteres*; x_{ik} é uma variável binária que assume valor 1, se o vértice *i* pertence ao *cluster k* e 0, caso contrário; y_{ij} é uma variável real que assume o valor 1 se os vértices *i* e *j* pertencem ao mesmo *cluster* e 0, caso contrário.

A função objetivo busca minimizar a distância entre os vértices que pertencem ao mesmo *cluster*. A restrição (1) garante que o vértice *i* pertence a somente um *cluster*. A restrição (2) garante que o *cluster k* possui pelo menos um vértice. As restrições (4) e (5) garantem que y_{ij} assume o valor 1 se ambos x_{ik} e x_{jk} forem iguais a 1.

Como o objetivo dos algoritmos empregados nesse estudo é de minimizar o somatório das distâncias dentro dos *clusteres* e maximizar o somatório delas entre *clusteres*, o modelo apresentado acima cumprirá com este objetivo ao fazer o agrupamento das redes que serão testadas.

Capítulo V - Descrição da Base de dados DBLP

A base de dados utilizada abrange apenas publicações referente a ciência da computação, focando em publicações internacionais e em inglês. Como há limitações de recursos para que todas as publicações realizadas na área sejam indexadas, há uma prioridade na inclusão das Conferências, Periódicos, etc, buscando-se aqueles com maior mérito e de importância para a comunidade científica dessa área (DBLP, 2014). Para a realização dos experimentos computacionais foram extraídos alguns dados de publicações dessa base para montagem de seis redes de coautoria.

Inicialmente a sigla significava, "***D**ata**b**ase systems and **L**ogic **P**rogramming"*, pois a base de dados focava publicações desse campo de pesquisa, porém com o passar dos anos expandiu-se as indexações para todos os campos de pesquisa relacionados à ciências da computação, e adotou-se "*Digital Bibliography & Library Project*" como novo significado para a sigla. Essa base de dados foi escolhida para análise, pois é uma das bases mais completas que contém informações bibliográficas sobre os principais periódicos, publicações em conferências, teses e livros de ciências da computação, desde o ano de 1936 até os dias de hoje. Atualmente, existem mais de 2,3 milhões de publicações indexadas, publicadas por mais de 1,2 milhões de autores. Essas publicações provêm de mais de 18.000 volumes de periódicos; aproximadamente, 20.000 conferências e *workshops*; e mais de 15.000 monografias.

Percebe-se que, nas últimas décadas, houve um crescimento considerável do número de publicações na área de ciências da computação, e um dos fatores que deve ter contribuído para tal fato é a maior disponibilidade de tecnologias e capacidade dos computadores. E também, o aumento da necessidade para lidar com a crescente quantidade e variedade de dados, além da complexidade de tratar tais volumes de dados, o que deve ter incentivado o maior número de trabalhos na área e o desenvolvimento de novas técnicas que auxiliam a resolução de diversos problemas em sistemas reais.

A Figura V.1 mostra esse aumento no número de publicações com dados extraídos das publicações indexadas na base DBLP. A variação, em números absolutos, foi de 12 publicações em 1936 a 192.664 no ano de 2013. As publicações estão divididas de acordo com o tipo, se foram publicadas em Conferências, Periódicos, no Repositório de Pesquisa da Computação, Editoriais ou Livros. Nota-se que a maior parte das publicações indexadas refere-se a Conferências e Periódicos, e que nos anos de 2011 e 2012 o número de publicações passou de 200.000. Com esse grande aumento do número de publicações, essas redes também devem ter sofrido diversas modificações ao longo do tempo, sendo significantes os estudos relacionados a redes de colaboração científica nessa área.

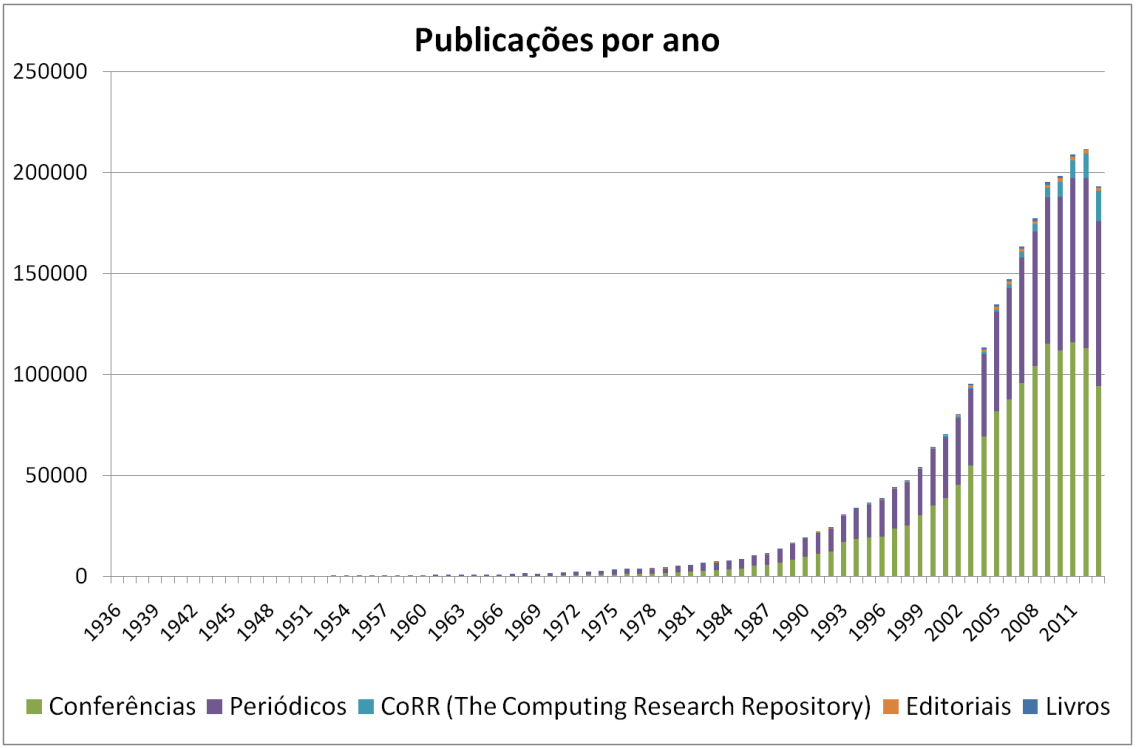


Figura V.1 Publicações indexadas por ano

Através da Figura V.2, que mostra os tipos de publicações indexadas na base durante o período de 1936 a 2013, fica mais clara a diferença nas quantidades de publicações de cada tipo. Aquelas referentes a Conferências e Periódicos são responsáveis por mais de 95% de todas as indexações da base de dados. A menor quantidade de publicações refere-se aos Livros. É esperado que esses estejam em menor número que os artigos, visto que existem diversos Periódicos e estes publicam muitos artigos, assim como as Conferências que publicam os trabalhos que são apresentados, enquanto os Livros demoram um tempo maior e necessitam de mais conteúdos para publicação.

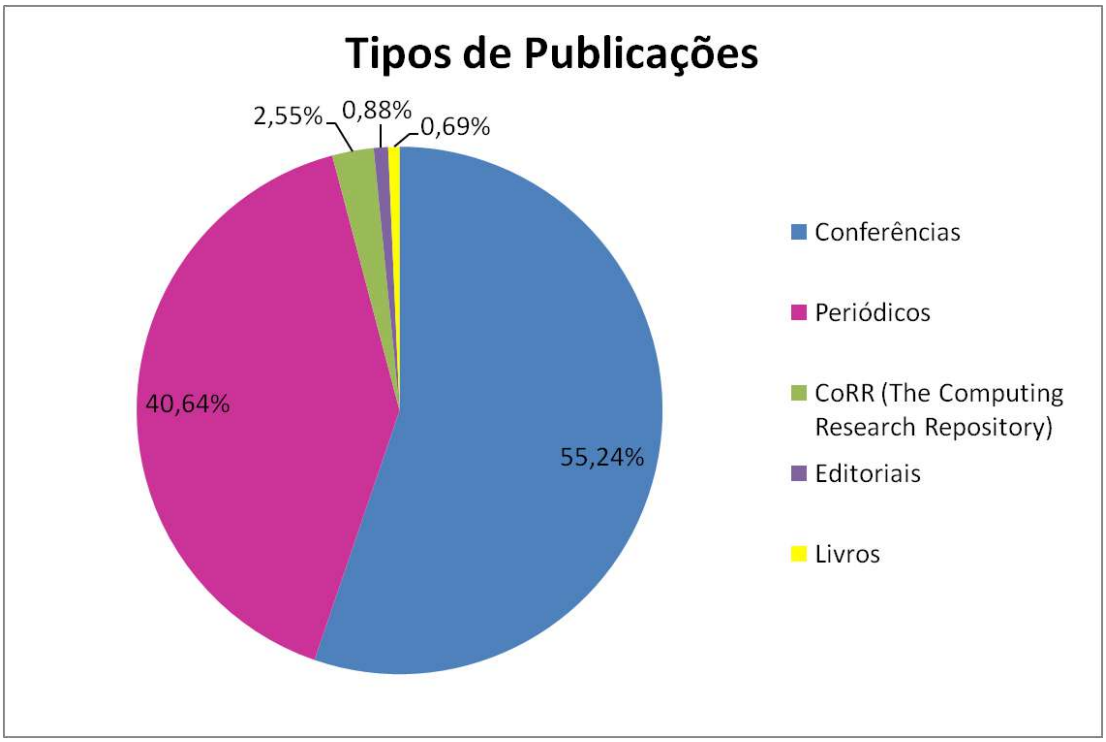


Figura V.2 Tipos de publicações indexadas na base DBLP

O site da DBLP apresenta uma lista com diversas informações, como nome dos autores, título do artigo, periódico, volume, páginas, ano de publicação, entre outras, que podem variar de acordo com o arquivo escolhido. O site também disponibiliza o arquivo com a base de dados completa em *xml*. E, além disso, é possível realizar buscas refinadas e gerar arquivos *xml* com tais resultados.

Dado que a base conta com mais de 2 milhões de publicações indexadas, lidar com todos esses dados se tornaria uma tarefa que despenderia muito tempo, pois são necessários alguns tratamentos. Como o objetivo principal não é aplicar os algoritmos em redes muito grandes, mas sim testar seus resultados para serem implementados, além de algumas limitações que foram impostas pelos *softwares* utilizados, a restrição no tamanho das redes foi uma solução escolhida, optando-se por um refinamento para seleção dos dados na base DBLP.

O site conta com alguns motores de busca, dentre eles escolheu-se utilizar o *CompleteSearch*. A seleção foi realizada com base na escolha de alguns critérios para refinamento: em tipo, escolheu-se *Journals* (Periódicos), e em autores foram selecionados os 25 com mais publicações. A partir dessa seleção foram gerados os arquivos em *xml* de cada autor. Os dados extraídos continham diversas informações como, nomes dos autores, título do artigo, nome do periódico, páginas, número, volume, ano, endereço no site, e cada artigo possui um id diferente.

Para formação das redes de autoria utilizou-se o *Microsoft Excel* para edição e tratamento dos dados. Selecionaram-se então as colunas id correspondente aos artigos e a coluna com os nomes dos autores participantes de cada artigo. Durante a montagem da rede, notou-se que a autora Diane Crawford, a 24ª da lista dos 25 autores, possuía apenas artigos publicados sozinha, como o objetivo do trabalho é o estudo de redes de coautoria, excluiu-se essa autora da lista e selecionou-se o autor David J. Evans, que está na 26ª posição.

Além disso, percebeu-se que 7 autores estavam citados com nomes diferentes na lista de artigos, como Thomas D. Wilson, que além desse nome também é citado como T. D. Wilson e Tom Wilson. Assim, foi necessário um tratamento dos nomes desses autores, padronizando os nomes para que fossem contados como apenas um. Como a seleção no site foi feita autor a autor, algumas publicações estavam repetidas, visto que alguns autores haviam publicado com outros dessa rede de 25 autores, dessa forma, foi necessário tratar esses dados excluindo-se os dados duplicados.

Após esses tratamentos, foi montada a rede de autoria contando com 5099 autores (inclui-se nessa lista os coautores desses 25 autores) e 7481 artigos, abrangendo o período de tempo de 1965 a 2013. Na rede de autoria os vértices representam autores e artigos, nelas os autores estão ligados a artigos que publicaram e não há arestas ligando artigo com artigo, nem autor com autor. A partir da rede de autoria foi gerada a rede de coautoria, onde os vértices passam a representar os autores e há arestas entre eles caso possuam alguma publicação em coautoria. A Figura V.3 mostra o total de publicações de cada autor da lista dos 25 durante o período de 1965 a 2013.

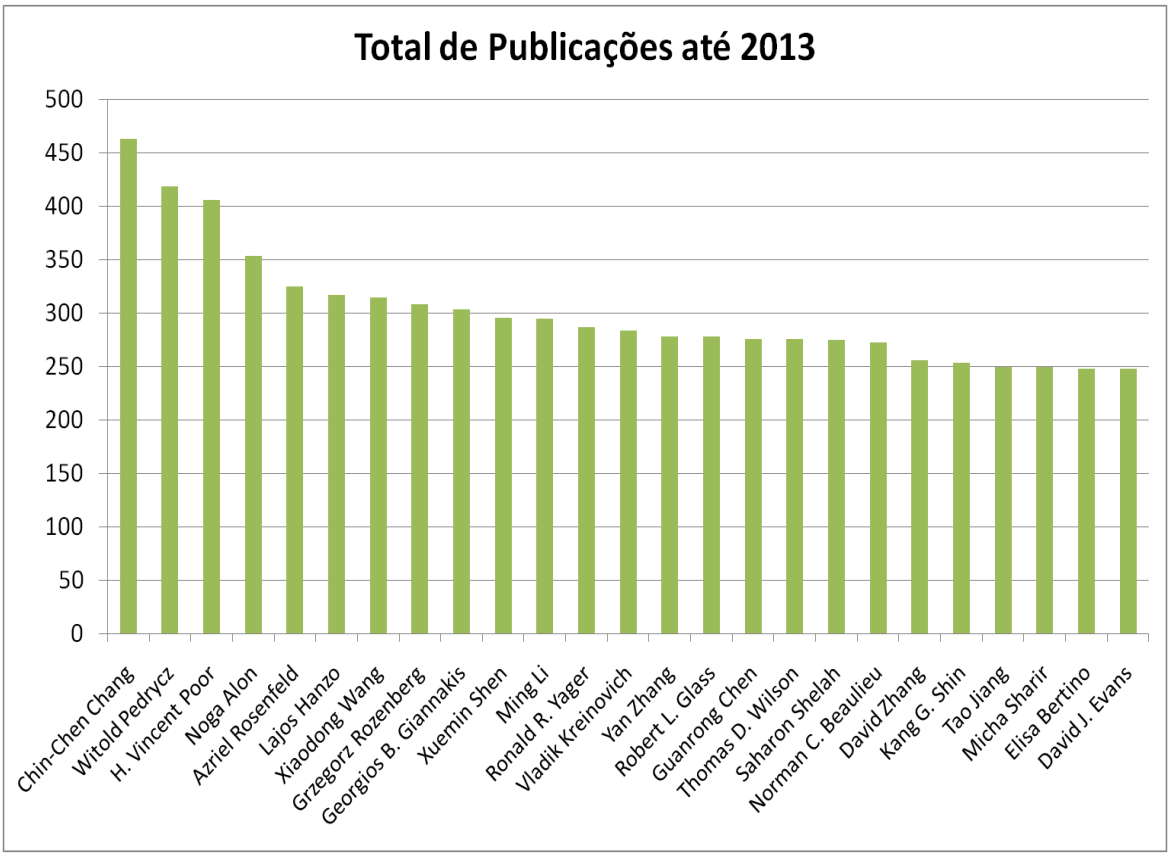


Figura V.3 Total de publicações de 1965 a 2013 por autor

A Figura V.4 apresenta o número de publicações por ano, é notável o crescimento nas publicações a partir do ano 2002. É interessante notar que não somente o número de publicações cresceram nos últimos anos, mas também o número de autores que publicam artigos como pode ser observado na Figura V.5, com isso há um aumento de contribuições em pesquisas na área.

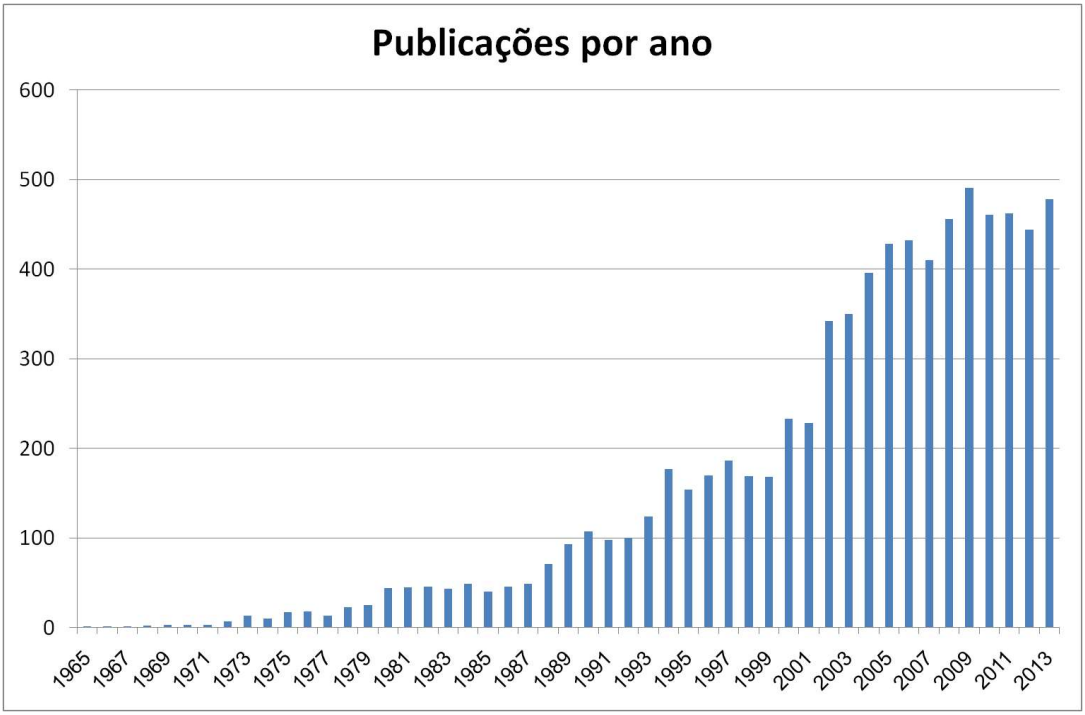


Figura V.4 Número de publicações por ano

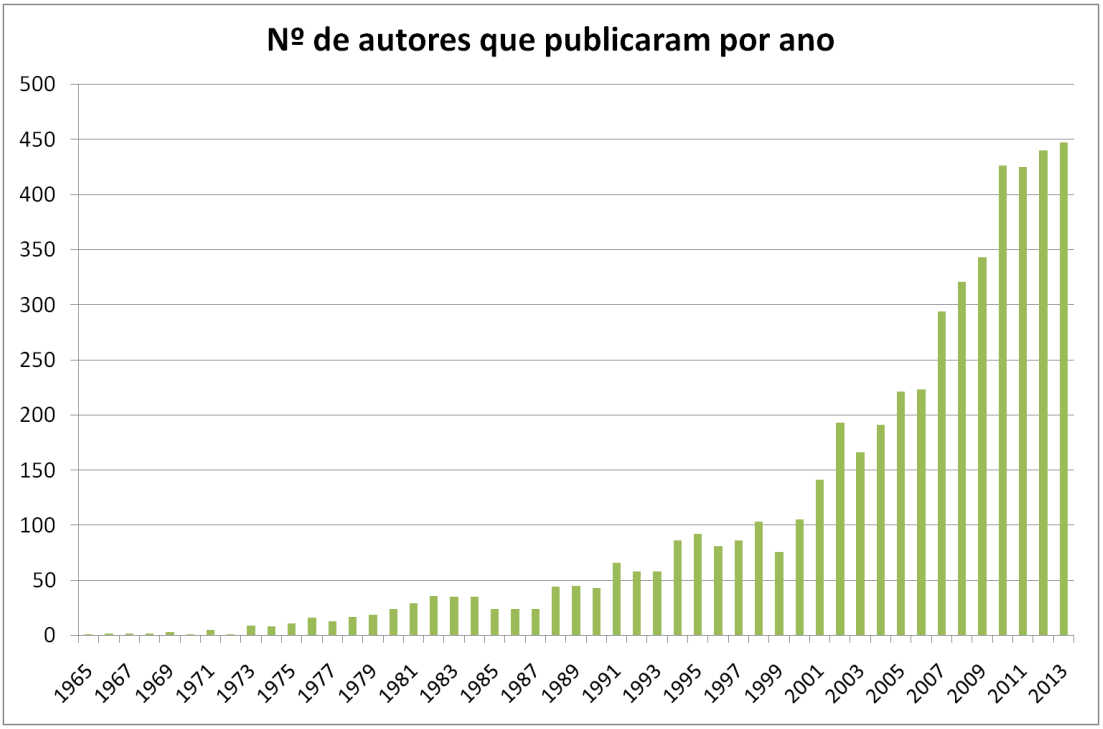


Figura V.5 Número de autores que publicaram em cada ano de 1965 a 2013

Analisando as publicações, observou-se que a maioria delas são em coautoria, mais de 80%, e o número de autores participantes em um artigo pode ser maior do que 20, todavia a

maioria delas, mais de 80%, possui entre 2 a 5 autores. Essas informações podem ser melhor visualizadas nas figuras V.6 e V.7.



Figura V.6 Porcentagem de publicações em coautoria e individual

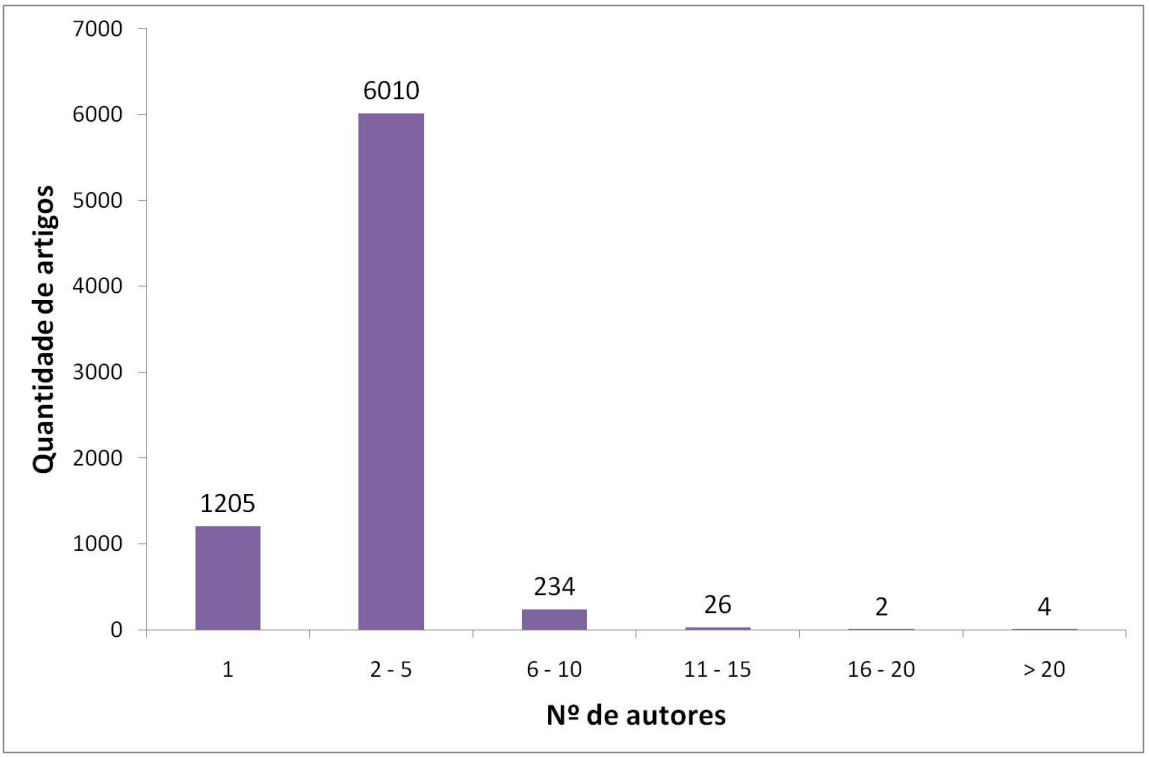


Figura V.7 Quantidade de artigos com diferentes números de autores

Na Figura V.8 pode-se visualizar os autores que mais publicam em coautoria e aqueles em que a maioria de suas publicações é feita de forma individual.

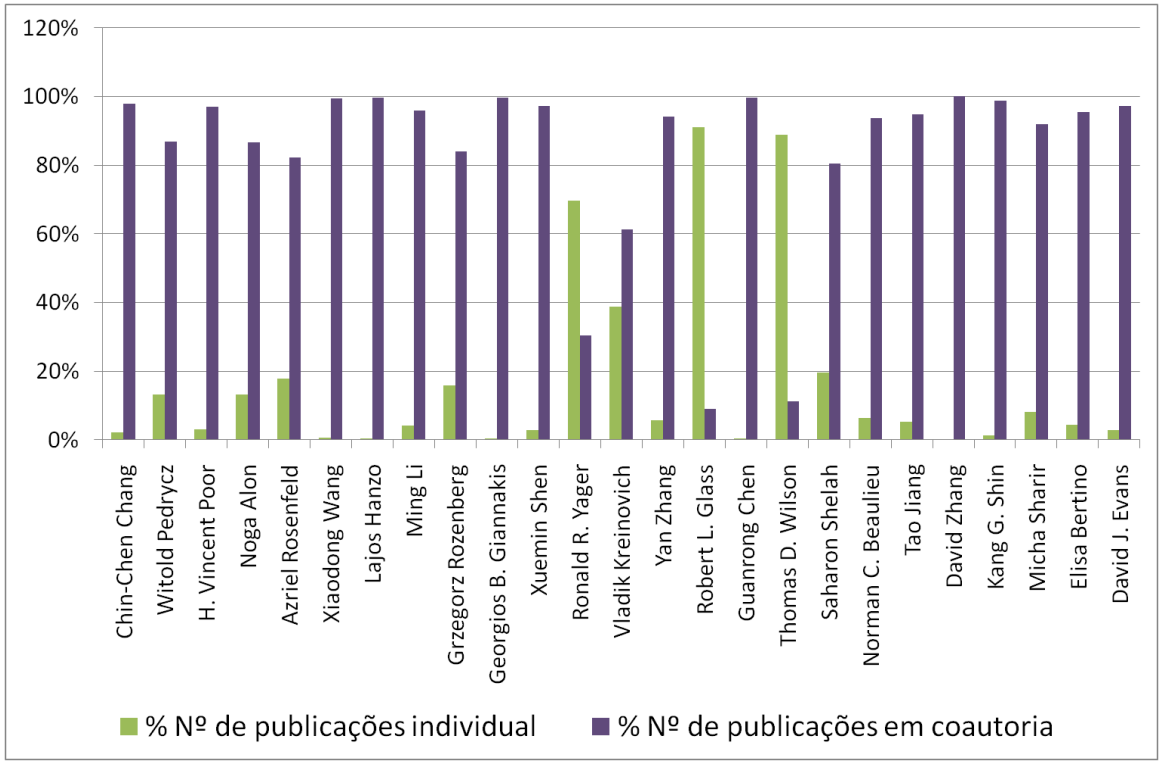


Figura V.8 Porcentagem de publicações individuais e em coautoria de cada autor

A Figura V.9 apresenta a quantidade de periódicos em que os autores publicaram, Ming Li foi o que mais publicou em periódicos diferentes, porém não é o que mais possui publicações, já os autores Robert L. Glass e Thomas D. Wilson foram os que menos publicaram em diferentes periódicos.

Na Figura V.10 estão os 31 periódicos que os autores selecionados mais publicaram, as publicações nesses periódicos somam 50% de todas as publicações feitas por eles.

