AGRUPAMENTOS MÚLTIPLOS NÃO-REDUNDANTES EM GRAFOS COM ATRIBUTOS

Gustavo Paiva Guedes e Silva

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia de Sistemas e Computação.

Orientador: Geraldo Bonorino Xexéo

Rio de Janeiro
Setembro de 2015

AGRUPAMENTOS MÚLTIPLOS NÃO-REDUNDANTES EM GRAFOS COM ATRIBUTOS

Gustavo Paiva Guedes e Silva

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

Prof. Geraldo Bonorino Xexéo, D.Sc.

Prof. Jano Moreira de Souza, Ph.D.

Prof. Abílio Pereira de Lucena Filho, Ph.D.

Prof. Eduardo Soares Ogasawara, D.Sc.

Prof. Carlos Eduardo Ribeiro de Mello, Ph.D.

RIO DE JANEIRO, RJ – BRASIL
SETEMBRO DE 2015

Silva, Gustavo Paiva Guedes e

Agrupamentos Múltiplos Não-Redundantes em Grafos com Atributos/Gustavo Paiva Guedes e Silva. – Rio de Janeiro: UFRJ/COPPE, 2015.

XVII, 104 p.: il.; 29, 7cm.

Orientador: Geraldo Bonorino Xexéo

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2015.

Referências Bibliográficas: p. 82 – 104.

1. Agrupamentos múltiplos. 2. Grafos com atributos. 3. Agrupamentos em grafos. I. Xexéo, Geraldo Bonorino. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

A verdadeira viagem de descobrimento não consiste em procurar novas paisagens, mas em ter novos olhos.

Marcel Proust

A meus pais, por tudo.

Agradecimentos

À Deus, por permitir a conclusão dessa tese;
À minha família, por tudo;
À Geraldo Xexéo, por tornar essa tese possível, por sua orientação e paciência;
À Eduardo Bezerra, por ter contribuído em meus estudos e vida acadêmica, pela ajuda nos formalismos e por ter me apresentado ao tema base dessa tese;
À Eduardo Ogasawara, imprescindível na realização dessa tese e dos artigos mais importantes. Também pelos ensinamentos, incentivo e pelas estratégias valiosas.
À Rafael Cotta, peça fundamental na minha trajetória acadêmica e profissional;
À Fellipe Duarte, Filipe Braida e Fabrício Pereira pelo incentivo;
Aos professores do CEFET/RJ pelo apoio;
À Lilian Ferrari pela amizade, por ter me iniciado na vida acadêmica e por seus valiosos conselhos durante a vida;
À Fernanda Rosa pelo total acolhimento no CEFET/RJ;
À Leandro Diniz por ter me ensinado a programar e por ser responsável pelo meu primeiro emprego na área de programação;
À Thereza Nascimento pela indicação para meu primeiro emprego com Java.
Aos meus amigos de tantos anos, Aline Aurora, Anne Sueli, Henrique Moreno, José Jorge, Leandro Diniz, Maria Izabel e Rodrigo Vinícius pelos momentos que desfrutamos nos momentos em que consegui dar pausa nos estudos;
Aos professores Jano Moreira, Abílio Pereira, Eduardo Ogasawara e Carlos Eduardo por participarem da banca de defesa.
Por fim, aos meus mestres de toda vida, os da faculdade de letras, faculdade de ciência da computação, mestrado e doutorado. Todos contribuíram para a minha formação;

agradeço.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

AGRUPAMENTOS MÚLTIPLOS NÃO-REDUNDANTES EM GRAFOS COM ATRIBUTOS

Gustavo Paiva Guedes e Silva

Setembro/2015

Orientador: Geraldo Bonorino Xexéo

Programa: Engenharia de Sistemas e Computação

Muitos algoritmos de agrupamento em grafo destinam-se a gerar uma única partição (agrupamento) dos dados. No entanto, um mesmo conjunto de dados pode produzir diferentes agrupamentos. De uma perspectiva exploratória, dado um conjunto de dados, a possibilidade de se gerar agrupamentos alternativos e não-redundantes é importante para a compreensão dos dados. Cada um desses agrupamentos poderia proporcionar um ponto de vista diferente sobre esses dados. Muitas áreas demandam a exploração de diversas soluções de agrupamento, como a área de marketing em redes sociais. É nesse contexto que esse trabalho se insere, apresentando um novo algoritmo para gerar agrupamentos múltiplos a partir de um grafo com atributos. Nesse tipo de grafo, cada vértice está associado a uma n-tupla de atributos (por exemplo, em uma rede social, os usuários têm interesses, sexo, idade, etc.). A abordagem concebida adiciona arestas artificiais entre vértices semelhantes do grafo utilizando a similaridade entre os atributos, o que resulta em um grafo com atributos aumentado. Em seguida, é aplicado um algoritmo de agrupamento nesse novo grafo. Dessa maneira, diversos agrupamentos são gerados utilizando diferentes combinações de atributos. Os resultados experimentais indicam que a abordagem é eficaz na tarefa de produzir agrupamentos múltiplos em grafos com atributos. No entanto, o problema não é apenas produzir os agrupamentos múltiplos, mas produzi-los de maneira que não sejam redundantes. Nesse cenário, esse trabalho também contribui com um novo algoritmo para selecionar os *top-k* agrupamentos não-redundantes. Resultados experimentais utilizando uma rede social *online* real e uma rede de co-autoria mostram a eficácia dos algoritmos propostos.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

MULTIPLE NON-REDUDANT CLUSTERINGS IN ATTRIBUTED GRAPHS

Gustavo Paiva Guedes e Silva

September/2015

Advisor: Geraldo Bonorino Xexéo

Department: Systems Engineering and Computer Science

Many graph clustering algorithms aim at generating a single partitioning (clustering) of the data. However, the same set of data may produce different clusterings. From a exploratory perspective, given a dataset, the availability of many different and non-redundant clusterings is important for data understanding. Each one of these clusterings could provide a different insight about the data. Many areas demands exploring multiple clustering solutions, such as marketing in social networks. This work is immersed in this context, presenting a novel algorithm that generates multiple clusterings from an attributed graph. In this type of graph, each vertex is associated to a n-tuple of attributes (e.g., in a social network, users have interests, gender, age, etc.). The approach adds artificial edges between similar vertices of the graph using similarity between attributes, which results in an augmented attributed graph. Then a clustering algorithm is applied in this new graph. Thus, multiple clusterings are produced by using different combinations of attributes. Experimental results indicate the algorithms are effective in providing multiple clusterings in attributed graphs. Indeed, the problem is not only to provide multiple solutions, but multiple non-redundant solutions. In this scenario, this work also contributes with a novel algorithm that discovers the top-k non-redundant clusterings. Experimental results using a real online social network and a co-authorship network show the effectiveness of proposed algorithm.

Sumário

Lista de Figuras **xi**

Lista de Tabelas **xiv**

Lista de Símbolos **xv**

Lista de Abreviaturas **xvii**

1 Introdução **1**

1.1 Exemplo motivador 5

1.2 Definição do problema 7

1.3 Contribuições 8

1.4 Organização da tese 9

2 Fundamentação Teórica **11**

2.1 Agrupamentos 11

2.2 Agrupamentos múltiplos 15

2.3 Agrupamentos em grafos 17

2.3.1 Agrupamentos em grafos com atributos nas arestas 21

2.3.2 Agrupamentos em grafos com atributos nos vértices 21

2.4 Comparação entre dois agrupamentos 22

2.5 Comparação entre k agrupamentos 25

3 Trabalhos relacionados **27**

3.1 Produção de agrupamentos múltiplos em grafos 28

3.2 Agrupamentos de consenso em grafos 30

3.3 Combinação da estrutura topológica com os atributos dos vértices de um grafo 32

3.4 Considerações 33

4 Agrupamentos múltiplos em grafos com atributos **35**

4.1 Algoritmo para agrupamento em grafos com atributos por adição de arestas artificiais 36

4.2 Algoritmo para agrupamentos múltiplos em grafos com atributos . . . 38
4.3 Medidas de comparação entre agrupamentos múltiplos 40
4.3.1 Baseada em média simples 42
4.3.2 Baseado em média e variância 42
4.3.3 Baseada na média dos quadrados 43
4.4 Algoritmo para ranquear agrupamentos não-redundantes 43
4.5 Efeito das medidas de comparação de agrupamentos no ranqueamento 45
4.6 RM-CRAG e o problema de máxima diversidade 46

5 Avaliações experimentais 50
5.1 Conjuntos de dados 51
5.1.1 MQD500b 51
5.1.2 MQD500c 52
5.1.3 DBLP3000 52
5.2 Medidas de avaliação 52
5.3 CRAG e M-CRAG 54
5.3.1 MQD500c 55
5.3.2 MQD500b 56
5.3.3 DBLP3000 57
5.4 RM-CRAG 58
5.4.1 MQD500c 58
5.4.2 MQD500b 61
5.4.3 DBLP3000 62
5.4.4 Análise por nuvem de palavras 63
5.4.5 Análise das diferentes medidas 67
5.5 Discussão 72

6 Conclusões 75
6.1 Sumário de contribuições 76
6.2 Limitações 77
6.3 Trabalhos futuros 78

Referências Bibliográficas 82

Lista de Figuras

1.1 Exemplo de uma rede social representada como um grafo. Os vértices representam os usuários. As arestas entre os vértices representam a relação de amizade. 2
1.2 Exemplo de usuários em uma rede social em que as arestas representam o contato físico. As arestas de contato físico estão representadas na cor verde. 3
1.3 Exemplo ilustrando múltiplas soluções de agrupamento em um grafo com atributos. 6

2.1 Agrupamento hierárquico aglomerativo 12
2.2 Agrupamento particional 13
2.3 Diferentes soluções de agrupamento utilizando o *k-means* com escolha aleatória dos centróides iniciais. 15
2.4 Representação de quatro objetos a serem agrupados. Dois quadrados (um azul e um vermelho) e dois círculos (um azul e um vermelho). . 15
2.5 Soluções de agrupamento 16
2.6 Exemplo de grafo não-dirigido e grafo dirigido. 18
2.7 Matriz de adjacência para o grafo representado na Figura 2.6a. 18
2.8 Matriz de grau para o grafo representado na Figura 2.6a. 19
2.9 Matriz de grau para o grafo representado na Figura 2.6a. 20
2.10 Representação de um multigrafo com as relações de amizade, trabalho e academia. 21
2.11 Matriz de NMI calculada através da comparação entre cada par de agrupamentos. 24

3.1 Síntese dos trabalhos relacionados. 34

4.1 Exemplo de grafo com atributos nos vértices: (a) sem arestas artificiais. (b) com arestas artificiais. 37
4.2 Produção de agrupamentos múltiplos pelo algoritmo M-CRAG. . . . 40
4.3 Representação gráfica das medidas μ, $\mu + \sigma^2$ e q^2 em 3D e 2D. 41

4.4 Modelo representando o funcionamento dos algoritmos CRAG, M-CRAG e RM-CRAG. 44
4.5 Representação de três conjuntos de agrupamentos dados por $\mathcal{C}_1$, $\mathcal{C}_2$ e $\mathcal{C}_3$. 46
4.6 Ilustração do MDP em um grafo. 47
4.7 Vetor representando a escolha de m soluções de agrupamento. 48

5.1 Avaliação da entropia, densidade e NMI dos agrupamentos gerados pelo algoritmo CRAG no conjunto de dados `MQD500c`. 55
5.2 Avaliação da entropia, densidade e NMI dos agrupamentos gerados pelo algoritmo CRAG no conjunto de dados MQD500b. 56
5.3 Avaliação da entropia, densidade e NMI dos agrupamentos gerados pelo algoritmo CRAG no conjunto de dados DBLP3000. 58
5.4 Avaliação do tempo e quantidade de combinações para a seleção dos *top-k* agrupamentos gerados pela abordagem de força bruta (`FB`) no conjunto de dados `MQD500c` para $m = 5$. 59
5.5 Comparação entre `RM-CRAG`, `ITS(20s)` e `ITS(1s)`, `FB` e Aleatório no conjunto de dados `MQD500c` para $m = 5$. 60
5.6 Comparação entre o `RM-CRAG`, `ITS(20s)` e `ITS(1s)` e Aleatório para o conjunto de dados `MQD500b` para $m = 5$. 62
5.7 Comparação entre `RM-CRAG`, `ITS(20s)` e `ITS(1s)` e Aleatório para o conjunto de dados `DBLP5000` para $m = 3$. 63
5.8 Comparação entre as nuvens de palavras geradas para os top-2 agrupamentos selecionados pelo RM-CRAG com $m = 5$ para o conjunto de dados `MQD500b`. 64
5.9 Matriz de similaridade entre as palavras presentes nos grupos dos agrupamentos selecionados pelo RM-CRAG para o conjunto de dados `MQD500b`. 65
5.10 Comparação entre as nuvens de palavras geradas para os top-2 agrupamentos selecionados pelo RM-CRAG com $m = 3$ para o conjunto de dados `DBLP3000`. 66
5.11 Comparação entre as nuvens de palavras geradas para os top-2 agrupamentos selecionados pelo RM-CRAG com $m = 3$ para o conjunto de dados DBLP3000. 67
5.12 Média e variância da NMI encontrada entre os *top-k* agrupamentos selecionados pelo RM-CRAG utilizando as diferentes medidas GANMI, MVNMI e QNMI para o conjunto de dados `MQD500b`. 68

5.13 Quantidade de agrupamentos diferentes selecionados pelo RM-CRAG utilizando as diferentes medidas de média para o conjunto de dados `MQD500b`. 69
5.14 Densidade e entropia dos k agrupamentos selecionados pelo RM-CRAG utilizando as diferentes medidas: GANMI, MVNMI e QNMI. 70
5.15 Média e variância da NMI encontrada entre os k agrupamentos selecionados pelo RM-CRAG utilizando as diferentes medidas: GANMI, MVNMI e QNMI. 71
5.16 Quantidade de agrupamentos diferentes selecionados pelo RM-CRAG utilizando as diferentes medidas de média. 71
5.17 Densidade e entropia dos k agrupamentos selecionados pelo RM-CRAG utilizando as diferentes medidas: GANMI, MVNMI e QNMI. 72
5.18 GANMI dos agrupamentos produzidos pelo RM-CRAG e por ITS aplicado em `strut-a`. 74

Lista de Tabelas

2.1 Probabilidade conjunta. 23
2.2 Representação dos agrupamentos das Figuras 1.3b, 1.3c, 1.3d e 1.3e pela perspectiva dos elementos. 24
2.3 Representação dos agrupamentos das Figuras 1.3b, 1.3c, 1.3d e 1.3e pela perspectiva dos rótulos. 24

4.1 Similaridade para o atributo a_1 entre os vértices (u_i,u_j) à distância δ, $1 < \delta \leq d$. 38
4.2 Representação de valores no gráfico de contorno da média, média + variância e média dos quadrados. 41
4.3 Tipo de cada elemento na fila de prioridades. 44
4.4 Representação dos valores das médias para o conjunto de agrupamentos da Figura 4.5. 46

5.1 Propriedades do conjunto de dados MQD500b. 51
5.2 Propriedades do conjunto de dados DBLP5000. 53
5.3 Tempo para a geração dos agrupamentos `Atr`, `Strut`, `CRAG` e da distância 2 para os conjuntos de dados `MQD500b`, `MQD500c` e `DBLP3000`. 54
5.4 Agrupamentos identificados pela abordagem de força bruta para $m = 5$. 60
5.5 Teste da hipótese nula. 74

Lista de Símbolos

A	Matriz de adjacência, p. 16
D	Densidade de um agrupamento, p. 52
R	Matriz de grau, p. 17
S	Entropia de um agrupamento, p. 52
Λ	Conjunto de atributos associados aos vértices, p. 20
δ	$1 < \delta \leq d$, p. 35
$\mathcal{C}$	Conjunto de agrupamentos, p. 22
$\mathcal{D}$	Conjunto de dados, p. 12
$\mathcal{F}$	conjunto não-vazio de Λ, p. 38
$\mathcal{N}_{u_i}$	conjunto de todos os vértices à distância δ de u_i, p. 36
$\mathcal{P}$	Subconjunto não-vazio de Λ, p. 38
$\mathcal{Q}$	Fila de prioridade, p. 43
μ	média, p. 39
σ^2	variância, p. 39
$attrSet$	Conjunto de atributos, p. 35
c	Um agrupamento de $\mathcal{C}$, p. 38
$cent_i$	i-ésimo centróide, p. 13
d	Distância dos vizinhos dos vértices, p. 35
d_i	Grau do vértice u_i, p. 17
g_i	i-ésimo grupo, p. 13

m	Número de grupos de um agrupamento, p. 35
q^2	média dos quadrados, p. 39
s	Limiar de similaridade para adicionar arestas artificiais, p. 36
u_q	Um vértice, p. 16
E	Conjunto de arestas, p. 16
I(X;Y)	Informação mútua entre as variáveis X e Y, p. 21
L	Matriz Laplaciana, p. 18
V	Conjunto de vértices, p. 16
k	Número de soluções de agrupamento, p. 43
m	Número de grupos, p. 12
n	Número de objetos, p. 11

Lista de Abreviaturas

ANMI	Average Normalized Mutual Information, p. 24
BFS	*Brief First Search*, p. 31
CRAG	Clustering in Attributed Graphs, p. 35
CkC	*Connected k Centers*, p. 31
FB	Abordagem de Força Bruta, p. 57
GANMI	Global Average Normalized Mutual Information, p. 41
GVNMI	Global Variance Normalized Mutual Information, p. 41
ITS	Iterative Tabu Search, p. 48
KS	Teste de normalidade Kolmogorov-Smirnov, p. 73
M-CRAG	Multiple Clusterings in Attributed Graphs, p. 37
MDP	Maximum Diversity Problem, p. 45
MI	Informação Mútua, p. 21
MPT	Moderna Teoria de Portifólio, p. 41
MQD	Meu Querido Diário, p. 50
NMI	Informação Mútua Normalizada, p. 22
QNMI	sQuared Normalized Mutual Information, p. 42
RM-CRAG	Ranked Multiple Clustering Algorithm, p. 42
SSE	Sum of Squared Error, p. 13
TS	Tabu Search, p. 48

Capítulo 1

Introdução

O grande número de redes sociais *online* que surgiram nos últimos anos apresenta um substrato bastante relevante para o estudo dos indivíduos que as compõem. Esses indivíduos possuem diferentes características (e.g., idade, gênero, classe social) e constroem uma complexa estrutura de relações nessas redes, nas quais os limites geográficos não estabelecem fronteiras. O que vemos através dessas relações é um alto tráfego de informações e conhecimento, além de um rápido acesso aos eventos de última hora. O número crescente de usuários nessas redes vem atraindo interesses econômicos entre os investidores e pesquisadores (HEIDEMANN *et al.*, 2012). As empresas perceberam a importância desses veículos em campanhas de marketing, dado que em 2011 foram gastos $3,08$ bilhões de dólares pelas empresas americanas apenas em propaganda nas redes sociais (WANG *et al.*, 2012). Os membros das equipes de campanhas publicitárias também perceberam as vantagens que possuem as campanhas publicitárias políticas em redes sociais (VERGEER *et al.*, 2013; WILLIAMS e GIRISH, 2012).

O sociólogo Zygmunt Bauman destaca que os seres humanos estão dando mais importância a relacionamentos em "rede" (BAUMAN e MEDEIROS, 2004), o que podemos observar com os altos números de amigos em redes sociais como Facebook[1]. Bauman concebe uma crítica, observando que a internet ajuda a enfraquecer e tornar mais superficiais as relações construídas na vida "*offline*" (BAUMAN e MEDEIROS, 2004). As argumentações decorrem da análise de que embora o alto número de conexões de amizade em uma rede seja "sólido", também pode ser bastante volúvel, visto que um novo amigo pode ser facilmente desconectado.

Essa facilidade de se conectar e desconectar a outros indivíduos nas redes sociais configuram uma estrutura complexa e bastante dinâmica, uma vez que, a cada instante, novas conexões podem ser feitas ou desfeitas. A análise dessas estruturas pode conduzir a estudos em diversas áreas, como na detecção de usuários

[1] http://www.facebook.com

confiáveis (SHEN *et al.*, 2012) ou influentes (SATHANUR *et al.*, 2013) bem como o estudo de terrorismo e violência política (PERLIGER e PEDAHZUR, 2011), dentre outros. Entretanto, para que isso possa ser feito, é necessária a compreensão das características e estrutura das redes sociais.

Uma rede é, em sua mais simples forma, uma coleção de pontos ligados em pares por linhas (NEWMAN, 2010). Há diversos tipos de redes, entretanto, no presente trabalho, iremos focar em redes sociais. Redes sociais são redes nas quais os vértices representam indivíduos, ou, algumas vezes, grupos de indivíduos e as ligações entre esses vértices, as arestas, representam alguma forma de interação social entre esses indivíduos, como a relação de amizade (NEWMAN, 2010). A Figura 1.1 ilustra uma pequena rede social com relacionamento de amizade entre os usuários, representada por um grafo. Nesse grafo, os vértices representam os usuários e as arestas retratam a relação de amizade entre os usuários. O usuário u_1 possui apenas um amigo (u_7), o usuário u_2 possui dois amigos (u_3 e u_7) e assim por diante.

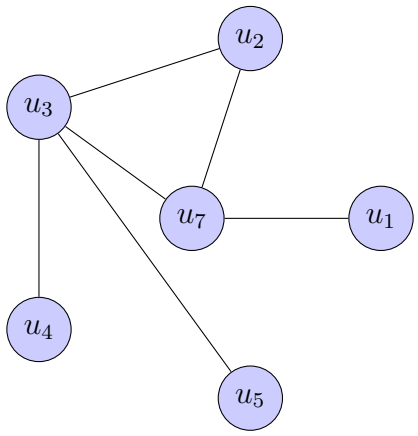


Figura 1.1: Exemplo de uma rede social representada como um grafo. Os vértices representam os usuários. As arestas entre os vértices representam a relação de amizade.

A Figura 1.1 caracteriza a relação de amizade entre os indivíduos da rede. Entretanto, em redes sociais típicas existem diversos tipos de relações entre os indivíduos, como as relações de afetividade, de negócio, relações econômicas, políticas, etc. Observando esse mesmo grafo ilustrado na Figura 1.1 sob a ótica do contato físico, pode-se observar uma estrutura distinta, conforme ilustra a Figura 1.2.

Com isso, os mesmos vértices são representados com arestas diferentes. As relações, agora representadas pelo contato físico, originam um grafo com uma nova estrutura. Pode-se perceber que o vértice u_1 possui relação de amizade com u_7 (Figura 1.1), mas não houve contato físico entre ambos (Figura 1.2). Com esse novo grafo seria possível estudar, por exemplo, a disseminação de um surto epidêmico.

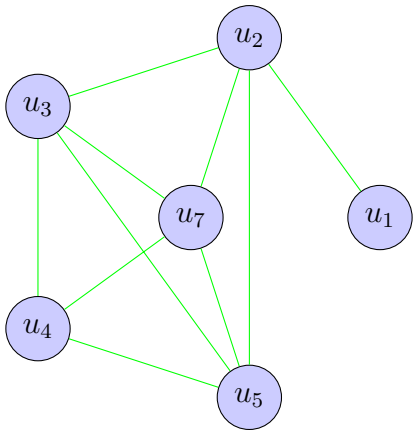


Figura 1.2: Exemplo de usuários em uma rede social em que as arestas representam o contato físico. As arestas de contato físico estão representadas na cor verde.

É importante destacar que além das múltiplas possibilidades de representar os relacionamentos em redes sociais, podem existir diferentes perspectivas de representação dos usuários. Assim, um usuário poderia ser representado por meio de seu perfil psicológico ou emotivo para decifrar padrões de comportamento. Esses padrões permitiriam, por exemplo, auxiliar em tomadas de decisões ou na escolha de um júri, como descreve Jo-Ellan Dimitrius em seu livro *Reading people* (DIMITRIUS e MAZZARELLA, 1998). A autora estudou durante anos os perfis de personalidade de testemunhas, advogados e juízes para a escolha de júris, utilizando os padrões encontrados para escolher os jurados "corretos" para os julgamentos.

Seguindo essa linha, Pennebaker descreve que diferentes padrões de palavras funcionais (*function words*) revelam importantes aspectos da personalidade dos indivíduos e a forma com que eles pensam (PENNEBAKER, 2013). Estudos recentes indicaram ser possível prever a personalidade dos usuários do Twitter[2] utilizando as abordagens de Pennebaker (GOLBECK *et al.*, 2011). Prosseguindo nesse raciocínio, alguns estudos mostram a existência de relação entre o gosto musical e a personalidade (RENTFROW e GOSLING, 2003), assim como entre o comportamento de compras e a personalidade de consumidores (WHELAN e DAVIES, 2006).

Essas diferentes perspectivas podem ser reveladas com técnicas provenientes da área de reconhecimento de padrões (*pattern recognition*). A tarefa de agrupamento, por exemplo, é amplamente utilizada nessa área (DU, 2010). Agrupar os indivíduos através de suas características e/ou relações pode permitir a revelação de subgrupos que compartilham os mesmos padrões. Pode-se, então, evidenciar subgrupos com interesses em comum (e.g., esportes, novelas, filmes) ou com características em comum (e.g., personalidade, perfil emotivo). Nesse domínio, (WASSERMAN e

[2]http://www.twitter.com

FAUST, 1994) relatam como um problema interessante o agrupamento dos usuários de uma rede social para a descoberta de padrões associados ao comportamento, características, interesses, etc.

A tarefa de agrupamento é de suma importância na área de análise de dados exploratórios e consiste em particionar um conjunto de objetos em grupos. Cada um desses grupos contém objetos que são similares entre si e dissimilares a objetos de outros grupos. Essa tarefa torna mais eficiente o entendimento de grandes conjuntos de dados. Existem diversos algoritmos de agrupamento, como os particionais, hierárquicos e baseados em grafos. Cada um desses explora os dados de acordo com um modelo de dados particular. Esse trabalho se insere no contexto de agrupamentos em grafos.

Os grafos são modelos estruturais frequentemente utilizados para modelar estruturas sociais. Um exemplo prototípico de um grafo é uma rede social (NETTLETON, 2013), em que a estrutura topológica representa relacionamentos entre os indivíduos e as propriedades dos vértices descrevem características e papeis de cada indivíduo (ZHOU *et al.*, 2009). Aplicações típicas de agrupamento em grafos incluem detecção de comunidades, partição de grafos, predição de *links* e agrupamento de documentos (NEWMAN, 2010; SRIVASTAVA *et al.*, 2013).

Nesse trabalho é feita a distinção entre algoritmos de agrupamento baseados em atributos, que exploram a estrutura relacional dos dados e os algoritmos baseados na topologia dos grafos, que consistem em encontrar grupos densamente conectados de vértices (SCHAEFFER, 2007). A maioria dos algoritmos de agrupamento em grafos considera apenas a estrutura topológica do grafo durante o processo de agrupamento, ignorando os atributos associados aos vértices. Entretanto, as propriedades dos vértices são relevantes em muitas aplicações, como as redes sociais (ZHOU *et al.*, 2009). Os grafos em que cada vértice possui dados associados (i.e., uma n-tupla para guardar propriedades de um vértice) são denominados *grafos com atributos* (ZHOU *et al.*, 2009). Na presente tese, a existência dos atributos dos vértices foi levada em consideração durante o processo de agrupamento.

Embora nas redes sociais os indivíduos possam ter várias conexões de relacionamento, pode-se verificar que muitas dessas relações tendem a ser homofílicas (MCPHERSON *et al.*, 2001). Homofilia é a tendencia que indivíduos possuem de se associar com outros indivíduos similares em relação a alguma característica (NEWMAN, 2010). Isso é muito atrativo do ponto de vista do marketing, visto que indivíduos similares tendem a ter interesses e comportamentos similares em padrões de compra (GUPTA *et al.*, 2013). Além disso, dependendo do conjunto de propriedades utilizado para representar cada usuário na rede social, diferentes configurações de grupos de usuários podem ser obtidas, como já observado no contexto de busca em redes sociais (WATTS *et al.*, 2002).

Por outro lado, muitos algoritmos de agrupamento em grafo identificam apenas uma partição dos dados (SCHAEFFER, 2007). Entretanto, uma única partição pode não prover conhecimento suficiente sobre os dados subjacentes (MÜLLER *et al.*, 2012). Nesse caso, seria interessante examinar soluções alternativas de agrupamentos. Isso motiva o presente trabalho, cujo objetivo é explorar o espaço de busca formado por diversos subconjuntos de propriedades (atributos) de cada vértice do grafo para produzir novas soluções de agrupamento.

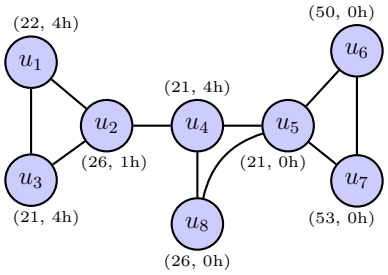
1.1 Exemplo motivador

Recentemente as redes sociais têm se tornado cada vez mais importantes na área de marketing (IACOBUCCI, 1996). Nessa área, a análise de agrupamentos é utilizada nas pesquisas para segmentar o mercado (*marketing segmentation*) e determinar os mercados-alvo (*target marketing*) (GAN *et al.*, 2007a). Segundo (MOOI e SARSTEDT, 2011), a segmentação de marketing é uma das atividades mais fundamentais na área de marketing. Isso é conseqüência das empresas não conseguirem contatar todos os consumidores potenciais e, portanto, necessitam dividir o mercado em grupos (segmentos) de consumidores ou clientes com necessidades e vontades similares (MOOI e SARSTEDT, 2011). Além disso, os consumidores são céticos com relação às empresas, mas acreditam em seus amigos (WEBSTER e MORRISON, 2004).

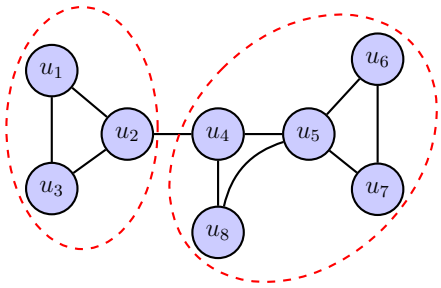
Segundo GE *et al.* (2008), um grupo de consumidores com atributos similares possui mais chances de pensarem da mesma maneira com a propagação do boca-a-boca. Ainda segundo o autor, em alguns casos, as relações sociais podem ser vitais na formação de segmentos e as intenções de compra podem depender da interação entre os clientes, como exemplo, clientes cautelosos que pretendem realizar uma cirurgia estética arriscada. Nesse contexto, a Figura 1.3 ilustra um exemplo de uma rede social a ser explorada por analistas de marketing. Essa rede social é representada como um grafo com atributos e ilustra a geração de agrupamentos alternativos através da inclusão de arestas artificiais entre usuários similares a distância 2. Essa abordagem gera os agrupamentos alternativos, onde cada vértice possui uma 2-tupla (i,h): idade e tempo gasto com *hobbies* em uma semana.

A produção de um agrupamento utilizando apenas a estrutura topológica (i.e., ignorando os atributos dos vértices) apresentaria os grupos $\langle[u_1,\ u_2,\ u_3],\ [u_4,\ u_5,\ u_6,\ u_7,\ u_8]\rangle$, conforme apresentado na Figura 1.3b. Um agrupamento baseado na estrutura topológica levando em consideração os atributos de *hobbies*, resultaria também no clustering $\langle[u_1, u_2, u_3], [u_4, u_5, u_6, u_7, u_8]\rangle$, demonstrado na Figura 1.3c. Além disso, um agrupamento levando em conta a estrutura topológica e o atributo

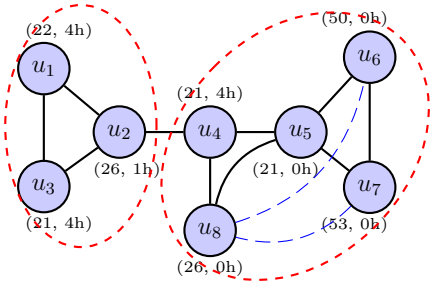
idade resultaria em $\langle[u_1,\ u_2,\ u_3,\ u_4,\ u_8],\ [u_5,\ u_6,\ u_7]\rangle$, como ilustrado na Figura 1.3d. Por fim, a Figura 1.3e apresenta um agrupamento realizado com a estrutura topológica e ambas as informações relacionais, o que resulta no agrupamento $\langle[u_1,\ u_2,\ u_3,u_4],\ [u_5,\ u_6,\ u_7,\ u_8]\rangle$.



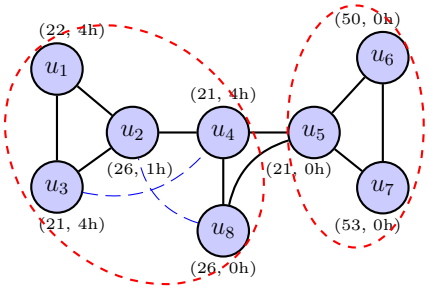
(a) Grafo com atributos original



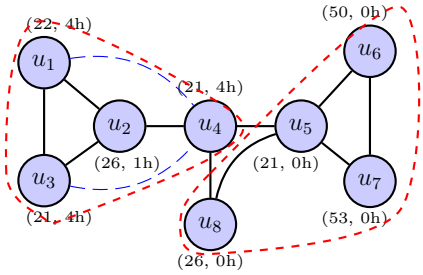
(b) Agrupamento apenas pela estrutura



(c) Agrupamento pela estrutura + atributo *hobbies*



(d) Agrupamento pela estrutura + atributo idade



(e) Agrupamento pela estrutura + atributos idade e *hobbies*

Figura 1.3: Exemplo ilustrando múltiplas soluções de agrupamento em um grafo com atributos.