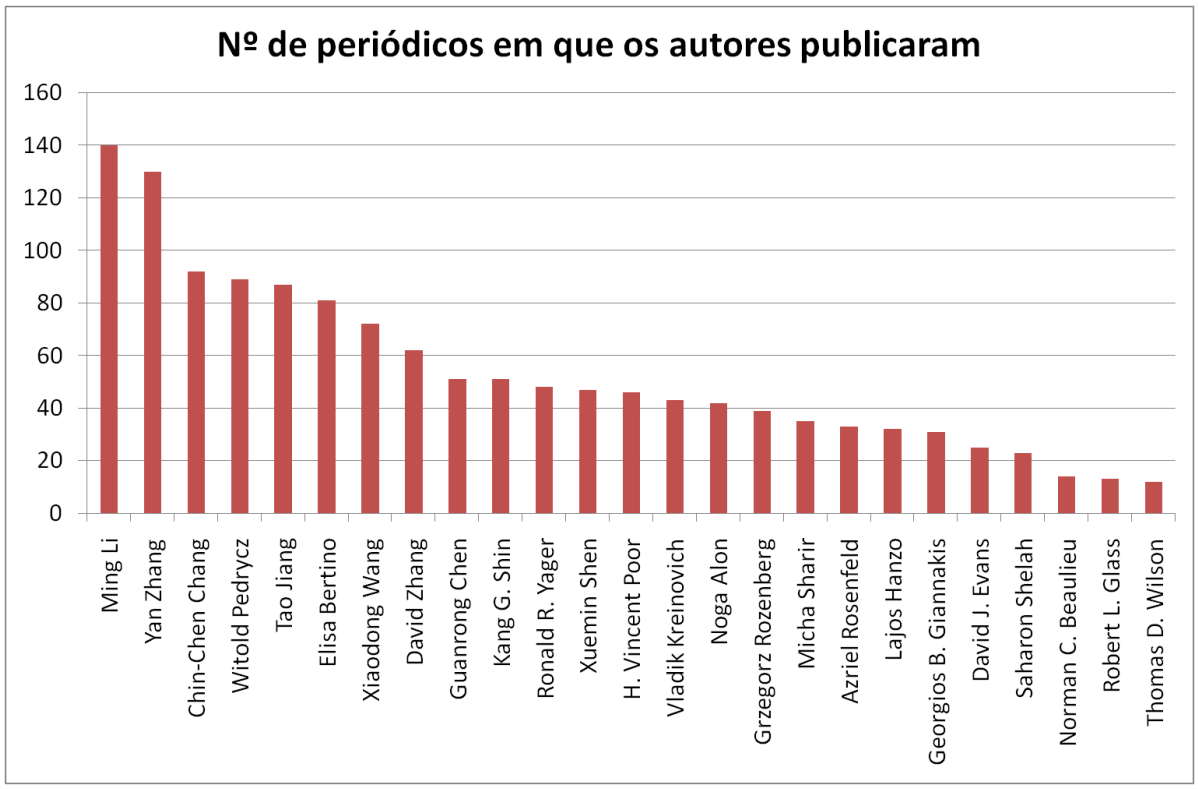


Figura V.9 Número de periódicos em que os autores publicaram de 1965 a 2013

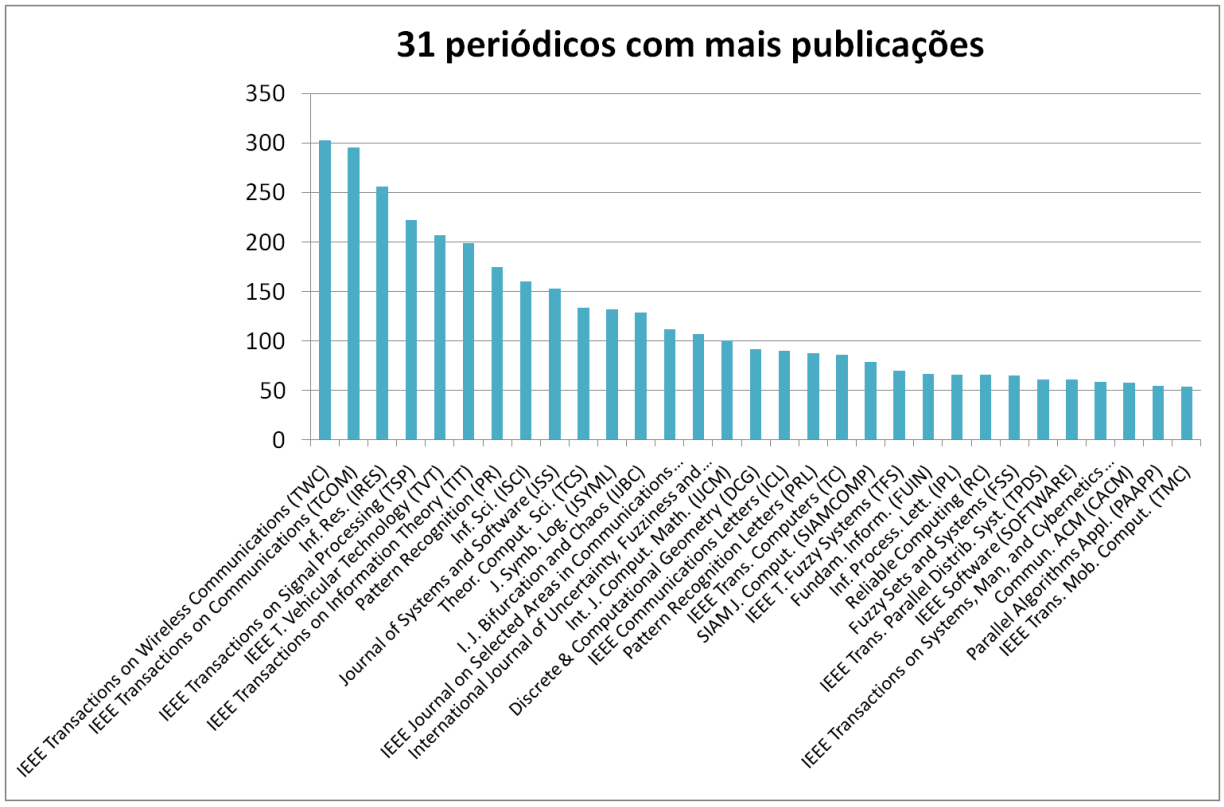


Figura V.10 Os 31 periódicos em que os autores mais publicaram

Capítulo VI - Experimentos computacionais

Neste capítulo, os resultados computacionais obtidos a partir da aplicação dos algoritmos espectrais e *k-means* sobre redes de coautoria de tamanhos diversos são analisados com o uso da modularidade. Em particular, foram selecionadas e analisadas redes provenientes da base DBLP (*Digital Bibliography & Library Project*). Um modelo de programação linear inteira foi utilizado com o objetivo de definir agrupamentos ótimos em redes pequenas (até 80 vértices) e esses resultados foram comparados com os gerados pelos algoritmos espectrais. Como os algoritmos espectrais funcionam sob a hipótese de conexão da rede, todas as análises dos algoritmos espectrais foram realizadas sobre as componentes gigantes das redes de coautoria. A implementação dos algoritmos considerou as seguintes ferramentas: *Matlab*, *OCTAVE*, *XPRESS* e *Pajek*.

Este capítulo está dividido em três seções. Na Seção VI.1, são apresentadas as 100 redes geradas seguindo a Lei de Potências, com a finalidade de comparar os resultados obtidos pelas heurísticas espectrais com as soluções exatas fornecidas pelo modelo de PLI aplicados sobre essas redes. A Seção VI.2 apresenta três redes de coautoria utilizadas na literatura, algumas medidas relacionadas a elas e os resultados de agrupamentos obtidos pelos algoritmos e a modularidade decorrente de cada rodada realizada. A Seção VI.3 apresenta seis redes de coautoria formadas a partir de dados retirados da base DBLP, assim como algumas medidas relacionadas a elas e os resultados de agrupamentos obtidos pelos algoritmos e a modularidade decorrente de cada rodada realizada.

VI.1 Caso 1: redes seguindo a Lei de Potência

Este primeiro experimento considera 100 redes geradas seguindo a Lei de Potências (ou livres de escala) a partir da ferramenta *Pajek*. Optou-se por gerar redes livres de escala porque nessas redes a probabilidade de dois vértices estarem ligados não é a mesma para todos os pares de vértices, o que favorece a formação de grupos. As redes geradas pelo *Pajek* seguem a distribuição de probabilidades dada pela equação (12):

$$\mathbf{Pr}(v) = \alpha \frac{d_v}{m} + \gamma \frac{1}{n} \qquad (12),$$

onde $\alpha+\gamma=1$ (BATAGELJ; MRVAR, 2011). Todas as redes geradas são não direcionadas e possuem grau médio igual a dois, a probabilidade inicial de linhas é igual a 0,2 e $\alpha=0,25$. O objetivo desta primeira bateria de testes é gerar redes pequenas, aplicar um método exato, a partir do *XPRESS*, e dois métodos heurísticos espectrais para definir os agrupamentos. Os resultados obtidos pelos métodos são comparados como uma forma preliminar de avaliar o desempenho das heurísticas.

A Tabela VI.1 apresenta a ordem de cada rede na primeira coluna e o total de redes geradas na segunda coluna. Vale ressaltar que os algoritmos exatos e heurísticos foram aplicados sobre a componente conexa de cada rede, o que apresentou uma variação na ordem de, aproximadamente, 20 a 80 vértices.

Tabela VI.1 Redes geradas segundo Lei de Potências

# de Redes	Nº de vértices	Nº de vértices na componente gigante
20	50	19-27
20	75	32-41
20	100	41-56
20	125	50-68
20	150	62-83

Para as 20 redes menores, em que a dimensão das componentes conexas variou de 19 a 27 vértices, o Modelo de PL atingiu a solução ótima em poucos segundos (por volta de 5s), porém quando o número de vértices passou a aumentar, a partir de 32 vértices, o modelo passou a ter um tempo de execução muito alto, levando horas, isso porque ao aumentar o número de vértices aumenta-se o número de restrições. Assim, para as 80 redes que possuem número de vértices maior ou igual a 32, fixou-se um tempo de 10 minutos para a execução do modelo, utilizando-se a solução encontrada nesse tempo para comparação com os algoritmos espectrais. Esses algoritmos tiveram um tempo de execução inferior a 10 minutos para todas as 100 redes. A Figura VI.1 apresenta o valor da função objetivo obtido a partir do modelo exato e dos algoritmos espectrais laplaciano e laplaciano sem sinal.

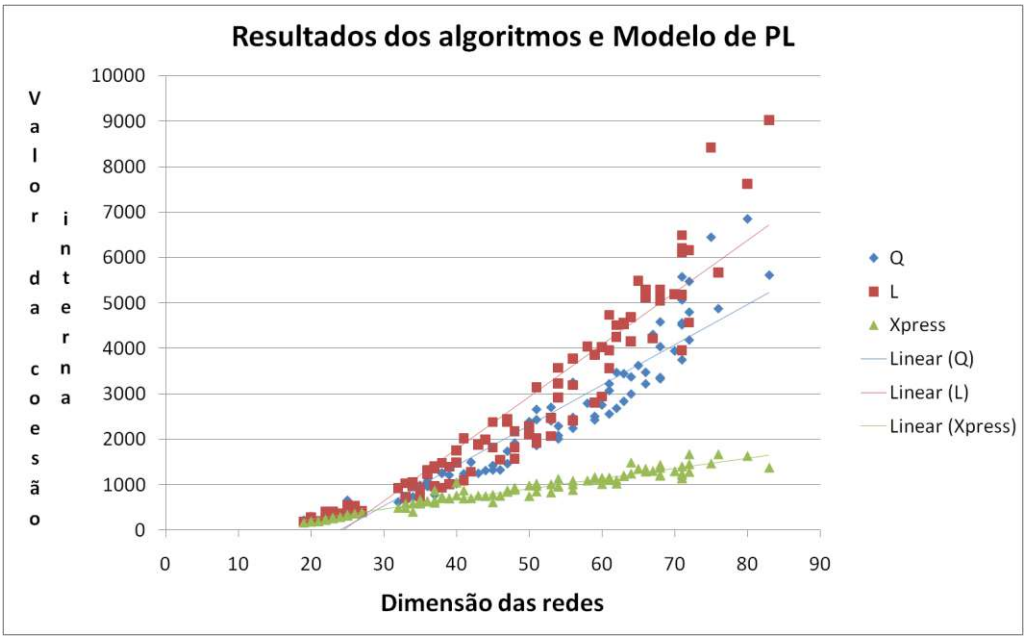


Figura VI.1 Resultados dos algoritmos espectrais e Modelo de Programação Linear Inteira

A partir da Figura VI.1, observa-se que para as 20 menores redes os algoritmos espectrais não apresentaram a solução ótima, mas obtiveram resultados próximos. Para as 80 redes maiores a solução obtida como resultado do modelo de PL foram melhores que aquelas apresentadas pelos algoritmos espectrais. Como a função objetivo do modelo de PLI buscava minimizar o somatório das distâncias dentro dos *clusteres*, calculou-se esse somatório para os *clusteres* gerados pelos algoritmos espectrais para comparação. O eixo vertical do gráfico mostra os valores dessa coesão interna, assim quanto maior esse valor pior o resultado do agrupamento, com isso é bem notável que o modelo de PL foi o melhor para todas as redes.

Todavia, esse modelo de programação linear inteira não pôde ser aplicado em redes com número de vértices da ordem de 10^3, pois além de o tempo de execução ser ainda mais alto, no computador em que se testou, o *software XPRESS* não foi capaz de ler os dados de entrada e rodar o modelo. As especificações do computador utilizado para rodar o modelo de PL no *XPRESS* são as seguintes: Processador: Intel Core i5-3300; Memória Instalada (RAM): 8,00 GB; Sistema Operacional de 64 Bits.

VI.2 Caso 2: 03 redes de coautoria extraídas da literatura

Foram analisadas três redes de coautoria retiradas de trabalhos publicados recentemente. Nesta seção serão apresentadas algumas análises referentes às componentes gigantes dessas redes e os resultados obtidos pela implementação dos algoritmos espectrais e *k-means*.

Três redes de coautoria são apresentadas nesta seção:

Rede 1: A rede formada por pesquisadores do CNPq, obtida por BARBOSA *et al.* (2011), é uma rede conexa que contém 207 vértices e 520 arestas e que pode ser visualizada na Figura VI.2. Cada vértice representa um pesquisador com publicação de artigos relacionados à área de Teoria dos Grafos. Dois pesquisadores são conectados por uma aresta caso possuam pelo menos uma publicação em coautoria que esteja informada em seus currículos Lattes. A informação foi extraída da plataforma Lattes no período de Outubro a Novembro de 2010, conforme informado em BARBOSA *et al.* (2011). Como essa é uma rede conexa, foi utilizada a rede original.

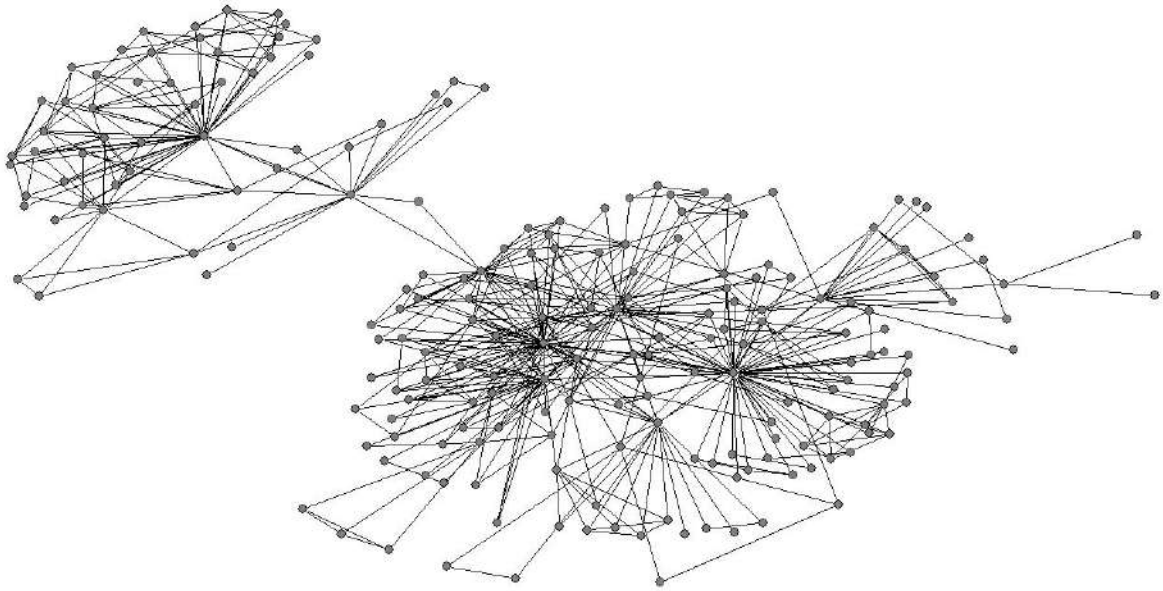


Figura VI.2 Rede de pesquisadores do CNPq

Rede 2: rede de coautoria obtida por SOUZA e BARBASTEFANO (2011) a partir dos autores que publicaram artigos sobre o tema Análise do Ciclo de Vida constantes na base *ISI/Web of Science.* Trata-se de uma rede com componente gigante contendo 955 vértices e 2810 arestas apresentada na Figura VI.3.

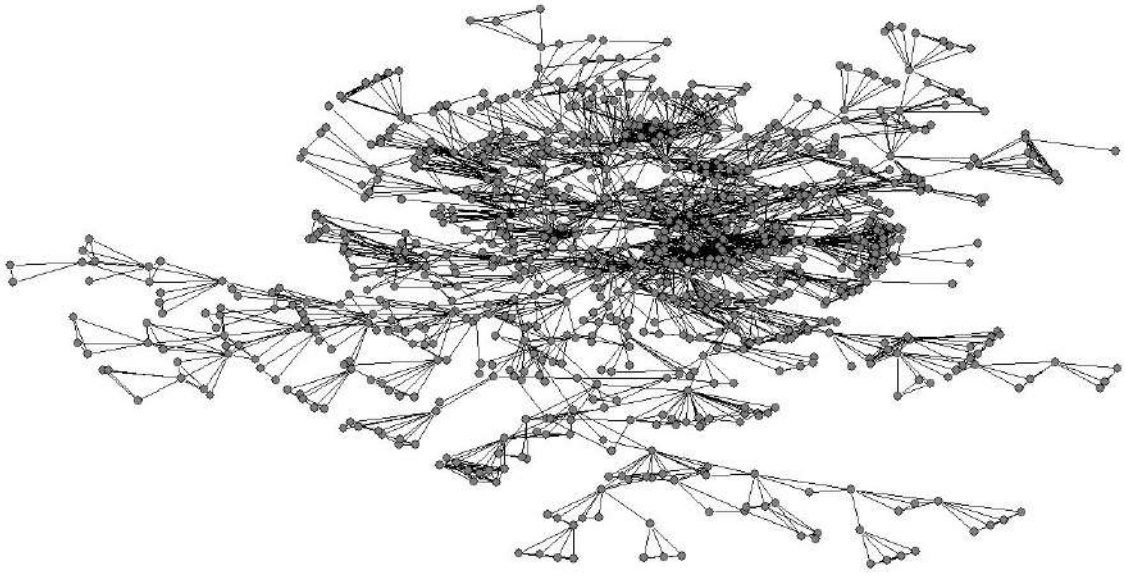


Figura VI.3 Componente gigante da rede de coautoria de publicações sobre Análise do ciclo de vida

Rede 3: rede de coautoria obtida por SOUZA et al. (2012) e cuja componente gigante contendo 2171 vértices e 6350 arestas pode ser visualizada na Figura VI.4. Este fragmento da

rede representa aproximadamente 50% da rede inicial e é composta por autores que publicaram na revista Química Nova desde sua indexação na base ISI/Web of Science.

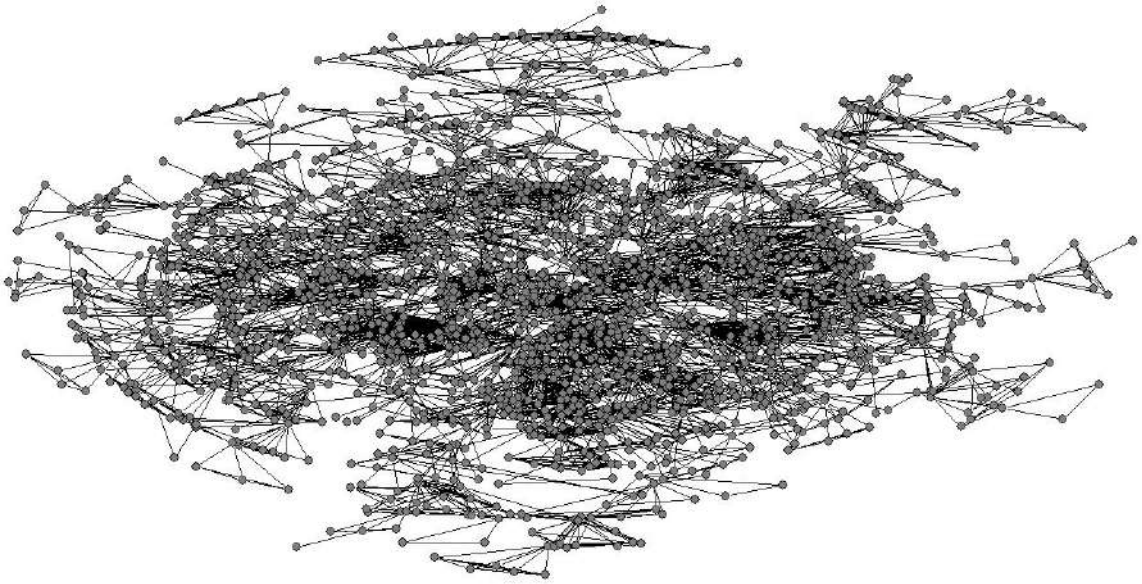


Figura VI.4 Componente gigante da rede de coautoria da revista Química Nova

Foram feitas algumas análises das propriedades dessas redes de coautoria relacionadas à medidas de centralidade e coesão.

A Tabela VI.2 apresenta alguns resultados relativos às propriedades das Redes 1, 2 e 3. Através dos valores das distâncias médias das redes, percebe-se que as redes CNPQ e ACV seguem um comportamento que se ajusta à hipótese de mundo pequeno, pois possuem em média seis graus de distância ou menos entre os autores. Já a rede formada pelos autores da revista Química Nova possui uma distância média um pouco maior de 8,196. A maior distância entre dois vértices também foi detectada na rede Química Nova, sendo de 19, e a menor refere-se à rede CNPQ, igual a 7.

A rede CNPQ é a rede mais densa das três analisadas na Tabela VI.2. Enquanto a rede ACV possui, aproximadamente 0,62% das possíveis arestas da rede, e a rede Química Nova possui, aproximadamente, 0,27%, a rede CNPQ possui mais de 2% das possíveis arestas da rede. Porém, em relação ao grau médio, a rede ACV é a que apresenta o maior valor de 5,8848 e a CNPQ o menor, 5,0242. Análises relacionadas aos graus das redes serão expostas mais à frente, comparando-se o grau normalizado dos principais autores de cada rede, pois como as redes possuem tamanhos bem diferenciados a comparação das medidas de centralidade, para ser mais criteriosa, deve utilizar valores normalizados, de acordo com o tamanho das redes analisadas.

Tabela VI.2 Propriedades das redes CNPq, ACV e Química Nova

	Rede		
Propriedade	**CNPQ**	**ACV**	**Química Nova**
Número de Autores	207	955	2171
Número de arestas	520	2810	6350
Distância média	3,601	6,592	8,196
Maior distância	7	17	19
Densidade	2,44E-02	6,17E-03	2,69E-03
Grau Médio	5,024	5,885	5,850
Coeficiente α da Lei de potências	1,217	1,763	1,979
Coeficiente de Regressão da Lei de potências - R^2	0,773	0,812	0,835
Coeficiente de Clusterização de Watts-Strogatz	0,8300	0,8333	0,8437

Uma observação dos resultados na Tabela VI.2 mostra que as três redes possuem valores próximos de grau médio, todavia nota-se que o tamanho da rede de pesquisadores do CNPq corresponde a apenas 9,5% do tamanho da rede de autores que publicaram na revista Química Nova. Essa observação é importante, pois uma rede com 207 vértices e grau médio igual a 5 é bem mais densa que uma rede com 2171 vértices e grau médio também igual a 5, pois como visto a primeira possui 2% de todas as possíveis arestas e a última, apenas 0,27%. Sendo assim, apesar de as redes possuírem valores parecidos de grau médio, a interpretação desses resultados leva a conclusões bem diferentes que poderão ser melhor visualizadas ao se comparar os valores normalizados.

O coeficiente de clusterização de Watts-Strogatz mede a razão média entre o número de pares de vértices conectados por arestas e o número de pares de vizinhos. É uma medida do quanto um grafo se aproxima localmente de um grafo completo ou clique (WATTS; STROGATZ, 1998). Como em uma rede de coautoria, cada artigo é uma clique, pois todos os autores do mesmo artigo estão ligados, em uma rede menos densa a probabilidade de um vértice estar ligado a apenas cliques é maior do que em uma rede mais densa. Isso é notado nos resultados da Tabela VI.2, a rede Química Nova é a menos densa das três e a que possui maior valor para o coeficiente de clusterização.

Através das figuras que apresentam a distribuição de graus das redes, percebe-se que apesar de a Rede CNPQ ser a mais densa, ela apresenta uma distribuição de graus mais suave (Figura VI.5a). Já as redes ACV e Química Nova, apesar de serem menos densas, são redes maiores, em que poucos autores possuem graus elevados, ou seja, estão ligados a

muitos outros autores, e a maioria possui um valor de grau pequeno, ou seja, publicam com um grupo mais limitado de outros autores (Figura VI.5b e Figura VI.5c). Percebe-se também que a distribuição dos graus da Rede Química Nova é a que mais se adéqua a lei de potências, visto que seu coeficiente de regressão da reta de tendência linear foi o maior das três, R^2=0,8353. Essas redes, entretanto, seguem lei de potências com corte exponencial, pois como se pode observar na Figura VI.5 as distribuições seguem uma curva na escala logarítmica utilizada e não uma linha reta como seria esperado caso seguissem uma lei de potências pura.

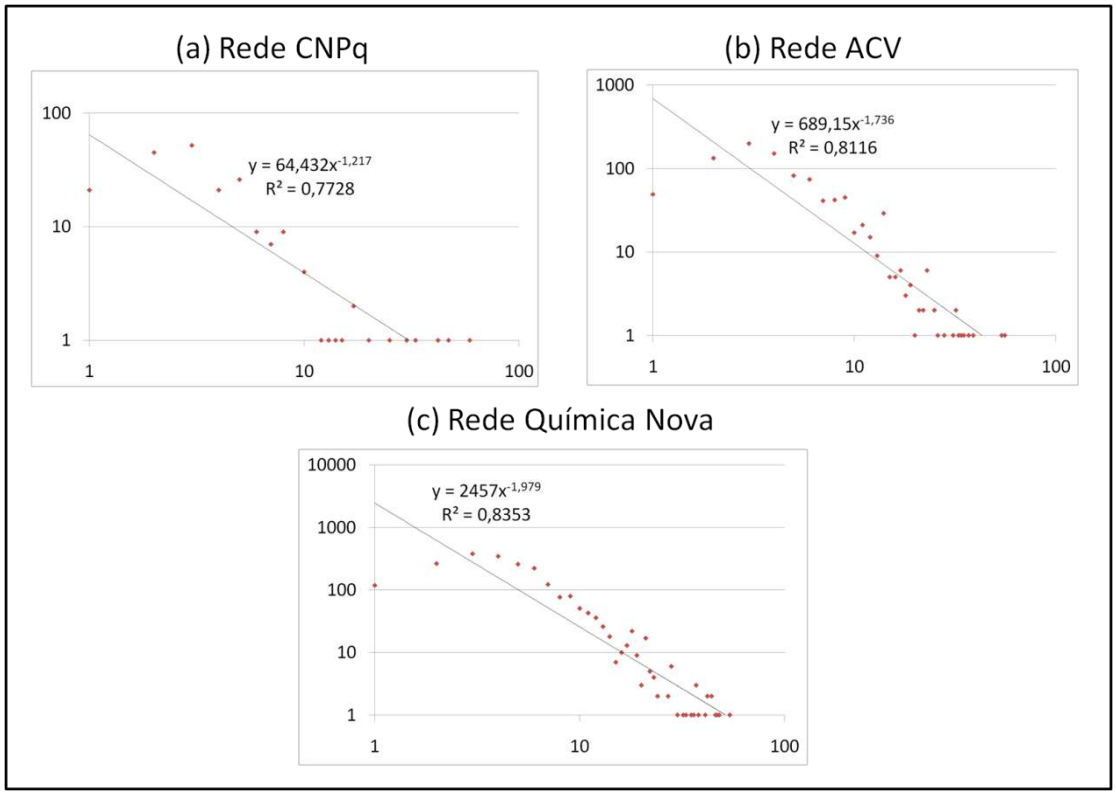


Figura VI.5 Gráfico log-log dos graus dos vértices versus a frequência dos graus da (a) Rede do CNPq; (b) Rede de Análise do Ciclo de Vida; e (c) Rede Química Nova

A taxa de decaimento da função de cada uma das redes é menor que aquela encontrada nas redes estudadas por ALBERT *et al.* (1999), isso pode ser porque nas presentes redes mesmo que seja maior o número de autores com grau pequeno e à medida que o valor do grau aumenta nota-se uma diminuição na quantidade de autores, essa queda é mais suave.

A centralidade de grau é uma das medidas mais simples e fácil de avaliar quão central é um vértice na rede. Entretanto, como observado anteriormente, não faz sentido comparar os graus de vértices de redes diferentes com tamanhos variados. Por exemplo, não se pode comparar a rede do CNPq em que o maior grau é igual a 59 e o total de vértices é igual a 207 com a rede de Química Nova onde o maior grau é igual a 54 e o total de vértices é igual a 2171. A comparação de valores absolutos nos diria que os vértices que possuem o maior grau

em cada rede possuem valor semelhante, porém na realidade, o vértice pertencente à rede do CNPq está ligado a, aproximadamente, 29% da rede, enquanto o pertencente à Química Nova conecta-se com apenas, aproximadamente, 2,5% dos vértices da rede. Assim, através dessa análise mais aprofundada, percebe-se que a rede do CNPq é mais conexa que a Química Nova. Para que fosse possível analisar os valores de centralidade dos principais autores de cada rede foi utilizada a medida normalizada, proposta por FREEMAN (1978). Escolheu-se os 25 autores com valores mais altos de grau de cada rede para compor essa comparação que pode ser vista na Tabela VI.3.

Tabela VI.3 Resultados para os 25 autores com maiores valores de grau normalizado de cada rede

	Grau normalizado					
	CNPQ		ACV		Química Nova	
1	SZWARCFITER,J.L.	0,2864	JOLLIET, O	0,0587	PINTO, AC	0,0249
2	FIGUEIREDO,C.M.H.	0,2282	HUIJBREGTS, M	0,0566	DE SOUZA, AO	0,0221
3	ABREU,N.M.M.	0,2039	HAUSCHILD, M	0,0409	BERLINCK, RGS	0,0217
4	PROTTI,F.	0,1602	VAN DE MEENT, D	0,0388	BRAZ, R	0,0212
5	KLEIN,S.	0,1456	MARGNI, M	0,0367	GALETTI, FCS	0,0203
6	FARIA,L.	0,1214	HUNGERBUHLER, K	0,0356	SILVA, CL	0,0203
7	MELLO,C.P.	0,0971	CANALS, L	0,0346	AFONSO, JC	0,0194
8	REED,B.	0,0825	HELLWEG, S	0,0335	BOLZANI, VD	0,0194
9	LUCCHESI,C.L.	0,0825	SEPPALA, J	0,0335	FERREIRA, AG	0,0189
10	MARKENZON,L.	0,0728	MCKONE, T	0,0325	YOUNG, MCM	0,0175
11	DANTAS,S.	0,0680	GUINEE, J	0,0294	BARREIRO, EJ	0,0171
12	CERIOLI,M.R.	0,0631	CLIFT, R	0,0273	CHAVES, MH	0,0171
13	OLIVEIRA,C.S.	0,0583	COWELL, S	0,0262	LONGO, E	0,0171
14	GRAVIER,S.	0,0485	HOSPIDO, A	0,0262	VIEIRA, PC	0,0166
15	DOURADO,M.C.	0,0485	BACHMANN, T	0,0241	SILVA, M	0,0161
16	MAFFRAY,F.	0,0485	DE KONING, A	0,0241	HAJDU, E	0,0152
17	KAWARABAYASHI,K.	0,0485	FINNVEDEN, G	0,0241	NOBREGA, JA	0,0147
18	GUTIERREZ,M.	0,0388	KEOLEIAN, G	0,0241	REZENDE, CM	0,0138
19	FONSECA,G.D.	0,0388	ROSENBAUM, R	0,0241	FERNANDES, JB	0,0129
20	STOLFI,J.	0,0388	SUH, S	0,0241	MORAES, MO	0,0129
21	LIN,M.C.	0,0388	SCHOLZ, R	0,0231	PEIXINHO, S	0,0129
22	DURAN,G.	0,0388	SONESSON, U	0,0231	PESSOA, CO	0,0129
23	MORGANA,A.	0,0388	MUNOZ, I	0,0220	POPPI, RJ	0,0129
24	NOGUEIRA,L.T.	0,0388	PENNINGTON, D	0,0220	TEMPONE, AG	0,0129
25	FREITAS,M.A.	0,0388	HEIJUNGS, R	0,0210	CAVALCANTI, BC	0,0124

Através dos valores de graus obtidos pode-se afirmar que os autores da rede do CNPq possuem mais conexões uns com os outros que aqueles autores pertencentes às redes ACV e Química Nova. Assim, na rede CNPq deve ser mais fácil a comunicação entre os autores, e as distâncias entre eles são menores facilitando a possibilidade de publicações ou pesquisas mesmo entre autores que ainda não estão conectados. A Tabela VI.2 já havia mostrado que a distância média entre os autores era menor na primeira rede apresentada e maior na terceira, aumentando o número de intermediários entre os autores nessa última.

A medida de proximidade calcula o quão próximo o vértice está dos demais vértices da rede. Essa medida é importante, pois quanto mais próximo um autor está dos demais, maior facilidade ele terá para obter informações, participar de pesquisas e até publicar com novos autores que possuem intermediários em comum com ele. Autores com maiores valores de proximidade são mais independentes em relação à influência dos demais e possuem maior

capacidade de ação, visto que conseguem se comunicar facilmente com diferentes autores. Assim como foi feito com o grau, calculou-se a proximidade normalizada para os 25 autores que apresentaram os maiores valores para tal medida em cada rede.

Tabela VI.4 Resultados para os 25 autores com maiores valores de proximidade normalizada de cada rede

	Proximidade normalizada					
	CNPQ		ACV		Química Nova	
1	FIGUEIREDO,C.M.H.	2,05E-03	HUIJBREGTS, M	2,57E-04	CADORE, S	9,04E-05
2	FARIA,L.	2,03E-03	HAUSCHILD, M	2,50E-04	PINTO, AC	8,92E-05
3	KLEIN,S.	2,02E-03	DE KONING, A	2,49E-04	POPPI, RJ	8,76E-05
4	PROTTI,F.	1,96E-03	JOLLIET, O	2,49E-04	BOLZANI, VD	8,72E-05
5	SZWARCFITER,J.L.	1,94E-03	GUINEE, J	2,48E-04	ARRUDA, MAZ	8,57E-05
6	DANTAS,S.	1,94E-03	SUH, S	2,47E-04	NETO, WB	8,56E-05
7	CERIOLI,M.R.	1,90E-03	ROSENBAUM, R	2,44E-04	VIEIRA, PC	8,38E-05
8	REED,B.	1,82E-03	BACHMANN, T	2,43E-04	YOUNG, MCM	8,27E-05
9	GUTIERREZ,M.	1,78E-03	HUNGERBUHLER, K	2,42E-04	BARREIRO, EJ	8,26E-05
10	FONSECA,G.D.	1,78E-03	VAN DE MEENT, D	2,41E-04	DE ANDRADE, JB	8,23E-05
11	MELLO,C.P.	1,77E-03	MOLANDER, S	2,40E-04	FERREIRA, VF	8,22E-05
12	GRAVIER,S.	1,72E-03	MARGNI, M	2,39E-04	PARDINI, VL	8,22E-05
13	STOLFI,J.	1,70E-03	MCKONE, T	2,37E-04	MOZETO, AA	8,22E-05
14	MARKENZON,L.	1,69E-03	PENNINGTON, D	2,37E-04	ZUCCO, C	8,20E-05
15	GIMBEL,J.	1,69E-03	HELLWEG, S	2,36E-04	CURI, LRL	8,20E-05
16	LUCCHESI,C.L.	1,67E-03	SLEESWIJK, A	2,34E-04	DE SOUSA, RA	8,19E-05
17	DOURADO,M.C.	1,67E-03	HUPPES, G	2,31E-04	BACCAN, N	8,18E-05
18	MENDONCANETO,C.F.X.	1,65E-03	VAN OERS, L	2,31E-04	FURLAN, M	8,18E-05
19	ESCHEN,E.	1,64E-03	FINNVEDEN, G	2,30E-04	JARDIM, WD	8,18E-05
20	SCHAFFER,K.	1,64E-03	LARSEN, H	2,29E-04	REZENDE, CM	8,13E-05
21	XAVIER,E.F.	1,64E-03	PAYET, J	2,29E-04	KUBOTA, LT	7,99E-05
22	ORTIZ,C.	1,64E-03	GOLD, L	2,28E-04	SILVA, JRD	7,98E-05
23	SRITHARAN,R.	1,64E-03	JURASKE, R	2,28E-04	FERREIRA, AG	7,96E-05
24	VILLANUEVA,M.I.	1,63E-03	KOEHLER, A	2,28E-04	ROCHA, JC	7,94E-05
25	FERREIRA,T.O.	1,63E-03	MACLEOD, M	2,28E-04	PINHEIRO, MLB	7,93E-05

Nota-se que os autores da rede CNPq também possuem maior valor de proximidade que os das outras duas redes, e assim, podem ser mais eficientes em obter informações na rede. Apesar de os autores da rede CNPq possuírem valores mais altos de proximidade facilitando a obtenção de novas informações e acesso a pesquisas, como esta rede é bem menor que as demais, uma análise pertinente é que a informação obtida por tais autores pode ser bem mais limitada, pois considerando apenas o Universo de cada rede eles possuem acesso a, no máximo, 206 autores. Já os autores pertencentes à rede Química Nova, ainda que estejam próximos a poucos autores relativamente, as informações que chegam a eles podem ser provenientes de até 2170 autores, ainda que indiretamente, o que pode facilitar o acesso a maior quantidade de informações. Comparando-se os autores das mesmas redes, entretanto, percebe-se que os valores não variam muito, o que para esses 25 principais quer dizer que nenhum deles possui grande vantagem em relação aos demais presentes na Tabela VI.4.

A medida de intermediação de um vértice mostra o quão ele é necessário como elo entre grupos, sendo importante no fluxo de informações que circulam na rede. Essa medida calcula os caminhos mais curtos de todos os vértices para todos os outros que passam pelo

vértice analisado. A Tabela VI.5 mostra os 25 maiores valores de intermediação normalizada para cada rede. Diferentemente dos resultados apresentados pelas medidas de grau e proximidade, os valores de intermediação mais altos foram os dos autores da rede ACV, indicando que esses vértices possuem mais importância para o fluxo de informações dessa rede. Observa-se também que o valor de intermediação cai bastante se comparar os primeiros autores de cada rede com os últimos, mostrando que poucos são os mais eficientes, reforçando a sua condição de facilitadores no fluxo de informações entre autores da rede.

Tabela VI.5 Resultados para os 25 autores com maiores valores de intermediação normalizada de cada rede

	Intermediação normalizada					
	CNPQ		ACV		Química Nova	
1	FARIA,L.	2,01E-03	SUH, S	2,35E-04	POPPI, RJ	1,17E-04
2	MARKENZON,L.	1,95E-03	HUIJBREGTS, M	2,23E-04	PINTO, AC	1,12E-04
3	SZWARCFITER,J.L.	1,69E-03	EKVALL, T	1,77E-04	CADORE, S	1,06E-04
4	ABREU,N.M.M.	1,51E-03	HAUSCHILD, M	1,72E-04	BOLZANI, VD	7,91E-05
5	FIGUEIREDO,C.M.H.	8,82E-04	ANDRAE, A	1,59E-04	FERREIRA, AG	7,11E-05
6	PROTTI,F.	7,79E-04	JOLLIET, O	1,55E-04	NOBREGA, JA	6,65E-05
7	KLEIN,S.	7,20E-04	MOLANDER, S	1,45E-04	KUBOTA, LT	5,12E-05
8	LUCCHESI,C.L.	6,37E-04	MATTHEWS, H	1,42E-04	YOUNG, MCM	5,08E-05
9	MELLO,C.P.	4,68E-04	ITSUBO, N	1,31E-04	BRAZ, R	5,03E-05
10	REED,B.	3,47E-04	YAMAMOTO, R	1,29E-04	MONTANARI, CA	4,99E-05
11	STOLFI,J.	2,32E-04	MACLEAN, H	1,24E-04	ROCHA, FRP	4,79E-05
12	DANTAS,S.	2,01E-04	FINNVEDEN, G	1,20E-04	MOZETO, AA	4,51E-05
13	LEE,O.	1,51E-04	SONESSON, U	1,01E-04	ROCHA, JC	3,52E-05
14	NOGUEIRA,L.T.	1,42E-04	HUR, T	1,00E-04	PILO-VELOSO, D	3,52E-05
15	CERIOLI,M.R.	1,33E-04	CLIFT, R	9,96E-05	VIEIRA, PC	3,46E-05
16	KAWARABAYASHI,K.	1,24E-04	CANALS, L	8,24E-05	FURLAN, M	3,25E-05
17	FONSECA,G.D.	1,07E-04	HANSEN, T	8,11E-05	BENVENUTTI, EV	3,16E-05
18	GUTIERREZ,M.	1,04E-04	ZHANG, Y	7,81E-05	FERREIRA, VF	3,05E-05
19	CARVALHO,M.H.	9,40E-05	HUNGERBUHLER, K	7,73E-05	ARRUDA, MAZ	2,94E-05
20	ARAUJO,L.H.C.	8,80E-05	SCHOLZ, R	6,33E-05	KAMOGAWA, MY	2,93E-05
21	JUSTEL,C.M.	7,44E-05	SEPPALA, J	6,24E-05	BARREIRO, EJ	2,86E-05
22	MORGANA,A.	4,70E-05	INABA, A	6,20E-05	MORO, CC	2,82E-05
23	GRAVIER,S.	3,46E-05	SCHMIDT, S	5,98E-05	SENA, MM	2,80E-05
24	DOURADO,M.C.	3,11E-05	TILLMAN, A	5,87E-05	NETO, WB	2,52E-05
25	OLIVEIRA,C.S.	2,75E-05	WANG, H	5,61E-05	SILVA, LC	2,44E-05

Comparando-se os resultados apresentados pelas três Tabelas anteriores, percebe-se que na rede CNPq a maior parte dos principais autores (um total de 16) está presente nas três Tabelas, ou seja, estão entre os 25 que possuem os maiores valores das três medidas de centralidade analisadas. Esses autores são de grande importância para a rede, pois são facilitadores no acesso de informação, estão ligados a um grande número de outros autores e comunicam-se com grande parte deles. Já na rede ACV, apenas 6 autores estão presentes nas três Tabelas, indicando que existem nessa rede diferentes autores com grau de importância diferenciado, alguns sendo mais eficientes para troca de informações, outros na obtenção dessas. Na rede Química Nova, assim como na ACV, poucos autores possuem altos valores para as três medidas de centralidade analisadas, apenas 7. Essas duas redes (ACV e Química Nova) são redes maiores que a do CNPq e menos densas, indicando que os autores que possuem mais conexões não necessariamente possuem um papel importante no fluxo de

informações, pois ele pode estar ligado a um grupo que possui poucas conexões com os demais autores da rede, dificultando o acesso e disseminação das informações.

A etapa de emprego dos algoritmos para identificação de agrupamentos nas redes utilizou os *softwares Octave* e *Matlab*. Os algoritmos espectrais foram implementados no *Octave* e utilizou-se uma função já existente no *Matlab* para uso do *k-means.*

Os algoritmos espectrais com matriz Laplaciana e Laplaciana sem sinal e o algoritmo *k-means* foram testados em cada rede com diferentes quantidades *k* de *clusteres*. O objetivo da divisão em diferentes quantidades para *k* foi detectar o melhor agrupamento dos autores para cada rede, e a qualidade dessas divisões foi calculada através da modularidade. Portanto quando o valor da modularidade parava de crescer e começava a cair, parava-se de testar novas quantidades de *clusteres*.

A Tabela VI.6 apresenta os valores de *k* utilizados nas rodadas para a rede CNPq e os resultados da modularidade para cada algoritmo. Conforme pode ser observado na Tabela a seguir para alguns valores de *k* não há resultados para todos os algoritmos. Isso ocorreu, pois ao se escolher o tamanho mínimo de um *cluster* para ser subdividido, nesses casos mais de um *cluster* possuía o mesmo tamanho, assim diferentes *clusteres* foram subdivididos e não foi possível para tais algoritmos dividir a rede em certas quantidades. Isso também pode ser verificado nas Tabelas referentes às redes ACV e Química Nova.

Tabela VI.6 Resultados da modularidade para diferentes valores de *k* na rede CNPq

	Modularidade		
	Algoritmo espectral		
CNPq	Matriz Laplaciana sem sinal (Q)	Matriz Laplaciana com sinal (L)	*k-means*
k=2	0,2534	0,3633	0,3614
k=3	0,3681	0,4601	0,3763
k=4	0,4865	0,5767	0,3550
k=5	0,5371	0,5958	0,5039
k=6	0,5468	0,6187	0,2724
k=7		0,6171	0,3579
k=8	0,5378	0,6100	0,4246
k=11	0,5190	0,5808	0,2756
k=14	0,5309	0,5068	0,3287
k=15	0,5270	0,5060	0,2586

A Figura VI.6 apresenta os resultados de modularidade para cada algoritmo utilizado na rede CNPq, o gráfico ajuda a visualizar melhor o comportamento de cada algoritmo com diferentes valores de *clusteres*. Nota-se que os algoritmos espectrais com matrizes Laplaciana

e Laplaciana sem sinal possuem um comportamento parecido, os valores de modularidade começam a crescer de acordo com o aumento da quantidade de *clusteres*, atingem um valor máximo e começam a cair. Já o *k-means* apresenta um comportamento diferente em que a modularidade cai e sobe um pouco, mas após atingir o valor máximo, as subidas são menores, para as quantidades de *k* calculadas. Os rótulos dos resultados apresentados no gráfico correspondem ao valor máximo da modularidade para cada algoritmo. Os algoritmos espectrais apresentaram seus melhores resultados quando *k=6*, já o *k-means* apresentou melhor resultado para *k=5*. No entanto, o melhor resultado foi aquele gerado pelo algoritmo espectral com matriz Laplaciana.

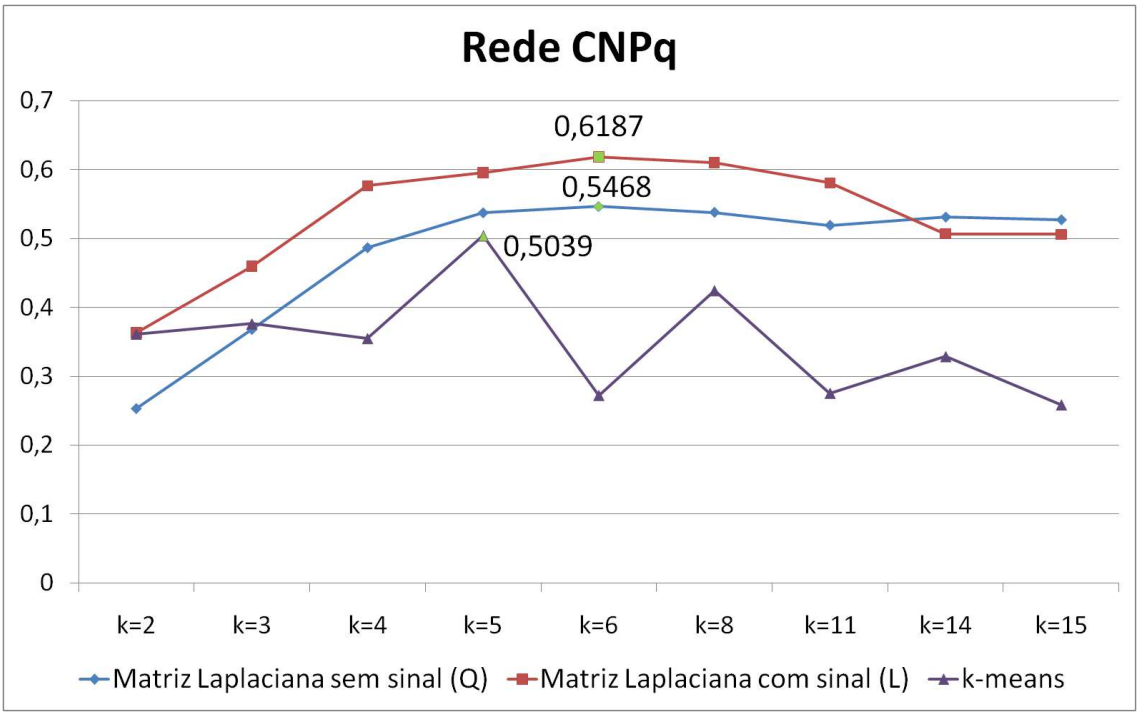


Figura VI.6 Valores da medida modularidade para diferentes quantidades de *clusteres* em cada algoritmo na Rede CNPq

A Tabela VI.7 apresenta os valores de *k* utilizados nas rodadas na rede ACV e os resultados da modularidade para cada algoritmo.

A Figura VI.7 apresenta os resultados de modularidade para cada algoritmo utilizado na rede ACV de acordo com os diferentes números de *clusteres k* gerados. Para quantidades de *clusteres* menores do que 5, o algoritmo com matriz Laplaciana sem sinal apresentou melhores resultados, porém com o aumento de *k* os resultados do algoritmo com a matriz Laplaciana foram melhorando. O algoritmo espectral com matriz Laplaciana sem sinal obteve o melhor valor com *k=22*; já com *k=25* obteve-se o melhor valor de modularidade com o uso da matriz Laplaciana, sendo este valor o melhor dentre todos os apresentados pelos três algoritmos. Para o *k-means* o melhor resultado foi para *k=35*.

Tabela VI.7 Resultados da modularidade para diferentes valores de *k* na rede ACV

ACV	Modularidade		
	Algoritmo espectral		
	Matriz Laplaciana sem sinal (Q)	Matriz Laplaciana com sinal (L)	*k-means*
k=2	0,3028	0,1974	0,2459
k=3	0,4458	0,3784	0,3509
k=4	0,5659	0,4905	0,5249
k=5	0,6631	0,6796	0,5551
k=11	0,7430	0,8032	0,6429
k=13	0,7620	0,8135	0,6573
k=17	0,7682	0,8151	0,6101
k=20	0,7737	0,8180	0,6524
k=22	0,7774	0,8195	0,6425
k=25	0,7730	0,8218	0,6333
k=27	0,7700	0,8201	0,6480
k=30	0,7685		0,6105
k=31	0,7671	0,8189	0,6136
k=32		0,8159	0,6077
k=35	0,7609	0,8162	0,6582

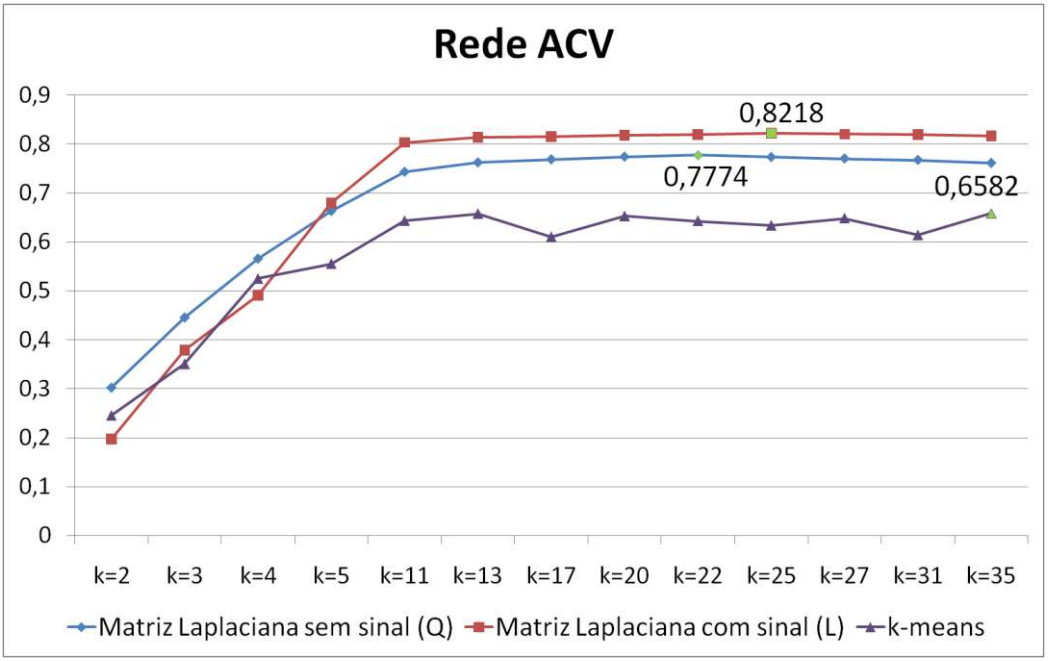


Figura VI.7 Valores da medida modularidade para diferentes quantidades de *clusteres* em cada algoritmo na Rede ACV

A Tabela VI.8 apresenta os valores de *k* utilizados nas rodadas para a rede Química Nova e os resultados da modularidade para cada algoritmo.

Tabela VI.8 Resultados da modularidade para diferentes valores de *k* na rede Química Nova

Química Nova	**Modularidade**		
	Algoritmo espectral		
	Matriz Laplaciana sem sinal (Q)	Matriz Laplaciana com sinal (L)	*k-means*
k=2	0,184	0,1556	0,2573
k=3	0,4695		0,4637
k=4	0,6659	0,6819	0,5761
k=5	0,7251	0,7435	0,613
k=11	0,7888	0,8539	0,6683
k=12	0,7962	0,8601	0,7325
k=14	0,8081	0,8655	0,717
k=15	0,8211		0,7358
k=16	0,8282		0,7343
k=17	0,8325	0,8798	0,6907
k=18	0,8394	0,8813	0,7247
k=20	0,8445	0,889	0,7318
k=21	0,8453	0,8905	0,707
k=22	0,8463	0,8913	0,7285
k=24		0,8948	0,7173
k=25	0,8505		0,7552
k=26	0,849	0,8982	0,7612
k=30	0,8498	0,8985	0,7013
k=32	0,8477	0,9008	0,7324
k=34	0,8469	0,8972	0,7446
k=35	0,8471		0,7119
k=36	0,8489	0,8977	0,7489
k=40		0,8947	0,7347
k=41	0,8499	0,8948	0,704

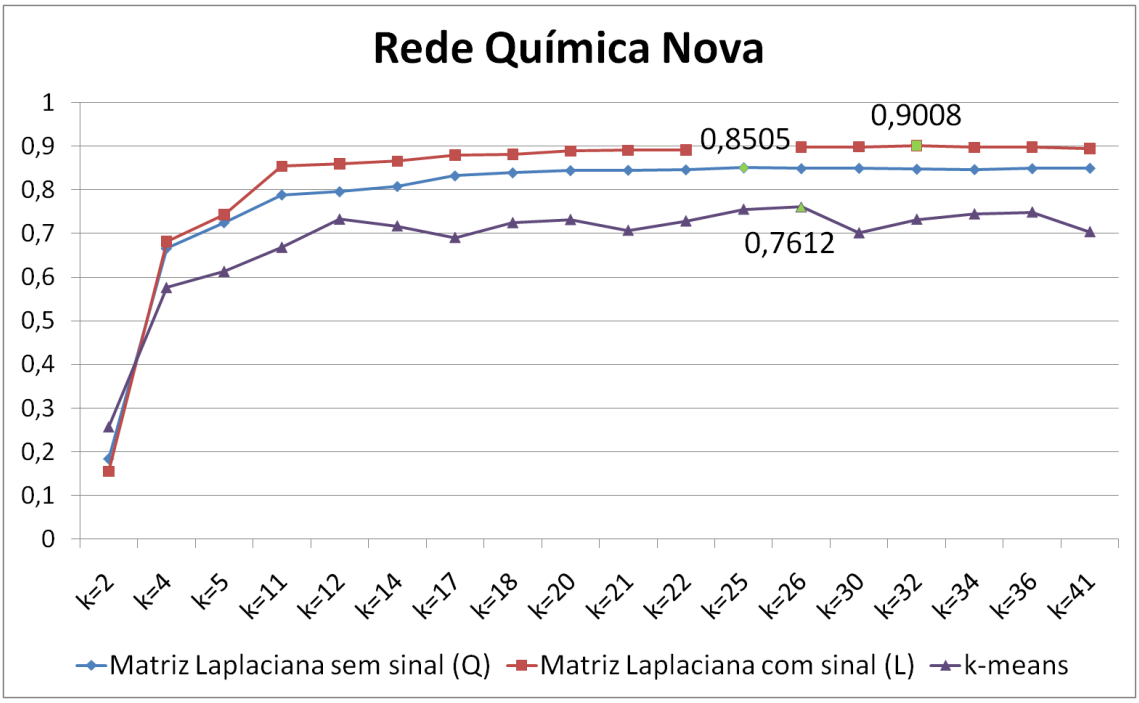


Figura VI.8 Valores da medida modularidade para diferentes quantidades de *clusteres* em cada algoritmo na Rede Química Nova

A Figura VI.8 apresenta os resultados referentes à rede Química Nova. Nota-se que os algoritmos apresentaram comportamentos semelhantes, com o valor da modularidade crescendo até *k=20* e depois o crescimento e queda foram pequenos quando se aumentava o número de *clusteres*. Novamente, o algoritmo espectral com a matriz Laplaciana foi o que apresentou o melhor resultado, chegando a 0,9008 para *k=32*, o algoritmo espectral com matriz Laplaciana sem sinal obteve uma modularidade de 0,8505 para *k=25*, e o *k-means*, 0,7612 para *k=26*.