Um exemplo de aplicação prática poderia advir da técnica de marketing viral. Essa técnica tem como meta atingir os indivíduos influentes de uma rede (KEMPE *et al.*, 2003b). Assim, é interessante que o analista de marketing possa selecionar os indivíduos mais adequedos. Como exemplo, podemos utilizar uma empresa de varejo que deseja enviar algumas amostras de produtos aos usuários de uma determinada rede social. Considerando os fenômenos que ocorrem em redes sociais, assim como a homofilia (MCPHERSON *et al.*, 2001; NEWMAN, 2010), é esperado que amigos de indivíduos influentes irão experimentar os produtos e recomendá-los a seus amigos

(KEMPE *et al.*, 2003a). Com isso, a chave para o sucesso é a identificação do conjunto adequado de indivíduos, de forma que as campanhas de marketing sejam direcionadas aos que apresentam comportamento similar com relação a alguma propriedade (BHATT *et al.*, 2010).

Suponhamos que o analista de marketing deseja oferecer amostras grátis de um produto a um usuário influente. Qual seria o melhor usuário a escolher? O analista poderia selecionar aleatoriamente, mas isso poderia resultar na escolha de um indivíduo com pouca influência no grafo correspondente à rede social. Assumindo que vértices com alto grau influenciam mais vizinhos (STONEDAHL *et al.*, 2010; WASSERMAN e FAUST, 1994) e considerando o exemplo da Figura 1.3b, o analista de marketing poderia escolher o vértice u_5 visto que é o vértice com maior grau. Entretanto, se o analista desejar direcionar o marketing para indivíduos que possuem *hobbies*, essa poderia ser uma má escolha, visto que o usuário representado pelo vértice u_5 não possui *hobbies*.

Com isso, podemos observar que a escolha pode ser obtida combinando os atributos dos indivíduos com as características sociais (e.g. a estrutura da rede local). Essa combinação pode levar a resultados melhores ao identificar futuros adotantes de algum produto (BHATT *et al.*, 2010). Desse modo, o desafio é prover ao analista formas alternativas de agrupamentos dos usuários, para que este possa fazer as melhores escolhas.

Um outro cenário é apresentado em redes de colaboração científica, nas quais os vértices podem representar os autores e os atributos dos vértices podem representar as áreas de interesse, número de publicações, entre outros. A estrutura pode caracterizar a relação de co-autoria, ou seja, existe uma aresta entre autores que publicaram conjuntamente. Agrupar os autores considerando os atributos e as relações pode ser útil para os pesquisadores identificarem os autores mais influentes por área, recomendando novas colaborações, dentre outros (GE *et al.*, 2008).

1.2 Definição do problema

O paradigma descrito na seção anterior pode ser especificado como um problema para se obter diferentes agrupamentos em grafos. Diversas propriedades estruturais podem ser calculadas nos grafos, como a *importância* de indivíduos na rede, o *grau*, dentre outras. Entretanto, o problema vai além das propriedades apresentadas pela estrutura dos grafos. As propriedades dos vértices (atributos) desempenham papel fundamental no estudo de algumas relações existentes entre os indivíduos. Diversos trabalhos evidenciam correlação entre os atributos e os indivíduos conectados (MCPHERSON *et al.*, 2001). A causa dessa correlação pode ser atribuída a *influência social* (MASON *et al.*, 2007), que é a tendência que os indivíduos

conectados possuem de serem influenciados pelos seus amigos e *homofilia*, em que as arestas são criadas com base na similaridade dos indivíduos. Um exemplo prototípico de influência social é apresentado quando valores nos atributos de um indivíduo são alterados para corresponder aos de um amigo. Por outro lado, a homofilia apresenta relações de amizade formadas com base em similaridade entre os atributos.

Nesse contexto, pode-se estabelecer que, em alguns casos, é necessário que se observe as informações da estrutura topológica e dos atributos dos vértices de um grafo. O enfoque desse trabalho se insere no paradigma dos grafos com atributos, mais especificamente em agrupamento de vértices em grafos com atributos. Portanto, além da estrutura topológica dos grafos, são considerados os atributos associados a cada vértice.

A hipótese geral desse trabalho é que dado um grafo com atributos, a combinação da estrutura topológica com os atributos dos vértices possibilita a produção de agrupamentos múltiplos não-redundantes. Esses agrupamentos devem apresentar grupos com vértices densamente conectados e homogêneos de acordo com alguma faceta do conjunto de dados de entrada. O objetivo de considerar a redundância é a obtenção de agrupamentos novos, ou seja, não-redundantes com respeito a alguma estrutura (BASU *et al.*, 2008).

As medidas de avaliação de redundância podem auxiliar na detecção de novos agrupamentos e algumas delas são empregadas no decorrer desse trabalho. Entretanto, existe uma lacuna no que diz respeito à medidas de avaliação na área de agrupamentos múltiplos (MÜLLER *et al.*, 2012). Com o intuito de contribuir nessa lacuna, nesse trabalho são propostas três medidas para avaliar a redundância em agrupamentos múltiplos. Essas medidas formam o núcleo de um dos algoritmos criados nessa tese.

1.3 Contribuições

Esta tese está contextualizada no âmbito da área de agrupamentos múltiplos em grafos com atributos, contribuindo nessa área por apresentar:

- a extensão do problema da produção de agrupamentos múltiplos em grafos para grafos com atributos.
- um algoritmo para gerar apenas um agrupamento combinando um atributo dos vértices com a estrutura topológica do grafo (CRAG).
- um algoritmo para gerar diversos agrupamentos combinando um atributo dos vértices com a estrutura topológica do grafo (M-CRAG).

- um algoritmo para selecionar os *top-k* agrupamentos não-redundantes (RM-CRAG).
- três medidas para avaliação de redundância em agrupamentos múltiplos (GANMI, MVNMI e QNMI).

Algumas dessas contribuições se encontram desmembradas nas seguintes publicações realizadas durante os estudos que compreendem o presente trabalho:

- GUEDES, G. P. ; BEZERRA, E. ; XEXÉO, G. B. . Multi-view Clustering in a Social Network. In: II Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), 2013, Maceió. Anais do 33o. Congresso da SBC, 2013. Porto Alegre: SBC, 2013.
- GUEDES. G. P. ; BEZERRA, E. ; OGASAWARA, E. ; XEXEO, G. B. MAM: Método para Agrupamentos Múltiplos em Redes Sociais Online Baseado em Emoções, Personalidades e Textos. iSys: Revista Brasileira de Sistemas de Informação, v. 7, p. 38-55, 2014.
- GUEDES, G. P. ; OGASAWARA, E. ; BEZERRA, E.; XEXEO, G. B. Exploring Multiple Clusterings in Attributed Graphs. In: ACM Symposium on Applied Computing, 2015, Salamanca, Spain. SAC 15, 2015. p. 915-918.
- GUEDES, G. P. ; OGASAWARA, E. ; BEZERRA, E.; XEXEO, G. B. Discovering top-k Non-Redundant Clusterings in Attributed Graphs. In: Neurocomputing (aceito).

1.4 Organização da tese

Essa tese é organizada em mais cinco capítulos. O Capítulo 2 fundamenta a teoria da presente tese, descrevendo, inicialmente, os conceitos de agrupamentos, agrupamentos múltiplos, agrupamentos em grafos e em seguida os conceitos relacionados a agrupamentos em grafos com atributos. Nesse capítulo também são descritas algumas medidas de comparação entre agrupamentos. O Capítulo 3 discute alguns trabalhos relacionados. O Capítulo 4 descreve as contribuições dessa tese. Inicialmente apresenta um novo algoritmo para combinar a estrutura topológica com os atributos dos vértices do grafo e, assim, gerar uma solução de agrupamento. Em seguida, especifica o algoritmo criado para gerar múltiplas soluções de agrupamentos. Esse capítulo também apresenta o algoritmo de seleção dos *top-k* agrupamentos não-redundantes, além de descrever as medidas propostas para avaliar a não-redundância dos agrupamentos selecionados. O Capítulo 5

apresenta a avaliação experimental das abordagens propostas em três conjuntos de dados. Por fim, o capítulo 6 discute os resultados alcançados e apresenta as conclusões, limitações e direções para trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Este capítulo aborda alguns conceitos fundamentais para o compreendimento da presente tese. Esses conceitos permeiam o conteúdo dos capítulos seguintes e estão divididos em 5 seções. A Seção 2.1 discute alguns conceitos referentes à tarefa de agrupamento. Na Seção 2.2 é apresentada a motivação para produção de agrupamentos múltiplos, além de alguns conceitos provenientes dessa área. Na Seção 2.3 é abordada a tarefa de agrupamento em grafos. A Seção 2.4 discute a medida utilizada nessa tese para a comparação entre dois agrupamentos. A Seção 2.5 descreve a medida existente na literatura para quantificar a informação compartilhada entre um agrupamento e um conjunto de agrupamentos.

2.1 Agrupamentos

A *mineração de dados* é o processo de descoberta de padrões e conhecimento a partir de grandes quantidades de dados (HAN *et al.*, 2011). Combina ferramentas de diferentes áreas, como aprendizagem de máquina, estatística, banco de dados, sistemas especialistas e visualização de dados (ANDA, 1999). Segundo DANIEL T. LAROSE (2014), esse processo pode ser dividido em algumas tarefas, dentre elas, a tarefa de classificação e a tarefa de agrupamento.

A tarefa de agrupamento (*clustering* – em inglês) é altamente popular na área de mineração de dados, muitas vezes utilizada como um passo inicial na análise exploratória de conjuntos de dados complexos. É intensamente estudada devido à sua aplicabilidade em diversas áreas de conhecimento (e.g., marketing, engenharia, medicina) e é apresentada como uma abordagem não-supervisionada, pois não se sabe a priori a que grupo cada objeto pertence.

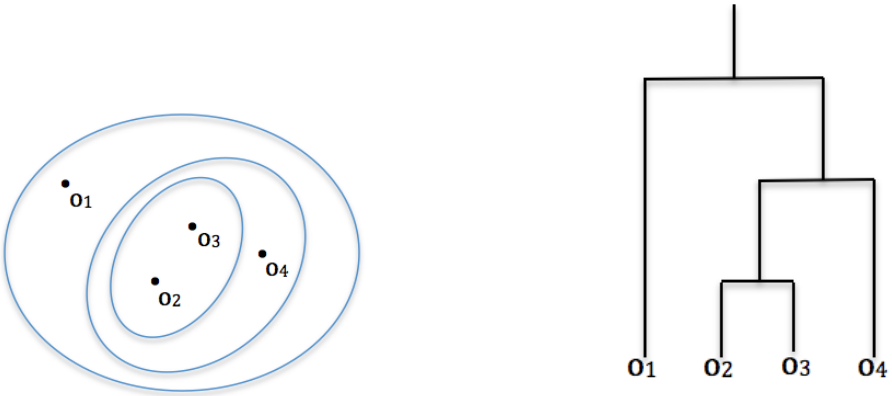
Muitos algoritmos de agrupamento podem ser considerados como procedimentos orientados por uma função objetivo. Tipicamente, o espaço de busca é bastante grande, posto que cada estado desse espaço corresponde a uma possível partição do conjunto de objetos. O procedimento de otimização procura encontrar uma partição

na qual os objetos de cada grupo sejam semelhantes e objetos em grupos distintos sejam dissimilares (HAN *et al.*, 2011).

Em geral, os algoritmos de agrupamento podem ser classificados em duas categorias amplas: algoritmos de agrupamento hierárquico (*hierarchical clustering* – em inglês) e algoritmos de agrupamento particional (*partitional clustering* – em inglês) (GAN *et al.*, 2007b). Com relação aos algoritmos de agrupamento hierárquico, os grupos de objetos são apresentados de forma aninhada, resultando em uma hierarquia de grupos, organizados como uma "árvore hierárquica" ou dendograma. O dendograma é um meio prático para apresentar um padrão de agrupamento.

Os algoritmos hierárquicos são subdivididos em duas abordagens: aglomerativos (*bottom-up* – em inglês) e divisivos (*top-down* – em inglês). Esses algoritmos têm seu funcionamento bastante semelhante. Os algoritmos divisivos consideram que, inicialmente, todo o conjunto de dados está em um grupo e, a cada iteração, particiona esse grupo em grupos menores.

A abordagem aglomerativa é iniciada com todos os indivíduos segmentados. Em seguida, os objetos são agrupados em pares até chegar a uma única raiz. A Figura 2.1 (a) ilustra um agrupamento hierárquico aglomerativo e a Figura 2.1 (b) ilustra a representação de um dendograma.



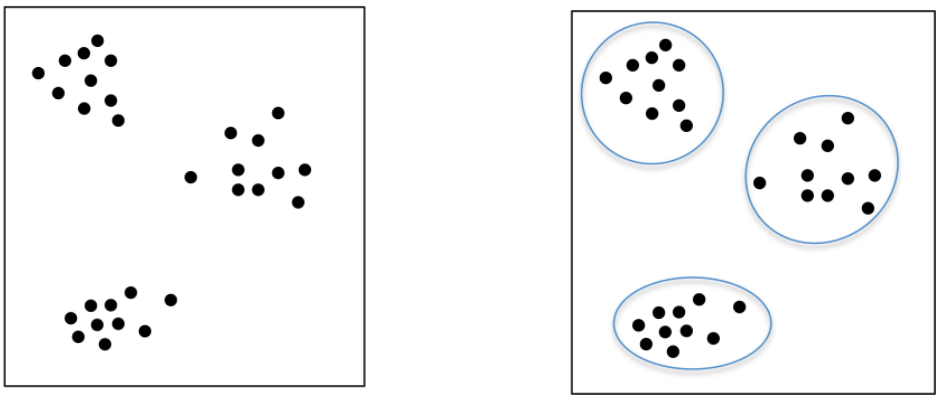
(a) Agrupamento hierárquico aglomerativo. (b) Dendograma.

Figura 2.1: Agrupamento hierárquico aglomerativo

Os algoritmos hierárquicos possuem algumas limitações, dentre elas: i) uma vez que a decisão de combinar dois grupos tenha sido tomada, não pode ser desfeita. ii) nenhuma função objetivo é minimizada diretamente. Além disso, podem ser utilizados apenas em conjuntos de dados relativamente pequenos (RAJARAMAN e ULLMAN, 2011).

Em contrapartida, os algoritmos de agrupamento particional dividem os dados diretamente em um número de grupos, não havendo uma estrutura hierárquica entre

os grupos (XU e WUNSCH, 2005). Assim, dado um número n de objetos, os dados são divididos em m grupos, considerando que cada objeto se encontre em exatamente um grupo (HAN *et al.*, 2011). A Figura 2.2b apresenta uma possível solução de agrupamento para os objetos não-agrupados apresentados na Figura 2.2a.



(a) Objetos não-agrupados. (b) Agrupamento particional.

Figura 2.2: Agrupamento particional

Os algoritmos de agrupamento particional podem ser aplicados em grandes quantidades de objetos e seus conceitos serão utilizados no decorrer dessa tese. O algoritmo *k-means* (MACQUEEN, 1967) é um dos algoritmos particionais mais usados e mais conhecidos (HAN *et al.*, 2011). É amplamente utilizado devido a sua simplicidade e competência (LAURENT *et al.*, 2014). O *k-means* é iniciado com a escolha dos centróides (centros dos grupos), que são pontos no espaço de objetos que representam uma posição média em cada grupo (SAJJA e AKERKAR, 2012). O Algoritmo 1 descreve os passos do *k-means*.

Algoritmo 1: Algoritmo *k-means*($\mathcal{D}$, m)

Input:

- $\mathcal{D}$ = um conjunto de dados contendo n objetos.
- m = número de grupos.

Output: Um conjunto de m grupos.

1: Escolher aleatoriamente m objetos de $\mathcal{D}$ como os centróides iniciais de cada grupo.
2: **repeat**
3: Calcular a distância entre cada objeto e os centróides, adicionando o objeto ao grupo que possuir menor distância.
4: Atualizar a média dos grupos, ou seja, calcular a média dos valores dos objetos para cada grupo (centróides).
5: **until** não haja mudança

O objetivo do *k-means* é buscar minimizar, de forma iterativa, a distância entre os n objetos e os m centros. Inicialmente, no Passo 1, o algoritmo escolhe

aleatoriamente m objetos para serem os centróides dos grupos. No passo 3, cada um dos objetos restantes é associado ao grupo ao qual mais se assemelha, baseado na distância euclidiana entre o objeto e o centróide dos grupos. A distância euclidiana entre dois pontos é calculada conforme a Eq. 2.1. Nessa equação, x_i e y_i são pontos no espaço Euclidiano e d é o número de dimensões.

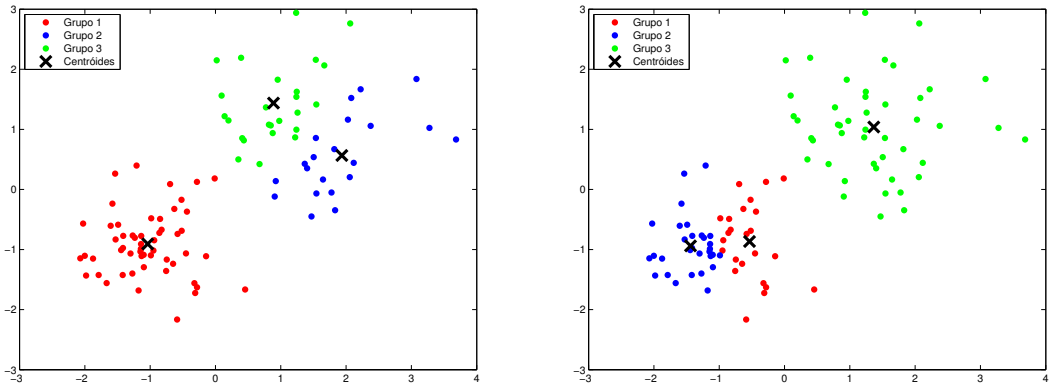
$$euc = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2} \tag{2.1}$$

Em seguida, no Passo 4, o algoritmo atualiza os valores dos centróides de cada grupo utilizando a média aritmética dos objetos assinalados a cada grupo. O *k-means* determina um número m de partições tentando minimizar uma função objetivo. A função objetivo mais comumente utilizada é a soma dos erros quadráticos (*Sum of Squared Error* – em inglês)(SSE) , conforme apresentado na Eq. 2.2. Nessa equação, m é o número de grupos, x é um ponto no grupo g_i e $cent_i$ é o centróide do grupo g_i.

$$SSE = \sum_{i=1}^{m}\sum_{x \in g_i} dist^2(cent_i, x) \tag{2.2}$$

Considerando que o *k-means* tradicional escolhe os centróides iniciais aleatoriamente, os resultados dos agrupamentos podem ter variações em sua qualidade (ISHIZUKA e SATTER, 2003). A Figura 2.3 ilustra um conjunto de objetos de tamanho $n = 100$. Deseja-se encontrar um número de grupos $m = 3$. Com o mesmo conjunto de objetos e sementes aleatórias, pode-se notar duas soluções distintas de agrupamento: na Figura 2.3a, a solução apresenta 1 grupo à esquerda e 2 à direita. Já na Figura 2.3b, observa-se 2 grupos à esquerda e 1 à direita. Os centróides são representados pela letra "x". Existem diversas estratégias para a inicialização dos centróides no algoritmo *k-means*, entretanto, no presente estudo, é utilizada a inicialização aleatória.

A maioria dos algoritmos de agrupamento apresenta uma única solução de agrupamento (JAIN *et al.*, 1999b). Consequentemente, cada objeto é assinalado a um único grupo. Conforme ilustrado na Figura 2.3, a simples modificação dos parâmetros de entrada (nesse caso, a mudança do centróides iniciais) pode produzir diferentes soluções de agrupamento. Com isso, é possível observar que, dado um conjunto de dados, múltiplas soluções de agrupamento podem ser produzidas. A interpretação dessas múltiplas soluções pode permitir que um analista evidencie características relevantes sobre o conjunto de dados, visto que dados multi-facetados são relativamente comuns. Essa perspectiva realça a necessidade de algoritmos capazes de produzir agrupamentos múltiplos.



(a) Uma possível solução de agrupamento. (b) Outra possível solução de agrupamento.

Figura 2.3: Diferentes soluções de agrupamento utilizando o *k-means* com escolha aleatória dos centróides iniciais.

2.2 Agrupamentos múltiplos

A área de agrupamentos múltiplos tem recebido bastante atenção nos últimos anos. Há diversas aplicações, como por exemplo na área de segmentação de clientes (marketing), onde os clientes podem ser agrupados de diferentes formas considerando subconjuntos dos atributos (MÜLLER *et al.*, 2012). Também é relevante em conjunto de dados de imagens de faces, propiciando, por exemplo, o agrupamento dos indivíduos com base em sua pose ou identidade (GUAN *et al.*, 2010). Um outro exemplo pode ser dado na área de análise de textos, em que documentos representam diferentes tópicos que são bem conhecidos, mas resultados alternativos podem ser desejados (MÜLLER *et al.*, 2012). O exemplo apresentado pela Figura 2.4 ilustra um conjunto de dados com 4 objetos: um círculo vermelho e um azul além de um quadrado vermelho e um azul.

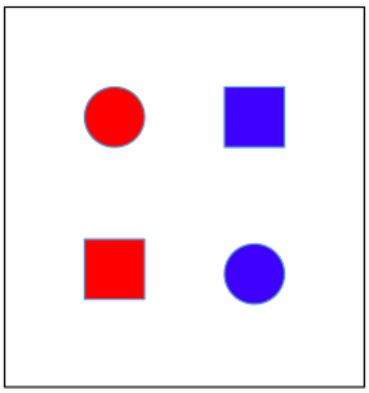
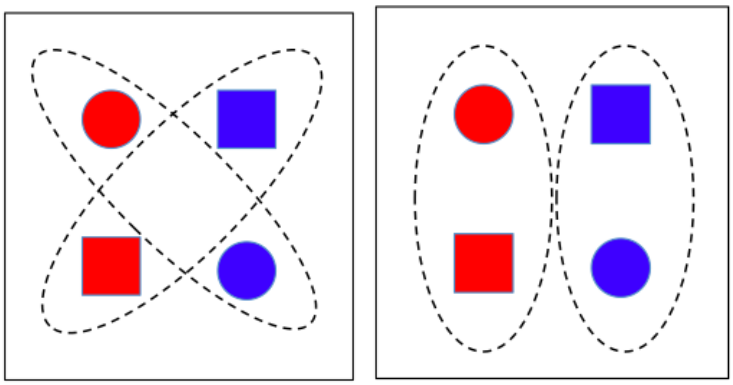


Figura 2.4: Representação de quatro objetos a serem agrupados. Dois quadrados (um azul e um vermelho) e dois círculos (um azul e um vermelho).

Esses objetos podem ser agrupados utilizando algumas perspectivas. Inicialmente, pode-se notar claramente duas soluções de agrupamento possíveis: uma que leva em consideração a forma dos elementos, conforme apresenta a Figura 2.5a e outra solução que considera a cor dos elementos conforme ilustra a Figura 2.5b.



(a) Agrupamento por forma. (b) Agrupamento por cor.

Figura 2.5: Soluções de agrupamento

Um algoritmo de agrupamento tradicional poderia convergir para uma das soluções mencionadas ou até mesmo uma diferente, entretanto, só convergiria para uma solução, omitindo as demais. Caso o algoritmo encontrasse a solução da Figura 2.5a, em que os objetos são agrupados pela forma, a informação de que eles poderiam ser agrupados pela cor poderia estar perdida. Neste cenário, nota-se que a produção de agrupamentos múltiplos pode fornecer uma nova percepção sobre os dados, visto que, em geral, o analista não sabe o que quer encontrar, necessita de opções. Com isso, muitas vezes a apresentação de mais de uma solução de agrupamento pode apresentar uma compreensão mais apurada sobre o conjunto de dados.

O exemplo ilustrativo na Figura 2.5 envolve apenas quatro objetos e dois grupos. Contudo, as soluções de agrupamentos com muitos objetos possuem uma complexidade muito mais alta. Por exemplo, para agrupar 30 objetos em 3 grupos, existem $2 * 10^{14}$ possíveis soluções de agrupamento. Entretanto, ao gerar todos os agrupamentos, muitos são redundantes e, dependendo do número de objetos, esse problema se torna computacionalmente intratável. Com isso, a busca por soluções significativas não contempla a geração de todos os agrupamentos possíveis.

Os algoritmos tradicionais de agrupamento, como o *k-means*, apresentam apenas uma solução de agrupamento para um conjunto de dados. É evidente que se pode executar esses algoritmos diversas vezes, obtendo-se diferentes soluções, conforme ilustrado na Figura 2.3. Entretanto, há de se considerar que elas podem ser bastante similares e em alguns conjuntos de dados, soluções praticamente idênticas poderiam ser apresentadas (i.e., alta redundância).

Em função disso, diversos algoritmos de agrupamentos múltiplos foram propostos recentemente (MÜLLER *et al.*, 2010) e o interesse da comunidade científica por essa área vem crescendo (CUI *et al.*, 2010; MÜLLER *et al.*, 2012; NIU *et al.*, 2010, 2014). É explícito que existem alguns desafios na área de agrupamentos múltiplos, dentre eles destaca-se o número de soluções de agrupamento a ser apresentado ao analista. Segundo MÜLLER *et al.* (2012), esse valor pode ser parametrizado ou obtido dentro do algoritmo de agrupamento múltiplo. Entretanto, a apresentação de um elevado número de soluções pode fornecer resultados redundantes. Assim, a eliminação de redundância é um desafio nessa área.

Há algumas medidas utilizadas para calcular a redundância entre dois agrupamentos (WAGNER e WAGNER, 2007). Na Seção 2.4 é apresentada a medida utilizada no presente trabalho. A medida descrita na Seção 2.5 calcula a redundância entre um conjunto de agrupamentos e um agrupamento. Essa medida serviu de inspiração para as medidas propostas nessa tese. No entanto, antes de descrever as medidas de redundância, a Seção 2.3 apresenta uma breve descrição sobre agrupamentos em grafos.

2.3 Agrupamentos em grafos

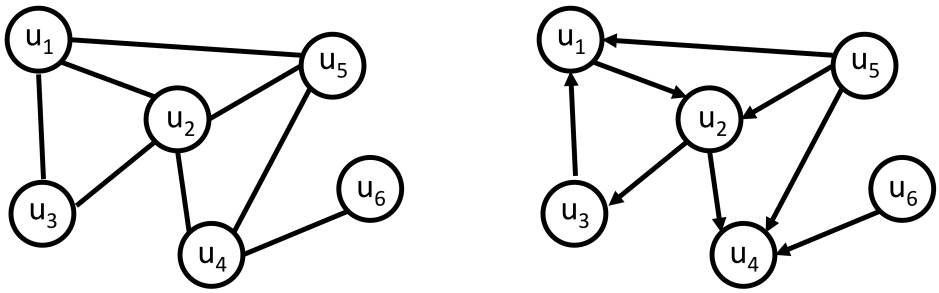
Um grafo é uma estrutura matemática que pode ser utilizada para representar os objetos de uma coleção e seus relacionamentos. Um grafo G pode ser descrito como um par $G(V, E)$, em que $V = \{u_1, u_2, ...u_q\}$ é um conjunto de q vertices e E é um conjunto de arestas que conecta pares de elementos de V.

Em um *grafo simples*, apenas uma aresta pode conectar dois vértices u_1 e u_2. Além disso, um mesmo vértice não pode ser origem e destino ao mesmo tempo. Caso ao menos uma dessas duas condições não seja respeitada, esse grafo é denominado um *multigrafo*.

Em um grafo *não-dirigido*, as arestas não possuem orientação, de forma que $\forall\, u_1, u_2 \in V$, se $(u_1, u_2) \in E$ então $(u_2, u_1) \in E$. Caso essa condição não seja respeitada, apresenta-se um grafo *dirigido*. É importante ressaltar que o foco dessa tese ocorre em grafos simples não-dirigidos. As Figuras 2.6a e 2.6b apresentam um grafo simples não-dirigido e um grafo simples dirigido respectivamente.

Os grafos podem ser representados através de suas matrizes de adjacência para análises matemáticas. Considerando um grafo simples não-dirigido, a matriz de adjacência A e seus elementos A_{ij} são representados da seguinte forma:

$$A = \begin{cases} 1, & \text{se existe uma aresta entre os vértices } u_i \text{ e } u_j \\ 0, & \text{caso contrário} \end{cases}$$



(a) Grafo simples não-dirigido sem pesos nas arestas.

(b) Grafo simples dirigido sem pesos nas arestas.

Figura 2.6: Exemplo de grafo não-dirigido e grafo dirigido.

A matriz de adjacência em grafos simples não-dirigidos sem pesos é simétrica ao longo da diagonal principal. A diagonal principal é preenchida por zeros, visto que nesse tipo de grafo uma aresta não pode ter um mesmo vértice como origem e destino simultaneamente (*self-loop*). A Figura 2.7 representa a matriz de adjacência do grafo ilustrado na Figura 2.6a.

$$A = \begin{array}{c} \\ u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{array} \begin{array}{c} \begin{array}{cccccc} u_1 & u_2 & u_3 & u_4 & u_5 & u_6 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

Figura 2.7: Matriz de adjacência para o grafo representado na Figura 2.6a.

Além da matriz de adjacência, pode-se calcular a matriz de grau, que representa um aspecto importante na análise de um grafo. É determinada pelo número de arestas conectadas a cada vértice (NEWMAN, 2010). Essa matriz pode ser calculada a partir da matriz de adjacência conforme a Eq. 2.3, em que d_i é o grau do vértice u_i.

$$d_i = \sum_{j=1}^{n} A_{ij} \tag{2.3}$$

A matriz de grau R pode ser calculada utilizando a Eq. 2.3 e possui a seguinte definiçao:

$$R_{i,j} = \begin{cases} d_i, & se\ i = j \\ 0, & \text{caso contrário} \end{cases}$$

Dada a definição acima, a matriz de grau referente à matriz de adjacência ilustrada na Figura 2.7 é apresentada na Figura 2.8. Pode-se notar que existem valores diferentes de zero apenas na diagonal principal. Essa é uma característica da matriz de grau.

$$R = \begin{array}{c} \\ u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{array} \begin{array}{c} \begin{array}{cccccc} u_1 & u_2 & u_3 & u_4 & u_5 & u_6 \end{array} \\ \begin{pmatrix} 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

Figura 2.8: Matriz de grau para o grafo representado na Figura 2.6a.

As duas matrizes supracitadas são de suma importância na tarefa de produção de agrupamentos no contexto desta tese. A tarefa de agrupamento em grafos consiste em agrupar vértices do grafo em grupos, considerando a estrutura das arestas, de forma que exista muitas arestas dentro de cada grupo e relativamente poucas entre os grupos ((SCHAEFFER, 2007)). Existem diversos algoritmos de agrupamento em grafos, dentre eles podemos destacar o algoritmo de agrupamento espectral.

A teoria espectral tem como um de seus principais objetivos deduzir as principais propriedades e a estrutura de um grafo a partir do seu espectro (CHUNG, 1997). Devido à sua eficiência e desempenho, o algoritmo de agrupamento espectral se tornou um dos mais populares métodos de agrupamento (CHUANG, 2012). A principal componente dos algoritmos de agrupamento espectral é a matriz Laplaciana (L), definida conforme a Eq. 2.4. Nessa equação, R representa a matriz de grau e A é a matriz de adjacência de um grafo.

$$L = R - A \tag{2.4}$$

A Figura 2.9 apresenta a matriz Laplaciana para o grafo ilustrado na Figura 2.6a. Existem algumas variantes das Laplacianas de um grafo, entretanto, essa versão é a utilizada no escopo dessa tese. Para maior aprofundamento, refira-se à VON LUXBURG (2007).

O agrupamento espectral tem o objetivo de particionar as linhas de uma matriz

$$L = \begin{array}{c} \\ u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{array} \begin{array}{c} \begin{array}{cccccc} u_1 & u_2 & u_3 & u_4 & u_5 & u_6 \end{array} \\ \begin{pmatrix} 3 & -1 & -1 & 0 & -1 & 0 \\ -1 & 4 & -1 & -1 & -1 & 0 \\ -1 & -1 & 2 & 0 & 0 & 0 \\ 0 & -1 & 0 & 3 & -1 & -1 \\ -1 & -1 & 0 & -1 & 3 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix} \end{array}$$

Figura 2.9: Matriz de grau para o grafo representado na Figura 2.6a.

de acordo com os seus componentes em alguns vetores associados à matriz. Assim, dada a matriz Laplaciana L associada a um grafo G, cada vértice é representado em um espaço vetorial usando alguns autovetores associados a essa matriz. Esses autovetores são apresentados em uma nova matriz, cujas linhas representam os pontos no espaço vetorial e as colunas representam os autovetores. O Algoritmo 2 descreve os passos necessários para o agrupamento espectral tradicional conforme VON LUXBURG (2007).

Algoritmo 2: Algoritmo tradicional de agrupamento espectral(G, m)

Input:

- $G(V, E)$ = grafo com um conjunto de vértices V e um conjunto de arestas E.
- m = número de grupos.

Output: $\{V_i\}_{i=1}^{m}$, uma partição de V.

1: Computar a matriz de adjacência A.
2: Computar a matriz de grau R.
3: Computar a matriz Laplaciana $L = R - A$.
4: Encontrar os menores autovetores de L e formar a matriz $Z = [z_1...z_k]$.
5: Considerando cada linha de Z como um ponto no espaço vetorial, agrupá-los em m grupos usando o algoritmo *k-means*.
6: Dado um vértice $u_i \in G$, assinalar u_i ao grupo j se e apenas se a linha i de Z é assinalada ao grupo j.

Muitos dos métodos de agrupamento em grafos focam apenas na estrutura topológica do grafo de forma que cada grupo contenha uma estrutura coesa (ZHOU *et al.*, 2009). Entretanto, conforme mencionado no Capítulo 1, em muitas aplicações, as propriedades associadas aos vértices do grafo também são importantes.