Nessas três redes de coautoria, o algoritmo espectral com matriz Laplaciana foi o que apresentou o melhor resultado de modularidade, o algoritmo espectral com matriz Laplaciana sem sinal veio em segundo lugar seguido do *k-means*. Assim, o uso de vetor de Fiedler como vetor de corte para agrupamento apresentou bons resultados para essas redes em comparação com os demais métodos testados.

Nota-se que mesmo o algoritmo espectral com matriz Laplaciana sendo o melhor método também na rede do CNPq, o valor da modularidade apresentado para o algoritmo nessa rede foi o menor se comparado ao valor de modularidade das demais redes. Isso pode ser devido ao fato de que como já observado essa rede é a mais densa entre as três, sendo assim, é a que possui menos característica de formação de *clusteres*. Já o melhor valor de modularidade foi aquele encontrado com o emprego do algoritmo da Laplaciana na rede Química Nova, chegando a um valor bem mais próximo de 1, como visto na Tabela VI.2, essa rede foi a que apresentou o maior coeficiente de clusterização, o que indica a existência de

cliques locais, que ajudam na formação de grupos. A rede de ACV também obteve um bom resultado de modularidade.

Como o valor da modularidade varia de -1 a 1, onde Q=0 indica que as arestas dentro dos grupos estão dispostas aleatoriamente e Q=1 indicando forte estrutura de comunidade, percebe-se que os resultados apresentados para essas redes indicam que esses algoritmos conseguiram detectar bem os grupos existentes nas redes. Além disso, segundo NEWMAN; GIRVAN (2004), os valores encontrados para a modularidade costumam estar entre 0,3 e 0,7, e valores maiores que esses são raros, o que reitera os bons resultados apresentados pelos algoritmos espectrais especialmente para as redes ACV e Química Nova.

A partir do melhor resultado de agrupamento encontrado para a rede CNPq foram feitas algumas análises desses grupos buscando detectar as linhas de pesquisa dos principais autores dessas partições e se há bastante comunicação entre esses grupos. Para a realização da análise foram consultados os currículos Lattes dos principais autores citados.

A Figura VI.9 mostra as partições da Rede CNPq para *k=6* gerada pelo algoritmo espectral com uso da matriz Laplaciana. A partição (2) é a maior de todas as partições, com 55 vértices, e nota-se que ela é a mais destacada, apresentando apenas duas ligações externas com a partição (10). A segunda maior partição é a (6), com 52 vértices, porém essa apresenta mais ligações externas comunicando-se com autores de diferentes grupos.

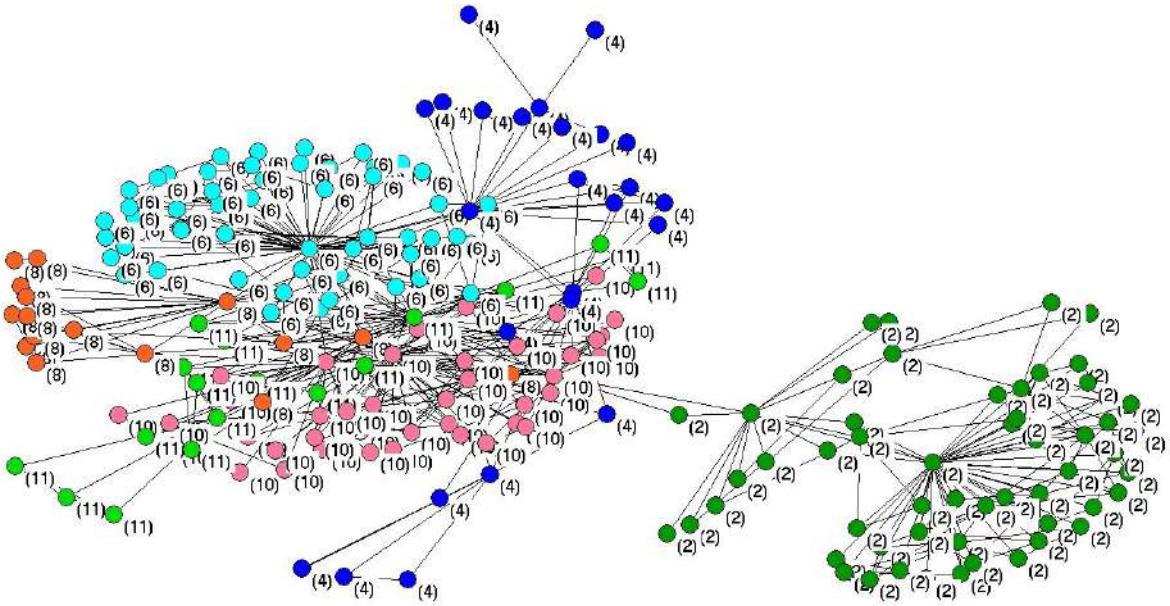


Figura VI.9 Partições geradas pelo algoritmo espectral com matriz Laplaciana para a Rede CNPq

A Figura VI.10 apresenta a maior partição encontrada no agrupamento realizado pela matriz Laplaciana para a Rede CNPq, contando com 55 vértices. Através de análises das medidas de centralidade dos autores dessa rede, nota-se que duas autoras são as mais

centrais, possuindo os maiores valores de grau, proximidade e intermediação, ABREU, N.M.M. (Nair Maria Maia de Abreu) e MARKENZON,L. (Lilian Markezon). Na Figura VI.10 percebe-se que há uma aglomeração maior ao redor delas. Uma análise de seus currículos Lattes mostra que ambas são pesquisadoras há muitos anos, sendo as linhas de pesquisa da Nair Maria Maia de Abreu: Otimização Combinatória e Teoria dos Grafos, e da Lilian Markezon: Estruturas de dados e algoritmos em grafos, Algoritmos em Grafos e Modelagem e Algoritmos em Redes.

Outras duas autoras que também ficaram entre os autores com maiores valores para as medidas de centralidade foram OLIVEIRA,C.S. (Carla Silva Oliveira) e JUSTEL,C.M. (Claudia Marcela Justel). OLIVEIRA,C.S. apresentou o terceiro maior valor de grau e intermediação e o quarto maior valor de proximidade, já JUSTEL,C.M. apresentou o sétimo maior valor para grau, o terceiro para proximidade e quarto para intermediação. É importante notar a presença dessas duas autoras na mesma partição que a autora Nair Maria Maia de Abreu, visto que as três autoras participam do mesmo grupo de pesquisa na COPPE/UFRJ.

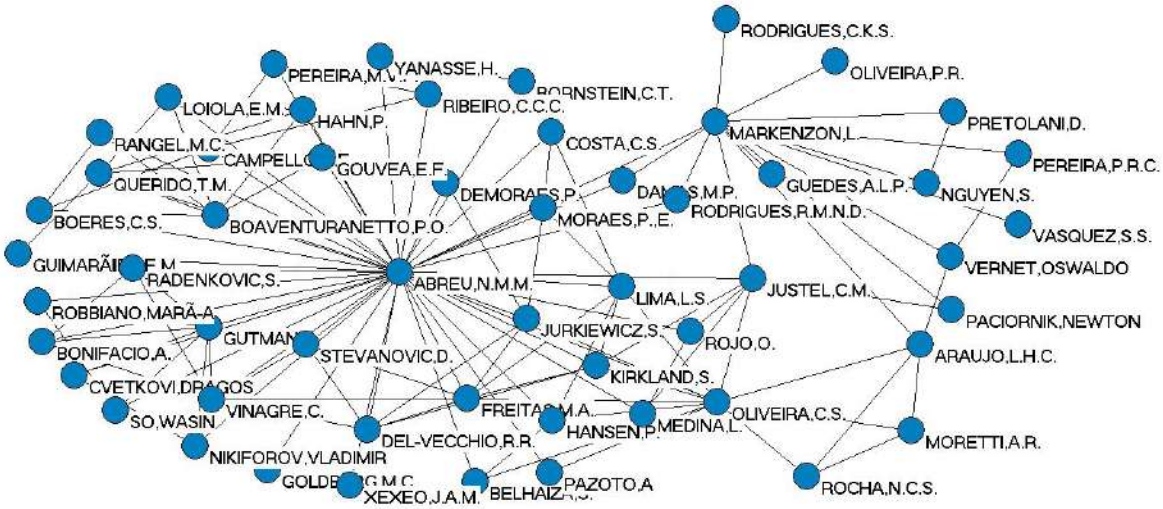


Figura VI.10 Resultado da maior partição encontrada para a Rede CNPq

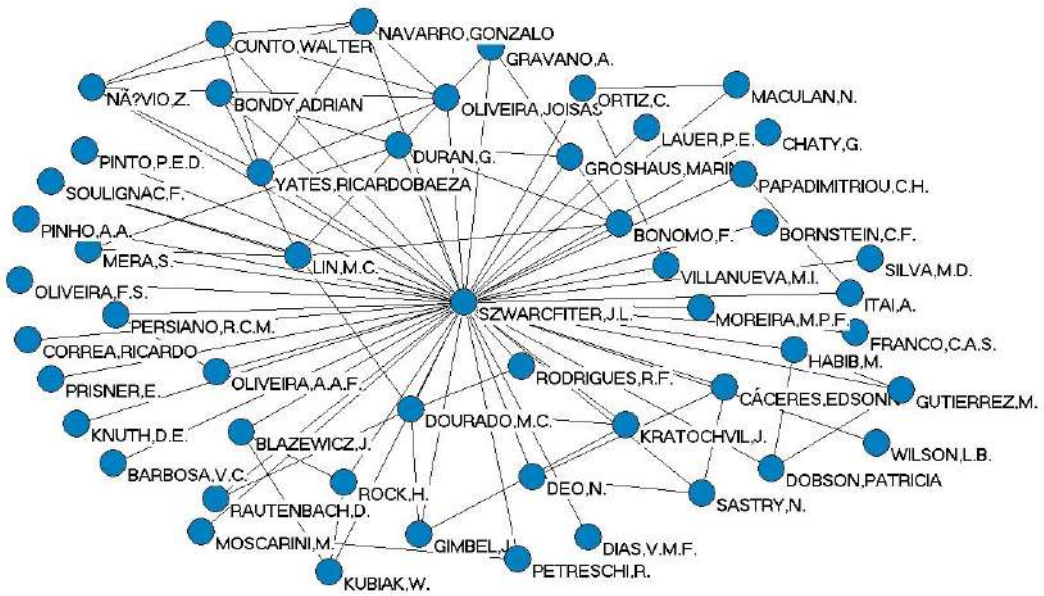


Figura VI.11 Resultado da segunda maior partição encontrada para a Rede CNPq

A Figura VI.11 apresenta o segundo maior grupo formado pela divisão na Rede CNPq. Nessa partição, é notável a importância do autor SZWARCFITER,J.L. (Jayme Luiz Szwarcfiter), ele é o autor mais central dessa partição possuindo ligação com todos os outros, sendo também o que possui maior valor para proximidade e intermediação. Outros três autores que também possuem os maiores valores para grau, proximidade e intermediação são: DURAN,G., LIN,M.C. e DOURADO,M.C.. Outros autores importantes na rede, que conforme as Tabelas VI.3, VI.4 e VI.5 apresentaram-se entre os 25 mais importantes, como FIGUEIREDO,C.M.H., FARIA,L., CERIOLI,M.R. e PROTTI,F. fazem parte de outras partições, sendo que FIGUEIREDO e C.M.H., FARIA,L. são os mais centrais em uma delas, a qual o CERIOLI,M.R. também pertence; e o PROTTI,F. é o mais central na partição da qual faz parte.

Essas análises demonstram a importância das partições para estudos em redes de coautoria, permitindo identificar grupos de pesquisadores que desenvolvem projetos em conjuntos, e aqueles que estão ligados a pesquisadores de outros grupos facilitando o fluxo de informações na rede, além de contribuir para uma maior divulgação de trabalhos e até mesmo descobertas científicas.

VI.3 Caso 3: redes extraídas da base DBLP

A seleção das redes retiradas da base de dados DBLP abrange publicações que datam de 1965 a 2013, porém nem todos os autores selecionados publicaram durante esse período de tempo, dessa maneira, foi selecionado o intervalo de tempo em que havia publicações dos 25 autores, compreendendo os anos de 1995 a 2007. Foram montadas seis redes, Rede I –

com as publicações do período de 1965 a 1994; Rede II – publicações de 1995 a 2001; Rede III – publicações de 2002 a 2007, Rede IV – publicações de 1995 a 2007; Rede V – publicações de 2008 a 2013; e Rede VI – publicações de 1965 a 2013. Optou-se por dividir em redes menores para avaliar os resultados e se o tamanho das redes poderá influenciar em melhorias nas comunidades geradas pelos algoritmos.

Nas figuras VI.12 a VI.17 é possível visualizar o desenho formado por essas redes, nota-se que não são redes conexas, possuindo diversas componentes. É interessante observar que apesar de as redes II, III, IV e V abrangerem intervalos de tempo de, em média, 6 anos, a Rede V possui um número maior de vértices bem aparente, isso se deve ao fato, já mencionado, de, no intervalo de tempo abrangido por essa rede, terem aumentado o número de publicações na área de ciências da computação.

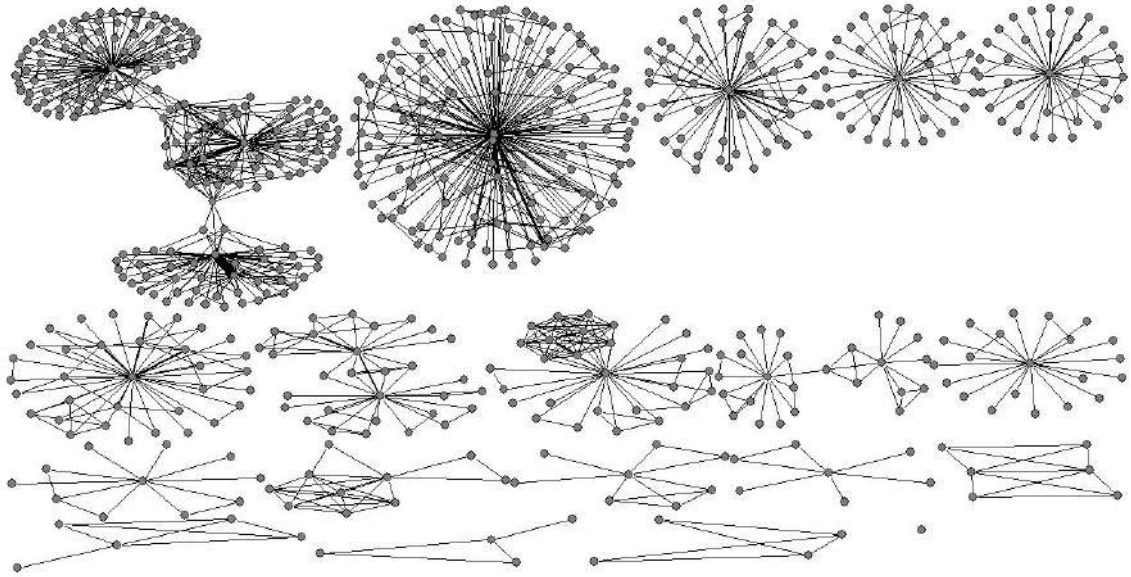


Figura VI.12 Rede I – Publicações de 1965 a 1994

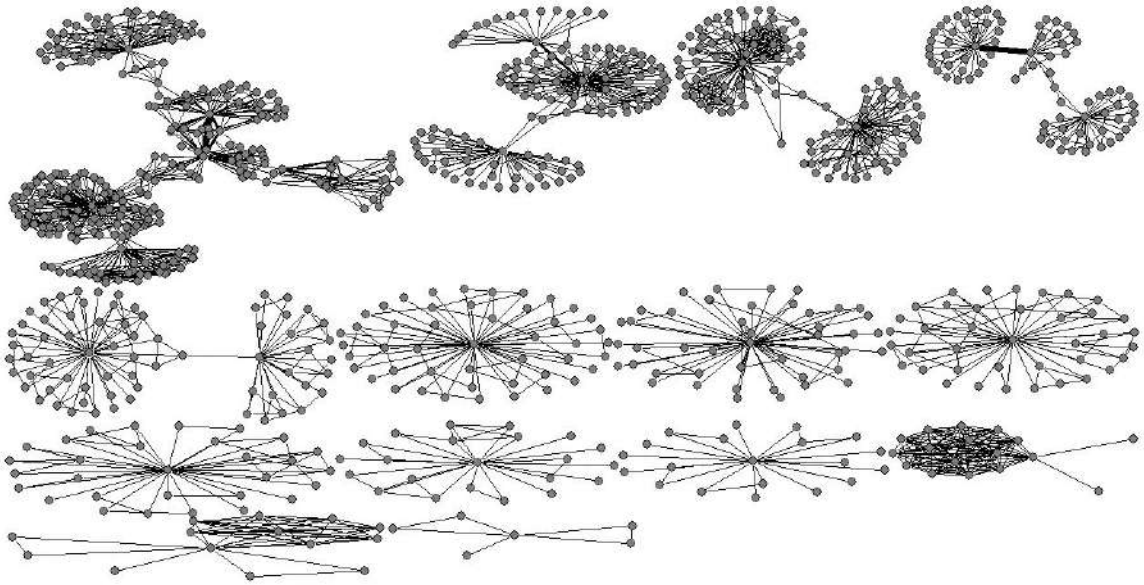


Figura VI.13 Rede II – Publicações de 1995 a 2001

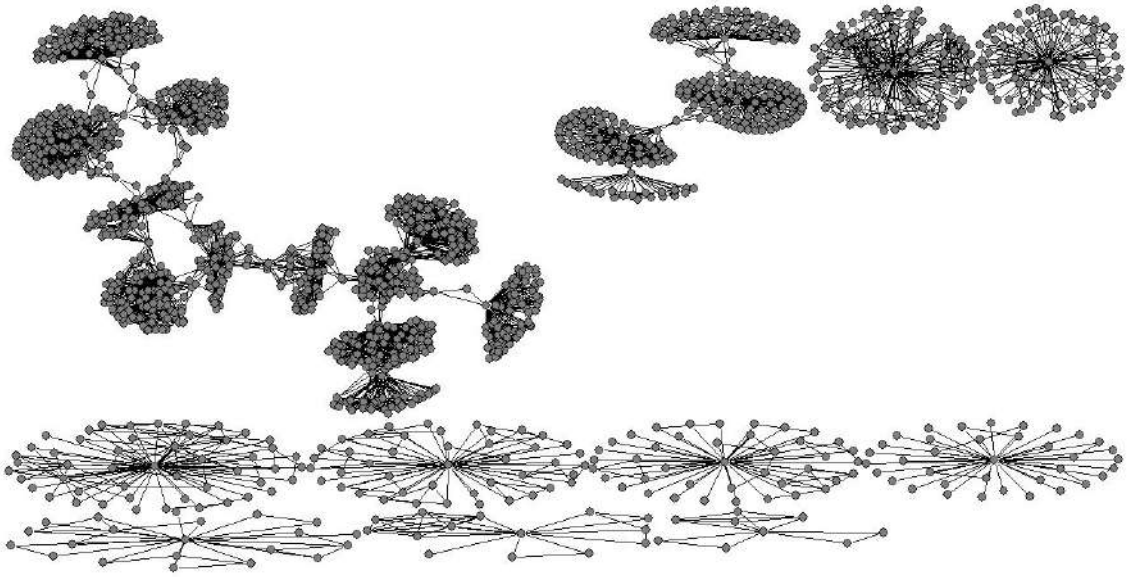


Figura VI.14 Rede III – Publicações de 2002 a 2007

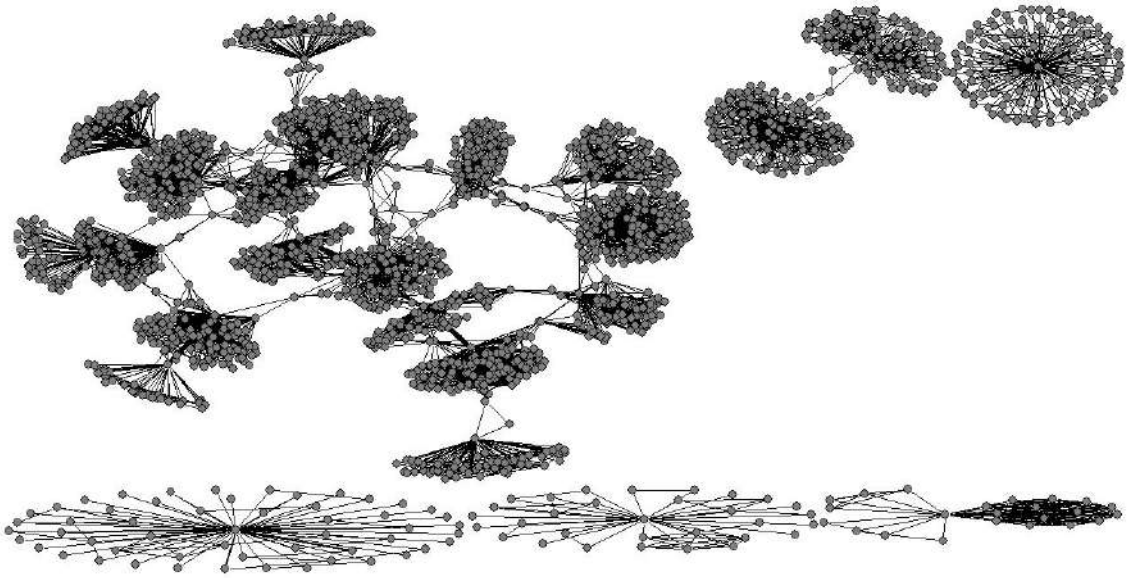


Figura VI.15 Rede IV – Publicações de 1995 a 2007

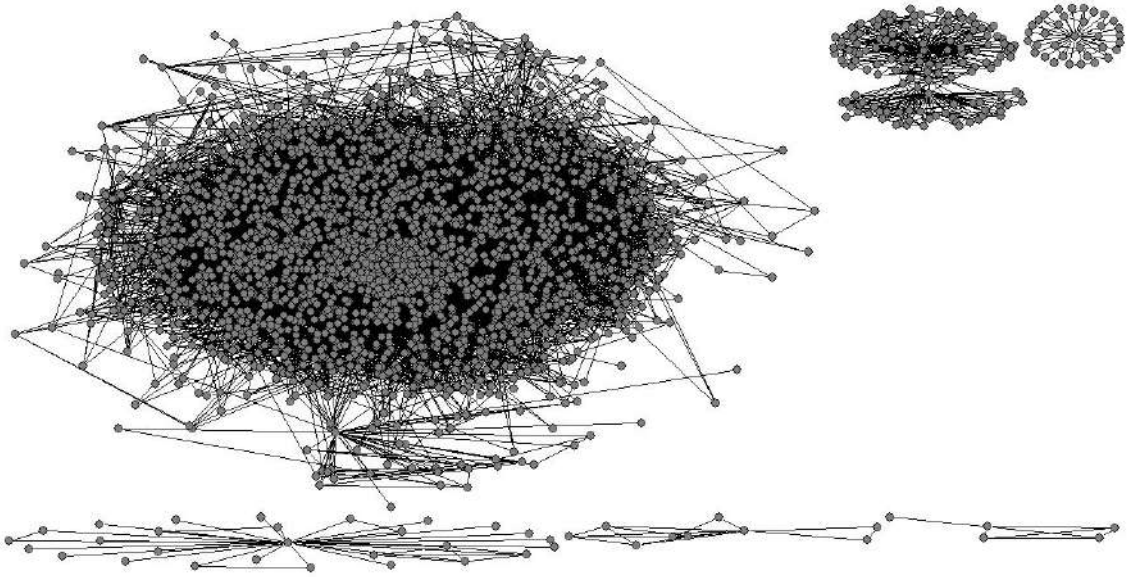


Figura VI.16 Rede V – Publicações de 2008 a 2013

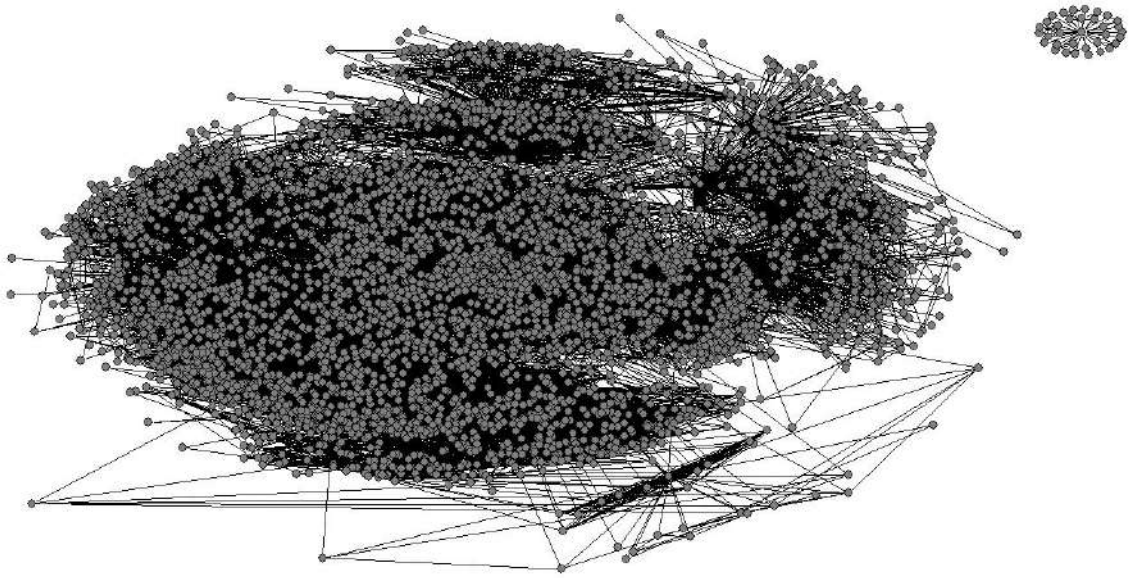


Figura VI.17 Rede VI – Publicações de 1965 a 2013

Tabela VI.9 Propriedades das redes extraídas da base de dados DBLP

	Rede					
Propriedade	1965 - 1994 **Rede I**	1995-2001 **Rede II**	2002-2007 **Rede III**	1995-2007 **Rede IV**	2008-2013 **Rede V**	1965-2013 **Rede VI**
Número de Autores	643	907	1.732	2.349	2.907	5.099
Número de arestas	1.153	1.899	4.685	6.228	9.690	15.661
Distância média	2,818	3,749	6,176	6,066	4,224	4,448
Maior distância	6	8	13	13	9	9
Densidade	5,56E-03	4,59E-03	3,12E-03	2,25E-03	2,29E-03	1,20E-03
Grau Médio	3,586	4,187	5,410	5,303	6,667	6,143
Número de vértices na Componente Gigante	184	285	1.094	1.895	2.680	5.065
Percentual de vértices na Componente Gigante	28,62%	31,42%	63,16%	80,67%	92,19%	99,33%
Coeficiente α da Lei de potências	1,363	1,494	1,419	1,399	1,306	1,311
Coeficiente de Regressão da Lei de potências - R^2	0,765	0,782	0,759	0,774	0,693	0,714
Coeficiente de Clusterização de Watts-Strogatz	0,8739	0,8971	0,8965	0,8932	0,8885	0,8783

Na Tabela VI.9 são apresentados os valores de algumas das propriedades dessas redes. Nota-se que nas redes III, IV, V e VI, as componentes gigantes abrangem mais de 60% dos vértices das redes, sendo que nas redes V e VI esse valor ultrapassa os 90%, assim é evidente que existe um grupo que se comunica com grande parte da rede direta ou indiretamente e que, provavelmente, há um grande fluxo de informações dentro dessas componentes. As distâncias médias também estão de acordo com a hipótese de mundo pequeno, ou seja, os autores possuem relações de proximidade mesmo na maior rede com

mais de 5000 vértices. Na Rede I, em que a distância média é de menos 3, nota-se através da Figura VI.12 que como existem diversas componentes de tamanhos pequenos, então as distâncias acabam por ser menores.

As Tabelas VI.10, VI.11, VI.12 apresentam os valores das medidas de centralidade de grau, proximidade e intermediação normalizados dos 25 autores que foram selecionados para geração das redes. Alguns autores estão sem valor para grau, proximidade e intermediação em alguma das Tabelas, como o Yan Zhang na Rede I da Tabela VI.10, pois eles não fazem parte de determinada rede, no caso do autor exemplificado este não possui publicações no período de 1965 a 1994. É interessante observar que na Rede I, o autor Xiadong Wang possui valores de centralidade iguais a zero, isso ocorre, pois no período de 1965 a 1994 ele só publicou um artigo e sozinho, assim não apresentou relação de coautoria com nenhum outro autor nessa rede.

Tabela VI.10 Valores de Grau, Proximidade e Intermediação normalizados dos 25 autores nas Redes I e II

	Rede I			**Rede II**		
Autores	Grau	Proximidade	Intermediação	Grau	Proximidade	Intermediação
1 Chin-Chen Chang	0,0514	8,24E-05	3,63E-06	0,0508	5,72E-05	2,68E-06
2 Witold Pedrycz	0,0234	3,75E-05	1,81E-06	0,0497	5,04E-05	5,01E-06
3 H. Vincent Poor	0,0312	5,09E-05	1,42E-06	0,0375	4,12E-05	5,13E-06
4 Noga Alon	0,1293	1,79E-04	8,26E-05	0,1159	1,19E-04	7,38E-05
5 Azriel Rosenfeld	0,2103	3,29E-04	6,73E-05	0,0596	7,87E-05	3,40E-05
6 Lajos Hanzo	0,0078	1,45E-05	1,51E-08	0,0331	3,69E-05	4,72E-06
7 Xiaodong Wang	0,0000	0,00E+00	0,00E+00	0,0210	4,49E-05	5,40E-06
8 Grzegorz Rozenberg	0,0717	1,35E-04	4,72E-05	0,0221	7,59E-05	5,79E-06
9 Georgios B. Giannakis	0,0156	2,66E-05	3,14E-07	0,0419	4,75E-05	1,76E-06
10 Xuemin Shen	0,0062	1,21E-05	2,27E-08	0,0055	7,30E-06	2,15E-08
11 Ming Li	0,0327	5,40E-05	2,24E-06	0,0519	1,10E-04	5,00E-05
12 Ronald R. Yager	0,0156	2,92E-05	1,09E-06	0,0210	5,97E-05	3,88E-06
13 Vladik Kreinovich	0,0109	1,94E-05	1,29E-07	0,0784	8,53E-05	1,65E-05
14 Yan Zhang	-	-	-	0,0143	1,70E-05	1,29E-07
15 Robert L. Glass	0,0140	2,42E-05	1,25E-07	0,0177	2,07E-05	7,81E-08
16 Guanrong Chen	0,0047	9,69E-06	1,51E-08	0,0320	3,65E-05	1,02E-06
17 Thomas D. Wilson	-	-	-	0,0232	2,68E-05	5,22E-07
18 Saharon Shelah	0,0826	1,31E-04	1,02E-05	0,0453	6,19E-05	1,07E-05
19 Norman C. Beaulieu	0,0093	1,70E-05	1,06E-07	0,0210	2,43E-05	4,46E-07
20 David Zhang	0,0047	9,69E-06	3,79E-09	0,0243	3,60E-05	3,11E-06
21 Kang G. Shin	0,0576	9,21E-05	4,89E-06	0,0695	7,36E-05	1,24E-05
22 Tao Jiang	0,0265	4,90E-05	1,66E-06	0,0519	1,23E-04	6,63E-05
23 Micha Sharir	0,0888	1,90E-04	5,50E-05	0,0442	9,49E-05	2,19E-05
24 Elisa Bertino	0,0436	7,03E-05	2,44E-06	0,0519	6,35E-05	1,01E-05
25 David J. Evans	0,0530	8,48E-05	4,14E-06	0,0497	5,60E-05	2,61E-06

Tabela VI.11 Valores de Grau, Proximidade e Intermediação normalizados dos 25 autores nas Redes III e IV

	Rede III			Rede IV		
Autores	Grau	Proximidade	Intermediação	Grau	Proximidade	Intermediação
1 Chin-Chen Chang	0,0451	2,64E-05	1,10E-06	0,0477	2,78E-05	1,26E-06
2 Witold Pedrycz	0,0560	3,59E-05	8,75E-06	0,0532	8,73E-05	6,43E-05
3 H. Vincent Poor	0,0433	8,46E-05	1,25E-04	0,0396	9,14E-05	9,23E-05
4 Noga Alon	0,0659	6,60E-05	5,66E-05	0,0809	1,03E-04	9,96E-05
5 Azriel Rosenfeld	0,0116	7,00E-06	6,58E-08	0,0260	7,59E-05	2,19E-05
6 Lajos Hanzo	0,0283	1,67E-05	3,94E-07	0,0281	6,22E-05	2,49E-05
7 Xiaodong Wang	0,0716	7,77E-05	4,59E-05	0,0562	8,27E-05	7,07E-05
8 Grzegorz Rozenberg	0,0306	2,40E-05	3,91E-06	0,0281	9,35E-05	3,07E-05
9 Georgios B. Giannakis	0,0451	8,02E-05	1,22E-04	0,0422	1,04E-04	1,05E-04
10 Xuemin Shen	0,0456	7,85E-05	1,16E-04	0,0349	8,20E-05	5,43E-05
11 Ming Li	0,0659	6,51E-05	3,75E-05	0,0626	1,09E-04	1,11E-04
12 Ronald R. Yager	0,0139	2,28E-05	1,87E-06	0,0162	6,98E-05	1,24E-05
13 Vladik Kreinovich	0,0422	2,96E-05	6,31E-06	0,0532	8,16E-05	6,07E-05
14 Yan Zhang	0,0439	5,60E-05	3,01E-05	0,0362	9,00E-05	5,61E-05
15 Robert L. Glass	0,0046	3,00E-06	7,72E-09	0,0094	5,66E-06	2,77E-08
16 Guanrong Chen	0,0924	6,57E-05	7,33E-05	0,0762	8,80E-05	7,35E-05
17 Thomas D. Wilson	0,0075	4,67E-06	2,20E-08	0,0128	7,62E-06	8,44E-08
18 Saharon Shelah	0,0208	1,23E-05	2,40E-07	0,0268	7,93E-05	2,92E-05
19 Norman C. Beaulieu	0,0231	1,37E-05	2,92E-07	0,0226	1,33E-05	2,83E-07
20 David Zhang	0,0572	5,19E-05	3,75E-05	0,0494	1,02E-04	1,18E-04
21 Kang G. Shin	0,0248	1,47E-05	3,35E-07	0,0417	2,67E-05	3,64E-06
22 Tao Jiang	0,0526	7,61E-05	1,32E-04	0,0537	1,17E-04	1,67E-04
23 Micha Sharir	0,0254	5,65E-05	1,12E-05	0,0285	8,54E-05	1,89E-05
24 Elisa Bertino	0,0572	3,34E-05	1,68E-06	0,0571	3,12E-05	4,41E-06
25 David J. Evans	0,0497	6,42E-05	4,74E-05	0,0494	8,69E-05	4,66E-05

Tabela VI.12 Valores de Grau, Proximidade e Intermediação normalizados dos 25 autores nas Redes V e VI

	Rede V			Rede VI		
Autores	Grau	Proximidade	Intermediação	Grau	Proximidade	Intermediação
1 Chin-Chen Chang	0,0554	8,81E-05	2,95E-05	0,0516	5,39E-05	1,80E-05
2 Witold Pedrycz	0,0716	1,04E-04	5,39E-05	0,0618	6,06E-05	3,03E-05
3 H. Vincent Poor	0,0905	9,98E-05	5,65E-05	0,0659	5,79E-05	2,19E-05
4 Noga Alon	0,0375	1,39E-05	7,72E-07	0,0632	5,60E-05	3,34E-05
5 Azriel Rosenfeld	-	-	-	0,0359	4,71E-05	1,33E-05
6 Lajos Hanzo	0,0482	9,17E-05	2,52E-05	0,0371	5,51E-05	1,19E-05
7 Xiaodong Wang	0,0850	8,67E-05	4,86E-05	0,0630	5,68E-05	2,21E-05
8 Grzegorz Rozenberg	0,0110	6,40E-05	5,86E-06	0,0222	4,63E-05	6,60E-06
9 Georgios B. Giannakis	0,0158	6,60E-05	7,01E-06	0,0259	5,67E-05	1,08E-05
10 Xuemin Shen	0,0678	1,04E-04	3,74E-05	0,0479	5,98E-05	1,82E-05
11 Ming Li	0,1046	1,04E-04	6,12E-05	0,0871	6,72E-05	4,10E-05
12 Ronald R. Yager	0,0138	6,36E-05	6,73E-06	0,0157	4,32E-05	4,70E-06
13 Vladik Kreinovich	0,0117	4,14E-06	4,37E-08	0,0306	4,78E-05	1,07E-05
14 Yan Zhang	0,1511	1,21E-04	1,15E-04	0,1004	6,78E-05	4,58E-05
15 Robert L. Glass	0,0028	1,07E-06	1,63E-09	0,0067	3,34E-05	1,53E-06
16 Guanrong Chen	0,0547	8,12E-05	3,10E-05	0,0583	5,37E-05	2,00E-05
17 Thomas D. Wilson	0,0014	5,92E-07	2,45E-10	0,0065	1,31E-06	7,44E-09
18 Saharon Shelah	0,0093	3,31E-06	2,78E-08	0,0216	3,89E-05	7,87E-06
19 Norman C. Beaulieu	0,0169	7,67E-05	1,02E-05	0,0179	4,77E-05	6,54E-06
20 David Zhang	0,0499	9,29E-05	2,64E-05	0,0451	5,88E-05	1,72E-05
21 Kang G. Shin	0,0162	7,87E-05	8,53E-06	0,0308	5,06E-05	1,37E-05
22 Tao Jiang	0,0685	1,05E-04	4,84E-05	0,0606	6,78E-05	4,45E-05
23 Micha Sharir	0,0175	1,06E-05	3,18E-07	0,0265	4,42E-05	5,72E-06
24 Elisa Bertino	0,0454	9,17E-05	2,53E-05	0,0514	5,18E-05	1,77E-05
25 David J. Evans	-	-	-	0,0253	4,93E-05	8,85E-06

Como neste trabalho optou-se pelo emprego de redes conexas, então foi realizado o processo de extração das componentes gigantes de cada uma das seis redes. Assim, a Tabela

VI.13 mostrará as propriedades relacionadas às componentes gigantes dessas redes que serão melhor discutidas a seguir.

Tabela VI.13 Propriedades das componentes gigantes das redes da base de dados DBLP

	Rede					
Propriedade	1965 - 1994 **Rede I**	1995-2001 **Rede II**	2002-2007 **Rede III**	1995-2007 **Rede IV**	2008-2013 **Rede V**	1965-2013 **Rede VI**
Número de Autores	184	285	1.094	1.895	2.680	5.065
Número de arestas	417	646	3.264	5.093	9.199	15.597
Distância média	3,520	4,290	6,383	6,133	4,231	4,448
Maior distância	6	8	13	13	9	9
Densidade	2,48E-02	1,57E-02	5,44E-03	2,83E-03	2,56E-03	1,21E-03
Grau Médio	4,5326	4,5333	5,9671	5,3752	6,8649	6,1587
Coeficiente α da Lei de potências	1,044	1,144	1,238	1,325	1,276	1,31
Coeficiente de Regressão da Lei de potências - R^2	0,6566	0,7296	0,7356	0,757	0,6763	0,7135
Coeficiente de Clusterização de Watts-Strogatz	0,8648	0,8890	0,9017	0,8918	0,8863	0,8779

Verifica-se através da Tabela VI.13 que as distâncias médias de todas as redes, exceto a Rede VI, aumentaram quando se considerou somente as suas componentes gigantes. Apesar disso todas continuaram se ajustando à hipótese de mundo pequeno possuindo distâncias médias por volta de 6 ou menos. Quanto à densidade, as redes I e II são as mais densas, possuindo 2,48% e 1,57% das arestas possíveis da rede, respectivamente, enquanto a Rede VI possui apenas 0,12%.

A Rede III é a que mais se aproxima localmente de uma clique, conforme o valor do coeficiente de clusterização de Watts-Strogatz. Como pode ser notado na Figura VI.14, a componente gigante dessa rede divide-se em diversos pequenos cliques com poucas ligações entre autores de diferentes grupos. De qualquer forma, os valores do coeficiente de clusterização para todas as redes é alto, o que é esperado em redes de coautoria, como citado anteriormente.

Quando foi feita a extração das componentes gigantes das redes percebeu-se que nem todos os 25 autores que foram selecionados inicialmente estavam presentes em todas as componentes, assim, para as análises das medidas de centralidade dos vértices das componentes gigantes, preferiu-se avaliar os resultados dos 10 autores que apresentaram os maiores valores de cada medida. As Tabelas VI.14, VI.15, VI.16 apresentam esses resultados para as medidas de grau, proximidade e intermediação normalizadas, respectivamente.

Tabela VI.14 Resultados dos 10 autores com maiores graus das componentes gigantes das Redes da DBLP

Grau normalizado

	Rede I		Rede II		Rede III		Rede IV		Rede V		Rede VI	
1	Noga Alon	0,4536	Noga Alon	0,3697	Guanrong Chen	0,1464	Noga Alon	0,1003	Yan Zhang	0,1639	Yan Zhang	0,1011
2	Micha Sharir	0,3115	Azriel Rosenfeld	0,1901	Xiaodong Wang	0,1134	Guanrong Chen	0,0945	Ming Li	0,1135	Ming Li	0,0877
3	Grzegorz Rozenberg	0,2514	Ming Li	0,1655	Ming Li	0,1043	Ming Li	0,0776	H. Vincent Poor	0,0982	H. Vincent Poor	0,0664
4	Andrzej Ehrenfeucht	0,0984	Tao Jiang	0,1655	Noga Alon	0,1043	Xiaodong Wang	0,0697	Xiaodong Wang	0,0922	Noga Alon	0,0636
5	Leonidas J. Guibas	0,0874	Micha Sharir	0,1408	David Zhang	0,0906	Tao Jiang	0,0665	Witold Pedrycz	0,0776	Xiaodong Wang	0,0634
6	Emo Welzl	0,0820	Grzegorz Rozenberg	0,0704	Tao Jiang	0,0833	Vladik Kreinovich	0,0660	Tao Jiang	0,0743	Witold Pedrycz	0,0622
7	Raimund Seidel	0,0820	Leonidas J. Guibas	0,0528	David J. Evans	0,0787	Witold Pedrycz	0,0660	Xuemin Shen	0,0735	Tao Jiang	0,0610
8	Richard Pollack	0,0820	Pankaj K. Agarwal	0,0493	Xuemin Shen	0,0723	David J. Evans	0,0612	Chin-Chen Chang	0,0601	Guanrong Chen	0,0586
9	János Pach	0,0765	Andrzej Ehrenfeucht	0,0387	Georgios B. Giann	0,0714	David Zhang	0,0612	Guanrong Chen	0,0594	Chin-Chen Chang	0,0519
10	Daniel J. Kleitman	0,0656	Paul M. B. Vitányi	0,0387	Yan Zhang	0,0695	Georgios B. Giann	0,0523	David Zhang	0,0541	Elisa Bertino	0,0517

Tabela VI.15 Resultados dos 10 autores com maiores proximidades das componentes gigantes das Redes da DBLP

Proximidade normalizada

	Rede I		Rede II		Rede III		Rede IV		Rede V		Rede VI	
1	Micha Sharir	2,33E-03	Tao Jiang	1,25E-03	H. Vincent Poor	2,12E-04	Tao Jiang	1,33E-04	Yan Zhang	1,43E-04	Yan Zhang	6,88E-05
2	Subhash Suri	2,25E-03	Douglas B. West	1,23E-03	Zhi-Quan Luo	2,06E-04	Ming Li	1,24E-04	Tao Jiang	1,24E-04	Tao Jiang	6,87E-05
3	Boris Aronov	2,25E-03	Dhruv Mubayi	1,23E-03	Sergio Barbarossa	2,06E-04	Georgios B. Giann	1,18E-04	Xuemin Shen	1,22E-04	Ming Li	6,81E-05
4	Pankaj K. Agarwal	2,24E-03	Richard M. Karp	1,23E-03	Jinjun Xiao	2,06E-04	Richard M. Karp	1,17E-04	Witold Pedrycz	1,22E-04	Qian Zhang	6,16E-05
5	Imre Bárány	2,22E-03	Zsolt Tuza	1,23E-03	Yingwei Yao	2,06E-04	Noga Alon	1,17E-04	Ming Li	1,22E-04	Witold Pedrycz	6,14E-05
6	Noga Alon	2,20E-03	Noga Alon	1,21E-03	Hisashi Kobayashi	2,02E-04	Xiaoyan Zhu	1,16E-04	Lei Shu	1,21E-04	Xuemin Shen	6,06E-05
7	Leonidas J. Guibas	2,12E-03	Ming Li	1,12E-03	Georgios B. Giann	2,01E-04	Rudolf Fleischer	1,16E-04	Lingyang Song	1,19E-04	Yong Zhang	6,04E-05
8	Raimund Seidel	2,11E-03	Vitaly I. Voloshin	1,06E-03	Xuemin Shen	1,97E-04	David Zhang	1,16E-04	Jiming Chen	1,19E-04	David Zhang	5,96E-05
9	János Pach	2,11E-03	Daniel P. Fasulo	1,05E-03	Xiaodong Wang	1,95E-04	Xin Chen	1,14E-04	Laurence Tianruo Y	1,19E-04	Yang Xiao	5,93E-05
10	Herbert Edelsbrunner	2,10E-03	Edward C. Thayer	1,05E-03	Tao Jiang	1,91E-04	Bin Ma	1,14E-04	Min Chen	1,18E-04	Naixue Xiong	5,89E-05

Tabela VI.16 Resultados dos 10 autores com maiores proximidades das componentes gigantes das Redes da DBLP

Intermediação normalizada

	Rede I		Rede II		Rede III		Rede IV		Rede V		Rede VI	
1	Noga Alon	3,58E-03	Noga Alon	2,40E-03	Tao Jiang	5,24E-04	Tao Jiang	2,34E-04	Yan Zhang	1,47E-04	Yan Zhang	6,88E-05
2	Micha Sharir	2,38E-03	Tao Jiang	2,16E-03	H. Vincent Poor	4,96E-04	David Zhang	1,66E-04	Ming Li	7,81E-05	Tao Jiang	6,87E-05
3	Emo Welzl	2,07E-03	Ming Li	1,63E-03	Georgios B. Giann	4,86E-04	Ming Li	1,56E-04	H. Vincent Poor	7,22E-05	Ming Li	6,81E-05
4	Grzegorz Rozenberg	2,05E-03	Azriel Rosenfeld	1,11E-03	Xuemin Shen	4,59E-04	Georgios B. Giann	1,47E-04	Witold Pedrycz	6,88E-05	Noga Alon	6,16E-05
5	Boris Aronov	1,08E-03	Bir Bhanu	1,09E-03	Hisashi Kobayashi	3,51E-04	Noga Alon	1,40E-04	Xiaodong Wang	6,21E-05	Witold Pedrycz	6,14E-05
6	Subhash Suri	6,82E-04	Micha Sharir	7,13E-04	Guanrong Chen	2,91E-04	H. Vincent Poor	1,30E-04	Tao Jiang	6,18E-05	Xiaodong Wang	6,06E-05
7	Pankaj K. Agarwal	4,37E-04	Richard M. Karp	4,43E-04	Zhi-Hong Guan	2,38E-04	Guanrong Chen	1,03E-04	Xuemin Shen	4,78E-05	H. Vincent Poor	6,04E-05
8	Imre Bárány	4,28E-04	Douglas B. West	4,41E-04	Noga Alon	2,25E-04	Xiaodong Wang	9,93E-05	Guanrong Chen	3,96E-05	Guanrong Chen	5,96E-05
9	Leonidas J. Guibas	2,01E-04	Dhruv Mubayi	4,24E-04	David J. Evans	1,89E-04	Xiaoyan Zhu	9,44E-05	Chin-Chen Chang	3,77E-05	Xuemin Shen	5,93E-05
10	Raimund Seidel	1,93E-04	Zsolt Tuza	4,18E-04	Xiaodong Wang	1,82E-04	Witold Pedrycz	9,04E-05	David Zhang	3,38E-05	Chin-Chen Chang	5,89E-05

Em resultados absolutos, o grau máximo da Rede I era 83, e o da Rede VI, 512, sendo os menor e maior valores ao se comparar os graus máximos de todas as redes. Entretanto, quando as medidas são normalizadas, os valores da Rede I passam a ser os maiores que os das demais redes em quase todos os casos. Isso mostra que mesmo que os autores da Rede VI estejam ligados a mais autores que aqueles da Rede I, os dessa rede possuem mais conexões levando-se em consideração o tamanho da rede que pertencem, ou seja, o número máximo de outros autores a que poderiam estar conectados.

Quanto à medida de proximidade os autores da Rede I apresentaram os maiores valores absolutos, assim quando foi feita a normalização dos valores, como essa era a menor rede, então continuaram a ser os maiores. Nessa rede os autores possuem mais facilidade para obtenção de informações que nas demais.

Na medida de intermediação, as redes I e II apresentaram os maiores valores absolutos, e ao normalizar a Rede I ficou com os maiores valores comparados aos das demais redes. Através desse resultado conclui-se que os autores dessa rede são mais eficientes no fluxo de informações.

Quanto à verificação se as redes da base DBLP seguiam lei de potência observou-se que as taxas de decaimento da função de cada uma dessas seis redes são menores que a apresentada por ALBERT *et al.*(1999), figuras VI.18 e VI.19. Além disso, como pode ser visualizado nessas figuras essas redes não seguem lei de potência.

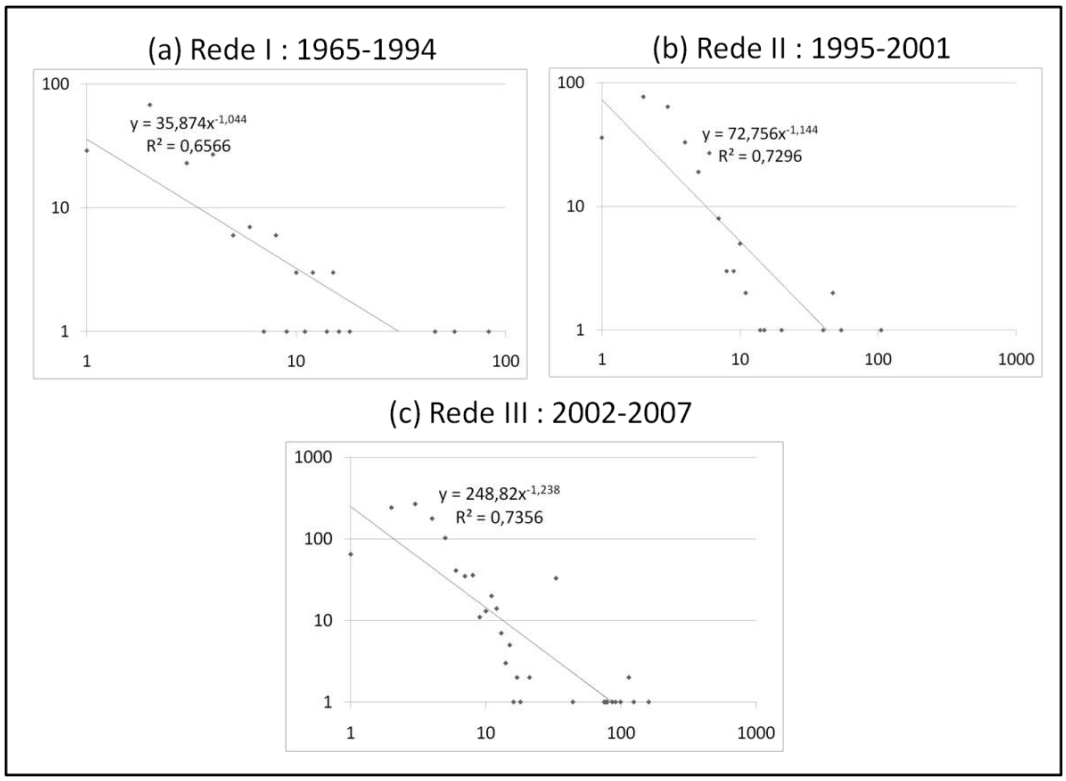


Figura VI.18 Distribuição dos graus da (a) Rede I, (b) Rede II, (c) Rede III

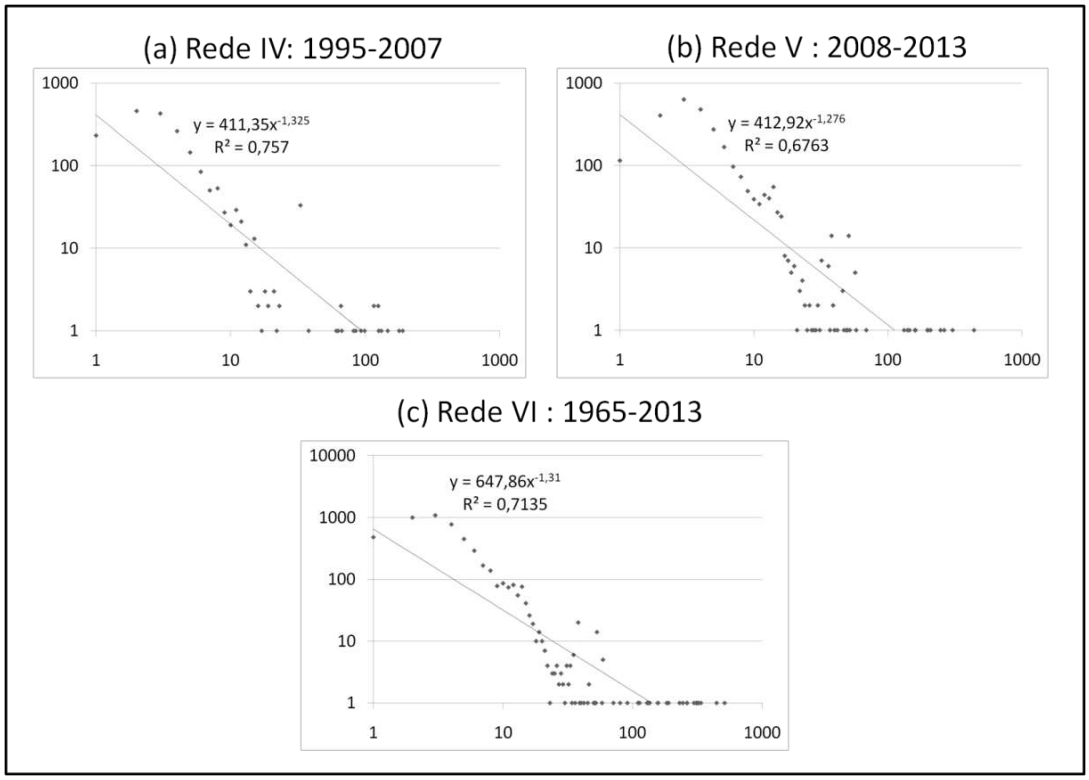


Figura VI.19 Distribuição dos graus da (a) Rede IV, (b) Rede V, (c) Rede VI

Para a implementação dos algoritmos nas Redes da base DBLP foi realizado o mesmo processo citado na Seção VI.2 para as redes de coautoria da literatura. Desse modo, utilizou-se os algoritmos espectrais através de um código implementado no *software OCTAVE*, e a função modificada do *k-means* no *Matlab*.