Segundo BOTHOREL *et al.* (2015), a taxonomia da tarefa de agrupamento em grafos com atributos pode ser dividida em duas categorias mais amplas: agrupamentos em grafos com atributos nas arestas e grupamentos em grafos com atributos nos vértices.

2.3.1 Agrupamentos em grafos com atributos nas arestas

Nos grafos com atributos nas arestas, os atributos indicam relacionamentos de natureza diferente, ou seja, os grafos permitem múltiplas arestas. Essa terminologia também é encontrada na literatura pela denominação de multigrafo.

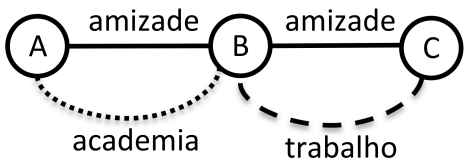


Figura 2.10: Representação de um multigrafo com as relações de amizade, trabalho e academia.

A Figura 2.10 ilustra um multigrafo representando um pequeno exemplo de uma rede social com três vértices e quatro arestas. Três relações de natureza distinta são apresentadas nesse grafo. A relação de amizade é representada com uma linha sólida, a relação de trabalho com uma linha pontilhada e se os usuários representados pelos vértices se exercitam na mesma academia, existe uma linha tracejada representando essa relação. Essa tese não utiliza multigrafos em sua abordagens. Caso haja necessidade de mais conhecimentos nessa área, favor referir-se à (BOTHOREL *et al.*, 2015).

2.3.2 Agrupamentos em grafos com atributos nos vértices

Os grafos com atributos nos vértices têm apresentado bastante interesse da comunidade acadêmica recentemente (BODEN *et al.*, 2013a; CHENG *et al.*, 2011; ZHOU *et al.*, 2009). É sabido que apenas a estrutura das redes sociais pode não ser suficiente parar determinar os grupos sociais (FREEMAN, 1996). Assim, a análise da estrutura juntamente com os atributos dos vértices pode evidenciar relações que também dependem das características dos indivíduos, como a homofilia.

Formalmente, um grafo com atributos nos vértices G é definido como uma 3-tupla $G(V, E, \Lambda)$, em que $V = \{u_1, u_2, ...u_q\}$ é um conjunto de q vertices, E é um conjunto de arestas que conecta pares de elementos de V e $\Lambda = \{a_1, a_2, \ldots, a_{|\Lambda|}\}$ é um conjunto de atributos associado aos vértices. Em um grafo com atributos G, cada vértice u_i é associado com um vetor de atributos de tamanho $|\Lambda|$.

A definição acima provê duas perspectivas possíveis para os dados subjacentes: topológica e relacional. A primeira corresponde à estrutura topológica do grafo. A última corresponde aos dados associados a cada vértice, em que cada vértice $u_i \in V$ é representado por $|\Lambda|$-tupla de valores de atributos.

Na teoria de agrupamentos em grafos (sem atributos), um agrupamento apresenta boa qualidade se os vértices dentro de um grupo são densamente conectados e poucas arestas existem entre os grupos (BODEN *et al.*, 2013a). Esse princípio não deve ser afetado no momento em que as informações dos atributos dos vértices são inseridas. Dessa forma, o agrupamento deve apresentar um balanceamento entre a estrutura topológica e os atributos dos vértices (ZHOU *et al.*, 2009). Consequentemente, os vértices dentro de um grupo devem ser estruturalmente coesos e homogêneos.

A informação dos atributos dos vértices no grafo pode ser adicionada na tarefa de agrupamento através da adição de arestas artificiais (GUEDES *et al.*, 2015). Nesse paradigma, se dois vértices de um grafo G são semelhantes, pode-se adicionar uma aresta artificial entre eles. Isso resulta em um novo grafo aumentado denominado G'. Esse novo grafo contém os vértices, as arestas originais e as arestas artificiais. Essa tese está contextualizada na área de grafos *com atributos nos vértices*. Existe uma vasta literatura que trata o problema de agrupamento de vértices em grafos (sem a utilização de atributos) e pode ser acessada em SCHAEFFER (2007).

2.4 Comparação entre dois agrupamentos

A comparação entre agrupamentos é formalmente definida como uma medida externa de avaliação (THEODORIDIS e KOUTROUMBAS, 2008). Há algumas medidas para comparar dois agrupamentos, dentre elas, pode-se destacar as medidas baseadas em contagem de pares (e.g., Índice de Jaccard[1]), medidas baseadas em entropia (e.g., Informação Mútua (COVER e THOMAS, 1991), Informação Mútua Normalizada (STREHL e GHOSH, 2003), entre outras). Para uma maior análise das medidas de comparação entre agrupamentos, por favor refira-se à WAGNER e WAGNER (2007).

A informação mútua (*mutual information* – em inglês)(MI) é uma medida que quantifica a dependência mútua (i.e., a quantidade de informação compartilhada) entre duas variáveis. Assim, dadas duas variáveis aleatórias X e Y com distribuição conjunta $p(x, y)$ e distribuições marginais $p(x)$ e $p(y)$, a MI é definida conforme a Eq. 2.5.

$$\mathrm{MI}(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{2.5}$$

O valor da MI é zero se as duas variáveis forem estatisticamente independentes. Dadas duas variáveis X = (1,1,0,0,1,1,1,0) e Y=(1,1,1,1,0,0,0,1), pode-se calcular a

[1]O índice Jaccard é definido pelo tamanho da interseção de dois documentos dividido pelo tamanho da união de dois documentos (JACCARD, 1901).

probabilidade conjunta conforme ilustra a Tabela 2.1. A distribuição marginal de X é $(\frac{3}{8}, \frac{5}{8})$ e a de Y é $(\frac{3}{8}, \frac{5}{8})$.

Tabela 2.1: Probabilidade conjunta.

x \ y	0	1
0	0	3/8
1	3/8	2/8

Dadas a probabilidade conjunta e a distribuição marginal, pode-se calcular a informação mútua entre as variáveis X e Y conforme apresenta a Eq. 2.6.

$$\mathrm{MI}(X;Y) = \left[0 + \frac{3}{8}\log\frac{\frac{3}{8}}{\frac{3}{8}\times\frac{5}{8}} + \frac{3}{8}\log\frac{\frac{3}{8}}{\frac{5}{8}\times\frac{3}{8}} + \frac{2}{8}\log\frac{\frac{2}{8}}{\frac{5}{8}\times\frac{5}{8}}\right] \approx 0.348 \tag{2.6}$$

A MI é uma medida que não possui limites superiores. Para melhores interpretações e comparações, uma versão normalizada pode ser utilizada para a obtenção de valores entre 0 e 1. Sendo assim, nesse trabalho é utilizada a informação mútua normalizada (*normalized mutual information* – em inglês)(NMI) para o cálculo da informação mútua.

Essa medida também pode ser utilizada para efetuar a comparação entre agrupamentos, conforme adotado em DANG e BAILEY (2014a); NIU *et al.* (2010); STREHL e GHOSH (2003). Nesses trabalhos, a NMI é utilizada para calcular quão semelhantes são dois agrupamentos. Seguindo STREHL e GHOSH (2003), a Eq. 2.7 apresenta a NMI para comparar dois agrupamentos c_i e c_j. Nessa equação, $H(c_i)$ e $H(c_j)$ representam a entropia (SHANNON, 1948b) de c_i e c_j respectivamente e $MI(c_i, c_j)$ é a informação mútua entre os agrupamentos c_i e c_j. $NMI(c_i, c_j)$ é igual a 1 se c_i e c_j são idênticos, e 0 se c_i e c_j são independentes.

$$\mathrm{NMI}(c_i, c_j) = \frac{MI(c_i, c_j)}{\sqrt{(H(c_i) \times H(c_j))}} \tag{2.7}$$

A entropia é uma medida de incerteza associada a uma variável aleatória e é representada conforme a Eq. 2.8. Caso não haja incerteza acerca de uma variável aleatória (e.g. uma moeda que dá sempre *cara*), a entropia é 0.

$$H(c_i) = -\sum_{i=0}^{i=n} p(c_i)\ log\ p(c_i) \tag{2.8}$$

A partir dessa definição, considerando um conjunto de agrupamentos $\mathcal{C}$, pode ser gerada uma matriz de NMI, denominada M_{NMI}, que possui valores de NMI para cada possível par de agrupamentos. Como exemplo, a Tabela 2.2 representa os agrupamentos ilustrados nas Figuras 1.3b, 1.3c, 1.3d e 1.3e. Esses agrupamentos

são denominados respectivamente c_1, c_2, c_3 e c_4.

Tabela 2.2: Representação dos agrupamentos das Figuras 1.3b, 1.3c, 1.3d e 1.3e pela perspectiva dos elementos.

Agrupamentos	Elementos dos grupos
c_1	$\langle[u_1, u_2, u_3], [u_4, u_5, u_6, u_7, u_8]\rangle$
c_2	$\langle[u_1, u_2, u_3], [u_4, u_5, u_6, u_7, u_8]\rangle$
c_3	$\langle[u_1, u_2, u_3, u_4, u_8], [u_5, u_6, u_7]\rangle$
c_4	$\langle[u_1, u_2, u_3, u_4], [u_5, u_6, u_7, u_8]\rangle$

A Tabela 2.2 também pode ser representada pela perspectiva dos rótulos dos grupos, conforme ilustrado na Tabela 2.3. Assim, cada elemento é representado com o rótulo do grupo em que foi enquadrado. Nesse caso, a representação é dada apenas por dois grupos: grupo 0 e grupo 1.

Tabela 2.3: Representação dos agrupamentos das Figuras 1.3b, 1.3c, 1.3d e 1.3e pela perspectiva dos rótulos.

Agrupamentos	Rótulos dos grupos
c_1	(0, 0, 0, 1, 1, 1, 1, 1)
c_2	(0, 0, 0, 1, 1, 1, 1, 1)
c_3	(0, 0, 0, 0, 1, 1, 1, 0)
c_4	(0, 0, 0, 0, 1, 1, 1, 1)

Dados os agrupamentos representados pela tabela 2.2 (ou pela tabela 2.3), pode-se calcular a matriz de NMI (M_{NMI}), representada na Figura 2.11. Cada elemento em M_{NMI} corresponde à NMI calculada entre um par de agrupamentos. No exemplo, quando c_1 é comparado com c_3, a NMI = 0.364. Pode-se notar também que c_1 e c_2 possuem a mesma configuração dos grupos, assim, seu valor é representado com 1 na matriz M_{NMI}, caracterizando que ambos os agrupamentos possuem o valor máximo de informação mútua normalizada. Os elementos da diagonal principal são representados com o valor 1, dado que representam a comparação entre cada agrupamento com ele mesmo.

$$M_{\text{NMI}} = \begin{array}{c} \\ c_1 \\ c_2 \\ c_3 \\ c_4 \end{array} \begin{array}{c} \begin{array}{cccc} c_1 & c_2 & c_3 & c_4 \end{array} \\ \begin{pmatrix} 1 & 1 & 0.364 & 0.562 \\ 1 & 1 & 0.364 & 0.562 \\ 0.364 & 0.364 & 1 & 0.562 \\ 0.562 & 0.562 & 0.562 & 1 \end{pmatrix} \end{array}$$

Figura 2.11: Matriz de NMI calculada através da comparação entre cada par de agrupamentos.

2.5 Comparação entre k agrupamentos

Na área de agrupamentos de consenso (*consensus clustering*) é possível combinar agrupamentos de forma que se obtenha uma solução única contendo as características de cada um dos agrupamentos iniciais (KUNCHEVA, 2004). Assim, dado um conjunto de agrupamentos $\mathcal{C}$, nesse problema chamados de agrupamentos-base, deseja-se encontrar uma única solução de consenso. Segundo GODER e FILKOV (2008), esses agrupamentos-base podem ser gerados por um algoritmo de agrupamento com variações nos parâmetros de entrada (e.g., variação na semente do k-*means*). É esperado que o agrupamento de consenso produzido a partir dos agrupamentos-base resulte em um agrupamento mais robusto. Isso também poderia levar à obtenção de um agrupamento com menor sensibilidade a *outliers* (KUNCHEVA e VETROV, 2006). Os agrupamentos de consenso ganharam popularidade inclusive na área médica (HAYES *et al.*, 2006).

Existem diversos algoritmos utilizados para a geração dos agrupamentos de consenso e podem ser examinados em NGUYEN e CARUANA (2007). Basicamente é utilizada uma *função de consenso* para gerar o agrupamento de consenso. Posteriormente são utilizadas medidas de avaliação nos agrupamentos de consenso para analisar a eficácia das funções utilizadas. Se considerarmos vários agrupamentos de consenso gerados por diferentes algoritmos, o melhor agrupamento de consenso pode ser avaliado por algumas medidas, dentre elas, a medida de informação mútua normalizada média (*Average Normalized Mutual Information* – em inglês)(ANMI) ((STREHL e GHOSH, 2003)).

A ANMI tem sido empregada em diversos trabalhos para avaliar a qualidade dos agrupamentos de consenso (HE *et al.*, 2008; IZAKIAN e PEDRYCZ, 2014; STREHL e GHOSH, 2003) . Essa medida calcula a quantidade de informação entre os agrupamentos de consenso e os agrupamentos-base. Altos valores de ANMI são desejáveis, dado que o melhor agrupamento de consenso é aquele que compartilha mais informação com os agrupamentos-base. Utilizando a matriz M_{NMI} como entrada, podemos calcular a ANMI. A ANMI é formalmente definida como uma medida entre um conjunto de agrupamentos-base $\{c_j\}$ e um agrupamento de consenso c (STREHL e GHOSH, 2003), conforme a Eq. 2.9, em que $c_j \in \mathcal{C}$.

$$\text{ANMI}(c, \{c_j\}) = \frac{1}{|\{c_j\}|} \sum_{j=1}^{|\{c_j\}|} M_{\text{NMI}}(c, c_j) \tag{2.9}$$

Essa medida pode ser utilizada para se obter um agrupamento que compartilhe menos informação com um conjunto de agrupamentos. No entanto, no problema da presente tese, dado um conjunto de agrupamentos $\mathcal{C}$, deseja-se encontrar os *top-k* agrupamentos não-redundantes. Essa tarefa necessita de uma medida capaz

de comparar os agrupamentos entre eles, todos com todos. Assim, o ANMI não pode ser utilizada para essa necessidade. Não foi encontrada na literatura de agrupamentos múltiplos uma medida capaz de comparar todos os agrupamentos entre eles. Conforme mencionado anteriormente, a presente tese também possui o objetivo de contribuir na lacuna de medidas de avaliação de agrupamentos múltiplos. No Capítulo 4 são apresentadas as medidas elaboradas nessa tese.

Capítulo 3

Trabalhos relacionados

Nesse capítulo são descritos os trabalhos relacionados à essa pesquisa. Ao proceder a revisão bibliográfica no ScienceDirect, não foram encontrados trabalhos que abordassem agrupamentos múltiplos não-redundantes em grafos com atributos. A busca realizada foi: ("multiple clustering" OR "multiple non-redundant clustering" OR "non-redundant clustering") AND "attributed graph". Foram encontrados dois trabalhos fracamente relacionados. Dessa forma, serão mencionados os trabalhos moderadamente relacionados ao tema dessa tese.

Os trabalhos descritos são relacionados de alguma maneira às seguintes áreas:

- produção de agrupamentos múltiplos em grafos
- consenso de agrupamentos múltiplos em grafos
- combinação entre a estrutura topológica e os atributos dos vértices de um grafo.

Embora sejam encontrados trabalhos na área de agrupamentos múltiplos em grafos (CARUANA *et al.*, 2006), tutoriais na área de agrupamentos múltiplos (MÜLLER *et al.*, 2010) e *surveys* na área de grafos com atributos (BOTHOREL *et al.*, 2015), não foram encontrados na literatura trabalhos que envolvam agrupamentos múltiplos em grafos com atributos, ou seja, trabalhos que envolvam os atributos dos vértices na produção de agrupamentos múltiplos. Algumas palavras-chave de áreas correlatas foram pesquisadas, combinadas com a palavra-chave *graph*, como *non-redundant clustering*, *alternative clustering*, além de, evidentemente, *multiple clusterings*. Entretanto, os algoritmos encontrados na literatura capazes de produzir agrupamentos múltiplos em grafos se concentram na estrutura topológica, ignorando as propriedades dos vértices.

É importante ressaltar que em alguns trabalhos na literatura de agrupamentos, um grupo (*cluster* – em inglês) é considerado uma comunidade (*community* – em inglês) (SCHAEFFER, 2007). Da mesma forma, um agrupamento também pode

ser referido como uma estrutura de comunidade (*community structure* – em inglês) (DEEPJYOTI e ARNAB). Segundo RIEDY *et al.* (2012), o problema de detecção de comunidades (*community detection* – em inglês) é um problema de agrupamento em grafo (*graph clustering* – em inglês). Consequentemente, os trabalhos relacionados podem mencionar o mesmo problema com diferentes terminologias. Portanto, nessa tese, não será feita distinção entre essas terminologias, conforme em DE MEO *et al.* (2012); NEUBAUER e OBERMAYER (2009); STAUDT e MEYERHENKE (2013).

Esse capítulo é dividido da seguinte forma: na Seção 3.1 são descritos alguns trabalhos com abordagens de produção de agrupamentos múltiplos em grafos. Na Seção 3.2 são mencionados trabalhos que se inserem na área de agrupamento de consenso em grafos. A Seção 3.3 expõe trabalhos que tratam o problema de gerar um agrupamento baseado na estrutura e nos atributos de grafos. Na Seção 3.4 são feitas as considerações finais do capítulo, com uma visão geral dos algoritmos propostos nessa tese.

3.1 Produção de agrupamentos múltiplos em grafos

A abordagem de *meta clustering* (CARUANA *et al.*, 2006) pode ser dividida em alguns passos. Inicialmente, o objetivo é produzir uma série de agrupamentos para encontrar conjuntos de agrupamentos potencialmente úteis. Em seguida, uma métrica de distância é utilizada para determinar a similaridade entre pares de agrupamentos com o objetivo de produzir agrupamentos de agrupamentos, tarefa denominada *meta clustering*. Após a produção dos agrupamentos de agrupamentos, os centróides dos grupos formados são apresentados ao usuário como candidatos potenciais. Se um desses agrupamentos (centróides) é apropriado para tarefa do usuário, os agrupamentos semelhantes podem ser examinados. Os autores utilizam um algoritmo de agrupamento hierárquico, entretanto mencionam que a abordagem de *meta clustering* pode ser executada utilizando qualquer algoritmo capaz de comparar pares de objetos. Uma vez que a abordagem de *meta clustering* utiliza o principio de agrupamentos de agrupamentos, os agrupamentos poderiam ser gerados a partir de grafos (ROCKLIN e PINAR, 2011). Entretanto, os autores não aplicam essa abordagem em conjuntos de dados de grafos.

Em ZHANG e LI (2011), os autores apresentam uma proposta baseada em *meta clustering* e agrupamentos de consenso (*consensus clustering*). Diferente da abordagem de *meta clustering* proposta por CARUANA *et al.* (2006), essa abordagem produz um conjunto de agrupamentos de consenso. Inicialmente são gerados diversos agrupamentos variando os parâmetros de entrada dos diversos

algoritmos de agrupamentos utilizados. Em seguida, os agrupamentos são comparados e agrupados utilizando a abordagem de *meta clustering*. Por fim, múltiplos agrupamentos de consenso são produzidos com a aplicação de algoritmos de agrupamento de consenso. Os autores utilizam quatro algoritmos diferentes para produzir os agrupamentos de consenso. Os experimentos mostram a eficácia dessa proposta, entretanto, não foram utilizados conjuntos de dados de grafos.

Na abordagem proposta em NIU *et al.* (2010) os agrupamentos múltiplos são produzidos a partir de um grafo utilizando projeção de subespaços (*views*) dos dados. Os agrupamento em projeções de subespaços funcionam de forma que os grupos são observados com perspectivas distintas, formadas por combinações arbitrárias de atributos (MÜLLER *et al.*, 2012). Assim, dadas as diferentes projeções, os objetos podem ser agrupados de diferentes maneiras (MÜLLER *et al.*, 2010). Essa abordagem utiliza o algoritmo de agrupamento espectral para reduzir a dimensionalidade do conjunto de dados. Os autores não conduziram experimentos em conjuntos de dados de grafos.

Na abordagem proposta em (NIU *et al.*, 2014) os agrupamentos múltiplos também são produzidos a partir de um grafo utilizando projeção de subespaços (*views*) dos dados. Entretanto, nesse caso, uma solução de agrupamento é apresentada como entrada. Assim, o objetivo é encontrar uma solução alternativa não-redundante. Isso se estende para o caso em que se provê mais de uma solução de agrupamento como entrada para os algoritmos, entretanto, nesse caso, o resultado é um agrupamento que seja não-redundante com relação a todos agrupamentos de entrada. Nesse trabalho também não são utilizados conjuntos de dados de grafos.

Em MANDALA *et al.* (2012), os autores abordam a produção de múltiplos agrupamentos próximos ao ótimo. Para atingir esse objetivo, são realizadas perturbações no grafo. As perturbações em grafos (*graph perturbations* – em inglês) são produzidas realizando pequenas modificações na estrutura do grafo (e.g., adicionando uma aresta ou um vértice) (CVETKOVIĆ *et al.*, 1997). Nesse trabalho, os autores adicionam nós isolados ao grafo original, de forma que novos grafos aumentados são gerados. Em seguida, é utilizada uma heurística de maximização da modularidade para produzir os agrupamentos, mais especificamente, o algoritmo de Louvain (BLONDEL *et al.*, 2008). É demonstrado numericamente que a metodologia proposta pode identificar agrupamentos bastante diferentes em diversos conjuntos de dados. Entretanto, a quantidade e diversidade os agrupamentos gerados parece ser dependente do domínio da rede representada pelo grafo. Os experimentos são realizados em grafos.

Essa tese se correlaciona aos trabalhos descritos nessa seção por produzir agrupamentos múltiplos em grafos. Entretanto, os agrupamentos não são produzidos com base em perturbações ou com a utilização de diversos algoritmos de

agrupamento com modificações nos parâmetros de entrada. São produzidos com modificações no grafo original, de forma que o grafo modificado reflita na estrutura topológica as informações dos atributos dos vértices.

3.2 Agrupamentos de consenso em grafos

A área de agrupamentos de consenso (GHOSH *et al.*, 2002) pode gerar resultados mais robustos e estáveis[1] quando comparados com a abordagens tradicionais de produzir um agrupamento (GHOSH e ACHARYA, 2011). Diversas terminologias são utilizadas na língua inglesa para se referir a esse mesmo problema: *clustering combination* (XU *et al.*, 2013), *cluster ensembles* (GHOSH e ACHARYA, 2011), clustering aggregation (ABU-JAMOUS *et al.*, 2015), dentre outras. Em geral, o problema consiste em combinar múltiplos agrupamentos produzidos a partir de um mesmo conjunto de dados, produzindo um único agrupamento com melhores resultados (LI *et al.*, 2010). Vale ressaltar que alguns autores diferenciam as abordagens de *ensemble clustering* e *consensus clustering*, conforme em (HAHMANN). Entretanto, essa distinção não será enfocada no presente trabalho.

Embora a literatura esteja repleta de trabalhos na área de agrupamento de consenso (STREHL e GHOSH, 2003; VEGA-PONS e RUIZ-SHULCLOPER, 2011), aqui são referidos os trabalhos que desenvolvem estudos aplicados em grafos ou redes (*networks* – do inglês) (i.e., dados estruturais), não um conjunto de dados vetoriais. Diversas abordagens poderiam ser aplicadas em grafos, entretanto, muitas não foram encontradas nesse contexto e, portanto, não são mencionadas no decorrer dessa seção.

Basicamente dois métodos têm sido empregados para combinar diferentes agrupamentos para a produção de agrupamentos de consenso em grafos: perturbações no grafo e a modificação inicial da configuração de algum algoritmo (SEIFI *et al.*, 2013). Dentre as abordagens que realizam perturbações, podem ser destacadas as obtidas com a modificação de arestas ou vértices (KARRER *et al.*, 2008) ou com a seleção de diferentes subespaços do conjunto de dados (MONTI *et al.*, 2003). Os métodos baseados na modificação inicial dos parâmetros de entrada dos algoritmos de agrupamento são subdivididos, essencialmente, em duas etapas: a produção de múltiplas soluções de agrupamento e, posteriormente, a combinação dessas soluções de forma a produzir uma solução mais robusta (HAHMANN).

Em KARRER *et al.* (2008), os autores propõem uma abordagem estocástica para encontrar comunidades de consenso utilizando perturbações no grafo original. Essas perturbações são realizadas com a modificação na posição de algumas arestas, removendo uma fração aleatoriamente e colocando de volta entre pares de vértices

[1] Conceito de estabilidade (*Stability* – em inglês) (KWAK *et al.*, 2009; TOPCHY *et al.*, 2005).

escolhidos com a abordagem de *random walk* (LOVÁSZ, 1996). Essa abordagem é um processo estocástico em que se caminha por vértices escolhendo, a cada passo, um vizinho aleatoriamente (NOH e RIEGER, 2004). Esse procedimento é realizado diversas vezes. A cada vez, o novo grafo perturbado é comparado com o grafo original utilizando a medida de variação da informação (MEILĂ, 2007). Os autores descrevem que o método proposto pode detectar estruturas de comunidades densas. HU *et al.* (2010) estendem essa abordagem para avaliar a estabilidade das estruturas de comunidades. MIRSHAHVALAD *et al.* (2012) também utilizam a abordagem de perturbar a rede adicionando arestas para encontrar comunidades significantes.

Em OVELGONNE e GEYER-SCHULZ (2012), os autores propõem uma abordagem denominada *Core Groups Graph Clustering* (CGGC). Inicialmente são gerados agrupamentos empregando alguns algoritmos de agrupamento determinísticos iniciais ou algum algoritmo não-determinístico iniciado varias vezes. Em seguida, os grupos que possuem mais sobreposições (denominados *core groups*) nas partições geradas são selecionados, formando uma nova partição. Essa nova partição induz a criação de um novo grafo em que cada novo vértice é composto pela união dos vértices de cada grupo. Por fim é aplicado um algoritmo de agrupamento nesse grafo induzido. Os autores mencionam que os algoritmos iniciais podem ser qualquer algoritmo apropriado para minimizar a função objetivo.

Em SEIFI *et al.* (2013), os autores executam o algoritmo de Louvain múltiplas vezes em um grafo G para produzir múltiplos agrupamentos. Em seguida, é calculada a proporção de vezes que cada par de vértices pertenceu a mesma comunidade. Um novo grafo G' com pesos é criado com todos os vértices do grafo original, de forma que o peso das arestas entre cada par de vértices é dado pela proporção de vezes que cada par pertenceu a mesma comunidade. Em seguida as arestas com valores de peso menores que o limiar α são removidas gerando um novo grafo G''. Os componentes conectados de G'' são os *community cores* (ou *consensual communities* (CAMPIGOTTO *et al.*, 2013)). Embora os experimentos tenham apresentado bons resultados, os autores relatam a necessidade de encontrar um meio de identificar o limiar.

As abordagens denominadas `CoPAM` (MOSER *et al.*, 2009) e `GAMer` (GÜNNEMANN *et al.*, 2010) combinam subespaços de atributos para detectar comunidades. Essas abordagens determinam conjuntos de vértices que apresentam alta similaridade em subconjuntos de suas dimensões e também são densamente conectados, considerando as dimensões relevantes e a densidade. A abordagem de GÜNNEMANN *et al.* (2010) considera a redundância e remove os grupos redundantes da solução final, diferente de MOSER *et al.* (2009).

Por fim, o trabalho concebido por ZHANG e LI (2011) também se enquadra nessa seção por produzir um consenso das soluções de agrupamento produzidas por

algoritmos com diferentes parâmetros de entrada. Cada um desses algoritmos produz diversos agrupamentos que, em seguida, são agrupados (*meta clustering*) conforme descrito na Seção 3.1.

Os trabalhos descritos nessa seção descrevem alguns algoritmos que produzem agrupamentos múltiplos com base na modificação da estrutura topológica do grafo ou geração de novas estruturas. Essa tese é correlacionada a esses trabalhos por alterar a estrutura inicial do grafo com a adição de arestas artificiais e se difere por adicionar essas arestas para produzir agrupamentos múltiplos, e não um consenso.

3.3 Combinação da estrutura topológica com os atributos dos vértices de um grafo

O problema de combinar a estrutura topológica e os atributos dos vértices de um grafo tem sido bastante estudado recentemente. A idéia geral é o desenho de uma medida de distância/similaridade entre pares de vértices que combinam a estrutura topológica e os atributos dos vértices (ABERER *et al.*, 2012). O algoritmo SA-cluster (ZHOU *et al.*, 2009) é considerado o estado da arte nessa abordagem (ABERER *et al.*, 2012).

O `SA-cluster` é baseado no princípio de que um processo ideal de agrupamento deve produzir grupos com estrutura coesa e vértices com atributos homogêneos. Isso é obtido balanceando a estrutura e a similaridade entre os atributos para gerar uma solução de agrupamento. Para esse fim, o algoritmo proposto adiciona vértices representando os atributos e arestas artificiais ao grafo original, o que produz um novo grafo aumentado. A partir desse momento, é utilizada uma medida de proximidade estrutural para calcular a proximidade entre os vértices utilizando *random walk*. Os pesos das arestas são iterativamente ajustados para balancear a importância entre a estrutura e os atributos dos vértices. Após o cálculo da proximidade, é aplicado o algoritmo *k-medoids* para produzir uma partição do grafo em m grupos. A saída do `SA-cluster` é uma única solução de agrupamento em que cada grupo é densamente conectado e possui valores homogêneos nos atributos dos vértices. Em trabalho posterior, os autores propuseram versões aprimoradas (CHENG *et al.*, 2011; ZHOU *et al.*, 2010) utilizando a mesma abordagem do `SA-cluster`.

No método *Connected k Centers* (CkC) proposto por GE *et al.* (2008), a estratégica de caminhada é baseada em busca em largura (*Brief First Search* – em inglês) (BFS). O CkC implementa um algoritmo baseado no *k-means*, utilizando a caminhada entre os vértices para calcular a distância entre os mesmos, considerando também um vetor de características de cada vértice nessa caminhada.

Primeiramente, m vértices são escolhidos ao acaso para serem os centróides dos grupos. Em seguida, cada um dos vértices restantes é associado a um dos grupos representados pelos centróides utilizando BFS. Em seguida, os centróides dos grupos são recalculados. Esses dois últimos passos são repetidos até que não haja modificação nos centróides dos grupos. Assim, esse método produz apenas um agrupamento combinando a estrutura topológica com os atributos dos vértices.

Em RUAN *et al.* (2013), os autores propõem um algoritmo denominado `CODICIL`. Esse algoritmo inicia criando um conjuntos de arestas artificiais com base nos *top-k* vizinhos mais similares de cada vértice. Essa similaridade é calculada utilizando a distância do coseno do vetor de termos associado a cada vértice. Em seguida, esse novo grafo é reprocessado e algumas arestas menos significantes são removidas. Por fim, um agrupamento é gerado utilizando algum algoritmo de agrupamento em grafos.

O trabalho de MOSER *et al.* (2009) também se enquadra nessa seção por utilizar os atributos dos vértices na tarefa de agrupamento do grafo, assim como ocorre em GÜNNEMANN *et al.* (2010). Ambos os trabalhos foram descritos na Seção 3.2.

Os trabalhos descritos nessa seção se aproximam da presente tese no que tange à utilização da estrutura topológica e valores dos atributos dos vértices para produzir uma solução de agrupamento. Entretanto, a idéia nessa tese não é gerar apenas um agrupamento e sim agrupamentos múltiplos não-redundantes combinando a estrutura topológica e os valores dos atributos. Essa combinação é feita com a criação de arestas artificiais entre vértices similares.

3.4 Considerações

Em grande parte das redes reais, existe um grande número de partições próximas ao ótimo (near-optimal), com algumas partições sendo muito distintas das outras e, portanto, escolher um agrupamento ótimo não apresenta informação completa sobre a estrutura da rede (MANDALA *et al.*, 2012). Assim sendo, essa tese se difere dos trabalhos que produzem um agrupamento ótimo combinando a estrutura topológica com atributos dos vértices. O objetivo não é produzir apenas um agrupamento ótimo utilizando a combinação entre a estrutura topológica e os atributos dos vértices, mas agrupamentos múltiplos não-redundantes com essa combinação.

Os trabalhos apresentados nesse capítulo possuem alguma relação com essa tese. Os trabalhos descritos foram divididos em três categorias. A Figura 3.1 ilustra uma síntese e a análise cronológica dos trabalhos. Pode-se notar que existe uma lacuna para investigação que combina as duas abordagens: *agrupamentos múltiplos* e *agrupamentos combinando a estrutura e atributos dos vértices do grafo.* Essa lacuna é explorada nos próximos capítulos.

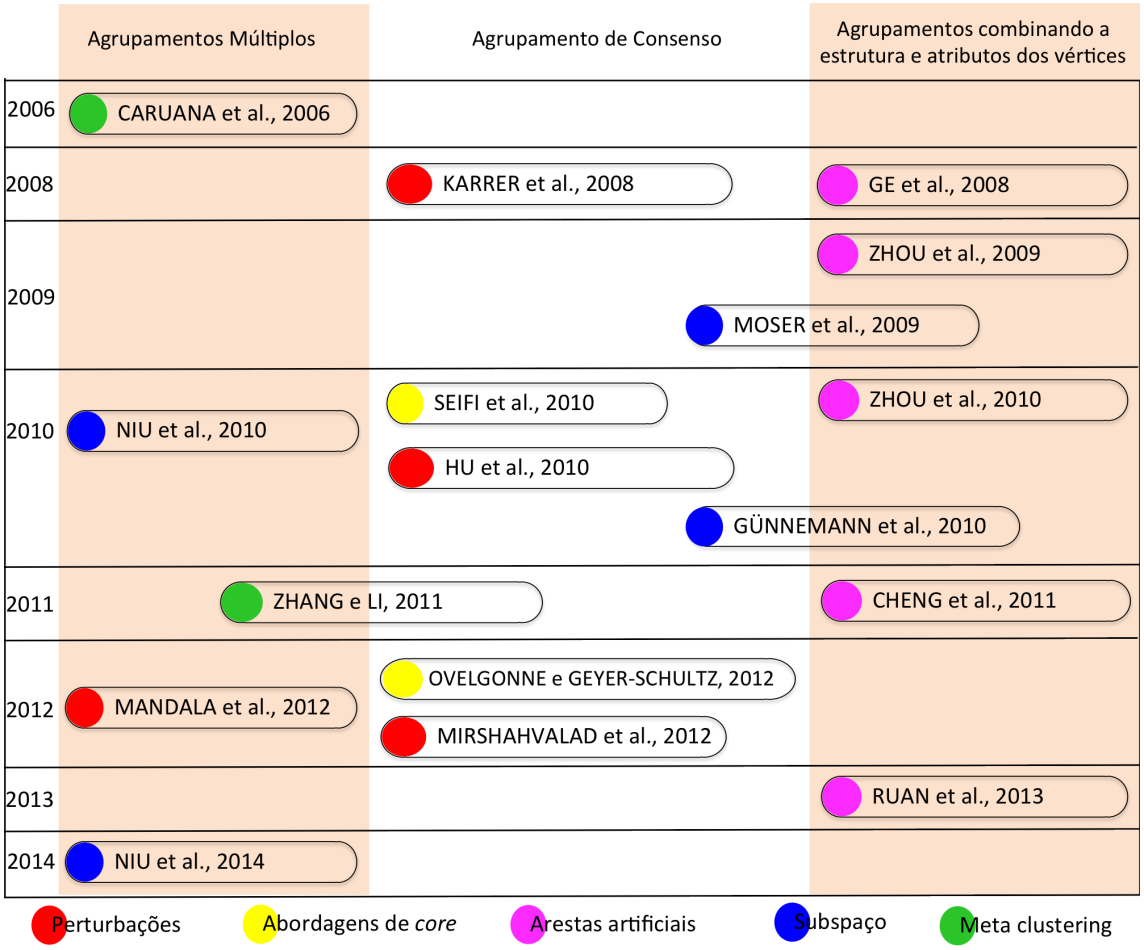


Figura 3.1: Síntese dos trabalhos relacionados.

Capítulo 4

Agrupamentos múltiplos em grafos com atributos

Conforme já descrito no Capítulo 2, o problema de agrupamento em grafos é bastante conhecido na literatura. Entretanto, não foram encontradas abordagens que produzam agrupamentos múltiplos em grafos com atributos nos vértices. Esse problema é importante em algumas áreas, por exemplo, na área de marketing, em que a combinação de características sociais com as propriedades de cada indivíduo pode conduzir a melhores resultados na identificação de futuros adotantes de um produto. Com isso, a abordagem aqui descrita se apresenta como uma das mais importantes contribuições desse trabalho.

As seções seguintes descrevem os algoritmos que, juntamente, produzem agrupamentos múltiplos não-redundantes em grafos com atributos. De forma geral, a abordagem proposta tem o objetivo de produzir um conjunto $\mathcal{C}$ de agrupamentos, em que cada um possui uma mesclagem entre a estrutura topológica e atributos dos vértices do grafo. Na Seção 4.1 encontra-se descrito o algoritmo para gerar um agrupamento combinando a estrutura topológica com os atributos dos vértices de um grafo. Na Seção 4.2 é apresentado o algoritmo para gerar agrupamentos múltiplos combinando a estrutura topológica com os atributos dos vértices de um grafo. A Seção 4.3 tem o objetivo de contribuir na lacuna existente na área de avaliação da redundância em agrupamentos múltiplos. São propostas três medidas baseadas na NMI para avaliar redundância em agrupamentos múltiplos. Em seguida, na Seção 4.4, é apresentado o algoritmo que utiliza as medidas de redundância para produzir um *ranking* dos agrupamentos não-redundantes. A Seção 4.5 apresenta uma intuição sobre o efeito de cada uma das medidas de redundância no ranqueamento dos agrupamentos. Por fim, a Seção 4.6 apresenta a formulação do algoritmo proposto como o problema de máxima diversidade.

4.1 Algoritmo para agrupamento em grafos com atributos por adição de arestas artificiais

Em síntese, o algoritmo proposto nessa seção utiliza o princípio da adição de arestas artificiais para incorporar informação dos atributos dos vértices na estrutura de um grafo G. Em seguida é utilizado um algoritmo de agrupamento espectral para produzir uma solução de agrupamento. Essa nova solução de agrupamento sugere que haja informações tanto da estrutura topológica quanto dos atributos dos vértices $V \in G$. Para esse propósito é utilizado o algoritmo CRAG (GUEDES *et al.*, 2015), acrônimo de ***C**luste**r**ing in **A**ttributed **G**raphs*, representado no Algoritmo 3.

Algoritmo 3: CRAG$(G, attrSet, d, m)$

Input:
- $G(V, E, \Lambda)$ = grafo com atributos
- $attrSet$ = conjunto de atributos
- d = distância dos vizinhos
- m = número de grupos

Output: uma solução de agrupamentos de vértices em G.

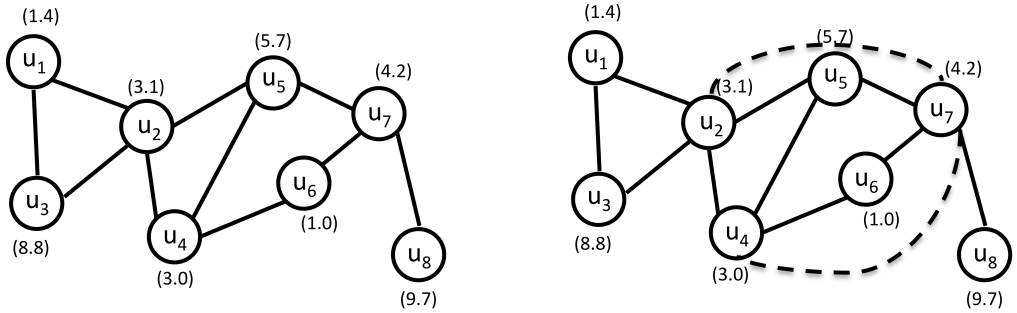
```
E' ← ∅
s ← getSimilarityThreshold(G, attrSet, d)
for all u_i ∈ G.V do
    N_{u_i} ← getNeighborhood(G, u_i, d)
    for all v_j ∈ N_{u_i} do
        if similarity(u_i, u_j, attrSet) ≥ s then
            E' ← E' ∪ {edge(u_i, u_j)}
        end if
    end for
end for
G.E ← G.E ∪ E'
return spectralClustering(G, m)
```

Esse algoritmo inicia com o recebimento de um grafo G, um conjunto de atributos $attrSet$ e dois inteiros positivos, d e m. O algoritmo CRAG insere informação dos atributos na estrutura do grafo com a inclusão de novas arestas artificiais. Essas novas arestas artificiais são criadas baseadas na similaridade entre os atributos $attrSet$ dos vértices do grafo G e correspondem aos vértices mais similares à distância δ, $1 < \delta \leq d$, de acordo com a distância de Shimbel (SHIMBEL, 1953). O número total de arestas artificiais criadas é limitado a 20% do número total de arestas originais em G. Esse limite segue o princípio de Pareto (KOCH, 1999) que indica que 80% dos efeitos são causados por 20% das causas, também conhecido como "a regra dos 80-20". Essa abordagem foi escolhida com o objetivo de preservar ao máximo a densidade da estrutura, e, com isso, não ir de encontro ao princípio dos

agrupamentos em grafos, que visam obter grupos densamente conectados.

No Passo 2, a função getSimilarityThreshold() computa a similaridade entre todos os pares de vértices a distância δ, para retornar um valor que corresponde ao limiar para se adicionar 20% de novas arestas artificiais. Em seguida, esse valor é atribuído à variável s. No Passo 4, a cada iteração, a função getNeighborhood() retorna $\mathcal{N}_{u_i}$, o conjunto de todos os vértices à distância δ de u_i. Então, no Passo 6, a função similarity() computa a similaridade entre u_i e cada $u_j \in \mathcal{N}_{u_i}$ a cada iteração. Se esse valor é maior que s, uma nova aresta artificial é criada entre os vértices u_i e u_j. Todas as arestas artificiais a serem criadas são adicionadas ao grafo G no Passo 11, o que produz um novo grafo aumentado.

O exemplo da Figura 4.1a ilustra um grafo não-direcionado com um valor para o atributo a_1 nos vértices. A distância euclidiana entre os valores do atributo a_1 dos vértices é utilizada para a determinação de quais arestas artificiais serão criadas. Para calcular os 20% de vértices mais similares, primeiramente deve ser calculada a distância euclidiana entre cada par de vértices (u_i, u_j) à distância δ, $1 < \delta \leq d$, conforme ilustra a Tabela 4.1a.



(a) Grafo sem arestas artificiais.

(b) Grafo com arestas artificiais.

Figura 4.1: Exemplo de grafo com atributos nos vértices: (a) sem arestas artificiais. (b) com arestas artificiais.

A Tabela 4.1b é ordenada crescentemente pela distância euclidiana. Visto que o grafo da Figura 4.1a possui 10 arestas, podem ser criadas no máximo 2 arestas artificiais (20%). Essas arestas são criadas entre os vértices mais similares, i.e., aqueles com distância igual ou superior à 0.99 ($s = 0.99$). A ordenação na Tabela 4.1b evidencia que os vértices mais similares são (u_2, u_7) e (u_4, u_7), indicando que devem ser criadas arestas artificiais entre esses vértices. A Figura 4.1b ilustra as duas arestas artificiais criadas, representadas em linha tracejada. As arestas originais são ilustradas em linha preenchida.

Em seguida, esse grafo aumentado é submetido ao processo de agrupamento utilizando um algoritmo de agrupamento espectral juntamente com o número de grupos m que a solução de agrupamento deve conter. Finalmente, a nova solução

Tabela 4.1: Similaridade para o atributo a_1 entre os vértices (u_i,u_j) à distância δ, $1 < \delta \leq d$.

(a) Vértices à distância 2 e a similaridade calculada para a_1.

Par de vértices	Similaridade
(u_1,u_4)	0.93
(u_1,u_5)	0.58
(u_2,u_6)	0.87
(u_2,u_7)	1.00
(u_3,u_4)	0.38
(u_3,u_5)	0.74
(u_4,u_7)	0.99
(u_5,u_6)	0.53
(u_5,u_8)	0.62
(u_6,u_8)	0.0

(b) Vértices à distância 2 e a similaridade calculada para a_1, ordenados.

Par de vértices	Similaridade
(u_2,u_7)	1.00
(u_4,u_7)	0.99
(u_1,u_4)	0.93
(u_2,u_6)	0.87
(u_3,u_5)	0.74
(u_5,u_8)	0.62
(u_1,u_5)	0.58
(u_5,u_6)	0.53
(u_3,u_4)	0.38
(u_6,u_8)	0.00

de agrupamento é retornada.

É importante destacar que o CRAG pode utilizar outros algoritmos de agrupamento, entretanto, os algoritmos de agrupamento espectral estão entre os algoritmos mais populares dentre os métodos de agrupamento baseados em grafo (HEIN e SETZER, 2011). Também é possível encontrar na literatura diversas implementações de algoritmos de agrupamento espectral, bem como variantes das matrizes Laplacianas (e.g., matriz Laplaciana normalizada). No entanto, o objetivo dessa tese não é explorar outros algoritmos de agrupamento ou diferentes matrizes Laplacianas, o que deixa margem para outros trabalhos.

4.2 Algoritmo para agrupamentos múltiplos em grafos com atributos

A Seção 4.1 descreve a abordagem utilizada para criar um agrupamento combinando a estrutura topológica e valores dos atributos dos vértices de um grafo. Utilizando essa abordagem é possível produzir agrupamentos múltiplos com base em diferentes conjuntos de atributos dos vértices. Nessa seção é apresentado o algoritmo M-CRAG (GUEDES *et al.*, 2015), acrônimo de ***M**ultiple **C**luste**r**ings in **A**ttributed **G**raphs*, conforme apresentado no Algoritmo 4. O M-CRAG tem o objetivo de produzir agrupamentos múltiplos utilizando o algoritmo CRAG.

Inicialmente, o M-CRAG recebe três parâmetros: um grafo com atributos G, a distância dos vizinhos d e o número de grupos a serem gerados pelo algoritmo de agrupamento espectral m. No Passo 2, o objetivo da função $\mathcal{P}$ é selecionar

Algoritmo 4: M-CRAG(G, d, t, m)

1 **Input**:
- $G(V, E, \Lambda)$ = grafo com atributos
- d = distância dos vizinhos
- t = limiar para a NMI
- m = número de grupos

Output: $\mathcal{C}$, um conjunto de soluções de agrupamento dos vértices em G.

1: $\mathcal{C} \leftarrow \emptyset$
2: $\mathcal{F} \leftarrow \mathcal{P}(G.\Lambda) - \{\emptyset\}$
3: **for all** $attrSet \in \mathcal{F}$ **do**
4: $\quad c \leftarrow \text{CRAG}(G, attrSet, d, m)$
5: $\quad \mathcal{C} \leftarrow \mathcal{C} \cup \{c\}$
6: **end for**
7: **return** $\mathcal{C}$

um subconjunto não-vazio de Λ. Em seguida, no Passo 3, o M-CRAG realiza uma iteração para cada subconjunto de atributos $attrSet$ em $\mathcal{F}$. O objetivo é invocar o algoritmo CRAG utilizando como argumento $attrSet$, de forma que um novo agrupamento c seja produzido com informações da estrutura topológica e dos atributos dos vértices. Assim, no Passo 5, c é adicionado a $\mathcal{C}$. Por fim, no Passo 7, o algoritmo retorna $\mathcal{C}$.

Note que no Passo 2, o conjunto não-vazio de Λ é atribuído a $\mathcal{F}$, ou seja, o *powerset* (DEVLIN, 2012). Por exemplo, se $\Lambda = \{a, b\}$, o M-CRAG invoca o CRAG três vezes, uma para cada combinação de atributos: (i) a; (ii) b; (iii) a e b. Assim, o M-CRAG gera três soluções de agrupamento: $\texttt{CRAG}_a$, $\texttt{CRAG}_b$ e $\texttt{CRAG}_{ab}$.

A Figura 4.2 ilustra a idéia geral do algoritmo M-CRAG, que recebe como entrada um grafo com atributos $\Lambda = \{a, b\}$ nos vértices. A função $\mathcal{P}$ produz todas as possíveis combinações entre os atributos em Λ. Em seguida, o algoritmo CRAG é invocado recebendo como entrada cada conjunto de atributos $attrSet$. Com isso, o CRAG produz uma nova estrutura topológica para cada conjunto de atributos em Λ. Essa nova estrutura possui informações do conjunto $attrSet$ em forma de novas arestas criadas (i.e., arestas artificiais). Em seguida, o CRAG produz um agrupamento com cada uma dessas estruturas.

Os algoritmos apresentados até esse momento produzem agrupamentos múltiplos em grafos com atributos nos vértices. Entretanto, é possível que haja redundância entre esses agrupamentos. No entanto, é importante que se obtenha múltiplas soluções de agrupamentos *não-redundantes* (NIU *et al.*, 2010). Assim, dado um conjunto de agrupamentos $\mathcal{C}$, o problema é definido pela seleção dos k agrupamentos não-redundantes. O espaço de busca desse problema é de tamanho $\binom{|\mathcal{C}|}{k}$. Dependendo dos valores de $|\mathcal{C}|$ e k, o processo de seleção pode se tornar

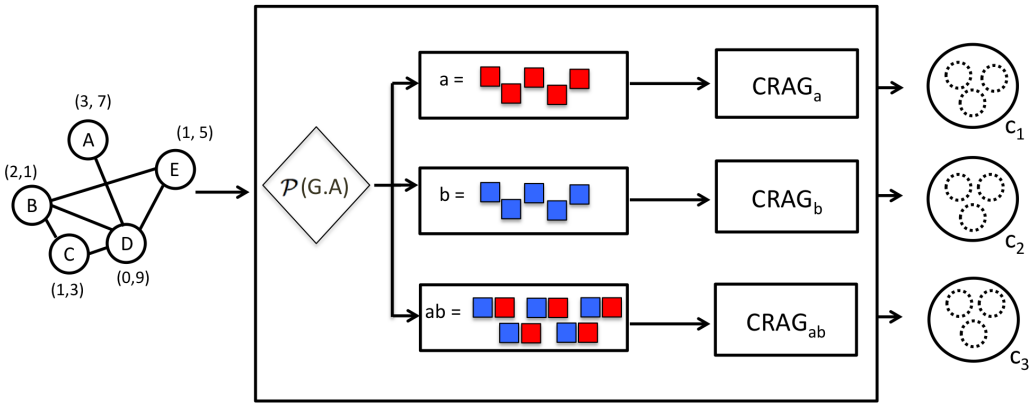


Figura 4.2: Produção de agrupamentos múltiplos pelo algoritmo M-CRAG.

computacionalmente intratável através do uso de abordagens de força bruta. Nesse cenário, é indispensável a utilização métodos que não sejam baseados na abordagem de força bruta.

4.3 Medidas de comparação entre agrupamentos múltiplos

A Seção 2.4 apresentou uma breve introdução sobre algumas medidas existentes para calcular a redundância entre agrupamentos. Entretanto, no problema de seleção dos *top-k* agrupamentos não-redundantes, é necessário que haja uma medida para comparar k agrupamentos entre eles. Segundo NIU *et al.* (2010), embora a literatura sobre a tarefa de agrupamento seja vasta, pouca atenção tem sido aplicada ao problema de encontrar múltiplas soluções não-redundantes. O objetivo do conteúdo dessa seção é contribuir nessa lacuna. Conforme descrito no Capítulo 2, esse problema também é destacado por MÜLLER *et al.* (2012), que evidencia uma lacuna na área de medidas de avaliação para agrupamentos múltiplos.

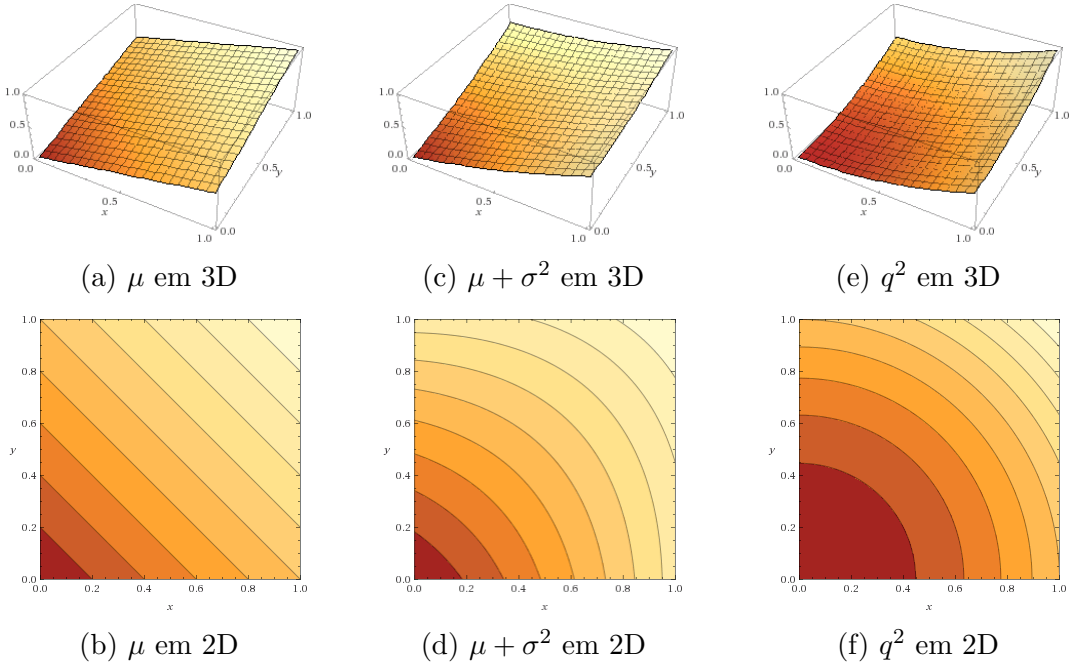
Para preencher essa lacuna, foram idealizadas três medidas. São baseadas na NMI e envolvem três medidas estatísticas: média (μ), exibida na Eq. 4.1; média + variância ($\mu + \sigma^2$), definida na Eq. 4.2 e média dos quadrados (q^2), apresentada (Eq. 4.3).

$$\mu = \frac{\sum_{i=1}^{n}(x_i)}{n} \tag{4.1}$$

$$\mu + \sigma^2 = \mu + \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n} \tag{4.2}$$

$$q^2 = \frac{\sum_{i=1}^{n}(x_i)^2}{n} \tag{4.3}$$

O comportamento apresentado por cada uma dessas medidas é ilustrado na Figura 4.3. Pode-se notar que cada uma delas se comporta de maneira diferente.



(a) μ em 3D (c) $\mu + \sigma^2$ em 3D (e) q^2 em 3D

(b) μ em 2D (d) $\mu + \sigma^2$ em 2D (f) q^2 em 2D

Figura 4.3: Representação gráfica das medidas μ, $\mu + \sigma^2$ e q^2 em 3D e 2D.

As Figuras 4.3b, 4.3d e 4.3f representam o gráfico de contorno (*contour plot*) das ilustrações em três dimensões das Figuras 4.3a, 4.3c e 4.3e respectivamente. Essa técnica permite visualizar valores nos eixos X e Y. Os valores no eixo Z representam os níveis de contorno. A Tabela 4.2 apresenta um exemplo ilustrando o comportamento das diferentes medidas.

Tabela 4.2: Representação de valores no gráfico de contorno da média, média + variância e média dos quadrados.

Exemplo	h_1	h_2	μ	$\mu + \sigma^2$	q^2
A	0.5	0.5	0.5	0.50	0.25
B	0.4	0.6	0.5	0.51	0.26
C	0.1	0.9	0.5	0.66	0.82

Nos exemplos `A`, `B` e `C`, h_1 e h_2 representam valores nos eixos X e Y das Figuras 4.3b, 4.3d e 4.3f. Tanto `A` quanto `B` e `C` apresentam o valor de 0.5 para μ. Observa-se que média dos quadrados se apresenta como a medida mais radical em termos de consideração da dispersão. No exemplo `A` pode-se notar que $\mu = \mu + \sigma^2 = 0.5$, entretanto, $q^2 = 0.25$, mesmo não havendo dispersão. O fato de não haver dispersão em q^2 faz com que esse valor seja inferior aos valores de μ e $\mu + \sigma^2$.

Por outro lado, no exemplo `C`, observa-se que a dispersão dada pelos valores $h_1 = 0.1$ e $h_2 = 0.9$ em q^2 gera valores superiores aos de μ e $\mu + \sigma^2$, o que indica que a dispersão das duas variáveis influi de forma mais significativa nos valores de q^2. As medidas baseadas μ, $\mu + \sigma^2$ e q^2 são descritas respectivamente nas subseções 4.3.1, 4.3.2 e 4.3.3.

4.3.1 Baseada em média simples

A medida baseada em média simples é inspirada na ANMI. Assim, a partir da Eq. 2.9, é intuitivo determinar a média simples entre os agrupamentos, aqui referida por `média global` para não haver confusão com a ANMI. A Eq. 4.4 define a medida *Global Average Normalized Mutual Information* (GANMI), em que os agrupamentos c_i e c_j são indexados por i e j, respectivamente. Essa medida computa a média da soma dos valores da NMI entre cada par de agrupamentos, utilizando a matriz de NMI M_{NMI}.

$$\text{GANMI}(\mathcal{C}) = \frac{1}{\binom{|\mathcal{C}|}{2}} \sum_{1 \leq i < j \leq |\mathcal{C}|} M_{\text{NMI}}(i, j) \tag{4.4}$$

4.3.2 Baseado em média e variância

No problema de encontrar *top-k* agrupamentos não-redundantes, deseja-se minimizar tanto a informação média compartilhada (redundância) entre os agrupamentos quanto a dispersão entre as médias. Com esse propósito, foi desenvolvida uma medida denominada Mean-Variance-NMI (MVNMI), baseada na média e na variância, inspirada pela Análise da Média-Variância (WANG, 2009), que por sua vez é baseada na Moderna Teoria de Portfolio (MPT) proposta pelo ganhador do Prêmio Nobel MARKOWITZ (1952).

A MPT é regularmente utilizada na área de finanças para encontrar um balanço entre a maximização de um potencial retorno de investimento e a minimização do risco. Baseado nesse princípio, WANG (2009) combina a média e a variância para gerar um ranking com os *top-k* documentos. Nesse cenário, foi definida uma medida baseada na variância.

A medida *Global Variance Normalized Mutual Information* (GVNMI) é definida na Eq. 4.5 e define a variância existente entre um conjunto de agrupamentos. Os agrupamentos c_i e c_j também são indexados por i e j, respectivamente. M_{NMI} é a matriz de NMI produzida com o cálculo da NMI entre cada par de agrupamentos.

$$\text{GVNMI}(\mathcal{C}) = \frac{1}{\binom{|\mathcal{C}|}{2}} \sum_{1 \leq i < j \leq |\mathcal{C}|} (GANMI(\mathcal{C}) - M_{\text{NMI}}(i, j))^2 \tag{4.5}$$

Assim, a MVNMI é apresentada na Eq. 4.6 como uma medida de avaliação

para selecionar os *top-k* agrupamentos não-redundantes. A MVNMI combina as medidas GANMI e GVNMI para avaliar e ranquear agrupamentos considerando a redundância.

$$\text{MVNMI}(\mathcal{C}) = \text{GANMI}(\mathcal{C}) + \text{GVNMI}(\mathcal{C}) \tag{4.6}$$

4.3.3 Baseada na média dos quadrados

A medida baseada na média dos quadrados é apresentada pela medida *sQuared Normalized Mutual Information* (QNMI) (Eq. 4.7), em que os agrupamentos c_i e c_j são indexados por i e j, respectivamente. A matriz de NMI é dada por M_{NMI}.

$$\text{QNMI}(\mathcal{C}) = \frac{1}{\binom{|\mathcal{C}|}{2}} \sum_{1 \leq i < j \leq |\mathcal{C}|} M_{\text{NMI}}(i,j)^2 \tag{4.7}$$

4.4 Algoritmo para ranquear agrupamentos não-redundantes

Essa seção apresenta uma alternativa para os casos em que grandes valores de $|\mathcal{C}|$ tornam intratáveis as abordagens de força bruta. Com essa finalidade, foi desenvolvido um algoritmo de busca denominado ***R**anking **M**ultiple **C**luste**r**ing **A**lgorithm*)(RM-CRAG), que constitui uma das contribuições da presente tese. O objetivo principal desse algoritmo é processar um conjunto de agrupamentos produzido pelo M-CRAG e selecionar os *top-k* agrupamentos não-redundantes.

O modelo apresentado na Figura 4.4 ilustra a arquitetura proposta para produção dos agrupamentos pelos algoritmos M-CRAG, CRAG e a posterior seleção dos *top-k* agrupamentos pelo RM-CRAG. O M-CRAG recebe como entrada um grafo com atributos nos vértices, processa esse grafo invocando o algoritmo CRAG, que produz um conjunto de agrupamentos. Esses agrupamentos resultantes são recebidos pelo RM-CRAG, que realiza a seleção dos *top-k* agrupamentos não-redundantes.

O RM-CRAG utiliza uma fila de prioridade $\mathcal{Q}$ no processo de seleção dos top-k agrupamentos de $\mathcal{C}$. Cada elemento de $\mathcal{Q}$ é internamente estruturado de acordo com a Tabela 4.3.

O RM-CRAG é inspirado no algoritmo A* (HART *et al.*, 1972) e possui uma função de custo, dada por $f(n) = g(n) + h(n)$. A função $g(n)$ estima a redundância e $h(n)$ estima a distância da posição atual até a posição final. Dessa forma, o custo para o elemento $e \in \mathcal{Q}$ é calculado utilizando a Eq. 4.8. Essa equação é decomposta em um custo do conjunto corrente de agrupamentos $e.\mathcal{C}$ e uma heurística que estima quantos passos são necessários para alcançar o número final de soluções k. Nessa

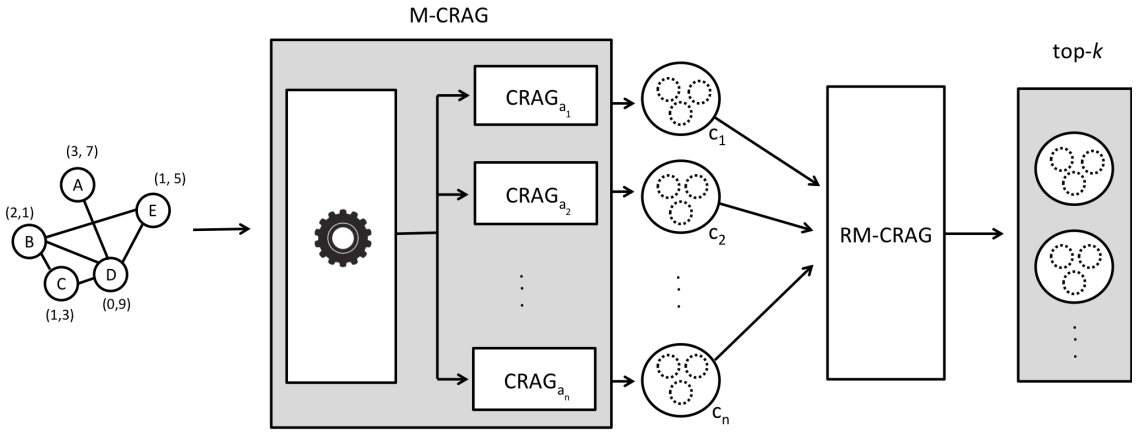


Figura 4.4: Modelo representando o funcionamento dos algoritmos CRAG, M-CRAG e RM-CRAG.

Tabela 4.3: Tipo de cada elemento na fila de prioridades.

Element {	
campo $\mathcal{C}$:	conjunto de agrupamentos
campo custo:	custo desse conjunto de agrupamentos (veja Eq. 4.8)
}	

equação, a função redundancy pode ser substituída por uma das equações propostas na seção 4.3.

$$\text{e.cost} = \underbrace{\text{redundancy}(e.\mathcal{C})}_{g(n)} + \underbrace{k - |e.\mathcal{C}|}_{h(n)} \tag{4.8}$$

No passo 1, o RM-CRAG calcula a matriz M_{NMI}. Do passo 2 ao passo 5, o algoritmo cria uma fila de prioridades de tamanho $|\mathcal{C}|$ em que cada elemento e corresponde a uma combinação de dois agrupamentos c_i e c_j. Isso é feito pela função appendTop, que lê a matriz M_{NMI} e escolhe o agrupamento que produz menos redundância quando combinado com c_i. Isso garante que cada $c_i \in \mathcal{C}$ apareça ao menos uma vez na fila de prioridades. Essa fila de prioridades é ordenada em ordem crescente de e.cost. Ao final do passo 5, o tamanho da fila de prioridade $\mathcal{Q}$ é dado por $|\mathcal{C}|$. No passo 6 o conjunto de $k = 2$ agrupamentos de menor custo é removido do topo da lista e atribuído à variável e.

Em seguida, do passo 7 ao 11, o RM-CRAG usa uma abordagem incremental que acrescenta um novo agrupamento ao conjunto de agrupamentos e, removido do topo da fila de prioridades. Novamente a função appendTop é invocada. Nesse momento, o primeiro parâmetro é dado por $e.\mathcal{C}$, ou seja, o subconjunto de agrupamentos de tamanho k que se encontrava no topo da lista. Em síntese, o principio é buscar por

Algoritmo 5: RM-CRAG($\mathcal{C}$, k)

1 **Input**:
- $\mathcal{C}$ = conjunto de soluções de agrupamento.
- k = número de agrupamentos múltiplos não-redundantes a serem retornados.

Output: $e.\mathcal{C}$ = *top-k* agrupamentos não-redundantes.

1: Create matrix M_{NMI}
2: **for** ($i = 1$ to $|\mathcal{C}|$) **do**
3: $\quad e.\,\text{appendTop}(\{c_i\}, \mathcal{C}, M_{\text{NMI}}, k)$
4: $\quad \mathcal{Q}.add(e)$
5: **end for**
6: $e \leftarrow \mathcal{Q}.head$
7: **while** $|e.\mathcal{C}| < k$ **do**
8: $\quad e.\,\text{appendTop}(e.\mathcal{C}, \mathcal{C}, M_{\text{NMI}}, k)$
9: $\quad \mathcal{Q}.add(e)$
10: $\quad e \leftarrow \mathcal{Q}.head$
11: **end while**
12: **return** $e.\mathcal{C}$

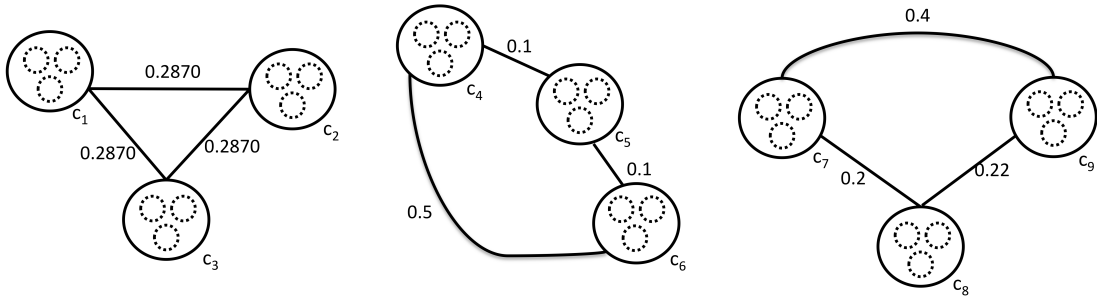
um agrupamento em $\mathcal{C}$ que acarrete menor valor de redundância quando combinado com $e.\mathcal{C}$. Esse valor de redundância é calculado pela Eq. 4.8. Em seguida, no passo 9, esse novo e é introduzido na fila de prioridades. Essa fila de prioridades é ordenada em ordem crescente de e. cost. Esse processo se repete até que o algoritmo obtenha o topo da fila e com tamanho igual a k. Por fim, o passo 12 retorna $e.\mathcal{C}$, que corresponde aos *top-k* agrupamentos não-redundantes.