Da mesma forma, os algoritmos espectrais com matriz Laplaciana e Laplaciana sem sinal e o algoritmo *k-means* foram testados em cada rede com diferentes quantidades *k* de *clusteres*. A qualidade dessas divisões foi calculada através da modularidade, logo foram testados diferentes valores de *k* enquanto o resultado para a modularidade melhorava, ou seja, o algoritmo conseguia uma melhor partição, e quando os resultados começavam a piorar, parava-se os testes.

A Tabela VI.17 apresenta os valores de *k* utilizados nas rodadas para a Rede I e os resultados da modularidade para cada algoritmo.

Tabela VI.17 Resultados da modularidade para diferentes valores de *k* na Rede I (1965-1994)

Rede I (1965 - 1994)	**Modularidade**		
	Algoritmo espectral		
	Matriz Laplaciana sem sinal (Q)	Matriz Laplaciana com sinal (L)	*k-means*
k=2	0,4108	0,4311	0,4274
k=3	0,5283	0,5865	0,6231
k=4	0,5318	0,6155	0,4580
k=5	0,5348	0,5537	-0,0298
k=7	0,5150	0,5478	-0,0285
k=8	0,5140	0,5335	-0,0241
k=10	0,5064	0,4771	-0,0244
k=15	0,4735	0,4311	0,2784
k=20	0,4796	0,4045	0,2234

Para a Rede I, como pode-se observar na Figura VI.20, o algoritmo *k-means* para *k=3* foi o que apresentou o melhor resultado para a modularidade, e com o aumento de *k* os resultados referentes ao uso do *k-means* pioraram bastante chegando a obter modularidade negativa, o que significa que os ramos gerados não apresentavam presença de mais comunidades e a divisão deles originou em *clusteres* ruins.O algoritmo espectral com uso da matriz Laplaciana obteve o segundo melhor resultado para *k=4*, e o melhor resultado obtido pelo uso da matriz Laplaciana sem sinal obteve um valor de modularidade mais baixo. Nota-se que mesmo pontos que não correspondem ao melhor resultado na reta da matriz Laplaciana foram melhores que a sem sinal, como se comparamos os casos em que *k=5* e *k=7* com o melhor valor da Laplaciana sem sinal.

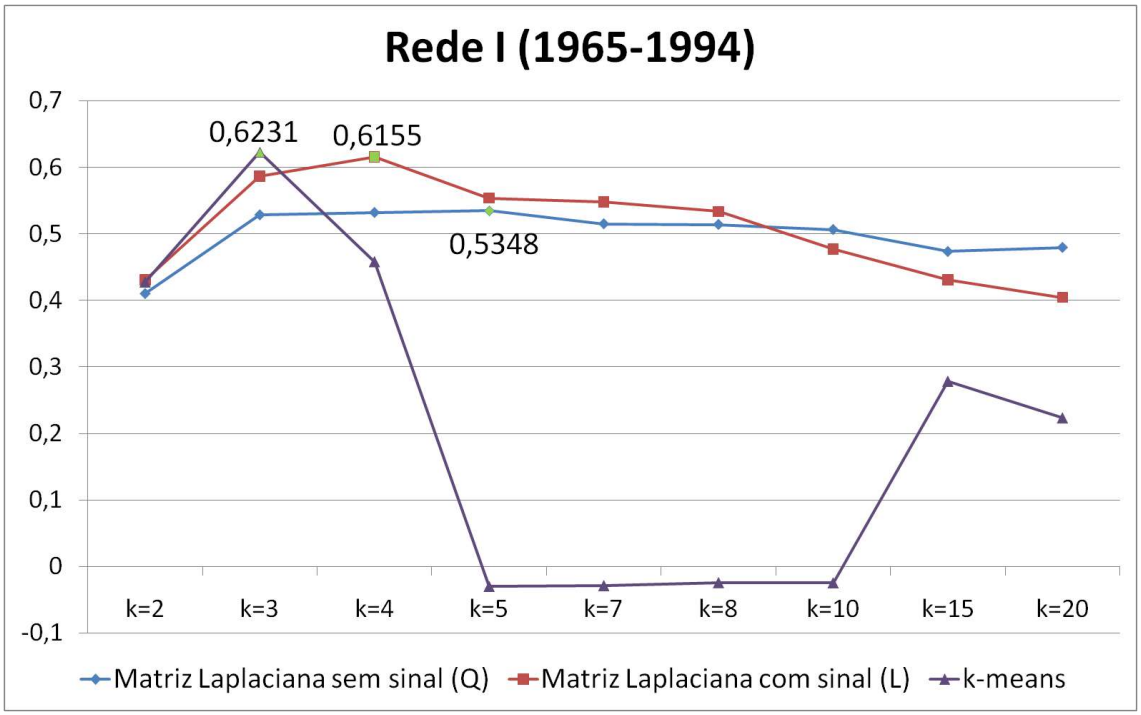


Figura VI.20 Valores da medida modularidade para diferentes quantidades de *clusteres* em cada algoritmo na Rede I (1965-1994)

A Tabela VI.18 apresenta os valores de *k* utilizados nas rodadas para a Rede II e os resultados da modularidade para cada algoritmo.

Na Figura VI.21 tem-se os resultados da modularidade para as diferentes quantidades de *clusteres* utilizadas em cada algoritmo na Rede II. O melhor resultado foi obtido pelo uso do algoritmo espectral com matriz Laplaciana sem sinal com *k=10*, no entanto, percebe-se que os resultados para esse algoritmo com *k=5, 6, 7, 9, 12, 13 e 15,* mesmo não sendo os melhores alcançaram modularidade maior que o melhor resultado obtido com o uso do algoritmo espectral com matriz Laplaciana e algoritmo *k-means*. Já o *k-means* foi o que apresentou valores mais baixos na medição da modularidade.

Tabela VI.18 Resultados da modularidade para diferentes valores de *k* na Rede II (1995-2001)

	Modularidade		
	Algoritmo espectral		
Rede II (1995-2001)	Matriz Laplaciana sem sinal (Q)	Matriz Laplaciana com sinal (L)	*k-means*
k=2	0,4959	0,4122	0,4945
k=3	0,6035	0,6157	0,5949
k=4	0,6868	0,6988	0,6292
k=5	0,7357	0,6544	0,6250
k=6	0,7390	0,6726	0,6448
k=7	0,7319	0,6772	0,4961
k=9	0,7223	0,6881	0,5889
k=10	0,7435	0,6767	0,4695
k=12	0,7385	0,6399	0,4472
k=13	0,7354	0,6211	0,4169
k=15	0,7109	0,5981	0,4784
k=27	0,6189	0,5241	0,3873
k=31	0,5962	0,5253	0,4237
k=45	0,5052	0,5024	0,3706

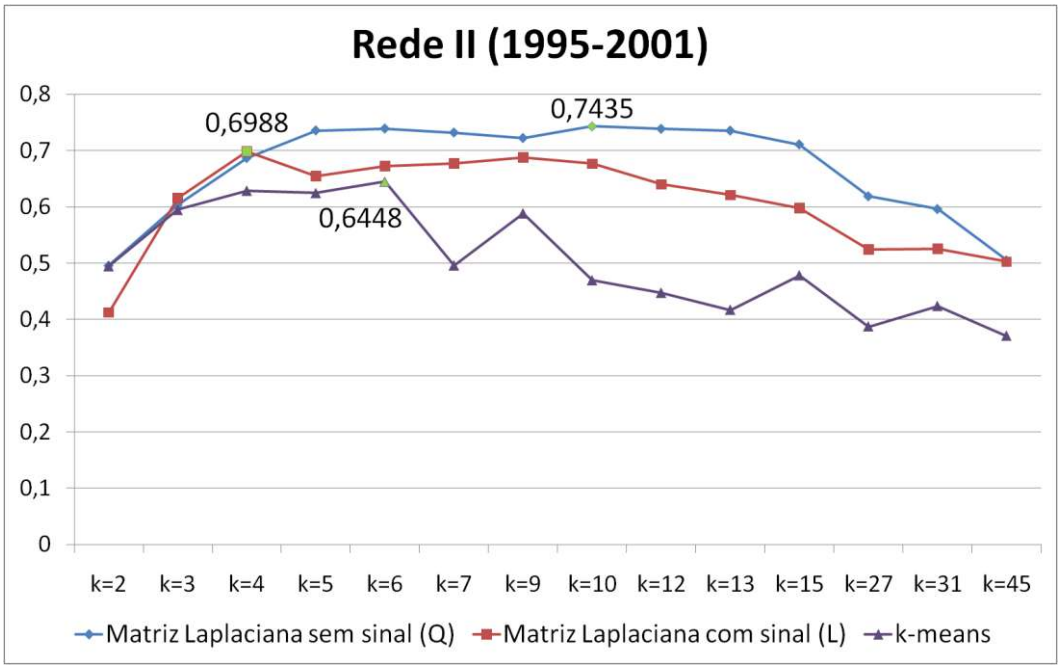


Figura VI.21 Valores da medida modularidade para diferentes quantidades de *clusteres* em cada algoritmo na Rede II (1995-2001)

A Tabela VI.19 apresenta os valores de *k* utilizados nas rodadas para a Rede III e os resultados da modularidade para cada algoritmo

Tabela VI.19 Resultados da modularidade para diferentes valores de *k* na Rede III (2002-2007)

	Modularidade		
	Algoritmo espectral		
Rede III (2002-2007)	Matriz Laplaciana sem sinal (Q)	Matriz Laplaciana com sinal (L)	*k-means*
k=2	0,3250	0,4844	0,4844
k=3	0,5696	0,6473	0,6455
k=4	0,6544	0,6921	0,6710
k=5	0,7256	0,7433	0,6788
k=7	0,7721	0,7868	0,7360
k=8	0,7916	0,8206	0,7355
k=9	0,7973	0,8274	0,6970
k=10	0,8010	0,8344	0,7713
k=11	0,8099	0,8566	0,7326
k=12	0,8094	0,8542	0,5743
k=13	0,8110	0,8554	0,7224
k=14	0,8099	0,8507	0,6402
k=15	0,8079	0,8421	0,7191
k=16	0,8221	0,8400	0,6158
k=18	0,8187	0,8151	0,6683
k=20	0,8108	0,7957	0,6058
k=22	0,8016	0,7730	0,6121
k=26	0,7924	0,7818	0,6082

Na Figura VI.22 os resultados da modularidade para a Rede III mostra que a matriz Laplaciana apresentou os melhores resultado, desde as primeiras quantidades de *k* sua reta está acima das demais, apenas em *k=2* que seu resultado foi o mesmo que o do *k-means* e para *k>17* o algoritmo com matriz Laplaciana sem sinal obteve melhores resultados. Entretanto, a melhor partição foi encontrada com o uso da matriz Laplaciana para *k=11*, o melhor resultado obtido pela matriz Laplaciana sem sinal foi com *k=16* e para o *k-means, k=10*.

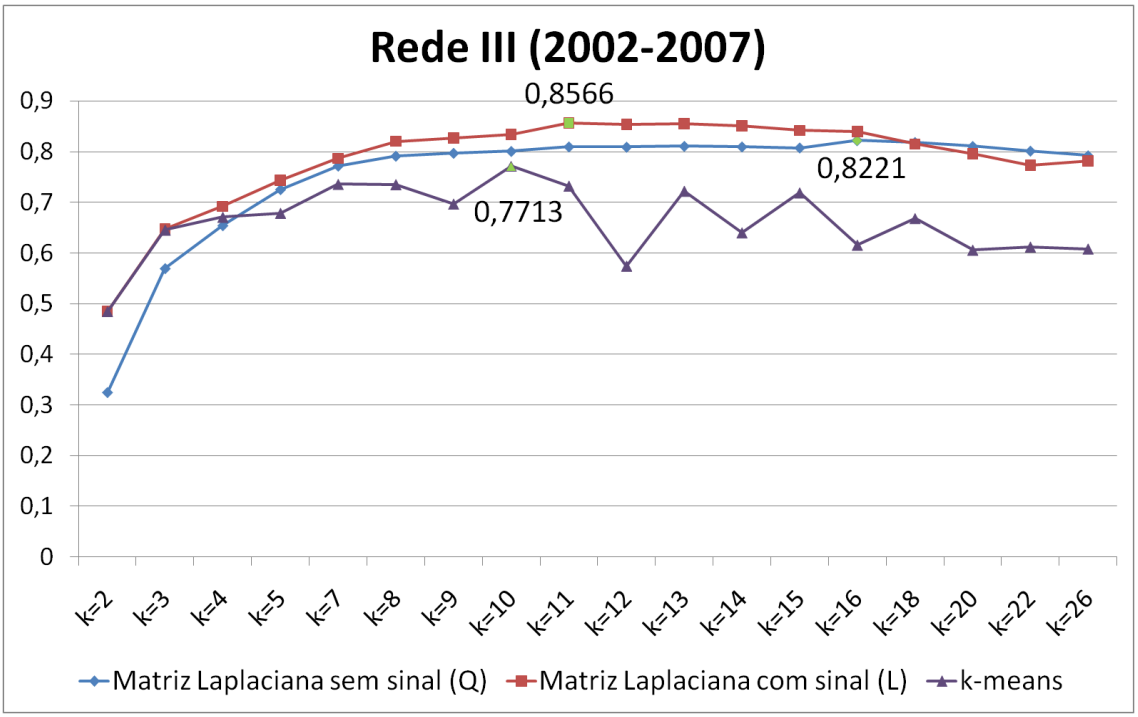


Figura VI.22 Valores da medida modularidade para diferentes quantidades de *clusteres* em cada algoritmo na Rede III (2002-2007)

A Tabela VI.20 apresenta os valores de *k* utilizados nas rodadas para a Rede IV e os resultados da modularidade para cada algoritmo

Tabela VI.20 Resultados da modularidade para diferentes valores de *k* na Rede IV (1995-2007)

	Modularidade		
	Algoritmo espectral		
Rede IV (1995-2007)	Matriz Laplaciana sem sinal (Q)	Matriz Laplaciana com sinal (L)	*k-means*
k=2	0,0055	0,5004	0,4734
k=4	0,2437	0,7333	0,5962
k=5	0,5683	0,7461	0,6726
k=6	0,6397	0,7696	0,7343
k=7	0,7000	0,7990	0,7076
k=9	0,7516	0,8242	0,7762
k=11	0,7985	0,8623	0,7925
k=13	0,8023	0,8710	0,7713
k=15	0,8045	0,8750	0,7277
k=16	0,7830	0,8875	0,7887
k=17	0,7812	0,8896	0,7817
k=18	0,7816	0,8915	0,8147
k=19	0,7996	0,8885	0,7848
k=20	0,7831	0,8678	0,7536
k=22	0,7776	0,8828	0,7374
k=24	0,7827	0,8574	0,6281
k=27	0,7832	0,8422	0,6795

O algoritmo espectral com matriz Laplaciana conseguiu gerar melhores partições para a Rede IV conforme pode ser observado na Figura VI.23, chegando ao melhor resultado quando *k=17*. Os resultados do algoritmo com matriz Laplaciana sem sinal foram melhorando com o aumento de *k*, chegando a ultrapassar o *k-means* em alguns momentos, porém o melhor resultado do *k-means* para *k=18* foi mais alto que o da matriz Laplaciana sem sinal com *k=15*.

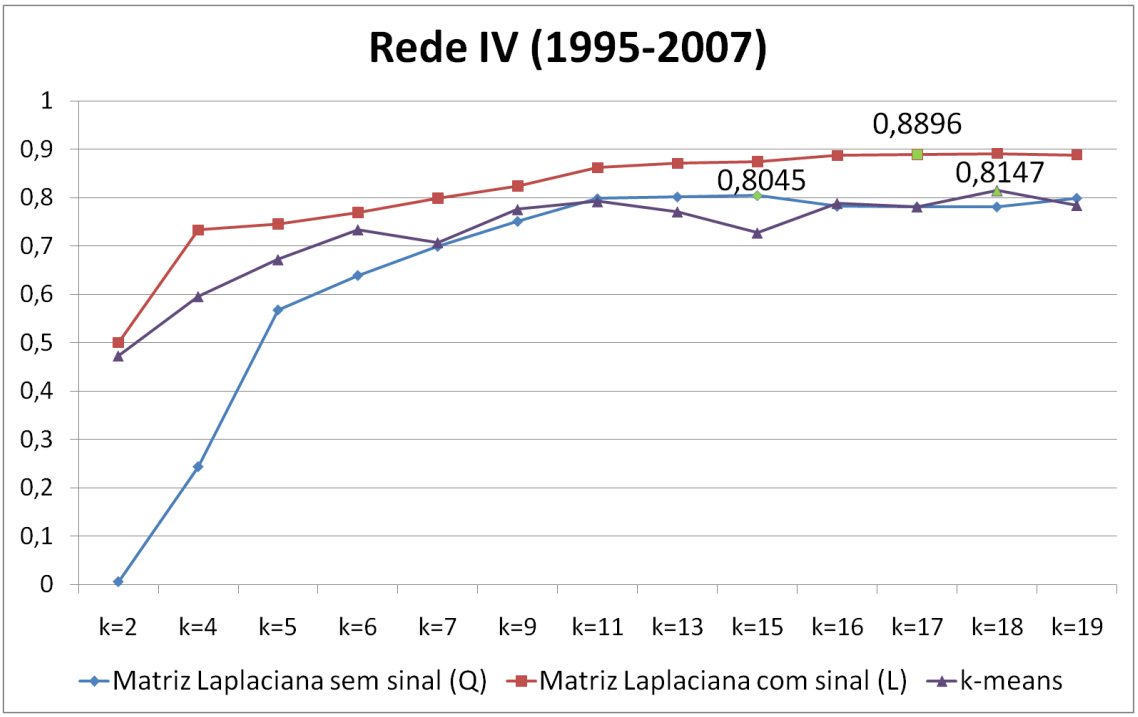


Figura VI.23 Valores da medida modularidade para diferentes quantidades de *clusteres* em cada algoritmo na Rede IV (1995-2007)

A Tabela VI.21 apresenta os valores de *k* utilizados nas rodadas para a Rede V e os resultados da modularidade para cada algoritmo

Tabela VI.21 Resultados da modularidade para diferentes valores de *k* na Rede V (2008-2013)

	Modularidade		
	Algoritmo espectral		
Rede V (2008-2013)	Matriz Laplaciana sem sinal (Q)	Matriz Laplaciana com sinal (L)	*k-means*
k=2	0,2124	0,0392	0,3492
k=3	0,4268	0,1327	0,4479
k=5	0,7130	0,6112	0,6799
k=7	0,7497	0,7531	0,6227
k=10	0,7863	0,8157	0,7523
k=11	0,7834	0,8208	0,6284
k=12	0,7836	0,8268	0,6295
k=13	0,7898	0,8288	0,6297
k=14	0,7935		0,7969
k=15	0,7937	0,8302	0,6491
k=17	0,7906	0,8272	0,6644
k=18	0,7907	0,8245	0,5696
k=20	0,8014	0,8250	0,6989
k=21	0,8015	0,8254	0,6400
k=22	0,8019		0,5897
k=23	0,8011	0,8266	0,6051
k=25	0,8005	0,8243	0,5493
k=27	0,7991	0,8191	0,5185
k=30	0,7970	0,8169	0,5724

O uso da matriz Laplaciana também gerou o melhor resultado de modularidade na Rede V como pode ser visto na Figura VI.24, esse resultado foi obtido para *k=15*. Para *k<7* os melhores valores de modularidade intercalaram entre o uso da matriz Laplaciana sem sinal e o *k-means*, já a partir de *k=7* a matriz Laplaciana sem sinal obteve os melhores valores. O algoritmo espectral com matriz Laplaciana sem sinal atingiu seu melhor resultado para *k=22* e o *k-means* para *k=14*.

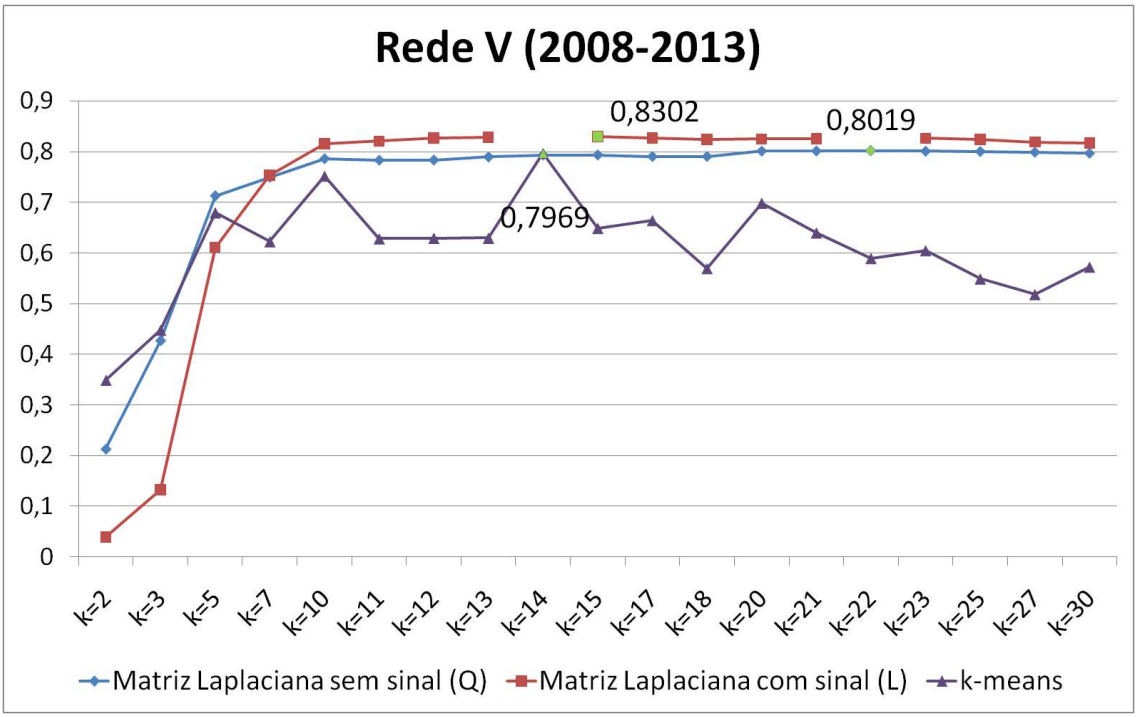


Figura VI.24 Valores da medida modularidade para diferentes quantidades de *clusteres* em cada algoritmo na Rede V (2008-2013)

A Tabela VI.22 apresenta os valores de *k* utilizados nas rodadas para a Rede VI e os resultados da modularidade para cada algoritmo.

A Figura VI.25 apresenta os resultados referentes à Rede VI. Nota-se que os algoritmos espectrais apresentaram comportamentos semelhantes, com o valor da modularidade crescendo até *k=15* e depois o crescimento e queda foram pequenos quando se aumentava o número de *clusteres*. Novamente, o algoritmo espectral com a matriz Laplaciana foi o que apresentou o melhor resultado, chegando a 0,865 para *k=22*, o algoritmo espectral com matriz Laplaciana sem sinal obteve uma modularidade de 0,8039 para *k=30*, e o *k-means*, 0,7087 para *k=15*.

Tabela VI.22 Resultados da modularidade para diferentes valores de *k* na Rede VI (1965-2013)

	Modularidade		
	Algoritmo espectral		
Rede VI (1965-2013)	Matriz Laplaciana sem sinal (Q)	Matriz Laplaciana com sinal (L)	*k-means*
k=2	0,1635	0,2292	0,4428
k=3	0,3158	0,5699	0,3489
k=4	0,5156	0,5990	0,4588
k=5	0,6703	0,6591	0,4936
k=6	0,7126	0,6945	0,4394
k=7	0,7313	0,7430	0,5900
k=8	0,7504	0,7677	0,6466
k=10	0,7774	0,8093	0,6736
k=11	0,7826	0,8200	0,6861
k=12	0,7854	0,8276	0,6680
k=13	0,7890	0,8477	0,6502
k=14	0,7914	0,8548	0,6922
k=15	0,7890	0,8612	0,7087
k=16	0,7915	0,8622	0,5484
k=17	0,7987	0,8637	0,5593
k=18	0,7987	0,8639	0,5744
k=20	0,8003	0,8638	0,5685
k=22	0,7960	0,8650	0,5313
k=25	0,7985	0,8644	0,5300
k=27	0,7993	0,8615	0,5516
k=30	0,8039	0,8444	0,5459
k=32	0,7955	0,8451	0,5596
k=35	0,7869	0,8417	0,5257
k=39	0,7844	0,8245	0,5119
k=40	0,7766	0,8231	0,5360
k=42	0,7764	0,8241	0,5103
k=45	0,7706	0,8134	0,5216
k=47	0,7643	0,8102	0,5164
k=50	0,7629	0,7996	0,5051

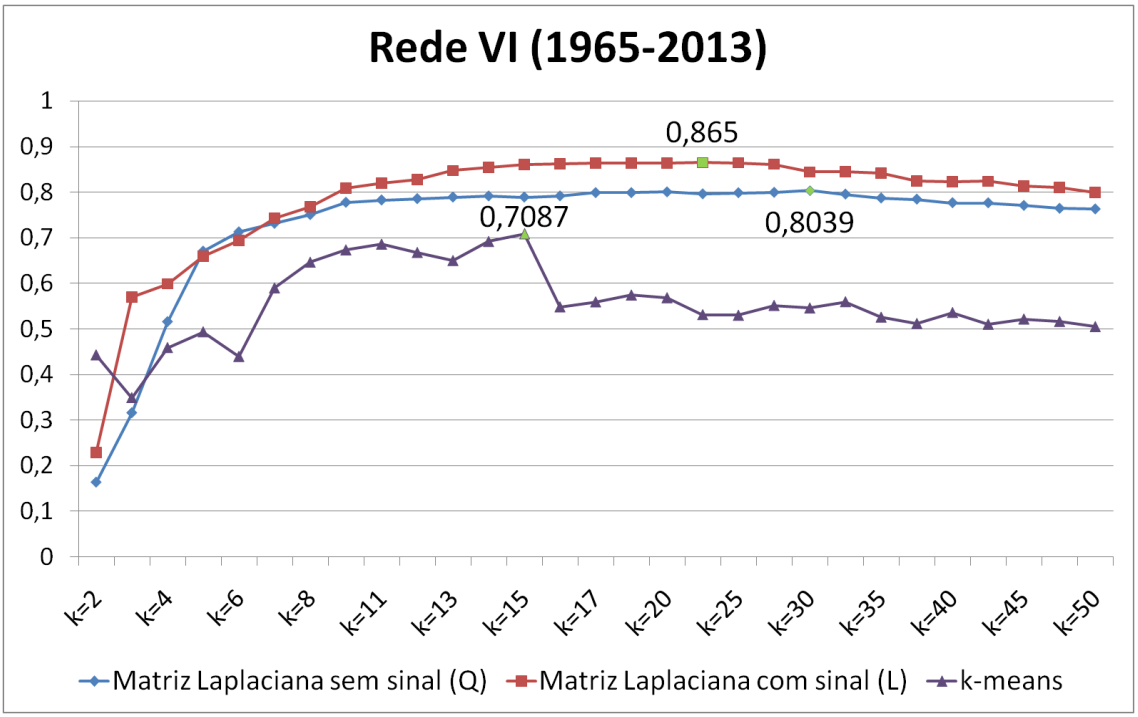


Figura VI.25 Valores da medida modularidade para diferentes quantidades de *clusteres* em cada algoritmo na Rede VI (1965-2013)

Através dos resultados apresentados nas figuras nota-se que para todas as redes, exceto a Rede I, os melhores resultados de modularidade obtiveram todos Q>0,7 indicando bons resultados de partições encontradas. Observa-se também que novamente o uso de vetor de Fiedler como vetor de corte para agrupamento apresentou bons resultados para a maioria dessas redes em comparação com os demais métodos testados. Entretanto, é bem notável, que apesar de a matriz Laplaciana ter apresentado os melhores resultados na maioria das vezes, os resultados da matriz Laplaciana sem sinal estiveram bem próximos daqueles.