Com relação à complexidade do algoritmo, o passo 1 é $O(p^2)$, os passos 2-5 são $O(p^2)$, o passo 6 é $O(1)$ e os passos 7-11 são $O(pk)$, $k \leq p$. Assim, a complexidade do RM-CRAG é $O(p^2)$.

4.5 Efeito das medidas de comparação de agrupamentos no ranqueamento

O RM-CRAG calcula o custo de k agrupamentos utilizando uma função de redundância denominada redundancy. Essa função pode utilizar diversas medidas de redundância, dentre elas, as propostas na Seção 4.3. O efeito prático das medidas estatísticas μ, $\mu + \sigma^2$ e q^2 pode ser ilustrado no exemplo da Figura 4.5. As Figuras 4.5a, 4.5b e 4.5c representam respectivamente três *conjuntos de agrupamentos* dados por $\mathcal{C}_1$, $\mathcal{C}_2$ e $\mathcal{C}_3$. Os pesos das arestas entre um par de agrupamentos c_i e c_j são os valores da NMI(c_i, c_j).

Dado esse exemplo, a proposta é selecionar o *conjunto de agrupamentos* $\mathcal{C}_i$ com menor redundância. A Tabela 4.4 apresenta os valores calculados para as diferentes



(a) $\mathcal{C}_1$. (b) $\mathcal{C}_2$. (c) $\mathcal{C}_3$.

Figura 4.5: Representação de três conjuntos de agrupamentos dados por $\mathcal{C}_1$, $\mathcal{C}_2$ e $\mathcal{C}_3$.

médias (μ, $\mu + \sigma^2$ e q^2) da NMI.

Tabela 4.4: Representação dos valores das médias para o conjunto de agrupamentos da Figura 4.5.

$\mathcal{C}_i$	$NMI(c_i, c_j)$	$NMI(c_i, c_j)$	$NMI(c_i, c_j)$	μ	$\mu + \sigma^2$	q^2
$\mathcal{C}_1$	0.2870	0.2870	0.2870	0.2870	0.2870	**0.0824**
$\mathcal{C}_2$	0.1000	0.5000	0.1000	**0.2333**	0.2867	0.0900
$\mathcal{C}_3$	0.2000	0.4000	0.2200	0.2733	**0.2855**	0.0828

Pode-se observar com base na Tabela 4.4 e na Figura 4.5 que dependendo da medida de média utilizada, o conjunto de agrupamentos menos redundante pode variar. Os valores na tabela destacados em negrito indicam os menores valores da média da NMI, apontando o conjunto de agrupamento com menor redundância utilizando a referida média.

Para o conjunto de agrupamentos dado $\mathcal{C}_1$, temos que $\mu = \mu + \sigma^2 = 0.2870$. Entretanto, esse agrupamento é o que produz o melhor valor para a NMI quando é utilizada a média dada por q^2. Pode-se também notar que o menor valor de média da NMI para μ é dado pelo conjunto de agrupamentos $\mathcal{C}_2$ e para $\mu + \sigma^2$, o conjunto de agrupamentos que apresenta menor redundância é dado por $\mathcal{C}_3$. Com isso, percebe-se que as diferentes medidas influenciam na seleção do conjunto dos *top-k* agrupamentos não-redundantes.

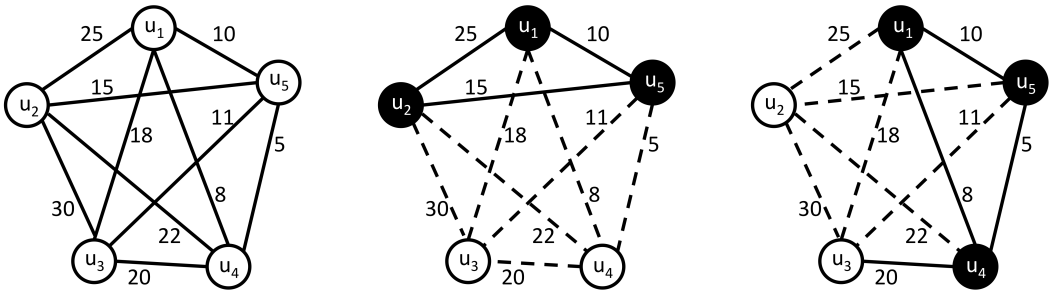
4.6 RM-CRAG e o problema de máxima diversidade

O problema de seleção dos top-k agrupamentos não-redudantes tratado pelo RM-CRAG pode ser caracterizado como o problema de máxima diversidade (MDP) (*Maximum Diversity Problem*, (GLOVER *et al.*, 1998)). O MDP é NP-difícil e e consiste em selecionar um subconjunto de m elementos a partir de um conjunto P

de n elementos ($m < n$) para maximizar a diversidade entre os elementos escolhidos (e.g., a soma das distâncias entre os elementos escolhidos é maximizada).

O MDP pode ser formulado como um problema de grafos (PALUBECKIS, 2007). Nessa formulação, cada elemento de P é representado como um vértice em um grafo não dirigido $G(V, E)$ em que $P = V$. Dados $p_1, p_2 \in P$, a distância entre p_1 e p_2 é utilizada como o peso ou custo associado à aresta $e = (p_1, p_2) \in E$. Assim, o MDP consiste em selecionar m vértices de forma que a soma das distâncias entre eles é maximizada. É importante destacar que a distância entre elementos de P pode ser formulada de diferentes maneiras.

Como exemplo, a Figura 4.6a apresenta um grafo com $n = 5$ vértices e dez arestas. Para selecionar um subgrafo de $m = 3$ vértices pode-se obter $\binom{5}{3} = 10$ configurações possíveis. As Figuras 4.6b e 4.6c retratam duas dessas configurações. As arestas possuem pesos que representam as distâncias (i.e. dissimilaridade, nesse caso) entre cada vértice. Os subgrafos de 3 vértices são ilustrados em preto, enquanto as arestas correspondentes são realçadas por linhas sólidas.



(a) Grafo com peso nas arestas.

(b) Possíveis três vértices selecionados.

(c) Outra configuração de vértices selecionados.

Figura 4.6: Ilustração do MDP em um grafo.

Formalmente, é apresentado um conjunto de n elementos, uma matriz simétrica $D = (d_{ij})$, onde d_{ij} é uma estimativa de diversidade (ou distância) entre os elementos i e j. O problema de máxima diversidade é dado pela seleção de ao menos m_1 e no máximo m_2 elementos do conjunto, de forma que a soma total das distâncias entre os elementos selecionados é máxima.

$$\text{maximizar } f(x) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij} x_i x_j \tag{4.9}$$

$$\text{sujeito à } m_1 \leqslant \sum_{i=1}^{n} x_i \leqslant m_2 \tag{4.10}$$

Quando os coeficientes $d_{ij} \geq 0$, $\forall\, i, j$, então a Eq. 4.10 pode ser substituida pela

seguinte equação:

$$\sum_{i=1}^{n} x_i = m \tag{4.11}$$

onde $m = m_2$.

Existem dois possíveis valores para x_i:

$$x_i = \begin{cases} 1, & \text{se a i-ésima solução é selecionada} \\ 0, & \text{caso contrário} \end{cases}$$

Dados os possíveis valores de x_i, o vetor representado na Figura 4.7 corresponde a um exemplo de $n = 10$ soluções de agrupamentos e a escolha de $m = 3$ soluções de agrupamento (x_0, x_1 e x_2).

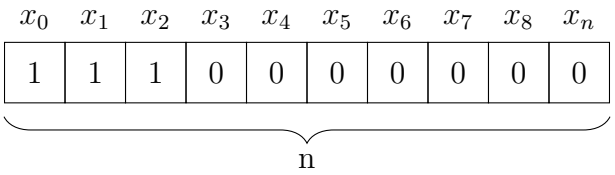


Figura 4.7: Vetor representando a escolha de m soluções de agrupamento.

No exemplo da Figura 4.6 pode-se calcular o somatório das distâncias entre os vértices selecionados empregando a Eq. 4.9. Para o grafo da Figura 4.6b, obtém-se o valor de f(x) = 25 + 15 + 10 = 50 e para o grafo da Figura 4.6c o valor de f(x) é f(x) = 8 + 5 + 10 = 23.

A escolha do subconjunto de vértices $\{u_1, u_2, u_5\}$ resulta em um valor de maior diversidade do que o subconjunto $\{u_1, u_4, u_5\}$. Entretanto, essas duas soluções foram selecionadas aleatoriamente a partir de um conjunto de soluções possíveis. Para a escolha de m vértices no problema do MDP, o subconjunto selecionado deve possuir máxima diversidade.

Na relação entre o MDP e o RM-CRAG, n é o número de agrupamentos gerados pelo $CRAG$, enquanto d_{ij} é a função de diversidade, sendo representada por $1 - NMI(c_i, c_j)$. O valor de m é dado pelo número de agrupamentos que deve ser obtido, ou seja k.

Conforme discutido no Capítulo 2, o problema da seleção de *top-k* agrupamentos se torna intratável com abordagens de força bruta à medida que n e m atingem valores consideráveis. Com isso, alguns métodos heurísticos e meta-heurísticos foram propostos para resolver o MDP. Esses métodos procuram soluções boas com um custo computacional aceitável, sem garantia de encontrar a solução ótima. Os métodos heurísticos procuram obter a solução de escolha dos m elementos

adicionando e removendo iterativamente um elemento a cada passo (MARTÍ *et al.*, 2013). Dentre esses métodos, podem ser destacados o ErkC (ERKUT, 1990) e I_LS (ARINGHIERI *et al.*, 2008).

Os métodos metaheurísticos pretendem obter soluções melhores que as obtidas pelas heurísticas tradicionais (MARTÍ *et al.*, 2013). Como exemplo, pode-se citar o Arrefecimento Simulado (*Simulated Anneling* (KIRKPATRICK *et al.*, 1983)), *Busca Tabu* (*Tabu Search*, TS (GLOVER e LAGUNA, 1997)) e Busca Tabu Iterativa (*Iterative Tabu Search*, ITS (PALUBECKIS, 2007)). Para uma discussão mais detalhada sobre esses e outros algoritmos propostos para o MDP, veja MARTÍ *et al.* (2013).

O Capítulo 5 realiza comparações entre o RM-CRAG e o ITS, descrito na literatura como um dos algoritmos com o melhor desempenho para resolver o problema do MDP (MARTÍ *et al.*, 2013; WANG, 2013; WU e HAO, 2013). O ITS é uma extensão do TS. Esse último é um método de busca metaheurístico para resolver o problema de análise combinatória, buscando explorar o espaço de solução além do ótimo local (GLOVER e LAGUNA, 1997). O ITS combina o TS com um mecanismo de perturbações para aprimorar a qualidade da solução.

Capítulo 5

Avaliações experimentais

Os experimentos apresentados nesse capítulo foram realizados em um computador Intel Core i5-3570 3.40GHz com 4 núcleos e com 16GB de memória RAM 1600 MHz. O sistema operacional utilizado foi o Linux Ubuntu 14.04.1 LTS x86_64. O código do ITS utilizado para produzir os experimentos foi a versão original proposta por PALUBECKIS (2007), implementada em linguagem C. Todos os algoritmos propostos nessa tese foram implementados em Java 6.

Nesse capítulo são apresentados os resultados encontrados com a utilização dos algoritmos CRAG e M-CRAG para a geração de agrupamentos múltiplos em três conjuntos de dados reais. Em seguida, são apresentados os resultados obtidos com o algoritmo RM-CRAG, cujo objetivo é selecionar os top-k agrupamentos não-redundantes a partir de um conjunto de agrupamentos $\mathcal{C}$. Esse capítulo se divide da seguinte maneira. Na Seção 5.1 são descritas as informações sobre os conjuntos de dados utilizados: `MQD500b`, `MQD500c` e `DBLP3000`. A Seção 5.2 descreve as medidas de avaliação utilizadas para analisar a qualidade dos grupos de cada agrupamento produzido pelos algoritmos CRAG e M-CRAG. Em seguida, a Seção 5.3 discute o comportamento dos algoritmos CRAG e M-CRAG nos conjuntos de dados supracitados. A Seção 5.4 analisa o comportamento do algoritmo RM-CRAG.

Um aspecto relevante na produção dos agrupamentos é o número de m grupos a serem formados em cada agrupamento c. O número de grupos apresentados nos experimentos realizados com o CRAG e o M-CRAG foi variado entre 2 e 20. Para determinar o número adequado de grupos, foi utilizado o joelho[1] obtido a partir da curva da função objetivo. O Joelho foi calculado utilizando os agrupamentos gerados utilizando apenas a estrutura topológica, ou seja, os atributos dos vértices não foram envolvidos. Os resultados apresentados para o RM-CRAG são discutidos com base nos agrupamentos com número de grupos m igual ao joelho da curva da função objetivo.

[1] O joelho de uma curva é definido como o ponto de máxima curvatura (MUNAGA *et al.*, 2012).

Ainda na Seção 5.4 é apresentada uma intuição visual sobre os agrupamentos selecionados pelo algoritmo RM-CRAG. É feita a apresentação de uma nuvem de palavras dos agrupamentos gerados para os conjuntos de dados `MQD500c` e `DBLP3000`, de forma que é possível comparar as diferenças entre os agrupamentos através das palavras que foram utilizadas pelos usuários e autores respectivamente. Ao fim dessa seção são discutidos os resultados apresentados para o RM-CRAG com as diferentes medidas baseadas na média (GANMI, MVNMI e QNMI). Por fim, a Seção 5.5 discute as considerações finais desse capítulo.

5.1 Conjuntos de dados

5.1.1 MQD500b

O conjunto de dados `MQD500b` (GUEDES, 2014) é formado por um grafo com atributos e foi gerado em 2014 a partir de uma rede social brasileira denominada (GUEDES, 2006). Esse conjunto de dados foi concebido empregando-se o conceito de *random walk* (WASSERMAN e FAUST, 1994) na componente principal da estrutura topológica do MQD para a obtenção de 500 vértices e suas arestas. Os vértices representam usuários anonimizados e cada aresta representa a relação de amizade entre dois usuários. A Tabela 5.1 apresenta algumas propriedades desse conjunto de dados.

Tabela 5.1: Propriedades do conjunto de dados MQD500b.

Propriedade	valor
Nós	500
Arestas	3.760
Atributos	12
Grau médio	15.040
Densidade	0.030
Diâmetro	4
Raio	4
Coeficiente de clusterização médio	$0,308$
Modularidade	$0,356$

Cada vértice possui um conjunto de 12 atributos: idade (I), extroversão (E), agradabilidade (A), neuroticismo (N), conscienciosidade (C), abertura (B), felicidade (F), tristeza (T), raiva (R), medo (M), nojo (O) e surpresa (S).

Os atributos relacionados à personalidade (idade, extroversão, agradabilidade, neuroticismo, conscienciosidade e abertura) foram extraídos de um teste de personalidade preenchido pelos usuários do MQD. Esse teste de personalidade foi baseado no modelo dos cinco grandes fatores da personalidade (PIEDMONT, 1998).

Os atributos relacionados às emoções (felicidade, tristeza, raiva, medo, nojo e surpresa) são provenientes das emoções que os usuários podem associar às entradas que escrevem no MQD. Essas emoções compreendem as 6 emoções básicas propostas por EKMAN (1992). Cada entrada pode ter apenas uma emoção associada.

5.1.2 MQD500c

O conjunto de dados `MQD500c` (GUEDES, 2015b) é composto por um subconjunto do `MQD500b` e foi gerado em 2015. A única diferença entre o `MQD500c` e o `MQD500b` é a quantidade de atributos em cada vértice. No `MQD500c`, os atributos relacionados às emoções e idade foram removidos. Assim, esse conjunto de dados possui apenas 5 atributos nos vértices, os relacionados à personalidade. Ambos os conjuntos de dados possuem a mesma estrutura topológica.

5.1.3 DBLP3000

Para a criação de um grafo com atributos a partir do conjunto de dados `DBLP3000` (GUEDES, 2015a) foi selecionado um subconjunto do conjunto de dados "DBLP four-area"[2] (SUN *et al.*, 2009).

DBLP four-area - Esse conjunto de dados é uma rede de co-autoria de pesquisadores em ciência da computação e contém 5 conferências representativas para cada uma das seguintes áreas: mineração de dados, banco de dados, recuperação da informação e aprendizado de máquina. Além disso, possui todos os 28.702 autores, suas publicações nessas conferências e os termos do extraídos dos títulos dos trabalhos, ou seja, possui quatro tipos de objetos: trabalho, conferência, autor e termos. Existem quatro tipos de relacionamento: autor-autor, autor-termo, conferência-autor e conferência-termo.

O `DBL3000` foi produzido em 2015 e foi gerado através da abordagem de *random walk* na componente principal do grafo até que fossem selecionados 3000 vértices. As arestas são representadas pela co-autoria dos autores. Foram selecionadas as três conferências menos esparsas de cada área, totalizando 12 conferências: *ijcai*, *aaai*, *icde*, *vldb*, *sigmodConference*, *sigir*, *icml*, *cikm*, *kdd*, *www*, *pakdd* and *icdm*. A Tabela 5.2 apresenta as propriedades desse conjunto de dados.

5.2 Medidas de avaliação

Para avaliar os agrupamentos gerados pelos algoritmos propostos na presente tese, foi utilizado a NMI conforme apresentado na Eq. 2.7, na Seçao 2.4. São desejáveis

[2]http://www.ccs.neu.edu/home/yzsun/data/DBLP_four_area.zip

Tabela 5.2: Propriedades do conjunto de dados DBLP5000.

Propriedade	Valor
Nós	3.000
Arestas	9.979
Atributos	12
Grau médio	6.653
Densidade	$0,002$
Diâmetro	15
Raio	8
Coeficiente de clusterização médio	$0,668$
Modularidade	$0,778$

baixos valores de NMI, visto que os baixos valores caracterizam agrupamentos mais distintos, ou seja, menos redundantes. É importante salientar que a NMI compara dois agrupamentos independentemente de quantos objetos existem em cada grupo.

Outro aspecto importante diz respeito à medida de qualidade de cada agrupamento produzido, ou seja, quão denso ou homogêneo são os grupos formados. Seguindo ZHOU *et al.* (2009), a densidade (Eq. 5.1) e a entropia (Eq. 5.2) são utilizadas para avaliar os agrupamentos produzidos pelo CRAG. A densidade representa a soma do número das arestas entre vértices presentes nos mesmos grupos dividido pelo número total de arestas. Isso resulta em valores entre 0 e 1, onde 0 é o pior valor, significando que todas as arestas são entre grupos. Por outro lado, 1 é o melhor valor, caracterizando que todas as arestas estão dentro dos grupos.

$$D(\{V_i\}_{i=1}^{m}) = \sum_{i=1}^{m} \frac{|\{(u_p, u_q)|u_p, u_q \in V_i, (u_p, u_q) \in E\}|}{|E|} \tag{5.1}$$

A entropia é uma medida de incerteza para uma distribuição de probabilidade. A Eq. 5.2 define a entropia normalizada do atributo a_i com relação a um conjunto de grupos $\{V_i\}$, onde $H(a_i, V_j)$ representa a entropia do atributo a_i no grupo V_j. O cálculo de $H(a_i, V_j)$ varia de acordo com o tipo de atributo. Os valores da entropia foram restringidos para o intervalo $[0, 1]$ atráves da constante de normalização $\mathcal{Z}$ Eq. 5.2.

$$\mathrm{S}(a_i, \{V_i\}_{i=1}^{m}) = \frac{1}{\mathcal{Z}} \sum_{j=1}^{m} \frac{|V_j|}{|V|} H(a_i, V_j) \tag{5.2}$$

Vale ressaltar que a densidade e a entropia são adequadas para as perspectivas topológica e relacional respectivamente. Assim, é esperado que os agrupamentos produzidos considerando apenas a estrutura topológica do grafo sejam bem avaliados com relação à densidade e os agrupamentos produzidos considerando apenas os atributos dos vértices seja bem avaliados com relação à entropia.

5.3 CRAG e M-CRAG

O objetivo do algoritmo M-CRAG é gerar múltiplos agrupamentos executando o algoritmo CRAG. Os resultados produzidos pelos algoritmos M-CRAG e CRAG são apresentados comparativamente com duas outras abordagens denomidadas `Strut` e `Atr`. Os agrupamentos representados por `Strut` foram produzidos utilizando um algoritmo de agrupamento espectral considerando apenas a estrutura topológica do grafo. Os agrupamentos retratados por `Atr` foram gerados apenas com as informações relacionais dos vértices (i.e., os atributos dos vértices). Nesse caso, foi calculada a distância euclidiana entre os valores dos atributos de cada vértice, sendo gerada uma matriz de tamanho $|V|x|V|$. Em seguida foi aplicado o algoritmo de agrupamento espectral nessa matriz. `Strut` e `Atr` podem ser apresentados como *baseline* das medidas de densidade e entropia respectivamente.

Na produção dos agrupamentos das abordagens `Atr` e `Strut`, os experimentos foram repetidos 20 vezes. Dessa forma, as curvas apresentadas nos gráficos refletem as médias obtidas em cada m. O tempo médio para a geração dos 19 agrupamentos (i.e., variando m entre 2 e 20) pode ser encontrado na Tabela 5.3.

Esses tempos foram calculados sem considerar a duração da tarefa de seleção dos vértices à distância dois. Para gerar os agrupamentos com o algoritmo CRAG, é necessário que sejam procurados, para cada vértice, os vértices à distância 2. Essa tarefa consome a maior parte do tempo necessário para o algoritmo convergir e por isso foi calculada uma vez para cada grafo, sendo gerado um arquivo com essa informação. Isso aumenta consideravelmente o desempenho obtido pelo algoritmo. Assim, como essa tarefa necessita ser realizada apenas uma vez para cada grafo, não foi incluída no tempo médio da produção de `CRAG`. O tempo do cálculo dos vértices à distância dois pode ser encontrado na última coluna da tabela.

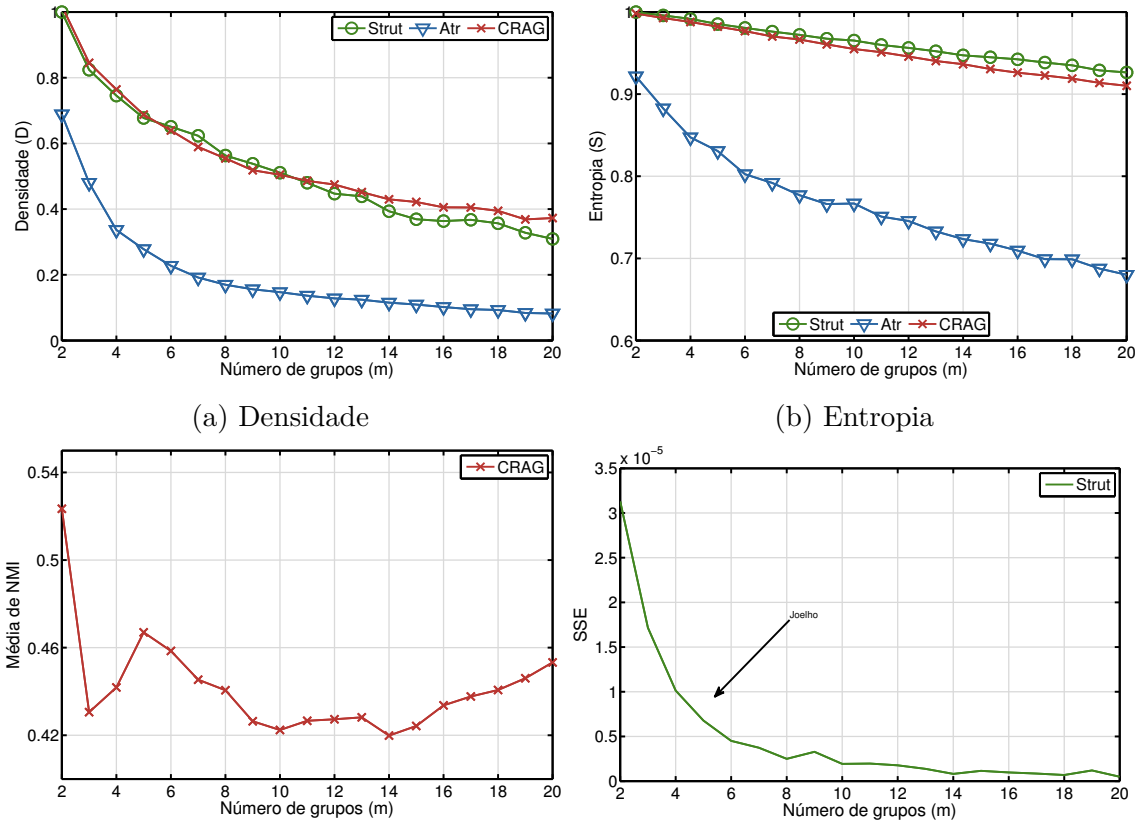
Tabela 5.3: Tempo para a geração dos agrupamentos `Atr`, `Strut`, `CRAG` e da distância 2 para os conjuntos de dados `MQD500b`, `MQD500c` e `DBLP3000`.

Conjunto de dados	Atr	Strut	CRAG	Distância 2
`MQD500b`	$0.09s$	$0.007s$	$0.041s$	$1.16m$
`MQD500c`	$0.09s$	$0.007s$	$0.032s$	$1.16m$
`DBLP3000`	$26.1m$	$21.3m$	$21.7m.$	$82m$

Pode-se notar que o tempo de processamento de `Atr` é superior ao tempo de processamento de `Strut` e `CRAG` nos três conjuntos de dados. Isso se deve ao cálculo da matriz de similaridade entre os objetos.

5.3.1 MQD500c

A Figura 5.1 apresenta os resultados obtidos pelo CRAG (`CRAG`) para o conjunto de dados `MQD500c`, assim como o joelho obtido para os valores da função objetivo dos agrupamentos produzidos apenas considerando a estrutura topológica (`Strut`).



(a) Densidade

(b) Entropia

(c) NMI

(d) Joelho

Figura 5.1: Avaliação da entropia, densidade e NMI dos agrupamentos gerados pelo algoritmo CRAG no conjunto de dados `MQD500c`.

Pode-se observar na Figura 5.1a que `CRAG` apresenta densidade semelhante a `Strut`, superando em alguns pontos (e.g., m=7). Isso evidencia que o algoritmo CRAG se comporta bem estruturalmente, visto que `Strut` é o *baseline* da avaliação estrutural. Em razão de `Atr` não levar em consideração a estrutura topológica, seus valores são os que apresentam pior densidade.

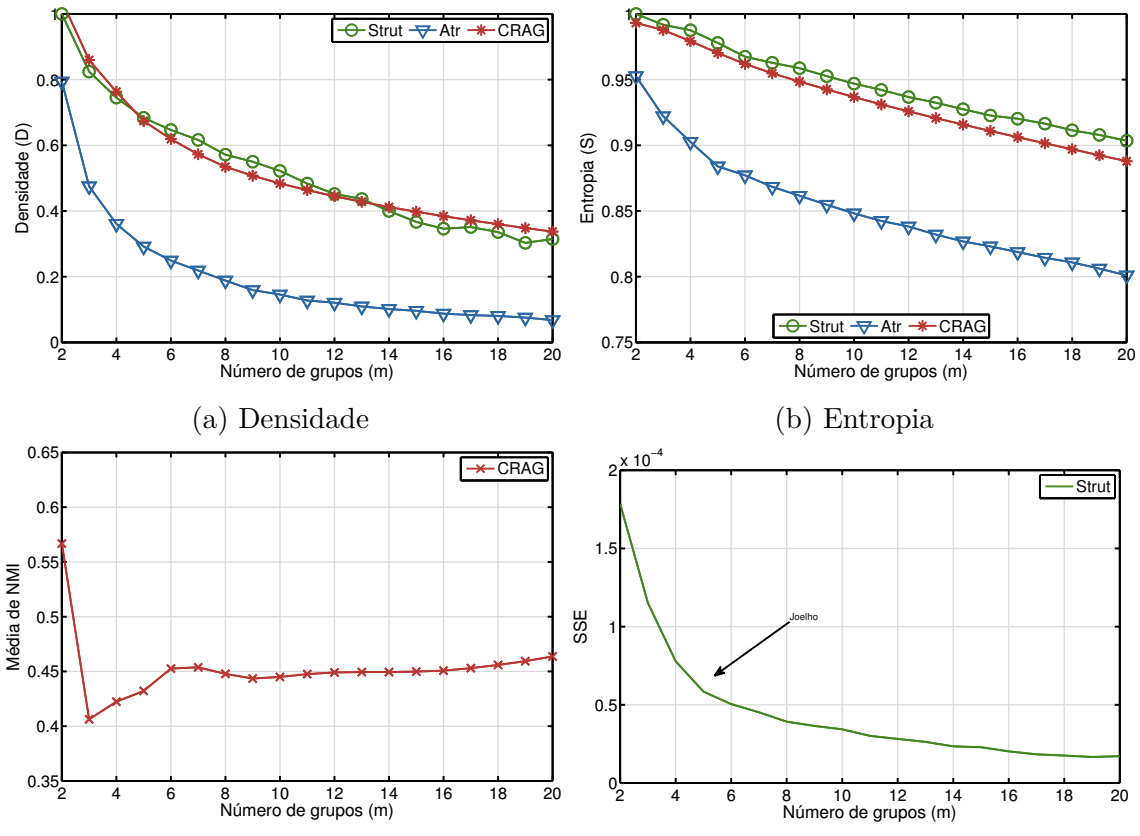
Na Figura 5.1b, nota-se que a entropia de `CRAG` apresenta valores intermediários entre `Strut` e `Atr`, revelando um resultado satisfatório. Isso indica que os grupos dos agrupamentos pelo algoritmo CRAG apresentam maior homogeneidade com relação aos grupos formados apenas utilizando a estrutura. A entropia de `Atr` apresentou o melhor valor, o que já era esperado, dado que `Atr` é o *baseline* para a entropia.

Uma característica significativa observada na Figura 5.1c é o baixo valor para a média do NMI computada para os agrupamentos gerados pelo algoritmo CRAG. Os valores se apresentam em um intervalo entre 0.43 e 0.53, apontando que os

agrupamentos possuem mais de 47% de diferença. Isso indica que a redundância entre os agrupamentos em alguns pontos é inferior a 58% ($m = 14$). Vale ressaltar que o número de grupos (m) identificado pelo joelho da curva da função objetivo é 5, conforme ilustra a Figura 5.1d.

5.3.2 MQD500b

Um comportamento semelhante foi observado nos resultados obtidos para o conjunto de dados `MQD500b`, confome ilustrado na Figura 5.2. A densidade de `CRAG` apresentou resultado comparável ao de `Strut`. Assim como os resultados para o conjunto de dados `MQD500c`, os resultados apresentados para o `MQD500b` apresentaram valores de entropia entre `Strut` e `Atr`, o que é considerado um resultado satisfatório.Com relação ao NMI médio, os valores se encontram entre 0.40 e 0.56, o que caracteriza uma média similaridade entre os agrupamentos gerados. No melhor caso ($m = 3$), o valor do NMI médio se aproxima de 0.4, indicando que os agrupamentos são em média 60% dissimilares. O joelho encontrado para esse conjunto de dados foi 5.



(a) Densidade

(b) Entropia

(c) NMI

(d) Joelho

Figura 5.2: Avaliação da entropia, densidade e NMI dos agrupamentos gerados pelo algoritmo CRAG no conjunto de dados MQD500b.

É interessante comparar os resultados produzidos com os conjuntos de dados `MQD500c` e `MQD500b`. A diferença entre ambos é a adição de 7 novos atributos

aos vértices do grafo (adicionados no MQD500b). Entretanto, os resultados dos agrupamentos produzidos em CRAG são bastante semelhantes. Para o MQD500c, o algoritmo CRAG produziu agrupamentos estruturalmente superiores (i.e., densidade superior) a Strut entre os valores de $m = 2$ e $m = 4$ assim como entre os valores de $m = 14$ e $m = 20$. Essa particularidade já foi evidenciada em ZHOU *et al.* (2009). Os autores apresentam resultados em que a combinação entre a estrutura topológica e os atributos dos vértices podem produzir agrupamentos com densidade superior aos agrupamentos realizados apenas com a estrutura topológica.