Nota-se também que mesmo os algoritmos que não apresentaram os melhores resultados, na maioria das vezes, conseguiram resultados de modularidade considerados bons (acima de 0,7), o que significa bons resultados de modularidade. O *k-means* nas melhores partições encontradas por ele em cada rede obteve Q>0,7 em quatro das seis redes analisadas; o algoritmo Laplaciano sem sinal espectral, em cinco dessas seis redes; e o algoritmo Laplaciano espectral em quatro das seis redes, sendo essas quatro as redes em que tal algoritmo apresentou os melhores resultados dentre todos. E mesmo para os resultados que não eram as melhores partições de cada algoritmo muitas vezes as partições encontradas também conseguiram valores de modularidade acima de 0,7, indicando um bom desempenho, principalmente para os algoritmos espectrais

Um dos objetivos em se identificar agrupamentos em redes de coautoria é para que seja possível a análise da formação de grupos de pesquisa entre os autores, a importância de certos autores na formação dessas partições e as linhas de pesquisa desses grupos. Como

para algumas redes o melhor resultado gerou um número grande de *clusteres*, optou-se por analisar apenas a maior partição do algoritmo que obteve o melhor resultado para cada rede. Com isso, foram feitas análises relacionadas à centralidade dos autores dentro dessa partição e as linhas de pesquisa deles.

A Figura VI.26 apresenta a Rede I dividida em três partições geradas pelo algoritmo *k-means*. Nota-se o número de ligações dentro dos *clusteres* é bem maior que entre os *clusteres*, sendo que os autores presentes na partição (3) nem se comunicam com os da partição (2), assim os vértices da partição (1) fazem essa ponte de transmissão de informações entre eles. Percebe-se também que apenas um vértice é responsável pela comunicação entre as partições (1) e (2).

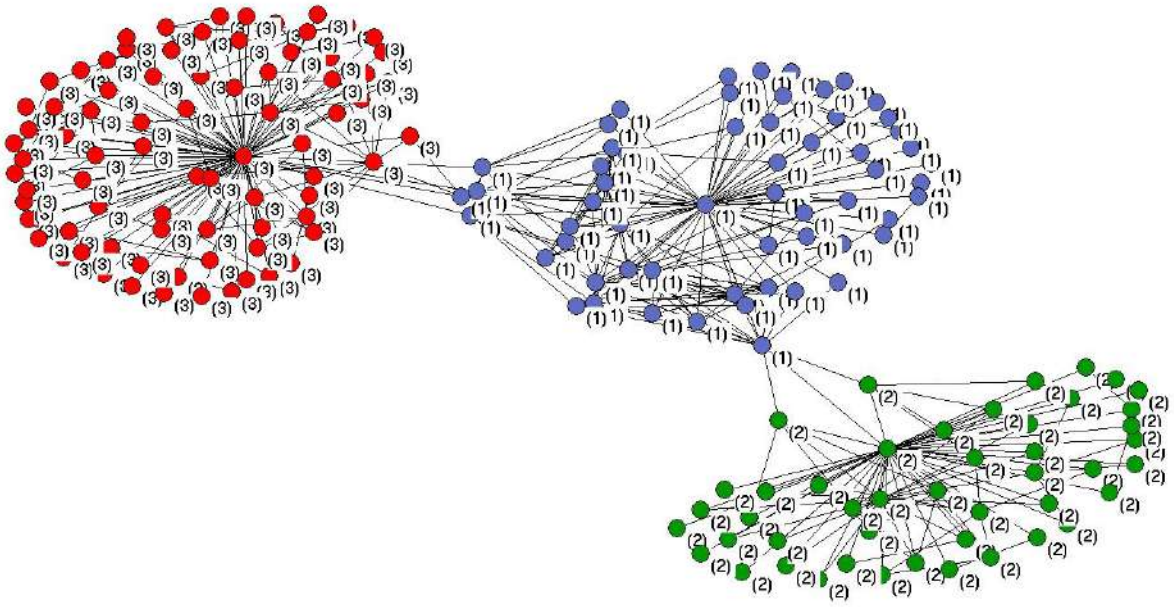


Figura VI.26 Partições geradas pelo algoritmo *k-means* para a Rede I

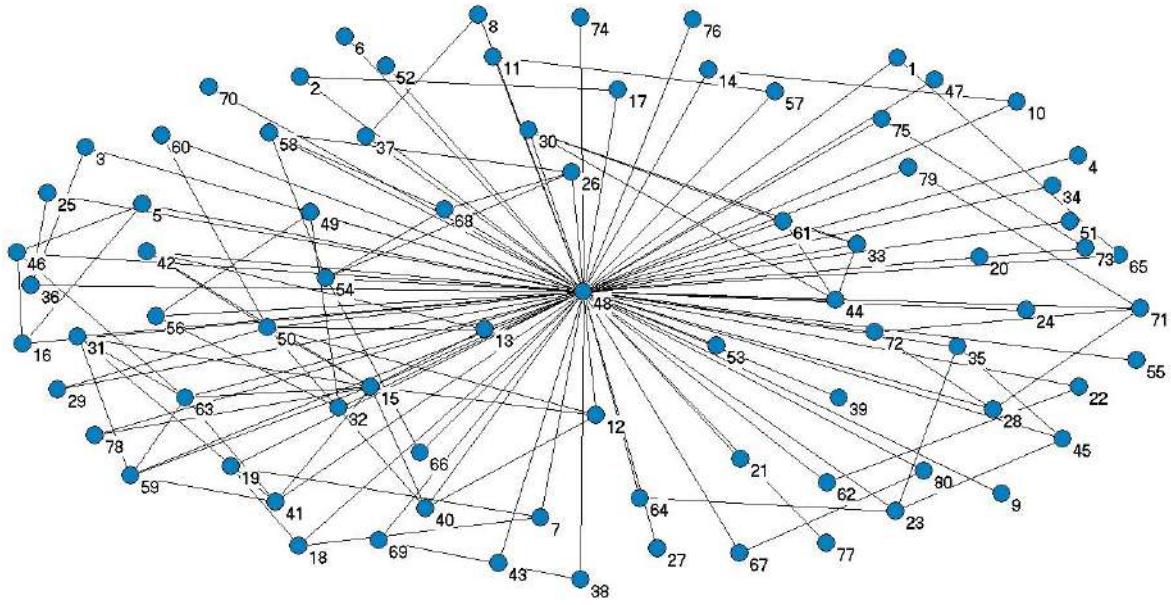


Figura VI.27 Resultado da maior partição encontrada para a Rede I

A Figura VI.27 apresenta a maior partição entre as três encontradas pelo algoritmo e possui 80 autores. Através dessa figura verifica-se que o vértice 48 é o que possui mais ligações com outros vértices nessa partição, ao analisar a centralidade desses autores, observou-se que o 48 está ligado a todos os demais na partição, sendo assim o mais central, com grau igual a 79, proximidade igual a 1 e intermediação igual a 0,9689. Esse vértice corresponde ao autor Noga Alon. Outros autores que também apresentaram os maiores valores de centralidade foram Daniel J. Kleitman, Paul D. Seymour e Roy Meshulam. Entretanto os valores de centralidade desses estão bem abaixo dos de Noga Alon. Daniel J. Kleitman, por exemplo, que ocupa a segunda posição obteve grau igual a 11, proximidade igual a 0,5374 e intermediação igual a 0,00514.

De acordo com as informações da base DBLP, Noga Alon possui 354 artigos publicados em periódicos de 1983 a 2013, sendo 101 deles (aproximadamente 29%) no intervalo de 1983 a 1994, pois o autor não possui trabalhos antes desse período. Dentre aqueles 101 artigos produzidos por Noga Alon no período representado pela Rede I, em 7 artigos o autor Daniel J. Kleitman também estava presente; 6, Paul D. Seymour; e 2, Roy Meshulam.

Os campos de interesse do autor são: Combinatória, Teoria dos Grafos e suas aplicações à Ciência da Computação Teórica; Algoritmos combinatórios e complexidade do circuito; Geometria combinatória e teoria dos números combinatórios; Álgebra e métodos probabilísticos em Combinatória. Segundo o índice de citações do *Google*, o autor possui um total de 29516 citações, sendo que 12800 delas (mais de 40%) ocorreram nos últimos cinco anos.

Noga Alon já publicou com 389 autores, os coautores que já publicaram mais artigos com ele foram: Michael Krivelevich, Benny Sudakov, Raphael Yuster, Asaf Shapira.

A Figura VI.28 apresenta as partições resultantes do emprego do algoritmo espectral com matriz Laplaciana sem sinal para *k=10*, cujo resultado apresentou a melhor modularidade.

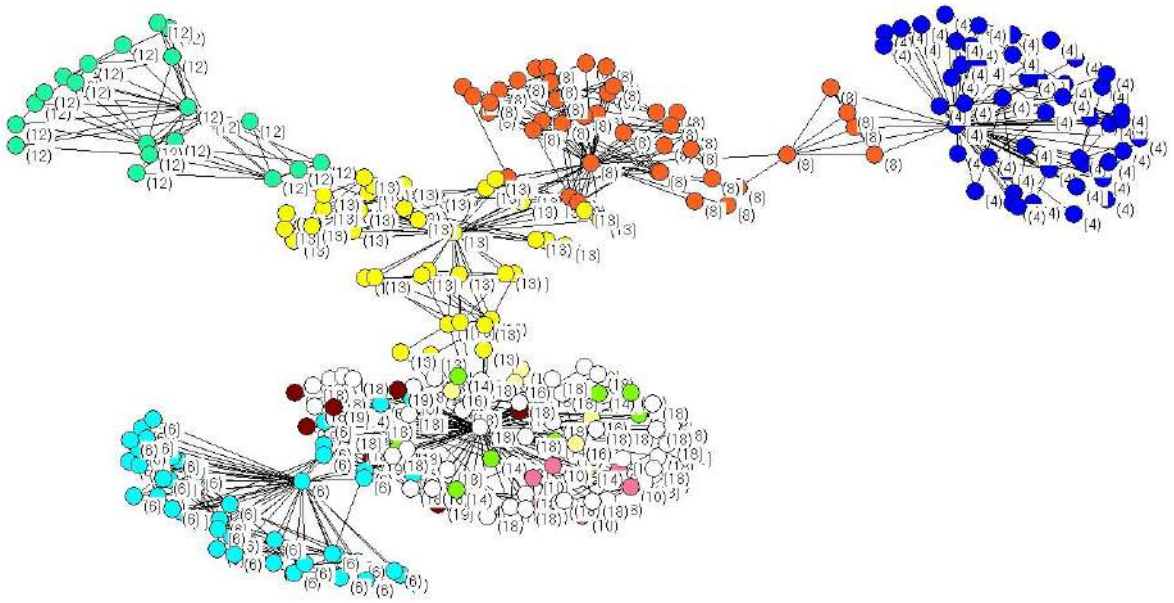


Figura VI.28 Partições geradas pelo algoritmo espectral com matriz Laplaciana sem sinal para a Rede II

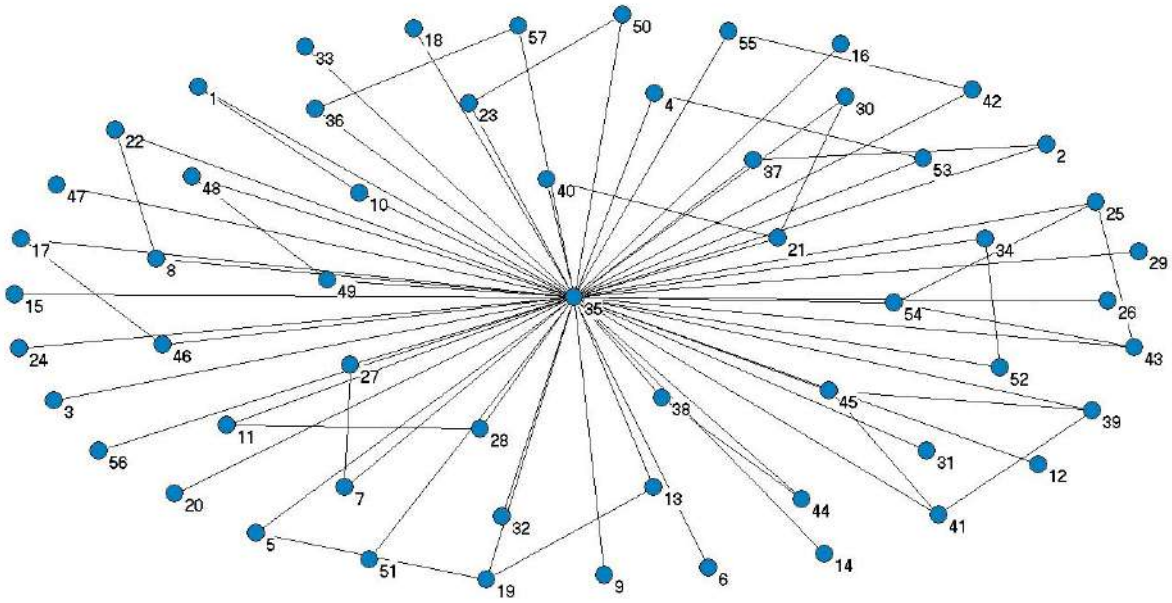


Figura VI.29 Resultado da maior partição encontrada para a Rede II

A partição da Figura VI.29 possui 57 autores e novamente há um autor mais central ligado a todos os demais e, como na partição da Rede I, esse autor também é Noga Alon. No

período de 1995 a 2001, ele publicou 88 artigos em periódicos. O grau desse autor na partição da figura é igual a 58, proximidade igual a 1 e intermediação igual a 0,9844. Os autores que ficaram em segundo lugar, apresentando os mesmos valores para as três medidas de centralidade analisadas foram: János Körner e Joel Spencer, ambos apresentaram grau igual a 3, proximidade igual a 0,5138 e intermediação igual a 0,000325. Os demais autores apresentaram valor de intermediação igual a zero. János Körner publicou dois artigos em coautoria com Noga Alon nesse período, e Joel Spencer publicou 2 com Noga Alon.

A Figura VI.30 apresenta as 11 partições geradas pelo algoritmo espectral com matriz Laplaciana, sendo essas partições correspondentes ao melhor resultado de modularidade para a Rede III. A maior partição da rede é composta por 161 vértices, o que corresponde a, aproximadamente, 14,71% dos autores da rede.

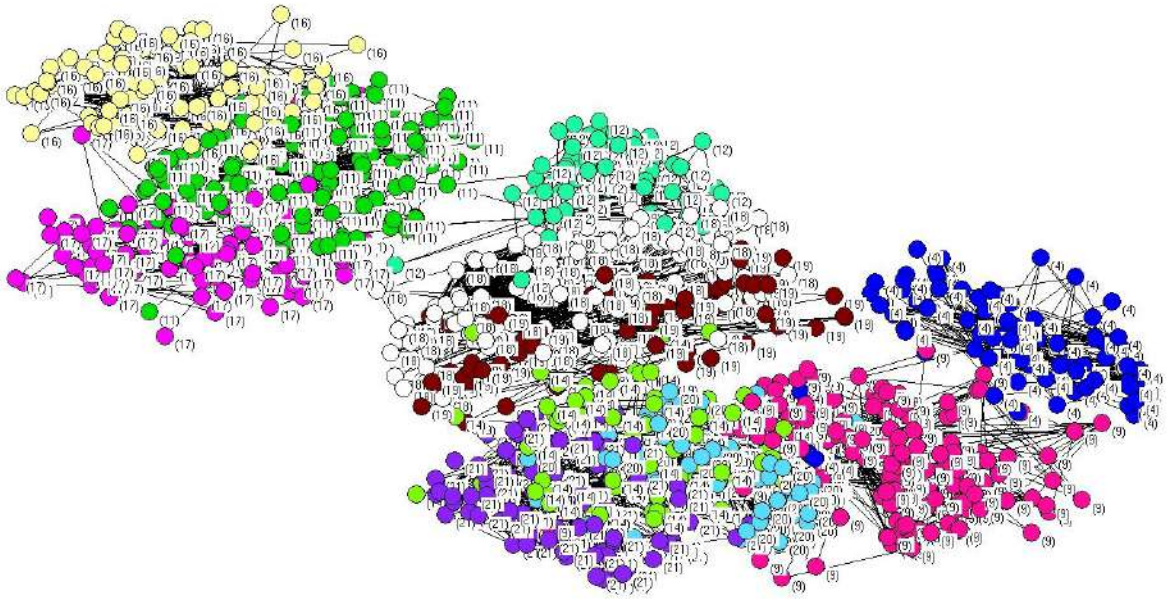


Figura VI.30 Partições geradas pelo algoritmo espectral com matriz Laplaciana para a Rede III

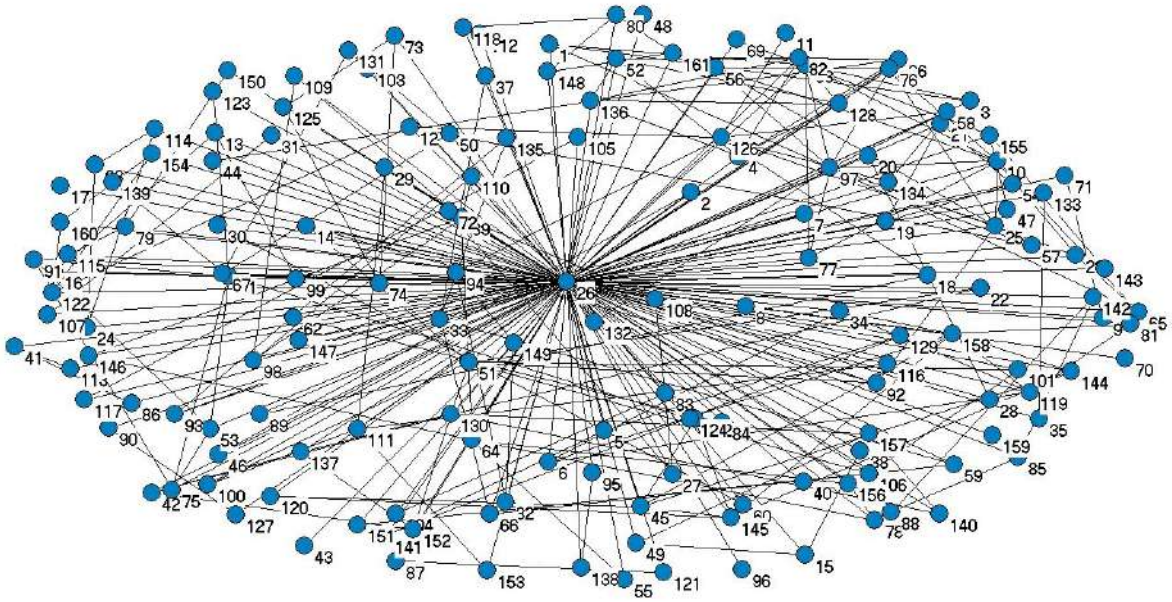


Figura VI.31 Resultado da maior partição encontrada para a Rede III

O autor mais central dessa maior partição, como pode-se verificar na Figura VI.31, é o autor 26 que corresponde a Guanrong Chen, ele possui grau igual a 159, proximidade igual a 0,9939 e intermediação igual a 0,9781. Outros autores que apresentaram as maiores medidas de centralidade foram: Shujun Li (grau=12, proximidade=0,5178, intermediação=0,0013), Chengqing Li (grau=9, proximidade=0,5128, intermediação=0,00042), Jinhu Lu (grau=8, proximidade=0,5112, intermediação= 0,0007075), Xiaofeng Liao (grau=8, proximidade=0,5128, intermediação=0,01301), Xinghuo Yu (grau=8, proximidade=0,5112, intermediação=0,000629).

Guanrong Chen possui 277 artigos publicados em periódicos até o ano de 2013, sendo que 135 (48,74%) foram publicados entre 2002 e 2007, período representado na Rede III. Segundo o índice de citações do *Google*, o total de citações desse autor é igual a 54536, e 33454 (61,34%) são citações referentes aos últimos 5 anos. Guanrong Chen já publicou artigos com 325 coautores diferentes. Desses, Jinhu Lu, Wenwu Yu, Shujun Li, Chengqing Li são os que mais publicaram com ele. Os principais campos de estudo de Chen são: Dinâmica não-linear, Redes complexas, a Teoria de sistemas de controle e Engenharia.

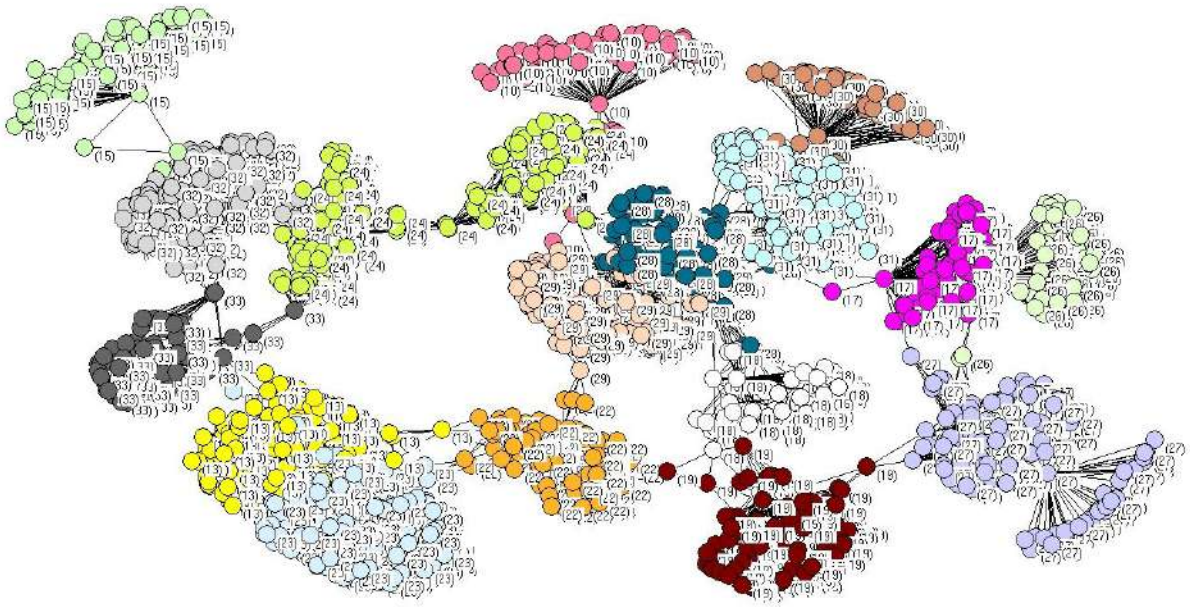


Figura VI.32 Partições geradas pelo algoritmo espectral com matriz Laplaciana para a Rede IV

A Figura VI.32 apresenta as 17 partições geradas pelo algoritmo espectral com matriz Laplaciana para *k=17*. O maior grupo possui 179 autores e pode ser melhor visualizado na Figura VI.33.

Os autores mais centrais nessa partição são: 49: Georgios B. Giannakis (grau=98, proximidade=0,5282, intermediação=0,6437), 54: H. Vincent Poor (grau=87, proximidade=0,5100, intermediação=0,6694), 130: Sergio Barbarossa (grau=12, proximidade=0,5174, intermediação=0,1652), 179: Zhi-Quan Luo (grau=9, proximidade=0,513, intermediação= 0,1187).

Georgios B. Giannakis publicou 304 artigos em periódicos até o ano de 2013, sendo que 192 foram no período de 1995 a 2007 representado pela Rede IV. Um total de 150 autores já publicaram com esse autor, sendo Shengli Zhou, Xiaoli Ma, Xiaodong Cai, Alejandro Ribeiro, Seung-Jun Kim e Gonzalo Mateos os que mais publicaram artigos em coautoria com ele. Georgios B. Giannakis possui um total de 42606 citações, segundo o *Google*, das quais 20595 (48,34%) são referentes aos últimos 5 anos. Os tópicos de pesquisa atuais do autor Georgios B. Giannakis são: Dispersão em sinais e sistemas, Aprendizagem *Big Data*, Rádios cognitivos sem fio, Energia renovável, *Power grid*, Redes Sociais e Regulação de gene.

H. Vincent Poor publicou 406 artigos em periódicos até o ano de 2013, e 120 no período representado pela Rede IV. Um total de 104 autores já publicaram com esse autor, sendo Shlomo Shamai, Sanjeev R. Kulkarni, Yingbin Liang, Lifeng Lai e Sergio Verdú, os principais coautores. A principais áreas de estudo de H. Vincent Poor são: Segurança, Computação e Redes, Ciências da Informação e Sistemas.

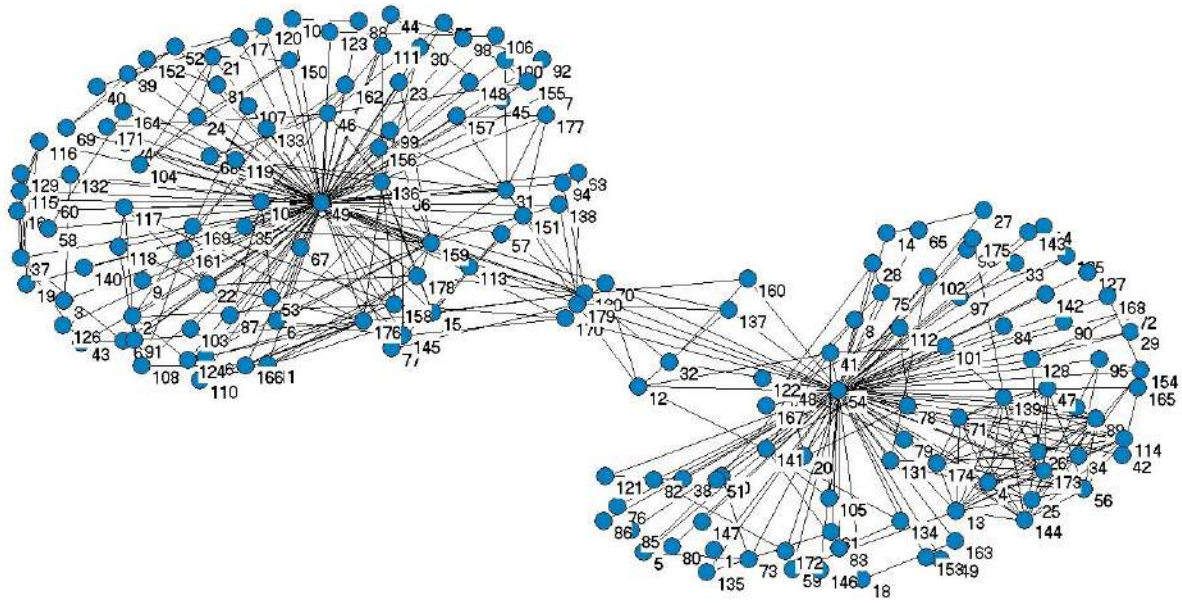


Figura VI.33 Resultado da maior partição encontrada para a Rede IV

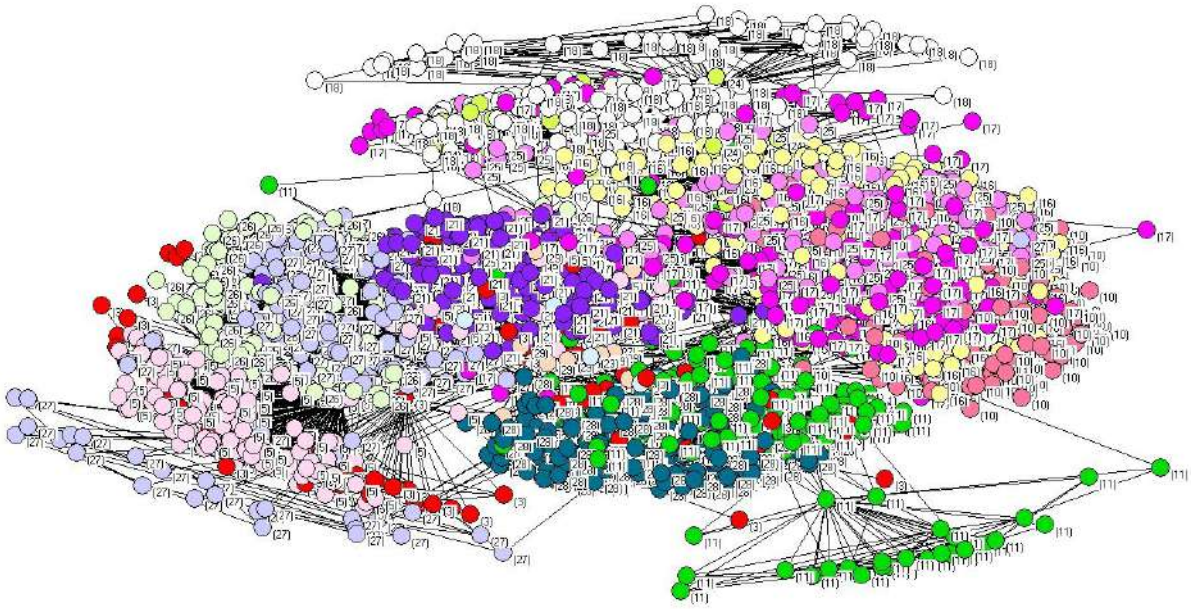


Figura VI.34 Partições geradas pelo algoritmo espectral com matriz Laplaciana para a Rede V

A Figura VI.34 mostra as 15 partições encontradas pela aplicação do algoritmo espectral com matriz Laplaciana na Rede V. Na Figura VI.35 tem-se a maior partição com 292 autores.