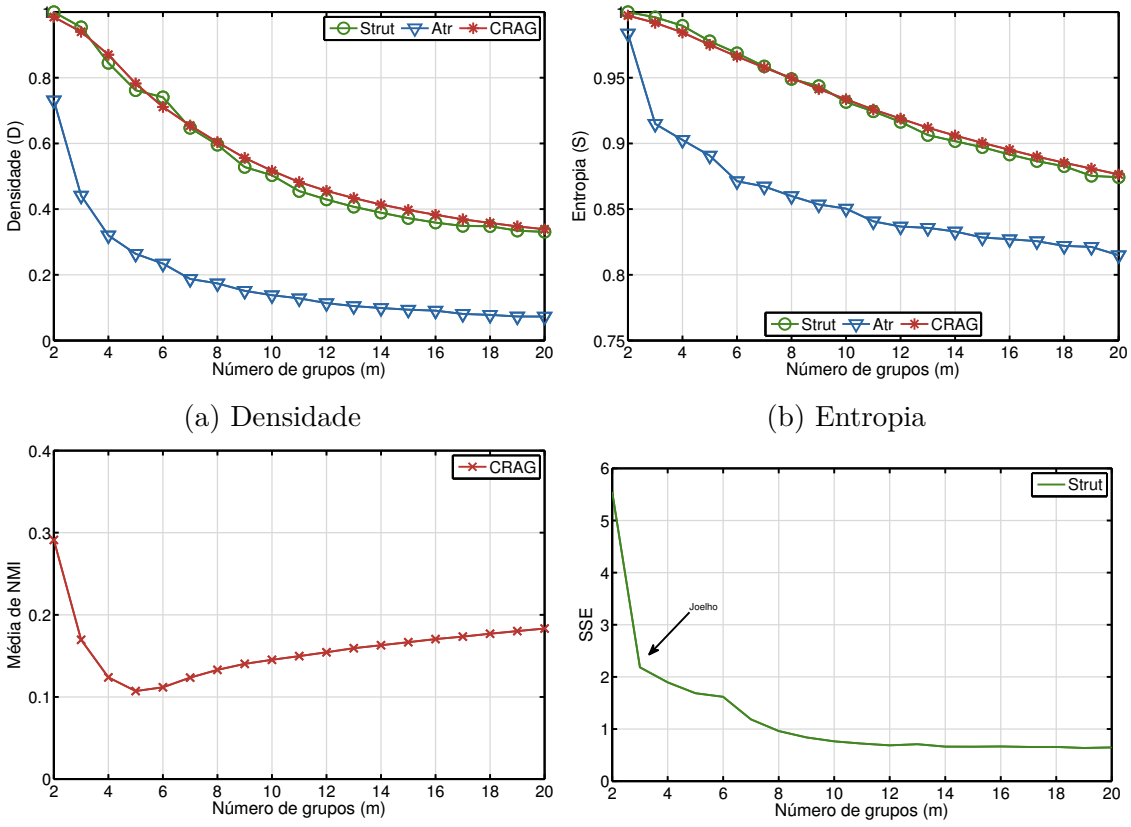
Com relação à entropia, a distância de CRAG para o Atr é similar para os dois conjuntos de dados (MQD500b e MQD500c), entretanto, para o conjunto MQD500b, apresenta valores ligeiramente mais próximos de Atr. Pode-se observar também uma pequena piora em alguns pontos das curvas da densidade do MQD500c para o MQD500b, por exemplo quando $m = 12$ e $m = 13$, entretanto, em termos gerais, ambos apresentam bons resultados. Vale à pena ressaltar que o joelho da curva se manteve em 5.

5.3.3 DBLP3000

O algoritmo CRAG também apresentou bons resultados para o conjunto de dados DBLP3000, conforme pode ser observado na Figura 5.3. Pode-se notar na Figura 5.3a que a densidade de CRAG supera em quase todos os momentos a densidade de Strut, apresentando resultado inferior apenas quando $m = 2$, $m = 3$ e $m = 6$. Como já era esperado, Atr apresentou os piores resultados.

Com relação à entropia (Figura 5.3b), a curva de CRAG apresenta valores intermediários entre Atr e Strut quando $m < 10$. Para os demais valores, houve uma leve piora na entropia com relação à Strut. Conforme já esperado, Atr apresentou o melhor resultado. Isso indica que a homogeneidade dos agrupamentos produzidos pelo algoritmo CRAG com $m < 10$ melhoraram com relação aos agrupamentos produzidos utilizando apenas a estrutura topológica.

Com relação ao NMI médio, os valores se encontram entre 0.1 e 0.3, o que caracteriza que os agrupamentos gerados pelo CRAG para esse conjunto de dados possuem baixa similaridade. No melhor caso ($m = 5$), nota-se que o valor do NMI médio se aproxima de 0.1, indicando que os agrupamentos são em média 90% dissimilares. O joelho para esse agrupamento foi encontrando em $k = 3$. É interessante ressaltar que para $k = 3$, os valores de CRAG se encontram entre Atr e Strut, tanto na densidade quanto na estrutura.



(a) Densidade

(b) Entropia

(c) NMI

(d) Joelho

Figura 5.3: Avaliação da entropia, densidade e NMI dos agrupamentos gerados pelo algoritmo CRAG no conjunto de dados DBLP3000.

5.4 RM-CRAG

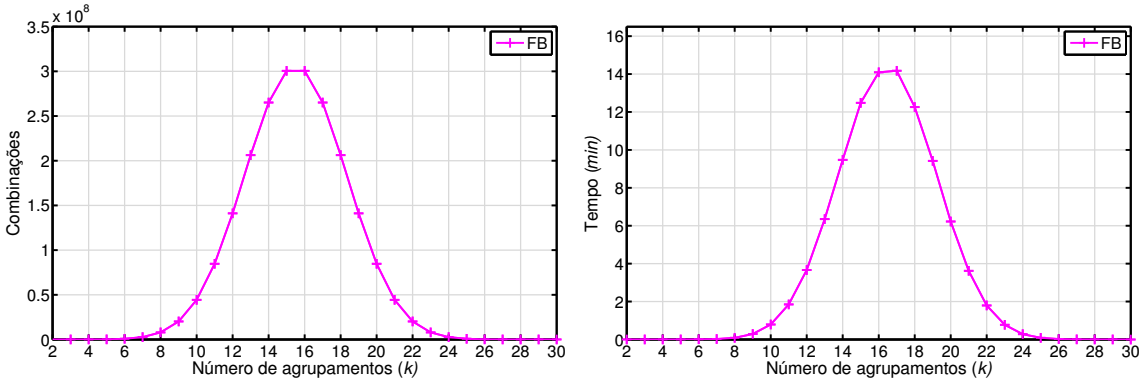
Essa seção apresenta os resultados obtidos utilizando o algoritmo RM-CRAG para a seleção dos top-k agrupamentos não-redundantes existentes em $\mathcal{C}$. Os resultados apresentados pelo RM-CRAG foram comparados com o estado da arte na área do MDP, o ITS e com a abordagem aleatória. Para essa análise foram observados a média do NMI e o tempo. Para plotar os gráficos, utilizamos valores de k variando entre 2 e 30.

5.4.1 MQD500c

Os estudos realizados com o `MQD500c` são considerados como *baseline* para os experimentos efetuados no decorrer desse trabalho, em razão do número reduzido de atributos existentes nesse conjunto de dados. Devido a essa característica, foi possível aplicar a abordagem de força bruta (FB) para a escolha dos *top-k* agrupamentos não-redundantes. Conforme o número de atributos cresce, a abordagem de força bruta tende a se tornar intratável, visto que o número de agrupamentos produzidos pelo CRAG cresce conforme $|\mathcal{C}| = 2^{|\lambda|} - 1$. Com isso, são

necessárias $\binom{|\mathcal{C}|}{k}$ operações de comparação para a escolha dos *top-k* agrupamentos não-redundantes.

A Figura 5.4a apresenta o número de operações necessárias para a seleção dos *top-k* agrupamentos não-redundantes utilizando a abordagem de força bruta e a Figura 5.4b apresenta o tempo consumido nessas operações. Pode-se notar que as duas figuras possuem aparência bastante semelhante, indicando que quanto maior o número de combinações, maior é o tempo necessário para a convergência do resultado.



(a) Número de combinações.

(b) Tempo.

Figura 5.4: Avaliação do tempo e quantidade de combinações para a seleção dos *top-k* agrupamentos gerados pela abordagem de força bruta (`FB`) no conjunto de dados `MQD500c` para $m = 5$.

A Tabela 5.4 apresenta os *top-k* agrupamentos não-redundantes identificados pela abordagem de força bruta para $2 < k <= 10$. Essa tabela reflete o comportamento do RM-CRAG em um conjunto de dados com a abordagem de força bruta, servindo de base para analisar o comportamento desse algoritmo nos demais conjuntos de dados. Os agrupamentos gerados com a combinação de atributos é caracterizada com os rótulos juntos (e.g., CA representa os agrupamentos gerados utilizando a similaridade de **c**onscienciosidade e **a**gradabilidade).

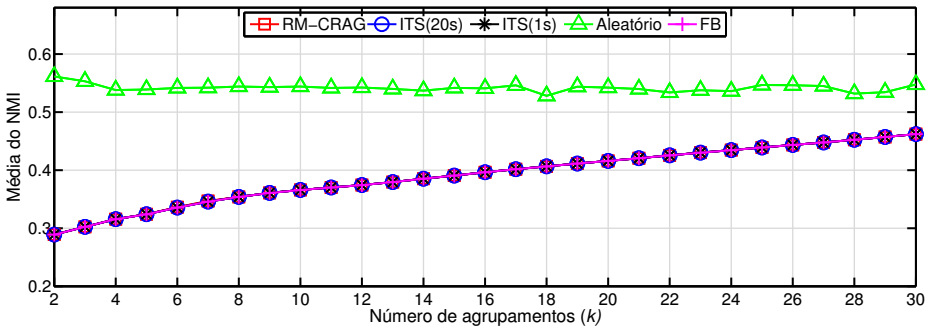
Os estudos conduzidos com o conjunto de dados `MQD500c` foram constuídos de maneira a comparar o algoritmo RM-CRAG com o ITS, a abordagem aleatória e a abordagem de força bruta (FB). O ITS foi configurado com os parâmetros descritos em PALUBECKIS (2007). Assim, o tempo máximo para convergência desse algoritmo foi determinado em $20s$, conforme utilizado pelos autores. Além disso, foi produzida uma curva limitando o tempo do ITS com o menor tempo aceitável para convergência, ou seja, $1s$. Para a abordagem aleatória, foram realizadas 20 execuções e a média dos resultados obtidos foi calculada.

A Figura 5.5a apresenta a média do NMI obtida pela escolha de k agrupamentos e a Figura 5.5b exibe o tempo necessário para as cinco abordagens realizarem a

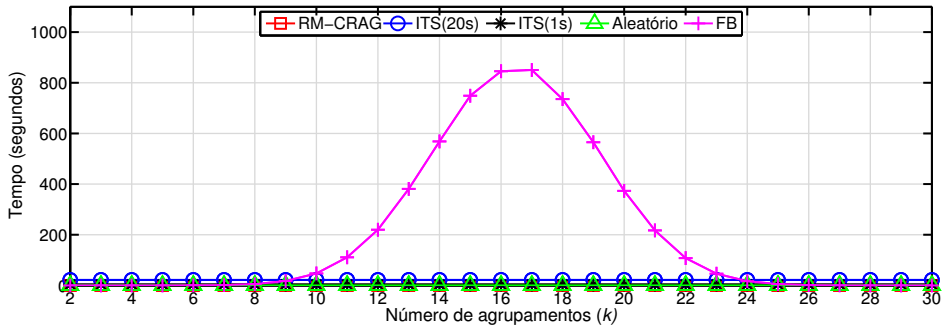
Tabela 5.4: Agrupamentos identificados pela abordagem de força bruta para $m = 5$.

k	Agrupamentos
2	A, CNA
3	A, E, CNA
4	C A E CNA
5	C A E CNA EN
6	C A E CNA ENB NB
7	C A E CNA ENB NB B
8	C A E CNA ENB NB B EN
9	C A E CNA ENB NB B EN CB
10	C A E CNA ENB NB B EN CB NA

escolha dos *top-k* agrupamentos para o `MQD500c`. Pode-se notar que a abordagem aleatória possui valores inferiores para a média do NMI (mais altos) e superiores no tempo de processamento (mais baixo). As curvas de `RM-CRAG`, `FB`, `ITS(20s)` e `ITS(1s)` são idênticas com relação à média do NMI, indicando que os agrupamentos selecionados pelas quatro abordagens em cada valor de k são os mesmos. Isso foi validado de acordo com os rótulos dos agrupamentos exibidos. Os agrupamentos da Tabela 5.4 também foram encontrados em `RM-CRAG`, `ITS(20s)` e `ITS(1s)`. Assim, pode-se afirmar que, para esse conjunto de dados, tanto o algoritmo RM-CRAG quanto o ITS encontraram os valores encontrados por `FB`.



(a) Média de NMI.



(b) Tempo.

Figura 5.5: Comparação entre `RM-CRAG`, `ITS(20s)` e `ITS(1s)`, `FB` e Aleatório no conjunto de dados `MQD500c` para $m = 5$.

Com relação ao tempo de processamento utilizado pelas abordagens comparadas, o RM-CRAG apresentou superioridade ao ITS(20s), ITS(1s) e força bruta. A abordagem aleatória apresentou o melhor tempo. A abordagem de força bruta apresenta bons resultados para k menores que 12 e maiores que 22. Por outro lado, produziu valores altos quando k variou de 12 a 22, chegando a consumir mais de 14 vezes o tempo das demais abordagens. Para esse conjunto de dados, o ITS(20s) apresentou o pior tempo enquanto o ITS(1s) superou parcialmente o tempo da abordagem de força bruta. Portanto, o RM-CRAG apresentou o melhor desempenho quando considerados o tempo e a média de NMI.

5.4.2 MQD500b

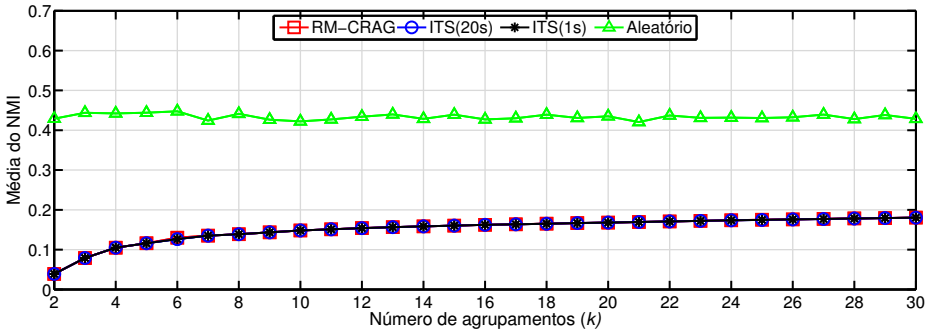
Nos experimentos realizados com o conjunto de dados `MQD500b` não são apresentadas as curvas referentes à abordagem de força bruta, visto que o número de atributos nos vértices desse conjunto de dados é 12. Isso faz com que a quantidade de agrupamentos gerados seja $\binom{12}{2} - 1$, isto é, 4.095. Por conseguinte, combinações envolvendo valores de $k > 3$ já começam a se tornar computacionalmente intratáveis, considerando que $\binom{4.095}{4} \approx 1.17E + 13$.

Os resultados apresentados na Figura 5.6a indicam que o RM-CRAG, ITS(20s) e ITS(1s) apresentaram resultado bastante semelhante com relação à média do NMI. Percebe-se uma pequena diferença para o valor de $k = 6$, em que houve uma modesta melhora do ITS(20s) e ITS(1s) com relação ao RM-CRAG. Para os demais valores de k, as curvas se sobrepõem. A abordagem aleatória apresentou o pior resultado.

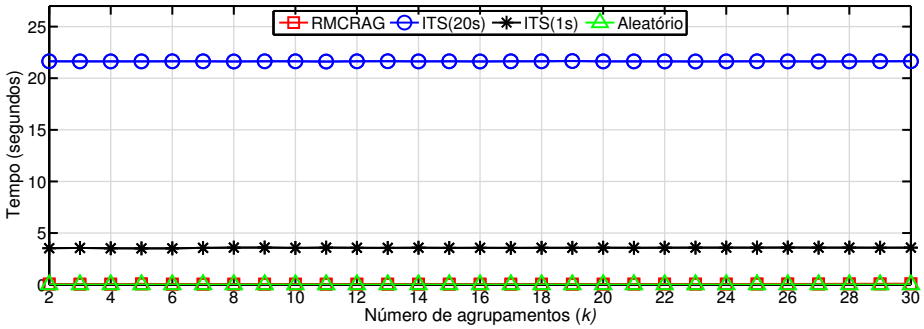
A Figura 5.6b ilustra o tempo de processamento consumido pelas quatro abordagens. Pode-se notar que a abordagem aleatória consumiu o menor tempo de processamento. `RM-CRAG` apresenta uma leve piora com relação à abordagem aleatória, entretanto consumiu consideravelmente menos tempo de processamento do que `ITS(20s)` e `ITS(1s)`.

Levando em conta o tempo de processamento e os resultados apresentados para o NMI, nota-se que o RM-CRAG é o algoritmo que obteve melhor desempenho. Os resultados apresentados pelo RM-CRAG, ITS(20s), ITS(1s) e a abordagem aleatória para o conjunto de dados `MQD500b` foram bastante semelhantes aos apresentados para o `MQD500c`. Em ambos os conjuntos de dados, o RM-CRAG foi superior em termos gerais. Pode-se notar que os valores de NMI médio alcançados para o `MQD500b` são consideravelmente inferiores aos apresentados para o `MQD500c`. É importante destacar que foram produzidos 31 agrupamentos para o `MQD500c` e 4.095 para o `MQD500b`. Com isso, espera-se que haja maior diversidade entre os agrupamentos produzidos e, consequentemente, entre os agrupamentos selecionados pelo RM-CRAG. Com isso, o número consideravelmente superior de

agrupamentos produzidos para o MQD500b permitiu que o RM-CRAG, ITS(20s) e ITS(1s) selecionassem agrupamentos significativamente menos redundantes quando comparados aos agrupamentos selecionados para o MQD500b.



(a) Média de NMI.



(b) Tempo.

Figura 5.6: Comparação entre o RM-CRAG, ITS(20s) e ITS(1s) e Aleatório para o conjunto de dados MQD500b para $m = 5$.

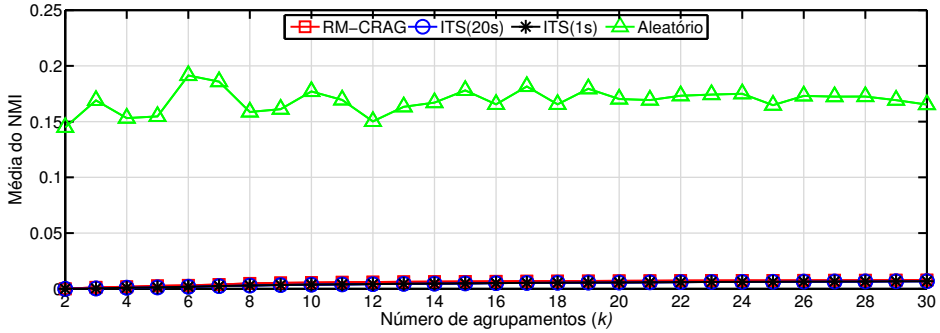
5.4.3 DBLP3000

Os experimentos realizados com o DBLP3000 possuem o objetivo de avaliar o comportamento do RM-CRAG em um conjunto de dados que não seja uma rede social. Esse conjunto de dados é uma rede de co-autoria modelada como uma rede social, na qual existe uma aresta entre dois vértices se os autores representados por esses vértices são co-autores em alguma publicação.

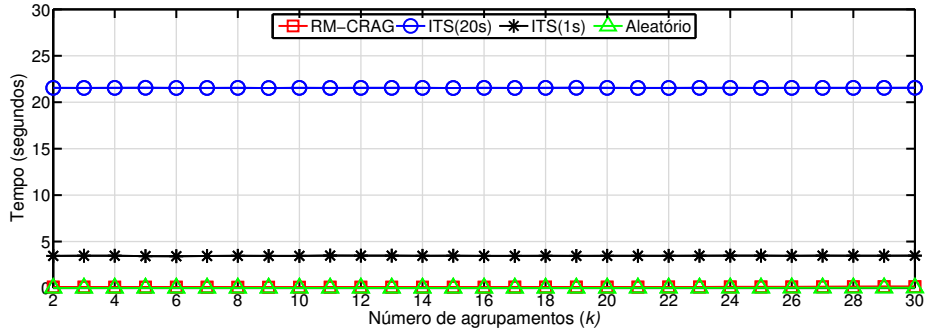
É possível notar pelo gráfico ilustrado na Figura 5.7a que as curvas do RM-CRAG, ITS(20s) e ITS(1s) praticamente se sobrepõem com relação à média do NMI. Pode-se notar uma ligeira superioridade de ITS com relação a RM-CRAG. Conforme já era esperado, a abordagem aleatória apresentou o pior resultado.

A Figura 5.7b ilustra o tempo de processamento das quatro abordagens. A curva RM-CRAG novamente supera o tempo de processamento de ITS(20s) e ITS(1s). A abordagem aleatória possui tempo ligeiramente inferior a RM-CRAG. Isso evidencia que o algoritmo RM-CRAG também apresenta performance superior na combinação

dos resultados do tempo e da média do NMI.



(a) Média de NMI.



(b) Tempo.

Figura 5.7: Comparação entre `RM-CRAG`, `ITS(20s)` e `ITS(1s)` e Aleatório para o conjunto de dados `DBLP5000` para $m = 3$.

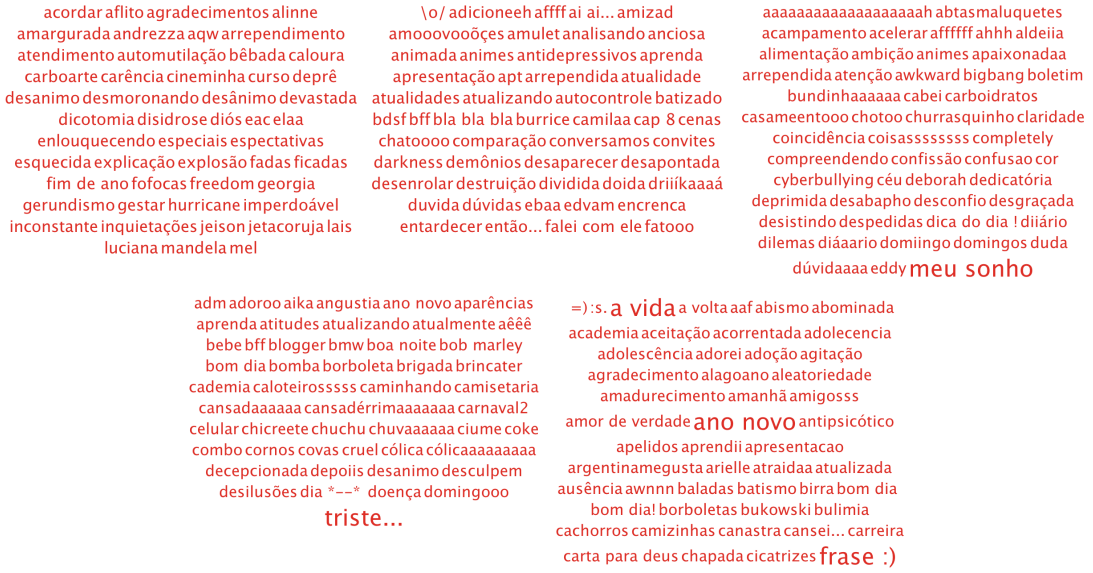
Os algoritmos propostos nessa tese como contribuições (CRAG, M-CRAG e RM-CRAG) apresentaram bons resultados na produção de agrupamentos em grafos com atributos e a posterior seleção dos *top-k* agrupamentos não-redundantes. Isso foi evidenciado pelas medidas de densidade, entropia e NMI. Entretanto, com essas medidas não é possível ter uma intuição visual com relação aos elementos dentro dos grupos. A seção seguinte apresenta uma análise das palavras utilizadas pelos indivíduos em cada grupo nos conjuntos de dados `MQD500b` e `DBLP3000`.

5.4.4 Análise por nuvem de palavras

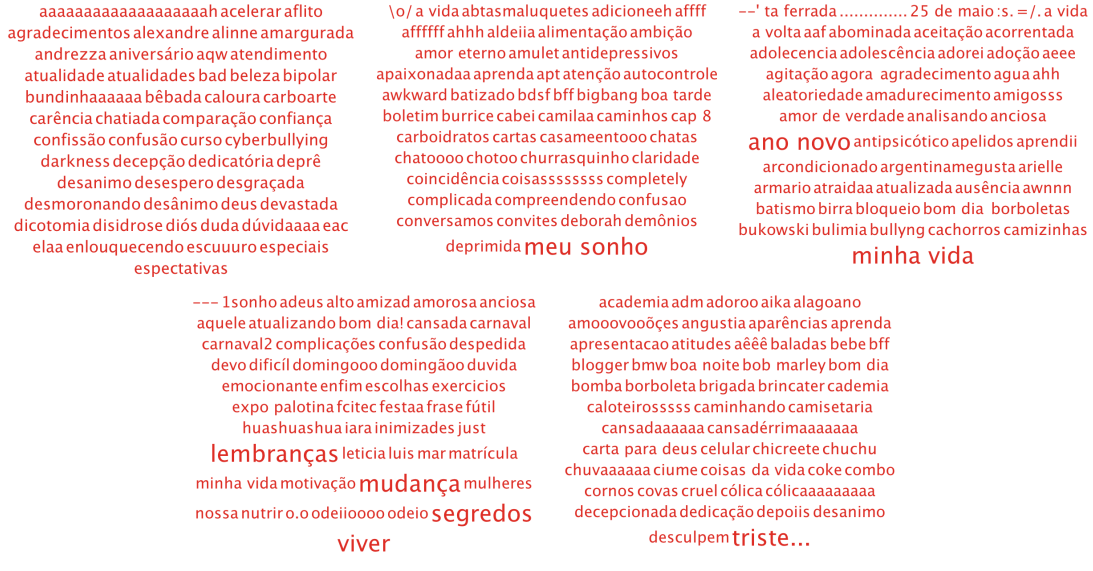
Os resultados apresentados pelos algoritmos propostos apontaram bons resultados em termos de produção dos *top-k* agrupamentos não-redundantes. O objetivo dessa seção é exibir uma intuição visual sobre cada grupo dos *top-k* agrupamentos utilizando a nuvem de palavras dos grupos. O tamanho que cada palavra aparece na nuvem é proporcional à quantidade de vezes que a mesma foi utilizada nos grupos. Além disso, foi produzida uma matriz de similaridade utilizando o índice de similaridade de Jaccard entre as palavras utilizadas em cada grupo.

A Figura 5.8 exibe os grupos gerados para os dois agrupamentos selecionados pelo

RM-CRAG para o conjunto de dados `MQD500b`. Para apresentar a nuvem de palavras desse conjunto de dados, foram selecionados todos os usuários de cada grupo, assim como os títulos das entradas inseridas no `MQD` por cada um deles. Em seguida, as palavras mais freqüentes (*stopwords*) foram removidas utilizando o teorema de Pareto, ou seja, 20% do total das palavras foram removidos. Para a exibição da nuvem de palavras, foram apresentadas as 50 palavras de maior ocorrência.



(a) Nuvem de palavras para o agrupamento gerado pelo CRAG com arestas artificiais produzidas com base nos atributos *idade*, *conscienciosidade*, *tristeza* e *agradabilidade*.



(b) Nuvem de palavras para o agrupamento gerado pelo CRAG com arestas artificiais criadas com base nos atributos *idade*, *extroversão*, *tristeza*, *nojo*, *abertura* e *agradabilidade*.

Figura 5.8: Comparação entre as nuvens de palavras geradas para os top-2 agrupamentos selecionados pelo RM-CRAG com $m = 5$ para o conjunto de dados `MQD500b`.

O agrupamento da Figura 5.8a foi gerado com a criação de arestas artificiais utilizando os atributos *idade*, *conscienciosidade*, *tristeza* e *agradabilidade*. O agrupamento da Figura 5.8b foi gerado com a criação de arestas artificiais utilizando os atributos *idade*, *extroversão*, *tristeza*, *nojo*, *abertura* e *agradabilidade*.

Pode-se observar que os agrupamentos possuem grupos com diferenças significativas entre as palavras utilizadas. No entanto, para apresentar um resultado mais apurado, foi calculada a similaridade entre os grupos dos agrupamentos, conforme ilustrado na Figura 5.9. Cada grupo de um agrupamento é comparado com os grupos do outro agrupamento, gerando uma matriz quadrada de $m \times m$.

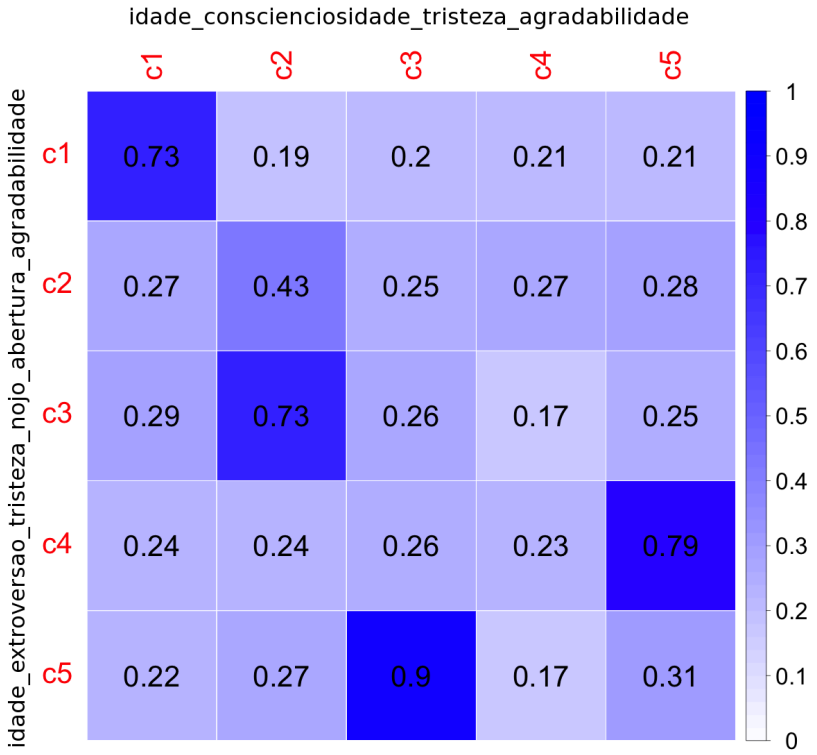


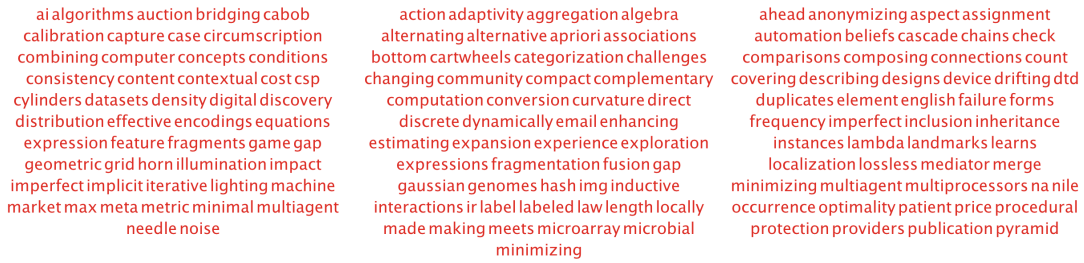
Figura 5.9: Matriz de similaridade entre as palavras presentes nos grupos dos agrupamentos selecionados pelo RM-CRAG para o conjunto de dados `MQD500b`.

É possível notar que a maioria dos grupos possui similaridade inferior a 0.3, o que indica uma baixa similaridade entre as palavras utilizadas pelos usuários nos agrupamentos. Apenas quatro comparações apresentaram valores superiores a 0.7. Considerando que apenas 20% de arestas artificiais foram criadas, esses valores representam um resultado satisfatório.

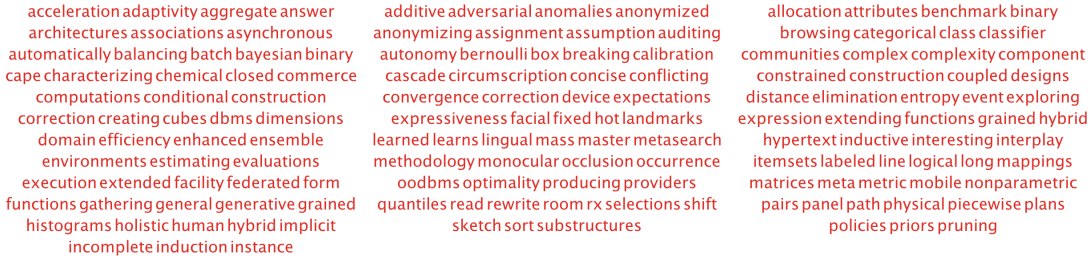
Para produzir a nuvem de palavras do conjunto de dados `DBLP3000`, os autores de cada grupo foram identificados e as palavras que os mesmos utilizaram nos títulos de seus trabalhos foram selecionadas. A Figura 5.10b ilustra a nuvem de palavras obtida para esse conjunto de dados para $m = 3$. É importante destacar que, nesse

conjunto de dados, também foram removidas 20% das palavras de maior ocorrência.

O agrupamento da Figura 5.10a foi gerado com a criação de arestas artificiais utilizando os atributos *ijcai*, *vldb*, *icml* e *icdm*. O agrupamento da Figura 5.10b foi gerado com a criação de arestas artificiais utilizando os atributos *aaai*, *icde*, *vldb*, *sigmodconference*, *sigir*, *icml*, *cikm*, *pakdd* e *icdm*. Assim como ocorreu no conjunto de dados `MQD500b`, nota-se a ocorrência de palavras distintas nos grupos dos diferentes agrupamentos. Embora algumas palavras podem ser encontradas em grupos dos dois agrupamentos, nenhum dos grupos exibidos na Figura 5.10a é exatamente igual a algum grupo da Figura 5.10b.



(a) Nuvem de palavras do agrupamento gerado pelo CRAG com arestas artificiais criadas com base nos atributos *ijcai*, *vldb*, *icml* e *icdm*.



(b) Nuvem de palavras do agrupamento gerado pelo CRAG com arestas artificiais criadas com base nos atributos *aaai*, *icde*, *vldb*, *sigmodconference*, *sigir*, *icml*, *cikm*, *pakdd* e *icdm*.

Figura 5.10: Comparação entre as nuvens de palavras geradas para os top-2 agrupamentos selecionados pelo RM-CRAG com $m = 3$ para o conjunto de dados `DBLP3000`.

A matriz de correlação entre as palavras utilizadas pelos autores de cada grupo é ilustrada na Figura 5.11. A maioria dos valores do índice de Jaccard se encontram abaixo de 0.25. Apenas a comparação entre os grupos c_3 e c_2 supera esse valor (i.e. 0.85). Esse resultado indica que existe uma diferença significante entre os grupos dos diferentes agrupamentos, o que é um resultado satisfatório. Dessa forma, além do NMI, os experimentos realizados com a nuvem de palavras e a matriz de similaridade entre as palavras evidenciam a diferença existente entre os agrupamentos produzidos pelos algoritmos elaborados nessa tese.

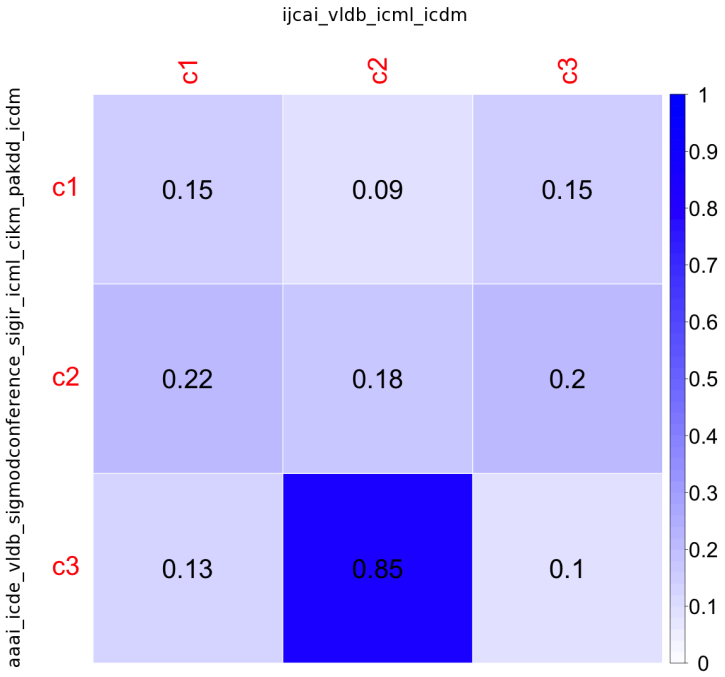


Figura 5.11: Comparação entre as nuvens de palavras geradas para os top-2 agrupamentos selecionados pelo RM-CRAG com $m = 3$ para o conjunto de dados DBLP3000.

.

5.4.5 Análise das diferentes medidas

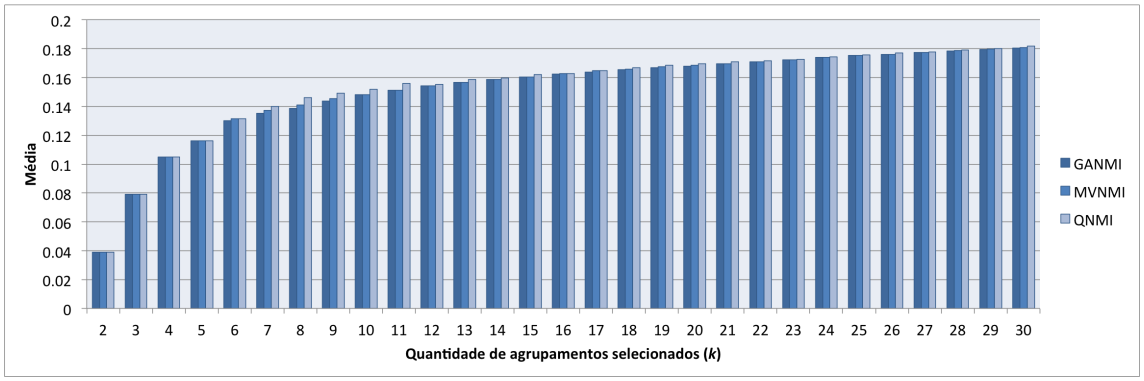
Nessa seção é feita uma análise comparativa entre os resultados apresentados pelas diferentes medidas de média apresentadas na Seção 4.3. O algoritmo RM-CRAG foi executado utilizando cada uma das três medidas: GANMI, MVNMI e QNMI. Essa análise foi realizada apenas para os conjuntos de dados `MQD500b` e `DBLP3000`. O conjunto de dados `MQD500c` possui uma quantidade pequena de atributos (5) e, portanto, poucos agrupamentos foram gerados (31). Isso inviabiliza uma análise mais detalhada a respeito dos agrupamentos selecionados pelo RM-CRAG nesse conjunto de dados.

Os resultados do algoritmo RM-CRAG com as diferentes medidas foram analisados em termos de duas medidas estatísticas: média e variância. Para cada valor de k foi calculado o valor da NMI entre todos os agrupamentos dois a dois. Em seguida, foi calculada a média e variância desses valores. As execuções do RM-CRAG com as medidas GANMI, MVNMI e QNMI são denominadas `RM-CRAG(GANMI)`, `RM-CRAG(MVNMI)` e `RM-CRAG(QNMI)` respectivamente. Também é analisado o número de agrupamentos distintos que foram selecionados com a utilização de cada uma das medidas.

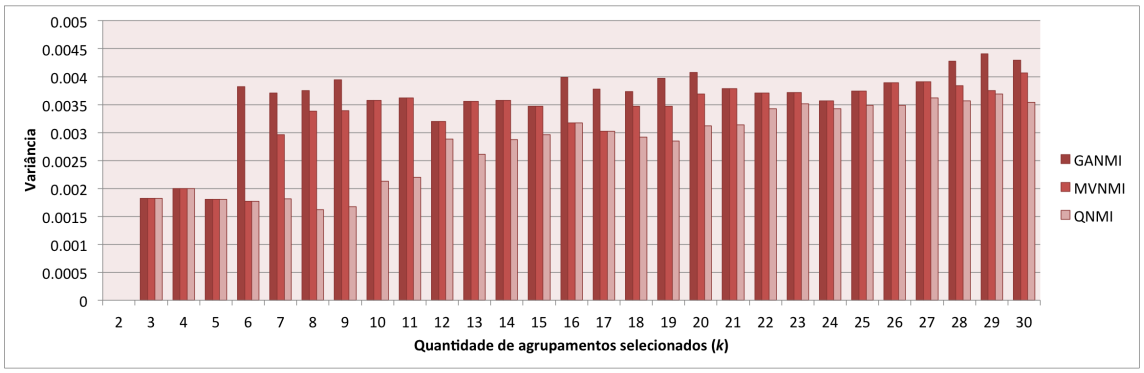
MQD500b

A Figura 5.12 apresenta a média e a variância entre os *top-k* agrupamentos selecionados pelo algoritmo RM-CRAG com cada uma das medidas apresentadas na Seção 4.3. Pode-se notar que quase não houve diferença entre as médias da NMI das três medidas. A medida QNMI apresentou uma pequena piora na média quando relacionada com as outras medidas. Por outro lado, a MVNMI e GANMI apresentam resultado praticamente igual para os diferentes valores de k.

Com relação à variância, as três medidas se comportam da mesma maneira no intervalo $2 < k < 6$. Para valores de $k > 5$, a QNMI apresenta sempre o menor valor de variância, o que indica que os agrupamentos são homogeneamente distintos. Em quase todos os valores de k, a medida MVNMI apresentou valores de variância entre GANMI e QNMI.



(a) Média



(b) Variância

Figura 5.12: Média e variância da NMI encontrada entre os *top-k* agrupamentos selecionados pelo RM-CRAG utilizando as diferentes medidas GANMI, MVNMI e QNMI para o conjunto de dados `MQD500b`.

A Figura 5.13 ilustra a quantidade de agrupamentos diferentes encontrados quando o RM-CRAG foi executado com a GANMI, MVNMI e QNMI. Tanto a utilização da MVNMI quanto a QNMI não alteraram a seleção dos agrupamentos selecionados para valores de $k < 6$. A MVNMI acarretou um número relativamente baixo de agrupamentos selecionados diferentes da GANMI, em seu caso mais

relevante ($k = 8$), ou seja, 25% dos agrupamentos. Para valores entre $6 <= k <= 30$, é possível notar que a utilização da medida QNMI selecionou diversos agrupamentos distintos. Esse valor chega a 4 quando $k = 9$, ou seja, 44% dos agrupamentos selecionados com a medida QNMI são diferentes dos agrupamentos selecionados com a GANMI.

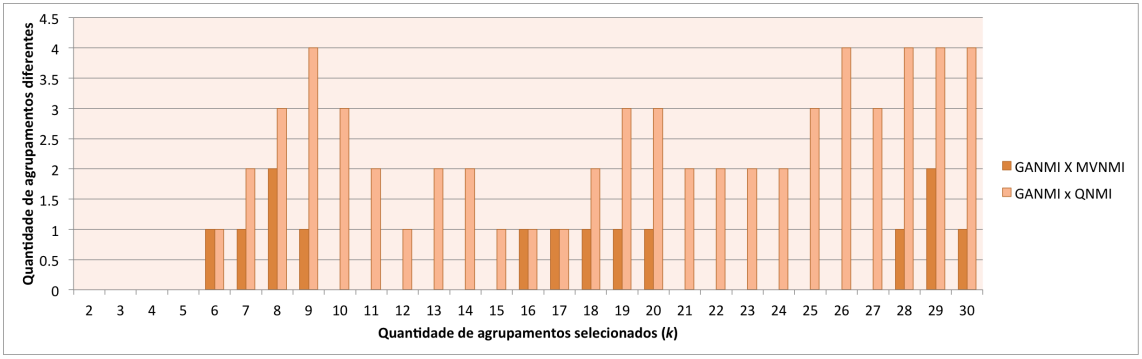


Figura 5.13: Quantidade de agrupamentos diferentes selecionados pelo RM-CRAG utilizando as diferentes medidas de média para o conjunto de dados `MQD500b`.

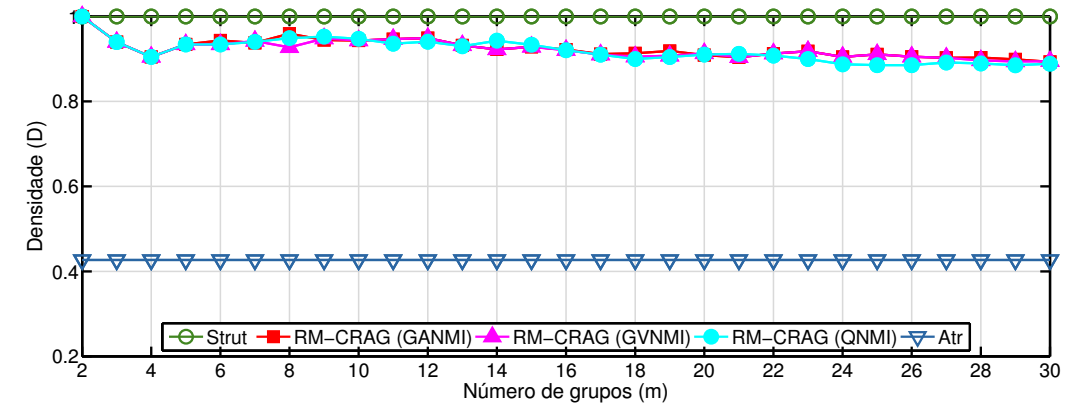
Como as medidas de média possuem características distintas, é interessante analisar a densidade e entropia dos agrupamentos selecionados pelo RM-CRAG. A Figura 5.14 ilustra a densidade e entropia de `Strut`, `Atr`, `RM-CRAG(GANMI)`, `RM-CRAG(MVNMI)` e `RM-CRAG(QNMI)` para $m = 5$.

Pode-se notar que tanto a densidade (Figura 5.14a) quanto a entropia (Figura 5.14b) dos agrupamentos selecionados pelo RM-CRAG com as diferentes medidas apresentam valores intermediários entre `Strut` e `Atr`. Com relação à densidade, o `RM-CRAG(QNMI)` selecionou agrupamentos com uma ligeira piora na densidade para valores de k maiores que 21 quando comparado a `RM-CRAG(GANMI)` e `RM-CRAG(MVNMI)`. Por outro lado, o `RM-CRAG(GANMI)` apresentou uma ligeira melhora. Com relação à entropia, `RM-CRAG(GANMI)`, `RM-CRAG(MVNMI)` e `RM-CRAG(QNMI)` apresentaram comportamento semelhante. Todas se encontram entre a curva de `Strut` e `Atr`, o que indica um bom resultado.

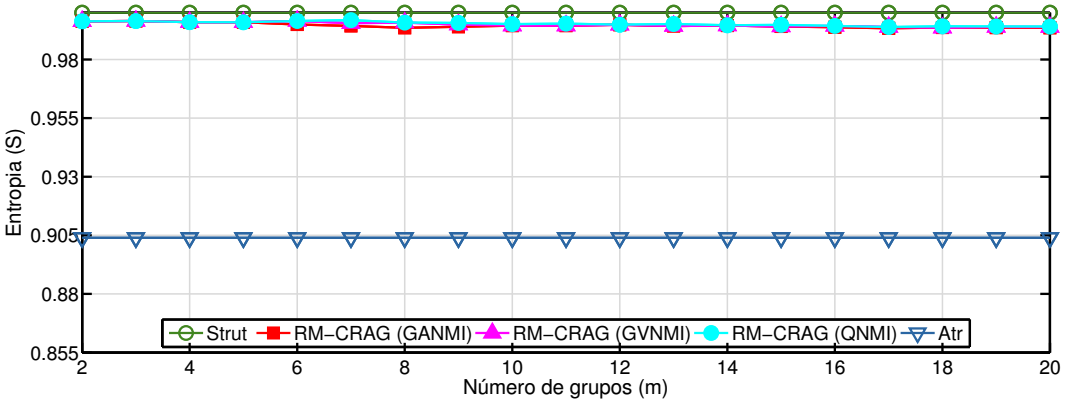
DBLP3000

A Figura 5.15 apresenta a média e a variância entre os k agrupamentos selecionados pelo algoritmo RM-CRAG. Pode-se notar que quase não houve diferença entre as médias da NMI quando foi usada a GANMI e a MVNMI. Para valores de k iguais ou maiores que 23, houve uma pequena alteração indicando que a MVNMI se comportou melhor que a GANMI. A medida QVNMI apresentou uma pequena piora na média quando relacionada com as outras medidas.

Com relação à variância, para $2 < k < 23$ não houve alteração entre as medidas GANMI e MVNMI. Para valores de k iguais ou maiores que 23, houve uma pequena



(a) Densidade



(b) Entropia

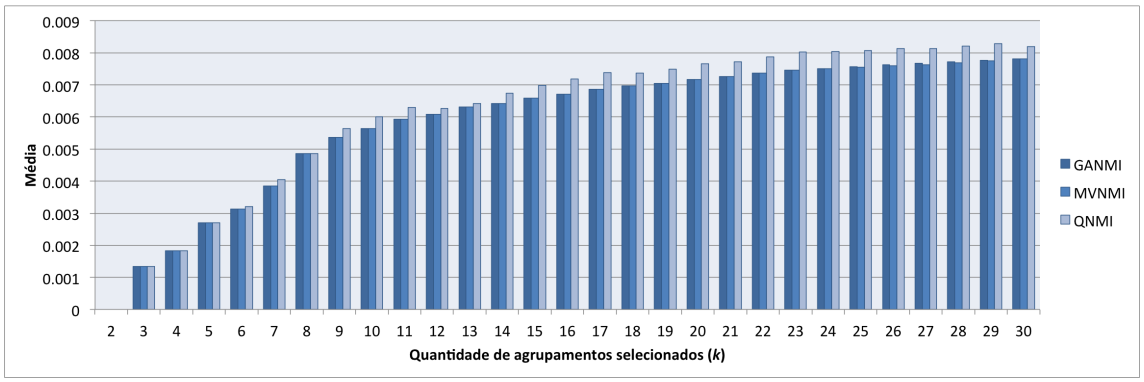
Figura 5.14: Densidade e entropia dos k agrupamentos selecionados pelo RM-CRAG utilizando as diferentes medidas: GANMI, MVNMI e QNMI.

alteração indicando que a MVNMI se comportou melhor que a GANMI. Em quase todos os valores de k, a medida QNMI apresentou valores de variância bem menores do que as demais medidas. Pode-se concluir, dessa forma, que a MVNMI se comportou melhor quando consideradas a média e a variância.

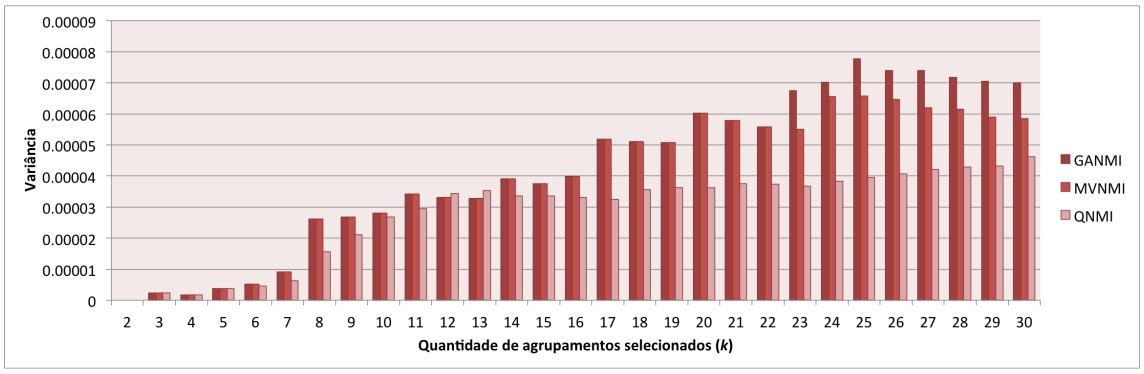
A Figura 5.16 ilustra a quantidade de agrupamentos diferentes encontrados quando o RM-CRAG foi executado com a GANMI, MVNMI e QNMI. Tanto a utilização da MVNMI quanto a QNMI não alteraram a seleção dos agrupamentos selecionados para valores de $k < 6$. A MVNMI acarretou um número relativamente baixo de agrupamentos selecionados diferentes da GANMI, em seu caso mais relevante ($k = 26$), ou seja, 11% dos agrupamentos.

Entretanto, para valores entre $6 <= k <= 30$, é possível notar que a utilização da medida QNMI levou à seleção de mais agrupamentos distintos. O número de agrupamentos distintos chega a 7 quando $k = 29$, ou seja, 24% dos agrupamentos selecionados com a medida QNMI são diferentes dos agrupamentos selecionados com a GANMI.

Como as medidas de média possuem características distintas, é interessante analisar a densidade e entropia dos agrupamentos selecionados pelo RM-CRAG.



(a) Média



(b) Variância

Figura 5.15: Média e variância da NMI encontrada entre os k agrupamentos selecionados pelo RM-CRAG utilizando as diferentes medidas: GANMI, MVNMI e QNMI.

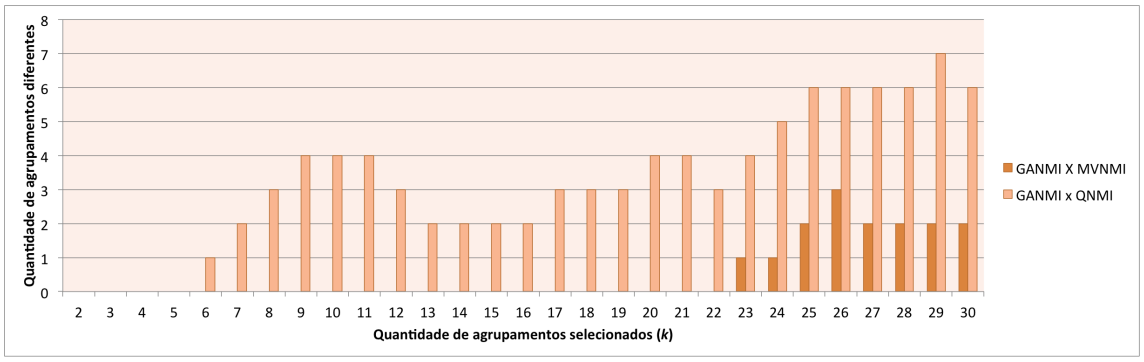
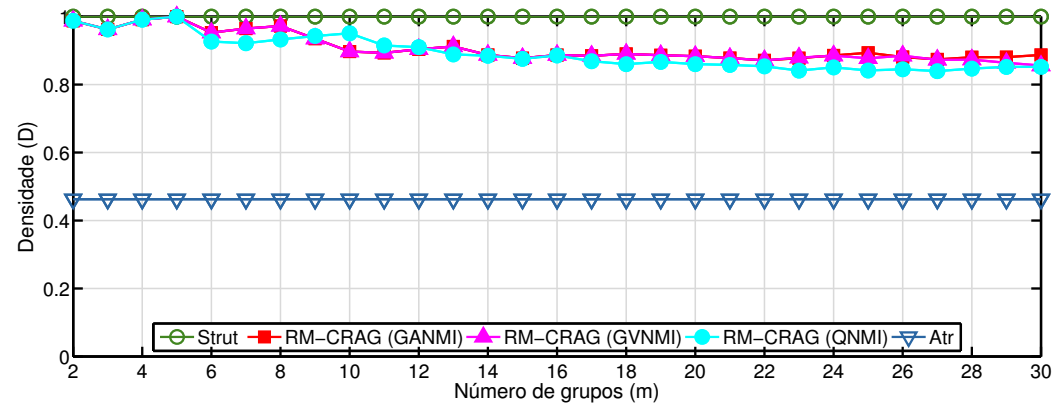


Figura 5.16: Quantidade de agrupamentos diferentes selecionados pelo RM-CRAG utilizando as diferentes medidas de média.

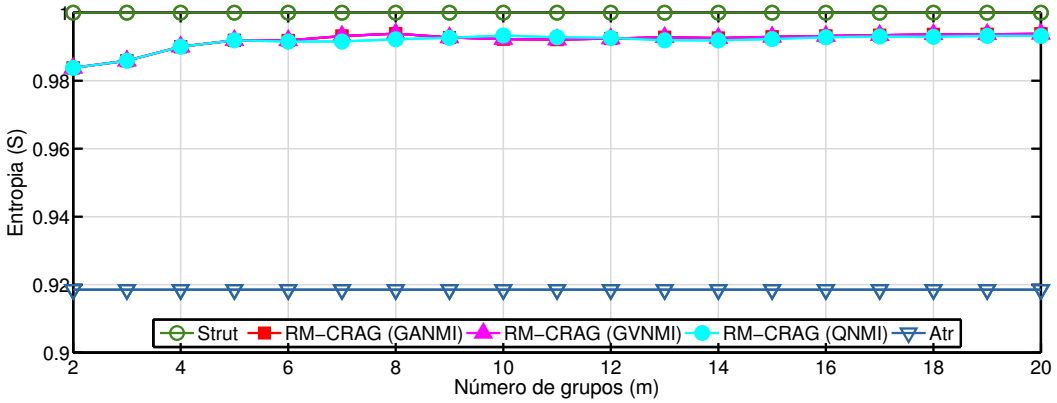
A Figura 5.17 ilustra a densidade e entropia de `Strut`, `Atr`, `RM-CRAG(GANMI)`, `RM-CRAG(MVNMI)` e `RM-CRAG(QNMI)` para $m = 3$.

Pode-se notar que tanto a densidade (Figura 5.17a) quanto a entropia (Figura 5.17b) dos agrupamentos selecionados pelo RM-CRAG com as diferentes medidas apresentam valores intermediários entre `Strut` e `Atr`, o que indica um bom resultado. Com relação à densidade, o `RM-CRAG(QNMI)` selecionou agrupamentos com uma ligeira piora na densidade. Por outro lado, o `RM-CRAG(GANMI)` apresentou uma

ligeira melhora. Com relação à densidade, o `RM-CRAG(GANMI)`, `RM-CRAG(MVNMI)` e `RM-CRAG(QNMI)` apresentaram comportamento semelhante.



(a) Densidade



(b) Entropia

Figura 5.17: Densidade e entropia dos k agrupamentos selecionados pelo RM-CRAG utilizando as diferentes medidas: GANMI, MVNMI e QNMI.

5.5 Discussão

Os experimentos realizados nesse capítulo foram divididos em duas fases. Na primeira fase foi analisado o comportamento dos algoritmos de produção de agrupamentos múltiplos em grafos com atributos nos vértices (i.e., CRAG e M-CRAG). Esses algoritmos apresentaram bons resultados para os conjutos de dados `MQD500b`, `MQD500c` e `DBLP3000`.

A segunda fase dos experimentos corresponde à avaliação do desempenho do RM-CRAG. O algoritmo se comportou bem nos experimentos realizados nos três conjuntos de dados (`MQD500b`, `MQD500c` e `DBLP3000`). O RM-CRAG apresentou os resultados em um tempo consideravelmente inferior à um dos algoritmos que integra o estado da arte na resolução do MDP, o ITS. Na avaliação da média do NMI dos agrupamentos selecionados por ambos, o resultado foi praticamente igual.

Em seguida, foi apresentada a nuvem de palavras dos grupos dos *top-2* agrupamentos selecionados. Vale à pena ressaltar que as nuvens de palavras e as matrizes de similaridade entre as palavras ilustradas correspondem aos resultados obtidos tanto pelo RM-CRAG quanto pelo ITS, visto que para $k = 2$, ambos selecionaram os mesmos agrupamentos. Em suma, o CRAG, M-CRAG e RM-CRAG apresentaram bons resultados trabalhando conjuntamente.

Por fim, foi realizada uma discussão sobre as diferentes medidas de médias propostas nessa tese. Foi possível notar que as três medidas se comportaram bem com relação à densidade e à entropia. A medida QNMI considerou de forma mais significativa a variância entre os agrupamentos, indicando que os agrupamentos selecionados pelo RM-CRAG possuem menos dispersão quando comparados uns com os outros. A QNMI não obteve valores muito inferiores de média do NMI quando comparada às demais medidas. Assim, visto que a medida QNMI apresentou baixa variância e média comparável às demais medidas (GANMI e MVNMI), é considerada a mais adequada para ser utilizada como medida de redundância no RM-CRAG.

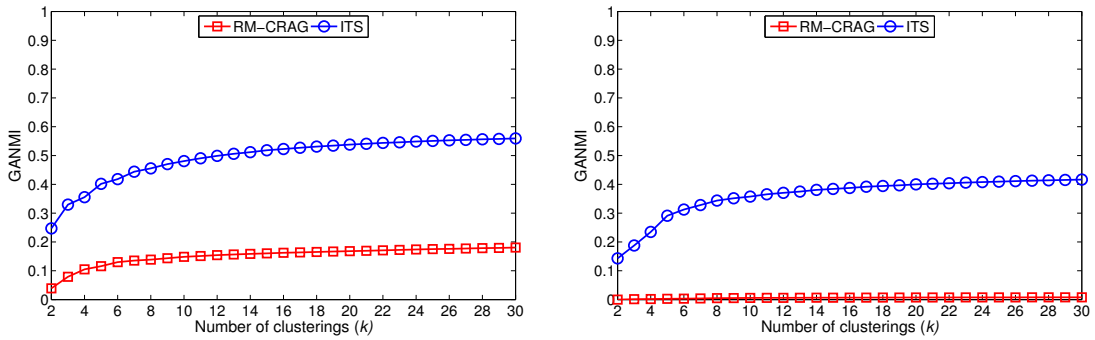
Os bons resultados experimentais corroboram para a retomada da hipótese geral desse trabalho, descrita no capítulo 1: dado um grafo com atributos, a combinação da estrutura topológica com os atributos dos vértices possibilita a produção de agrupamentos múltiplos não-redundantes. Para avaliar essa hipótese, considera-se a GANMI dos agrupamentos provenientes do RM-CRAG em comparação com o ITS aplicado a um conjunto de agrupamentos gerados apenas com a estrutura topológica, denominado `strut-a`. A escolha da GANMI se deveu ao fato de não ser necessária nenhuma modificação na implementação do ITS. Nesse cenário, para comprovar a hipótese geral do trabalho, estabelecem-se as hipóteses nula (H_0) e alternativa (H_1) abaixo:

- H_0, não há diferença entre a GANMI dos agrupamentos produzidos pelo RM-CRAG e a GANMI dos agrupamentos múltiplos produzidos utilizando apenas a estrutura topológica.
- H_1, a GANMI dos agrupamentos produzidos pelo RM-CRAG é menor que a GANMI dos agrupamentos múltiplos produzidos utilizando apenas a estrutura topológica.

Para a concepção de `strut-a`, foram produzidos 4.095 agrupamentos[3] aplicando o agrupamento espectral na estrutura topológica, variando a semente do *k-means*. Em seguida, foi executado o algoritmo ITS para selecionar os agrupamentos menos redundantes. Nesse cenário, a hipótese foi testada a partir dos valores de GANMI obtidos pelos algoritmos RM-CRAG e ITS executados nos conjuntos de dados

[3]Mesmo número de agrupamentos produzidos pelo M-CRAG.

MQD500b e DBLP3000 para 2 <= k <= 30. A Figura 5.18 ilustra o resultado obtido pelos algoritmos. Pode-se notar uma considerável diferença entre as curvas GANMI produzidas pelos algoritmos RM-CRAG e ITS.



(a) MQD500b (b) DBLP3000

Figura 5.18: GANMI dos agrupamentos produzidos pelo RM-CRAG e por ITS aplicado em strut-a.

Inicialmente, o teste de normalidade *Kolmogorov-Smirnov* (KS) foi adotado para analisar a normalidade dos valores de GANMI dos agrupamentos provenientes do ITS e RM-CRAG. A Tabela 5.5 apresenta os valores-p para os conjuntos de dados MQD500b e DBLP3000. Como esses valores são inferiores a 0.05, refuta-se a hipótese nula de normalidade do teste de *Kolmogorov-Smirnov*.

Tabela 5.5: Teste da hipótese nula.

–	MQD500b	DBLP3000
Agrupamentos	KS(valor-p)	KS(valor-p)
RM-CRAG	1.097e−07	3.145e−07
ITS	2.014e−10	5.412e−09

Dada a não-normalidade dos dados, foi utilizado o teste não-paramétrico pareado de *Wilcoxon signed rank test* comparando as amostras de GANMI produzidas pelo RM-CRAG e ITS entre $2 < k < 30$. Os resultados indicaram valores-p iguais a 1.863e-09 e 1.863e-09 para o MQD500b e DBLP3000, respectivamente. Isso indica que a hipótese nula (H_0) foi rejeitada, partindo-se para a hipótese alternativa (H_1) de que a GANMI dos agrupamentos produzidos pelo RM-CRAG é menor que a GANMI dos agrupamentos múltiplos produzidos utilizando apenas a estrutura topológica. Assim sendo, os resultados obtidos forneceram elementos para corroborar a hipótese dessa tese.

Capítulo 6

Conclusões

Nessa tese foi estudado o problema de agrupamentos múltiplos em grafos com atributos nos vértices. Diversas estruturas podem ser modeladas como grafos com atributos, como por exemplo as redes sociais, redes de co-autoria, etc. Ao representar uma rede social como um grafo com atributos nos vértices, esses atributos representam papéis e características dos indivíduos e as arestas representam a relação de amizade entre eles.

Os agrupamentos múltiplos em grafos com atributos nos vértices foram motivados a partir da área de marketing. Nessa área, as técnicas de agrupamento são utilizadas nas pesquisas para segmentar o mercado e determinar os mercados-alvo, visto que as empresas não conseguem estabelecer contato com todos os consumidores potenciais. Além disso, os consumidores tendem a ser céticos com relação às empresas, mas confiam em seus amigos. Nesse cenário, a apresentação de múltiplos agrupamentos aos analistas de marketing pode auxiliar à descoberta de diferentes padrões nos agrupamentos de vértices. Com isso, os analistas podem fazer escolhas mais conscientes e efetuar o marketing direcionado a um grupo específico de consumidores potenciais.

Foram encontrados diversos trabalhos no que concerne à área de agrupamentos múltiplos em grafos (sem atributos), assim como na área de agrupamentos em grafos com atributos, (i.e., combinando a estrutura com os atributos e gerando uma solução de agrupamento). Entretanto, não foram encontradas na literatura abordagens de agrupamentos múltiplos em grafos com atributos. Nesse cenário, essa tese introduziu um novo problema na área de agrupamentos múltiplos, caracterizado pela produção de agrupamentos múltiplos não-redundantes em grafos com atributos. Esse capítulo sumariza as contribuições, limitações e as direções dos trabalhos futuros dessa tese.

No Capítulo 4 foram apresentados dois algoritmos para produzir agrupamentos múltiplos em grafos com atributos. Esses algoritmos atuam conjuntamente para combinar a estrutura topológica com os atributos dos vértices de um grafo para produzir agrupamentos múltiplos. A abordagem utilizada envolve a criação de

arestas artificiais entre vértices similares à distância 2. O total de arestas artificiais criadas corresponde a 20% do total de arestas existentes no grafo original.

Após a apresentação dos algoritmos que produzem agrupamentos múltiplos, foram propostas três medidas para calcular a redundância: GANMI, MVNMI e QNMI. Essas medidas objetivam avaliar a redundância na seleção dos *top-k* agrupamentos e integram o conjunto de contribuições dessa tese. São aplicadas ao resultado do NMI calculado para todos os agrupamentos, dois a dois e são baseadas na média, média + variância e média dos quadrados, respectivamente representadas como μ, $\mu + \sigma^2$ e q^2.