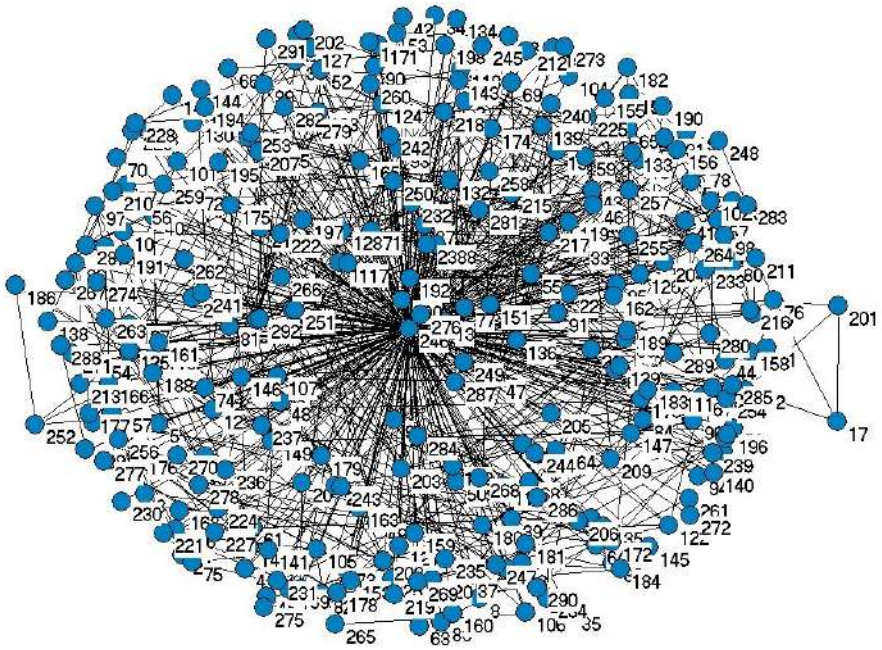


Figura VI.35 Resultado da maior partição encontrada para a Rede V

O autor mais central da partição da Figura VI.35 é o 246: Yan Zhang (grau=289, proximidade=0,9864, intermediação=0,9819). Outros autores que também apresentaram valores mais altos para as medidas de centralidade foram: Naixue Xiong (grau=23, proximidade=0,5169, intermediação=0,006063), Laurence Tianruo Yang (grau=17, proximidade=0,5114, intermediação=0,00498), Athanasios V. Vasilakos (grau=14, proximidade= 0,5087, intermediação=0,004645).

Yan Zhang publicou 282 artigos em periódicos até o ano de 2013 e 221 (78,34%) no período de 2008 a 2013, referente à Rede V. Esse autor já publicou com 1137 autores, sendo que os que mais publicaram artigos em coautoria com ele foram: Rong Yu, Hsiao-Hwa Chen, Shengli Xie, Laurence T. Yang e Lingyang Song.

O melhor resultado de partições encontrado para a Rede VI foi com 22 *clusteres* utilizando o algoritmo espectral com matriz Laplaciana, essa divisão pode ser visualizada na Figura VI.36.

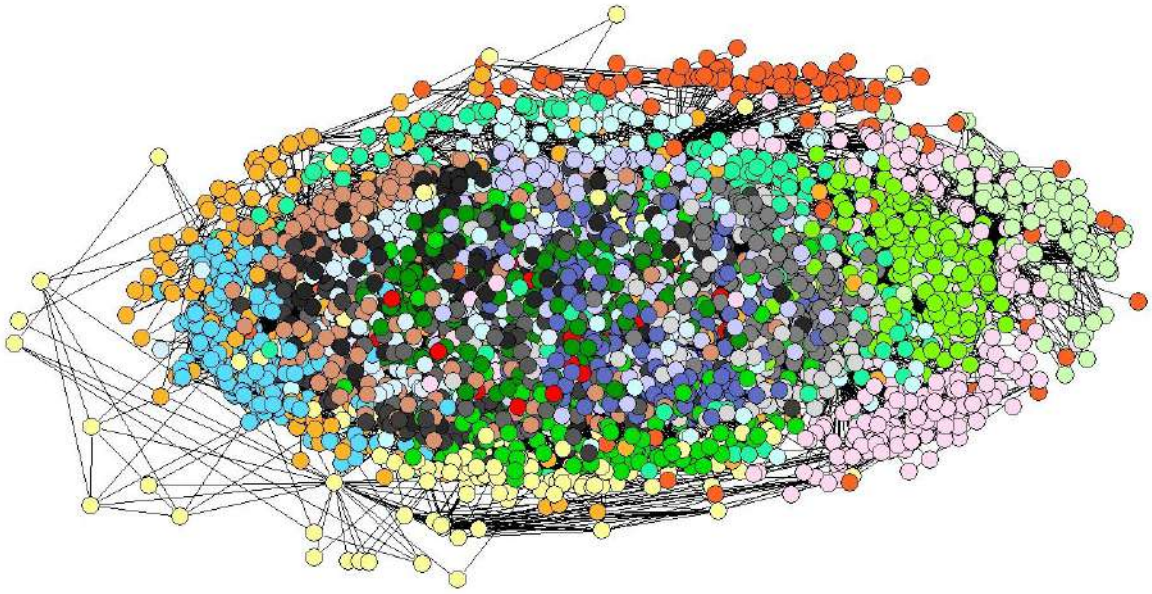


Figura VI.36 Partições geradas pelo algoritmo espectral com matriz Laplaciana para a Rede VI

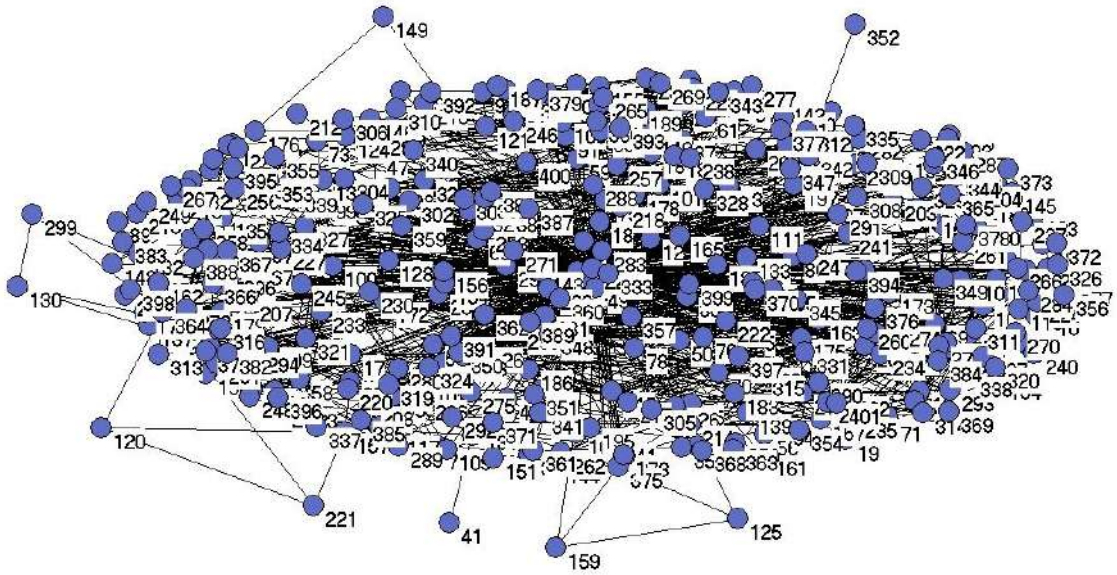


Figura VI.37 Resultado da maior partição encontrada para a Rede VI

A Figura VI.37 apresenta a maior partição da divisão, a qual contém 401 autores. O autor mais central nessa partição é Ming Li (grau=391, proximidade=0,9732, intermediação=0,9829). Outros autores mais centrais são: Yu Shyr (grau=21, proximidade=0,5063, intermediação=0,0004678), Yan Wu (grau=18, proximidade=0,5044, intermediação=0,0004367), Yong Qi (grau=6, proximidade=0,4988, intermediação=0,01492).

O autor Ming Li publicou 293 artigos em periódicos até 2013, compreendendo o período da Rede VI (1965-2013). O número de autores com os quais Ming Li já publicou conjuntamente é igual a 919, sendo Paul M. B. Vitányi, Tao Jiang, Bin Ma e Yan Wu os principais. Os projetos

de pesquisa recentes do autor estão relacionados a : Previsão da estrutura de proteínas, Consulta e resposta do motor de busca e Reconhecimento de imagem celular.

Como se pode perceber a maioria dos principais autores que fazem parte da maior partição das seis redes DBLP estão presentes nas Tabelas VI.14, VI.15 e VI.16, ou seja, fazem parte do grupo dos 10 autores mais centrais nessas redes. Alguns presentes nessas Tabelas e que não estão nessas maiores partições devem estar presentes como principais em outras partições não analisadas, visto que é esperado que os autores mais centrais na rede completa também devem aparecer como mais centrais nas partições em que fazem parte.

Através das análises dessas partições verificou-se também as diferentes áreas de pesquisa desses autores que inclui desde Teoria dos Grafos, Redes Sociais a pesquisas nas áreas de Biologia, como Previsão da estrutura de proteínas e Reconhecimento de imagem celular. Essas análises são importantes por auxiliarem a detectar as áreas que estão sendo mais pesquisadas em ciências da computação, visto que como esses autores principais demonstraram ter uma importância para a rede como um todo, suas pesquisas também contribuem para avanços da área.

Capítulo VII - Conclusões

Este trabalho buscou fazer um estudo sobre agrupamentos em redes sociais de coautoria utilizando algoritmos espectrais, com as matrizes Laplaciana e Laplaciana sem sinal, com o objetivo de verificar a qualidade dos resultados gerados para a clusterização dos autores em tais redes. Esses resultados foram comparados com os alcançados pelo algoritmo *k-means*. A medida de qualidade utilizada para avaliar as partições foi a modularidade.

Para isso foi feito um levantamento teórico sobre o assunto e uma fase de teste em cem redes geradas aleatoriamente, nas quais foram aplicados os algoritmos espectrais e um modelo de programação linear inteira que buscava otimizar a similaridade dentro dos *clusteres* formados. Essa fase de testes realizada em redes com número de vértices que variavam de 20 a 80, aproximadamente, apresentou melhores resultados para o uso do modelo de PL. Porém, com o aumento do número de vértices nas redes, esse modelo passava a ter um custo computacional muito elevado, sendo que, dependendo das especificações do computador devido ao grande número de restrições do modelo, não era sequer possível gerar os resultados.

Dessa forma, foram escolhidas nove redes de coautoria para aplicação dos algoritmos espectrais e *k-means*, três já utilizadas em outros trabalhos na literatura e seis retiradas de uma busca na base de dados de ciências da computação chamada DBLP. As componentes gigantes dessas redes possuem tamanhos bem diferentes, a menor tem 184 vértices e a maior, 5065. Foram feitas também análises quanto às propriedades e estrutura das redes. Para comparação dos resultados de agrupamentos foi escolhida a modularidade, pois essa medida é baseada na coesão dentro dos *clusteres* e é eficiente na avaliação da formação de grupos nas redes.

Através dos resultados e das análises feitas, concluiu-se que o uso do algoritmo espectral com matriz Laplaciana obteve melhor desempenho dentre os três algoritmos, conseguindo melhores resultados em sete das nove redes analisadas. O uso do algoritmo com matriz Laplaciana, como pode ser visualizado nas figuras dos resultados conseguiu, na maioria das vezes, os melhores resultados dentre os algoritmos, mesmo nas demais quantidades de clusteres que não eram a melhor. Deve-se considerar, entretanto, que, na maioria das vezes, a matriz Laplaciana sem sinal conseguiu resultados bem próximos aos da Laplaciana. E, além disso, tanto o algoritmo Laplaciano sem sinal espectral quanto o *k-means* mesmo não tendo alcançado os melhores resultados na maioria das redes, muitas vezes conseguiram obter partições com modularidade $Q>0,7$, valor este que indica um bom desempenho na clusterização realizada por tais algoritmos.

Portanto, essa dissertação buscou analisar e validar o uso das técnicas utilizadas em redes de coautoria verificando-se os resultados de partições obtidos, e concluiu-se que o uso

da matriz Laplaciana foi mais eficiente que os outros dois algoritmos, demonstrando melhor performance quando analisada a qualidade dos agrupamentos.

VII.1 Melhorias e Trabalhos Futuros

As melhorias propostas para pesquisas futuras estão relacionadas a um aprofundamento das análises realizadas nas partições. Nesse trabalho foram analisados apenas os principais autores das maiores partições encontradas e as análises foram feitas com enfoque na centralidade deles e nas linhas de pesquisa. Contudo, podem ser feitas análises das demais partições e das ligações existentes entre os grupos, quais autores estão ligados a outras partições e se participam de diferentes grupos de pesquisa.

Como trabalhos futuros serão enumeradas duas tarefas. Uma delas é a utilização das matrizes Laplacianas a partir da matriz de distâncias do grafo para verificação de possíveis melhorias nos resultados. Essa modificação no uso das matrizes Laplacianas pode ser vista em AOUCHICHE e HANSEN (2013). Outra está relacionada à construção de um *workflow* contendo todas as etapas de geração dos resultados, como o uso dos diferentes algoritmos, escolhas das quantidades de *clusteres*, cálculo da modularidade, etc. Esse *workflow* ajuda no entendimento de todo o processo para construção do trabalho e em implementações futuras em outras redes, além disso, ele indica a utilização de *e-Science* na construção do trabalho.

Referências Bibliográficas

AALST, W.M.P.van der; SONG, M. Mining Social Networks: Uncovering Interaction Patterns in Business Processes. In: DESEL, J.; PERNICI, B.; WESKE, M. (Eds.). *Business Process Management*. Lecture Notes in Computer Science. Berlin, Germany: Springer, 2004. p. 244–260.

ABREU, N.M.M. et al. Introdução à teoria espectral de grafos com aplicações. São Carlos: SBMAC, 2007.

ALBERT, R.; JEONG, H.; BARABÁSI, A.-L. Internet: Diameter of the World-Wide Web. *Nature*, v. 401, n. 6749, p. 130–131, 1999.

ALPERT, C.J.; KAHNG, A.B.; YAO, S.-Z. Spectral partitioning with multiple eigenvectors. *Discrete Applied Mathematics*, v. 90, n. 1, p. 3-26, 1999.

AMARAL, L.A.N. et al. Classes of small-world networks. *Proceedings of the National Academy of Sciences*, v. 97, n. 21, p. 11149–11152, 2000.

AOUCHICHE, M.; HANSEN, P. Two Laplacians for the distance matrix of a graph. *Linear Algebra and its Applications*, v. 439, n. 1, 1, p. 21-33, 2013.

BARABÁSI, A.L. et al. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, v. 311, n. 3–4, p. 590–614, 2002.

BARABÁSI, A.L.; ALBERT, R.; JEONG, H. Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A: Statistical Mechanics and its Applications*, v. 281, n. 1–4, p. 69–77, 2000.

BARBOSA, D. et al. Medidas de centralidade e detecção de comunidades em redes de co-autoria. In: XLIII SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL, 43., 2011, Ubatuba. *Anais do XLIII Simpósio Brasileiro de Pesquisa Operacional*. Ubatuba: 2011.

BATAGELJ, V.; MRVAR, A. Pajek - Program for Analysis and Visualization of Large Networks: *Reference Manual*, 2011. Disponível em: <http://vlado.fmf.uni-lj.si/pub/networks/pajek/doc/pajekman.pdf>. Acesso em março de 2012.

BAVELAS, A. A Mathematical Model for Group Structures. *Human Organization*, v. 7, n. 3, p. 16–30, 1948.

BOCALETTI, S.; LATORA, V.; MORENO, Y.; CHAVEZ, M.; HWANG, D.-U. Complex networks: Structure and dynamics. *Physics Resports*, v.424, n. 4, p. 175-308, 2006.

BOGINSKI, V.; BUTENKO, S.; PARDALOS, P. M. Mining market data: A network approach. Computers & Operations Research, *Part Special Issue: Operations Research and Data Mining*. v. 33, n. 11, p. 3171–3184, 2006.

BONACICH, P. Power and Centrality: A Family of Measures. *American Journal of Sociology*, v. 92, n. 5, p. 1170–1182, 1987.

BONDY, J.A.; MURTY, U.S.R. *Graph theory*. New York: Springer, 2008.

CHEN, J.; YUAN, B. Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics*, v. 22, n. 18, p. 2283–2290, 2006.

CHEONG, F.; CORBITT, B.J. A Social Network Analysis of the co-authorshipnetwork of the Pacific Asia Conference on Information Systems from 1993 to 2008. In: PACIS ’09. XVII PACIFIC ASIA

CONFERENCE ON INFORMATION SYSTEMS, 17., 2009. Hyderabad, India. *Proceedings of the Seventeenth Pacific Asia Conference on Information Systems*. Hyderab, India, jul. 2009.

CHOI, T.Y.; HONG, Y. Unveiling the structure of supply networks: case studies in Honda, Acura, and DaimlerChrysler. Journal *of Operations Management*, v. 20, n. 5, p. 469–493, 2002.

CHUNG, F. *Spectral Graph Theory.* CBMS Regional Conference Series in Mathematics, No. 92. American Mathematical Society, 1996.

CRANE, D. *Invisible colleges; diffusion of knowledge in scientific communities*. Chicago: University of Chicago Press, 1972.

CVETKOVIĆ, D.; SIMIĆ, S.K. Towards a spectral theory of graphs based on the signless Laplacian, I. *Publications de L'institut Mathématique*, v. 85, n. 99, p. 19–33, 2009.

DBLP. dblp: DBLP Bibliography. Disponível em: <http://dblp.uni-trier.de/db/>. Acesso em janeiro de 2014.

DING, C. H. Q. et al. A min-max cut algorithm for graph partitioning and data clustering. In: ICDM '01, Proceedings IEEE International Conference on Data Mining, 2001. Anais... In: ICDM 2001. INTERNATIONAL CONFERENCE ON DATA MINING, 2001. San Jose, CA. *Proceedings of the International Conference on Data Mining*. San Jose, CA: IEEE, 2001, p. 107-114.

DUAN, D. et al. Incremental K-clique clustering in dynamic social networks. *Artificial Intelligence Review*, v. 38, n. 2, p. 129–147, 2012.

FIEDLER, M. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, v. 23, n. 2, p. 298–305, 1973.

FORTUNATO, S. Community detection in graphs. *Physics Reports*, v. 486, n. 3–5, p. 75–174, 2010.

FORTUNATO, S.; BARTHÉLEMY, M. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, v. 104, n. 1, p. 36–41, 2007.

FREEMAN, L. C. A Set of Measures of Centrality Based on Betweenness. *Sociometry*, v. 40, n. 1, p. 35, 1977.

FREEMAN, L. C. Centrality in social networks conceptual clarification. *Social Networks*, v. 1, n. 3, p. 215–239, 1978.

GIRVAN, M.; NEWMAN, M. E. J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences,* v. 99, n. 12, p. 7821–7826, 2002.

GROSSMAN, J. W. The evolution of the mathematical research collaboration graph. *Congressus Numerantium*, p. 201–212, 2002.

HAGEN, L.; KAHNG, A.B. New spectral methods for ratio cut partitioning and clustering. *Computer-aided design of integrated circuits and systems, IEEE Council on Electronic Design Automation*, v. 11, n. 9, p. 1074-1085, 1992.

HAN, J.; KAMBER, M.; PEI, J. *Data mining concepts and techniques*, Third edition. Waltham, Mass.: Morgan Kaufmann Publishers, 2012.

HOFSTAD, R. VAN DER. Random Graphs and Complex Networks. *Technische Universiteit Eindhoven*, 2010.

IACOBUCCI, D.; OSTROM, A. Commercial and interpersonal relationships; Using the structure of interpersonal relationships to understand individual-to-individual, individual-to-firm, and firm-to-firm relationships in commerce. *International Journal of Research in Marketing*, v. 13, n. 1, p. 53–72, 1996.

JIANG, P.; SINGH, M. SPICi: a fast clustering algorithm for large biological networks. *Bioinformatics*, v. 26, n. 8, p. 1105–1111, 2010.

JIANG, Y. Locating active actors in the scientific collaboration communities based on interaction topology analyses. *Scientometrics*, v. 74, n. 3, p. 471–482, 2008.

KARINTHY, F. Chains. In: Everything is different. Budapeste, 1929.

KATZ, J. S.; HICKS, D. How much is a collaboration worth? A calibrated bibliometric model. *Scientometrics*, v. 40, n. 3, p. 541–554, 1997.

KERNIGHAN, B.W.; LIN, S. An Efficient Heuristic Procedure for Partitioning Graphs. *Bell System Technical Journal*, v. 49, n. 2, p. 291–307, 1970.

KLEINBERG, J. The Small-world Phenomenon: An Algorithmic Perspective. In: STOC'00. THIRTY-SECOND ANNUAL ACM SYMPOSIUM ON THEORY OF COMPUTING, 32., 2000, New York: USA. *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*. New York, USA: ACM, 2000. p. 163-170.

KRETSCHMER, H. Coauthorship networks of invisible colleges and institutionalized communities. *Scientometrics*, v. 30, n. 1, p. 363–369, 1994.

KRINGS, G. et al. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, v. 2009, n. 07, p. L07003, 2009.

LIU, N.; STEWART, W.J. Markov Chains and Spectral Clustering. In: HUMMEL, K. A.; HLAVACS, H.; GANSTERER, W. (Eds.). Performance Evaluation of Computer and Communication Systems. Milestones and Future Challenges. Lecture Notes in Computer Science. Berlin, Germany: Springer, 2011. p. 87–98.

LUXBURG, U.von. A tutorial on spectral clustering. *Statistics and Computing*, v. 17, n. 4, p. 395–416, 2007.

MILGRAM, S. The Small-World Problem. *Psychology Today*, v. 1, n. 1, p. 61–67, 1967.

MIZRUCHI, M.S. Cohesion, equivalence, and similarity of behavior: a theoretical and empirical assessment. *Social Networks*, v. 15, n. 3, p. 275–307, 1993.

MOHAR, B. Laplace eigenvalues of graphs—a survey. *Discrete Mathematics*, v. 109, n. 1–3, p. 171–183, 1992.

MOODY, J.; WHITE, D.R. Structural Cohesion and Embeddedness: A Hierarchical Concept of Social Groups. *American Sociological Review*, v. 68, n. 1, p. 103, 2003.

NASCIMENTO, M.C.V.; TOLEDO, F.M.B.; DE CARVALHO, A.C.P.L.F. Investigation of a new GRASP-based clustering algorithm applied to biological data. *Computers & Operations Research*, v. 37, n. 8, p. 1381–1388, 2010.

NEWMAN, M.E.J. Models of the Small World. *Journal of Statistical Physics*, v. 101, n. 3-4, p. 819–841, 1 2000.

NEWMAN, M.E.J. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, v. 98, n. 2, p. 404–409, 2001.

NEWMAN, M.E.J. Who is the best connected scientist? A study of scientific coauthorship networks. In: Ben-Naim, E.; Frauenfelder, H.; Toroczkai, Z. (Eds.), Complex networks. Berlin, Germany: Springer, 2004. p. 337–370.

NEWMAN, M.E.J. Detecting community structure in networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, v. 38, n. 2, p. 321–330, 2004b.

NEWMAN, M.E.J. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, v. 101, n. 1, p. 5200–5205, 2004c.

NEWMAN, M.E.J. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, v. 103, n. 23, p. 8577–8582, 2006.

NEWMAN, M.E.J.; GIRVAN, M. Finding and evaluating community structure in networks. *Physical Review E*, v. 69, n. 2, p. 026113, 2004.

NEWMAN, M.E.J.; STROGATZ, S. H.; WATTS, D. J. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, v. 64, n. 2, p. 026118, 2001.

NOOY, W. DE; MRVAR, A.; BATAGELJ, V. *Exploratory social network analysis with Pajek*. New York: Cambridge University Press, 2005.

PRICE, D.D.S. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, v. 27, n. 5, p. 292–306, 1976.

PRICE, D.J. . Networks of Scientific Papers. *Science*, v. 149, p. 510–515, 1965.

QIU, H.; HANCOCK, E.R. Graph matching and clustering using spectral partitions. *Pattern Recognition*, v. 39, n. 1, p. 22–34, 2006.

RAO, M.R. Cluster Analysis and Mathematical Programming. *Journal of the American Statistical Association*, v. 66, n. 335, p. 622–626, 1971.

RICHARDSON, T.; MUCHA, P.J.; PORTER, M.A. Spectral tripartitioning of networks. *Physical Review E*, v. 80, n. 3, p. 036111, 2009.

SARKAR, S.; DONG, A. Community detection in graphs using singular value decomposition. *Physical Review E*, v. 83, n. 4, p. 046114, 2011.

SCHAEFFER, S.E. Graph clustering. *Computer Science Review*, v. 1, n. 1, p. 27–64, 2007.

SCOTT, J. Social Network Analysis: a Handbook. 2nd ed. London: Sage Publications, 2000.

SERG. MIT Strategic Engineering Research Group: SERG. Disponível em: <http://strategic.mit.edu/downloads.php?page=matlab_networks>. Acesso em agosto de 2013.

SHI, J.; MALIK, J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, n. 8, p. 888–905, 2000.

SIN, S.-C.J. International coauthorship and citation impact: A bibliometric study of six LIS journals, 1980–2008. *Journal of the American Society for Information Science and Technology*, v. 62, n. 9, p. 1770–1783, 2011.

SOUZA, C.G.; BARBASTEFANO, R.G.; LIMA, L.S. Redes de Colaboração científica na área de Química no Brasil: um estudo baseado nas coautorias dos artigos da Revista Química Nova. *Química Nova*, v. 35, n. 4, p. 671–676, 2012.

SOUZA, C. G. DE; BARBASTEFANO, R. G. Knowledge diffusion and collaboration networks on life cycle assessment. *The International Journal of Life Cycle Assessment*, v. 16, n. 6, p. 561–568, 2011.

SPIELMAN, D. A.; TENG, S.H. Spectral Partitioning Works: Planar graphs and Ginite element meshes. In: FOCS '96. THIRTY-SEVENTH ANNUAL SYMPOSIUM ON FOUNDATIONS OF COMPUTER SCIENCE, 37., 1996, Burlington, VT. *Proceedings of the 37th Annual Symposium on Foundations of Computer Science.* Burlington, VT: IEEE, 1996. p. 96-105.

TAHA, H. A. *Operations research: an introduction*. 8th ed ed. Upper Saddle River, N.J: Pearson/Prentice Hall, 2007.

TOMASSINI, M.; LUTHI, L. Empirical analysis of the evolution of a scientific collaboration network. *Physica A: Statistical Mechanics and its Applications*, v. 385, n. 2, p. 750–764, 2007.

TRAVERS, J.; MILGRAM, S. An experimental study of the Small World problem. Sociometry, v. 32, n. 4, p. 425–443, dez. 1969.

VICINI, L.; SOUZA, A. M. *Análise multivariada da teoria à prática.* Santa Maria: UFSM, 2005.

WASSERMAN, S.; FAUST, K. *Social network analysis: methods and applications*. Cambridge ; New York: Cambridge University Press, 1994.

WATTS, D. J. *Small worlds: the dynamics of networks between order and randomness*. Princeton, N.J.; Woodstock: Princeton University Press, 2004.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of “small-world” networks. *Nature*, v. 393, n. 6684, p. 440–442, 1998.

WEBSTER, C. M.; MORRISON, P. D. Network Analysis in Marketing. *Australasian Marketing Journal (AMJ)*, v. 12, n. 2, p. 8–18, 2004.

WHITE, D. R.; HARARY, F. The Cohesiveness of Blocks In Social Networks: Node Connectivity and Conditional Density. *Sociological Methodology*, v. 31, n. 1, p. 305–359, 2001.

WU, X. et al. Top 10 algorithms in data mining. *Knowledge and Information Systems*, v. 14, n. 1, p. 1–37, 2008.