Ainda no Capítulo 4, foi descrito o RM-CRAG, um algoritmo desenvolvido para criar um ranking dos *top-k* agrupamentos não-redundantes. Esse algoritmo também integra o conjunto de contribuições dessa tese. O RM-CRAG pode utilizar a GANMI, MVNMI ou QNMI para selecionar os agrupamentos não-redundantes. Cada uma dessas medidas causa um efeito no resultado final da seleção dos *top-k* agrupamentos. Assim, foi efetuadas uma análise dos efeitos de cada uma dessas medidas na avaliação da redundância realizada pelo RM-CRAG.

Em seguida, o algoritmo de seleção dos *top-k* agrupamentos não-redundantes foi formulado como o MDP. É importante ressaltar que o algoritmo proposto é um algoritmo para resolver o MDP. Conforme já mencionado, existem diversos algoritmos para resolver esse problema.

O Capítulo 5 realizou uma avaliação experimental dos algoritmos e medidas propostos nessa tese. Os experimentos foram aplicados em três conjuntos de dados. Dois desses conjuntos de dados são derivados de uma rede social brasileira (`MQD500b` e `MQD500c`). O outro conjunto de dados é um subconjunto de uma rede de co-autoria (`DBLP3000`).

Os algoritmos M-CRAG e CRAG mostraram ser possível produzir agrupamentos múltiplos em grafos com atributos. Também foi possível selecionar os *top-k* agrupamentos não-redundantes utilizando o algoritmo RM-CRAG. Vale destacar que o RM-CRAG superou o tempo de execução de um dos algoritmos que compõem o estado da arte. Por conseguinte, foram realizados testes estatísticos para avaliar a hipótese geral desse trabalho. Os resultados obtidos forneceram elementos para corroborar a hipótese. Dessa maneira, dado um grafo com atributos, a combinação da estrutura topológica com os atributos dos vértices possibilita a produção de agrupamentos múltiplos não-redundantes.

6.1 Sumário de contribuições

Essa tese contribui na área de agrupamentos múltiplos em grafos com atributos. As principais contribuições podem ser sumarizadas da seguinte forma:

- **Algoritmo para produzir um agrupamento combinado:** Foi desenvolvido um algoritmo para produzir um agrupamento combinando a estrutura topológica e os atributos dos vértices de um grafo através da adição de arestas artificiais entre vértices semelhantes à distância 2. Essa abordagem para a criação de agrupamentos múltiplos não foi encontrada na literatura. O objetivo é que o agrupamento resultante dessa combinação tenha um balanceamento entre a estrutura topológica e os atributos do vértice. O novo grafo com arestas artificiais é então submetido a um algoritmo de agrupamento estrutural, sendo nesse trabalho o algoritmo de agrupamento espectral. Os resultados experimentais evidenciaram que o algoritmo procedeu bem, produzindo um balanceamento entre a estrutura e os atributos dos vértices.

- **Algoritmo para produzir agrupamentos múltiplos:** Foi desenvolvido um algoritmo para produzir agrupamentos múltiplos utilizando a abordagem anterior. Esse algoritmo produz um conjunto de agrupamentos que combinam a estrutura topológica com os atributos dos vértices do grafo. Resultados experimentais indicam bons resultados, conforme discutido no Capítulo 5.

- **Algoritmo para selecionar os *top-k* agrupamentos não-redundantes:** Foi elaborado um algoritmo para receber um conjunto de agrupamentos e produzir um ranking. Em seguida o algoritmo seleciona os *top-k* agrupamentos não-redundantes. É um algoritmo determinístico e resolve essa tarefa em pouco tempo. Se mostrou superior ao ITS, um dos algoritmos que integram o estado da arte na resolução desse problema.

- **Medidas para avaliação da qualidade de agrupamentos:** Foram concebidas três medidas para avaliação de redundância em agrupamentos múltiplos. Conforme discutido no Capítulo 2, existe uma carência de medidas de avaliação na área de agrupamentos múltiplos e o objetivo é contribuir nessa lacuna. As medidas foram baseadas na medida NMI e nas medidas estatísticas média, média + variância e média dos quadrados. Também foi apresentado um estudo sobre o impacto de cada uma dessas medidas estatísticas no cálculo da redundância dos agrupamentos. O comportamento de cada uma delas foi analisado. Cada uma delas atribui um peso diferente à dispersão dos dados.

6.2 Limitações

É importante ressaltar as limitações desse trabalho de forma que os trabalhos derivados possam evoluir da melhor forma possível. As limitações encontradas são

sumarizadas a seguir.

- **Tempo de processamento:** Os experimentos realizados nessa tese utilizaram conjuntos de dados relativamente pequenos. O maior conjunto de dados utilizado foi o `DBLP3000` que possui 3.000 vértices e 9.979 arestas. Um dos algoritmos propostos (CRAG) necessita calcular os vértices a distância dois para cada um dos vértices do grafo. O tempo demandado para essa tarefa foi de 82 minutos. É sabido que esse cálculo só necessita ser realizado uma única vez, entretanto, é preciso considerar que esse cálculo pode ter que ser realizado para milhares de vértices. É necessário que sejam utilizadas abordagens para escalar esse problema.

- **Dados esparsos:** Os atributos dos vértices do conjunto de dados `MQD500c` não possuem dados esparsos. Em contrapartida, o `MQD500b` possui os 6 atributos relacionados à emoções que podem ter dados esparsos, visto que se um usuário não criou entradas com determinada emoção, esse atributo tem valor 0. Da mesma forma, o `DBLP3000` contém dados esparsos, nesse caso, em seus 12 atributos. Isso ocorre porque diversos autores não publicaram em diversas conferências. Essa esparsidade nos dados pode influenciar o resultado dos agrupamentos, visto que se dois usuários não possuem publicação em determinada conferência, eles são considerados similares. Entretanto, observa-se que o algoritmo proposto considera que a ausência de um atributo caracteriza uma similaridade entre os vértices.

6.3 Trabalhos futuros

Durante o desenvolvimento da presente tese foram idealizados diversos desdobramentos para trabalhos futuros na área de agrupamentos múltiplos em grafos com atributos nos vértices. Esses trabalhos se encontram sumarizados a seguir.

- **Qualidade dos grupos:** Uma das contribuições desse trabalho é um algoritmo para selecionar os *top-k* agrupamentos não-redundantes. Junto com essa contribuição, foram apresentadas, também como contribuições, as medidas para calcular redundância em agrupamentos múltiplos. Essas medidas comparam os agrupamentos gerados pelos algoritmos CRAG e M-CRAG. No entanto, essas medidas comparam os agrupamentos apenas nos termos da redundância, não sendo relevantes, nessa comparação, medidas estruturais (e.g., densidade) e de homogeneidade (e.g., entropia). É um interesse envolver essas medidas na seleção dos *top-k* agrupamentos em um próximo trabalho.

- **Abordagem de adição de arestas artificiais:** O algoritmo proposto utiliza o teorema de Pareto para criar um número máximo de 20% de arestas artificiais entre os vértices. Seria relevante avaliar o comportamento decorrente da adição de outros limites de valores para a criação dessas arestas.

- **Número de arestas artificiais criadas nos agrupamentos:** O número de arestas artificiais criado na abordagem proposta nessa tese é compartilhado entre todos os agrupamentos gerados, ou seja, todos os agrupamentos possuem o mesmo número de arestas artificiais. Seria interessante realizar estudos que proporcionem que a solução final de agrupamentos possa ter agrupamentos com diferentes porcentagens de arestas artificiais (e.g. 20% de arestas, 50%, etc.), ou seja, misturar agrupamentos com essas variações na solução final.

- **Parametrizar medidas estruturais e de homogeneidade:** Uma abordagem interessante para a geração de agrupamentos alternativos poderia ser a parametrização das medidas de entropia e densidade, de forma que o número de arestas artificiais inseridas fosse baseado em limites dessas medidas. Isso poderia resultar em agrupamentos mais homogêneos e mais coesos. Um subtrabalho interessante seria o estudo do desempenho de um algoritmo com essa característica.

- **Número de grupos em cada agrupamento:** Os algoritmos propostos nessa tese são parametrizados para receber o número de grupos de cada agrupamento. Assim, se faz necessário que os analistas informem o número desejado de grupos. Posteriormente, os algoritmos retornam agrupamentos com o número de grupos inserido. São vislumbrados trabalhos que desenvolvam algoritmos capazes de retornar agrupamentos com número de grupos diferentes, ou seja, se o analista solicitar 4 agrupamentos, esses agrupamentos podem ter tamanho distintos de grupos (e.g., 2, 14, 20, 5). Por exemplo, a modularidade poderia ser utilizada como forma de produzir agrupamentos com diferentes grupos.

- **Processamento distribuído:** O desenvolvimento dos algoritmos propostos não levou em conta o processamento distribuído. Ao lidar com grafos contendo milhões de vértices e arestas, faz-se necessário que sejam implementadas técnicas de *Big Data*.

- **Agrupamento por homofilia:** Conforme descrito no Capítulo 1, a homofilia é a tendencia que indivíduos possuem de se associar com outros indivíduos similares em relação a alguma característica. Essa é uma tendência que vem sendo cada vez mais estudada. Isso traz a luz novas possibilidades na área

de agrupamentos múltiplos, visto que a homofilia poderia ser considerada no processo de agrupamento dos algoritmos. Por exemplo, produzindo agrupamentos múltiplos em que uma solução de agrupamento poderia conter vértices com relação de homofilia baseada na idade e outra solução poderia ser baseada na relação de homofilia de sexo, etc.

- **Agrupamentos múltiplos em multigrafos:** Uma abordagem interessante que pode ser considerada é a adaptação das abordagens propostas nessa tese para o domínio dos grafos com atributos nas arestas, os multigrafos. Nesse modelo, arestas com diferentes relacionamentos poderiam ser artificialmente criadas.

- **Problema da esparsidade:** Conforme discutido na Seção de limitações (6.2), o problema da esparsidade nos dados pode alterar o comportamento dos algoritmos propostos. Trabalhos futuros podem ser originados para resolver essa limitação, como por exemplo, simplesmente não considerando atributos com valores iguais a 0 na criação das arestas artificiais.

- **Pesos nas arestas do grafo:** Os algoritmos propostos nesse trabalho não operam sobre os pesos das arestas do grafo. Existem abordagens na literatura que resolvem o agrupamento de grafos com pesos nas arestas, dessa forma, pode-se considerar utilizar a similaridade entre os vértices para auxiliar no processo de agrupamento.

- **Seleção de atributos:** Para produzir agrupamentos múltiplos em grafos com atributos nos vértices, o algoritmo proposto nessa tese utiliza os atributos dos vértices para gerar os agrupamentos combinados com a estrutura. Com isso, se considerarmos um número relativamente grande de atributos, por exemplo, 100, o número de agrupamentos gerados será de $1.26E + 30$. Esse aspecto abre campo para estudos na área de seleção de atributos, de forma que haja uma redução no número de atributos e assim, essa abordagem seja computacionalmente tratável.

- **Agrupamentos utilizando restrições:** A área de agrupamento com restrições (*constrained clustering*) tem o objetivo de gerar restrições *must-link* e *cannot-link* entre objetos, o que poderia ser adaptado na abordagem proposta para sugerir que alguns vértices devem ficar no mesmo grupo ou não devem ficar no mesmo grupo. Isso poderia dar à luz a resultados interessantes.

- **Parametrizar atributos para gerar arestas artificiais:** É interessante ressaltar que na abordagem proposta para gerar agrupamentos múltiplos, todos os atributos dos vértices são utilizados. Isso poderia ser parametrizado,

de forma que o analista pudesse selecionar os atributos mais significantes para ele.

- **Identificação do número de grupos dos agrupamentos:** Na seção de experimentos, utilizamos o joelho da curva da função objetivo para determinar o número de grupos. Entretanto, existem outras técnicas na literatura que poderiam ser utilizadas pelos algoritmos propostos para identificar o número ideal de grupos de cada agrupamento, como por exemplo, a modularidade.

- **Abordagem estocástica do algoritmo proposto:** O algoritmo proposto para selecionar os *top-k* agrupamentos não-redundantes possui uma abordagem determinística. Entretanto, essa abordagem pode apresentar um alto custo para convergir em números muito grandes de agrupamentos. Assim, percebe-se uma necessidade de criar uma versão estocástica desse algoritmo, de forma que o ranking de agrupamentos gerado na primeira fase (i.e., quando o algoritmo compara os agrupamentos dois a dois) possa ser produzido com abordagens aleatórias.

- **Outros algoritmos de agrupamento:** Seria interessante avaliar se o RM-CRAG é capaz de selecionar os *top-k* agrupamentos não-redundantes utilizando agrupamentos gerados por outros algoritmos. Nessa tese o algoritmo RM-CRAG foi avaliado apenas com os agrupamentos gerados pelo CRAG e M-CRAG.

- **Grafos direcionados:** Os algoritmos propostos nessa tese poderiam ser adaptados para grafos direcionados. Essa adaptação é simples e pode ser interessante para produção de agrupamentos múltiplos em grafos direcionados com atributos nos vértices.

Referências Bibliográficas

ABDELHAQ, H., SENGSTOCK, C., GERTZ, M., 2013, "EvenTweet: Online Localized Event Detection from Twitter", *Proc. VLDB Endow.*, v. 6, n. 12, pp. 1326–1329. Disponível em: <http://dl.acm.org/citation.cfm?id=2536274.2536307>.

ABERER, K., FLACHE, A., JAGER, W., et al., 2012, *Social Informatics: 4th International Conference, SocInfo 2012, Lausanne, Switzerland, December 5-7, 2012, Proceedings*. Lecture Notes in Computer Science. Springer Berlin Heidelberg. ISBN: 9783642353864. Disponível em: <https://books.google.com.br/books?id=NVO5BQAAQBAJ>.

ABU-JAMOUS, B., FA, R., NANDI, A., 2015, *Integrative Cluster Analysis in Bioinformatics*. Wiley. ISBN: 9781118906569. Disponível em: <https://books.google.com.br/books?id=gYC4CAAAQBAJ>.

ANDA, C., 1999. "Data Mining Techniques in Supporting Decision Making". .

ARAMAKI, E., MASKAWA, S., MORITA, M., 2011, "Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter". In: *Proceedings of the 2011 Conference on Emperical Methods in Natural Language Processing - EMNLP '11*, pp. 1568–1576, Edinburgh, United Kingdom. Association for Computational Linguistics. ISBN: 978-1-937284-11-4. Disponível em: <http://dl.acm.org/citation.cfm?id=2145600>.

ARINGHIERI, R., CORDONE, R., MELZANI, Y., 2008, "Tabu Search versus GRASP for the maximum diversity problem", *4OR*, v. 6, n. 1, pp. 45–60. ISSN: 1619-4500. doi: 10.1007/s10288-007-0033-9. Disponível em: <http://dx.doi.org/10.1007/s10288-007-0033-9>.

BAE, E., BAILEY, J., 2006, "COALA: A Novel Approach for the Extraction of an Alternate Clustering of High Quality and High Dissimilarity." In: *ICDM*, pp. 53–62. IEEE Computer Society. Disponível em: <http://dblp.uni-trier.de/db/conf/icdm/icdm2006.html>.

BASU, S., DAVIDSON, I., WAGSTAFF, K., 2008, *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman & Hall/CRC. ISBN: 1584889969, 9781584889960.

BAUMAN, Z., MEDEIROS, C., 2004, *Amor líquido*. Jorge Zahar Editor. ISBN: 9788571107953. Disponível em: <http://books.google.com.br/books?id=GD8vAAAACAAJ>.

BECKER, H., NAAMAN, M., GRAVANO, L., 2011a, "Beyond Trending Topics: Real-World Event Identification on Twitter". In: *Fifth International AAAI Conference on Weblogs and Social Media*, pp. 1–17, Barcelona, Spain, a. AAAI Press. ISBN: 9781605588896. Disponível em: <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2745>.

BECKER, H., NAAMAN, M., GRAVANO, L., 2011b, "Beyond Trending Topics: Real-World Event Identification on Twitter", *Columbia University Computer Science Technical Reports*. Disponível em: <http://hdl.handle.net/10022/AC:P:10668>.

BELL, E. T., 1934, "Exponential Numbers", *The American Mathematical Monthly*, v. 41, n. 7, pp. 411–419. ISSN: 00029890. doi: 10.2307/2300300. Disponível em: <http://dx.doi.org/10.2307/2300300>.

BHATT, R., CHAOJI, V., PAREKH, R., 2010, "Predicting Product Adoption in Large-scale Social Networks". In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pp. 1039–1048, New York, NY, USA. ACM. ISBN: 978-1-4503-0099-5. doi: 10.1145/1871437.1871569. Disponível em: <http://doi.acm.org/10.1145/1871437.1871569>.

BIFULCO, I., FEDULLO, C., NAPOLITANO, F., et al., 2009, "Global optimization, Meta Clustering and consensus clustering for class prediction." In: *IJCNN*, pp. 332–339. IEEE.

BLONDEL, V., GUILLAUME, J., LAMBIOTTE, R., et al., 2008, "Fast unfolding of communities in large networks", *J. Stat. Mech*, p. P10008.

BODEN, B., HAAG, R., SEIDL, T., 2013a, "Detecting and Exploring Clusters in Attributed Graphs: A Plugin for the Gephi Platform". In: *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management*, CIKM '13, pp. 2505–2508, New York, NY, USA, a. ACM. ISBN: 978-1-4503-2263-8. doi: 10.1145/2505515.

2508200. Disponível em: <http://doi.acm.org/10.1145/2505515.2508200>.

BODEN, B., HAAG, R., SEIDL, T., 2013b, "Detecting and exploring clusters in attributed graphs: a plugin for the gephi platform". In: *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, CIKM '13, pp. 2505–2508, New York, NY, USA, b. ACM. ISBN: 978-1-4503-2263-8. doi: 10.1145/2505515.2508200. Disponível em: <http://doi.acm.org/10.1145/2505515.2508200>.

BONCHI, F., CASTILLO, C., GIONIS, A., et al., 2011, "Social Network Analysis and Mining for Business Applications", *ACM Trans. Intell. Syst. Technol.*, v. 2, n. 3 (maio), pp. 22:1–22:37. ISSN: 2157-6904. doi: 10.1145/1961189.1961194. Disponível em: <http://doi.acm.org/10.1145/1961189.1961194>.

BOTHOREL, C., CRUZ, J. D., MAGNANI, M., et al., 2015, "Clustering attributed graphs: models, measures and methods", *CoRR*, v. abs/1501.01676. Disponível em: <http://arxiv.org/abs/1501.01676>.

BRAHA, D., 2012, "Global Civil Unrest: Contagion, Self-Organization, and Prediction", *PLoS ONE*, v. 7, n. 10, pp. e48596. ISSN: 19326203. doi: 10.1371/journal.pone.0048596. Disponível em: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0048596>.

CAMPIGOTTO, R., GUILLAUME, J., SEIFI, M., 2013, "The power of consensus: random graphs have no communities". In: *Advances in Social Networks Analysis and Mining 2013, ASONAM '13, Niagara, ON, Canada - August 25 - 29, 2013*, pp. 272–276. doi: 10.1145/2492517.2492650. Disponível em: <http://doi.acm.org/10.1145/2492517.2492650>.

CARUANA, R., ELHAWARY, M., NGUYEN, N., 2006, "Meta clustering". In: *In Proceedings IEEE International Conference on Data Mining*.

CHAUDHURI, S., GRAVANO, L., 1999, "Evaluating Top-k Selection Queries". In: *Proceedings of the 25th International Conference on Very Large Data Bases*, VLDB '99, pp. 397–410, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. ISBN: 1-55860-615-7. Disponível em: <http://dl.acm.org/citation.cfm?id=645925.671359>.

CHEN, F., ARREDONDO, J., KHANDPUR, R. P., et al., 2012, "Spatial Surrogates to Forecast Social Mobilization and Civil Unrests", *Position Paper in CCC Workshop on "From GPS and Virtual Globes to Spatial*

Computing-2020", pp. 1–3. Disponível em: <http://people.cs.vt.edu/naren/papers/CCC-VT-Updated-Version.pdf>.

CHEN, L., ROY, A., 2009, "Event Detection from Flickr Data Through Wavelet-based Spatial Analysis". In: *Proceedings of the 18th ACM conference on Information and knowledge management - CIKM'09*, pp. 523–532, Hong Kong, China. ACM. ISBN: 9781605585123. doi: 10.1145/1645953.1646021. Disponível em: <http://dl.acm.org/citation.cfm?id=1646021>.

CHENG, D., VEMPALA, S., KANNAN, R., et al., 2005, "A Divide-and-merge Methodology for Clustering". In: *Proceedings of the Twenty-fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '05, pp. 196–205, New York, NY, USA. ACM. ISBN: 1-59593-062-0. doi: 10.1145/1065167.1065192. Disponível em: <http://doi.acm.org/10.1145/1065167.1065192>.

CHENG, H., ZHOU, Y., YU, J. X., 2011, "Clustering Large Attributed Graphs: A Balance Between Structural and Attribute Similarities", *ACM Trans. Knowl. Discov. Data*, v. 5, n. 2 (fev.), pp. 12:1–12:33. ISSN: 1556-4681. doi: 10.1145/1921632.1921638. Disponível em: <http://doi.acm.org/10.1145/1921632.1921638>.

CHRISTAKIS, N., FOWLER, J., 2009, *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*. Little, Brown. ISBN: 9780316071345. Disponível em: <https://encrypted.google.com/books?id=LXHi4wgIkzEC>.

CHUANG, Y.-Y., 2012, "Affinity Aggregation for Spectral Clustering". In: *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, pp. 773–780, Washington, DC, USA. IEEE Computer Society. ISBN: 978-1-4673-1226-4. Disponível em: <http://dl.acm.org/citation.cfm?id=2354409.2355074>.

CHUNG, F. R. K., 1997, *Spectral Graph Theory*. American Mathematical Society.

CLAUSET, A., NEWMAN, M. E. J., , et al., 2004, "Finding community structure in very large networks", *Physical Review E*, pp. 1– 6. doi: 10.1103/PhysRevE.70.066111. Disponível em: <www.ece.unm.edu/ifis/papers/community-moore.pdf>.

COMPTON, R., LEE, C., LU, T.-C., et al., 2013, "Detecting future social unrest in unprocessed Twitter data". In: *2013 IEEE International Conference*

on Intelligence and Security Informatics, pp. 56–60, Seattle, WA, USA. IEEE. ISBN: 9781467362139. doi: 10.1109/ISI.2013.6578786. Disponível em: <http://dx.doi.org/10.1109/ISI.2013.6578786>.

COVER, T. M., THOMAS, J. A., 1991, *Elements of Information Theory*. New York, NY, USA, Wiley-Interscience. ISBN: 0-471-06259-6.

CUI, Y., FERN, X. Z., DY, J. G., 2010, "Learning Multiple Nonredundant Clusterings", *ACM Trans. Knowl. Discov. Data*, v. 4, n. 3 (out.), pp. 15:1–15:32. ISSN: 1556-4681. doi: 10.1145/1839490.1839496. Disponível em: <http://doi.acm.org/10.1145/1839490.1839496>.

CVETKOVIĆ, D., ROWLINSON, P., SIMIC, S., 1997, *Eigenspaces of Graphs*. Encyclopedia of Mathematics and its Applications. Cambridge University Press. ISBN: 9780521573528. Disponível em: <https://books.google.com.br/books?id=fV3jjYyX_JwC>.

DALGIC, T., 2006, *Handbook of Niche Marketing: Principles and Practice*. Haworth series in segmented, targeted, and customized marketing. Best Business Books, Haworth Reference Press. ISBN: 9780789023308. Disponível em: <http://books.google.com.br/books?id=TWEqsqWTOq8C>.

DANG, X., BAILEY, J., 2014a, "Generating multiple alternative clusterings via globally optimal subspaces", *Data Mining and Knowledge Discovery*, v. 28, n. 3, pp. 569–592. ISSN: 1384-5810. doi: 10.1007/s10618-013-0314-1. Disponível em: <http://dx.doi.org/10.1007/s10618-013-0314-1>.

DANG, X., BAILEY, J., 2014b, "Generating multiple alternative clusterings via globally optimal subspaces", *Data Mining and Knowledge Discovery*, v. 28, n. 3, pp. 569–592. ISSN: 1384-5810. doi: 10.1007/s10618-013-0314-1. Disponível em: <http://dx.doi.org/10.1007/s10618-013-0314-1>.

DANIEL T. LAROSE, C. D. L., 2014, *Discovering Knowledge in data An Introduction to Data Mining*. Wiley Series on Methods and Applications in Data Mining. Wiley. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=2e7ec27c28edd858191bdc6a3d7a77e5>.

DE MEO, P., FERRARA, E., FIUMARA, G., et al., 2012, "A novel measure of edge centrality in social networks", *Knowledge-based systems*, v. 30, pp. 136–150.

DEEPJYOTI, C., ARNAB, P., "Community Detection in Social Networks: An Overview", Disponível em: <http://ijret.org/Volumes/V02/I14/IJRET_110214017.pdf>.

DEVLIN, K., 2012, *Fundamentals of Contemporary Set Theory*. Universitext. Springer New York. ISBN: 9781468400847. Disponível em: <https://books.google.com.br/books?id=4FHtBwAAQBAJ>.

DIMITRIUS, J., MAZZARELLA, M., 1998, *Reading People: How to Understand People and Predict Their Behavior– Anytime, Anyplace*. Random House. ISBN: 9780345425874. Disponível em: <http://books.google.com.br/books?id=JEdbT6IZvTYC>.

DU, K. L., 2010, "Clustering: A Neural Network Approach", *Neural Netw.*, v. 23, n. 1 (jan.), pp. 89–107. ISSN: 0893-6080. doi: 10.1016/j.neunet.2009.08.007. Disponível em: <http://dx.doi.org/10.1016/j.neunet.2009.08.007>.

DUGGAL, G., NAVLAKHA, S., GIRVAN, M., et al., 2010, "Uncovering Many Views of Biological Networks Using Ensembles of Near-Optimal Partitions". In: *1ST INTL WORKSHOP ON DISCOVERING, SUMMARIZING, AND USING MULTIPLE CLUSTERINGS, KDD*. ACM.

EKMAN, P., 1992, "An argument for basic emotions", *Cognition & Emotion*, v. 6, n. 3-4 (maio), pp. 169–200. doi: 10.1080/02699939208411068. Disponível em: <http://dx.doi.org/10.1080/02699939208411068>.

ERKUT, E., 1990, "The discrete p-dispersion problem", *European Journal of Operational Research*, v. 46, n. 1 (maio), pp. 48–60. ISSN: 03772217. doi: 10.1016/0377-2217(90)90297-o. Disponível em: <http://dx.doi.org/10.1016/0377-2217(90)90297-o>.

FA-LIANG, H., MING-XUAN, H., CHANG-AN, Y., et al., 2014, "Spectral clustering ensemble algorithm for discovering overlapping communities in social networks", *Control and Decision*, 29(4):713. doi: 10.13195/j.kzyjc.2012.1730. Disponível em: <http://www.kzyjc.net:8080/EN/abstract/article_12689.shtml>.

FÄRBER, I., GÜNNEMANN, S., KRIEGEL, H., et al., 2010, "On using class-labels in evaluation of clusterings". In: *MultiClust: 1st International Workshop on Discovering, Summarizing and Using Multiple Clusterings Held in Conjunction with KDD 2010, Washington, DC*. Disponível

em: <http://scholar.google.com.au/scholar.bib?q=info:d8mnI5RN02oJ:scholar.google.com/&output=citation&hl=en&as_sdt=0,5&ct=citation&cd=0>.

FREEMAN, L. C., 1996, "Cliques, Galois lattices, and the structure of human social groups", *Social Networks*, v. 18, n. 3, pp. 173–187.

GAN, G., MA, C., WU, J., 2007a, *Data Clustering: Theory, Algorithms, and Applications*. ASA-SIAM Series on Statistics and Applied Probability. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104). ISBN: 9780898718348. Disponível em: <https://books.google.com.br/books?id=HMfJHBW8x_EC>.

GAN, G., MA, C., WU, J., 2007b, *Data clustering - theory, algorithms, and applications*. SIAM.

GE, R., ESTER, M., GAO, B. J., et al., 2008, "Joint Cluster Analysis of Attribute Data and Relationship Data: The Connected K-center Problem, Algorithms and Applications", *ACM Trans. Knowl. Discov. Data*, v. 2, n. 2 (jul.), pp. 7:1–7:35. ISSN: 1556-4681. doi: 10.1145/1376815.1376816. Disponível em: <http://doi.acm.org/10.1145/1376815.1376816>.

GHOSH, J., ACHARYA, A., 2011, "Cluster ensembles", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 1, n. 4, pp. 305–315. ISSN: 1942-4795. doi: 10.1002/widm.32. Disponível em: <http://dx.doi.org/10.1002/widm.32>.

GHOSH, J., STREHL, A., MERUGU, S., 2002, "A Consensus Framework for Integrating Distributed Clusterings Under Limited Knowledge Sharing". In: *In Proc. NSF Workshop on Next Generation Data Mining*, pp. 99–108.

GIRVAN, M., NEWMAN, M. E. J., 2002, "Community structure in social and biological networks", *Proceedings of the National Academy of Sciences*, v. 99, n. 12 (jun.), pp. 7821–7826. ISSN: 1091-6490. doi: 10.1073/pnas.122653799. Disponível em: <http://dx.doi.org/10.1073/pnas.122653799>.

GLOVER, F., LAGUNA, M., 1997, *Tabu Search*. Norwell, MA, USA, Kluwer Academic Publishers. ISBN: 079239965X.

GLOVER, F., KUO, C.-C., DHIR, K. S., 1998, "Heuristic algorithms for the maximum diversity problem", *Journal of Information and Optimization*

Sciences, v. 19, n. 1, pp. 109–132. doi: 10.1080/02522667.1998.10699366. Disponível em: <http://dx.doi.org/10.1080/02522667.1998.10699366>.

GODER, A., FILKOV, V., 2008, "Consensus Clustering Algorithms: Comparison and Refinement". In: *Proc. 9th Workshop on Algorithm Engineering and Experiments (ALENEX'08)*.

GOLBECK, J., ROBLES, C., EDMONDSON, M., et al., 2011, "Predicting Personality from Twitter." In: *SocialCom/PASSAT*, pp. 149–156. IEEE. ISBN: 978-1-4577-1931-8. Disponível em: <http://dblp.uni-trier.de/db/conf/socialcom/socialcom2011.html#GolbeckRET11>.

GOODCHILD, M. F., 2007, "Citizens as sensors: The world of volunteered geography", *GeoJournal*, v. 69, n. November, pp. 211–221. ISSN: 03432521. doi: 10.1007/s10708-007-9111-y. Disponível em: <http://link.springer.com/article/10.1007%2Fs10708-007-9111-y>.

GRAHAM, M., HALE, S. A., GAFFNEY, D., 2014, "Where in the world are you? Geolocation and language identification in Twitter", *The Professional Geographer*, v. 66, n. 4 (aug), pp. 568–578. ISSN: 00330124. doi: 10.1080/00330124.2014.907699. Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/00330124.2014.907699>.

GUAN, Y., DY, J. G., NIU, D., et al., 2010, "Variational inference for nonparametric multiple clustering". In: *In Workshop on Discovering, Summarizing and Using Multiple Clustering (MultiClust) at the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*.

GUEDES, G. P., 2015a. "DBLP3000 dataset". http://sourceforge.net/p/gpca/wiki/DBLP3000/, a.

GUEDES, G. P., 2006. "Meu Querido Diário". http://www.meuqueridodiario.com.br.

GUEDES, G. P., 2014. "MQD500b dataset". http://sourceforge.net/p/gpca/wiki/MQD500B/.

GUEDES, G. P., 2015b. "MQD500c dataset". http://sourceforge.net/p/gpca/wiki/MQD500C/, b.

GUEDES, G. P., BEZERRA, E., OGASAWARA, E., et al., 2015, "Exploring Multiple Clusterings in Attributed Graphs". In: *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, SAC '15, pp.

915–918, New York, NY, USA. ACM. ISBN: 978-1-4503-3196-8. doi: 10.1145/2695664.2696008. Disponível em: <http://doi.acm.org/10.1145/2695664.2696008>.

GUHA, S., RASTOGI, R., SHIM, K., 1998, "CURE: An Efficient Clustering Algorithm for Large Databases". In: *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, SIGMOD '98, pp. 73–84, New York, NY, USA. ACM. ISBN: 0-89791-995-5. doi: 10.1145/276304.276312. Disponível em: <http://doi.acm.org/10.1145/276304.276312>.

GÜNNEMANN, S., FARBER, I., BODEN, B., et al., 2010, "Subspace clustering meets dense subgraph mining: A synthesis of two paradigms". In: *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pp. 845–850. IEEE.

GUPTA, S., JUNEJA, M., BATRA, D., 2013, "Article: Predictive Computational Model for Target Marketing using Social Network Analysis and Artificial Neural Network", *International Journal of Computer Applications*, v. 74, n. 21 (July), pp. 1–5. Full text available.

HAHMANN, MARTIN, H. D. L. W. "Large-Scale Data Analytics Using Ensemble Clustering". .

HAN, J., KAMBER, M., PEI, J., 2011, *Data Mining: Concepts and Techniques*. 3rd ed. San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. ISBN: 0123814790, 9780123814791.

HART, P. E., NILSSON, N. J., RAPHAEL, B., 1972, "A Formal Basis for the Heuristic Determination of Minimum Cost Paths", *SIGART Bull.*, , n. 37 (dez.), pp. 28–29. ISSN: 0163-5719. doi: 10.1145/1056777.1056779. Disponível em: <http://doi.acm.org/10.1145/1056777.1056779>.

HAUTAMÄKI, V., CHEREDNICHENKO, S., KÄRKKÄINEN, I., et al., 2005, "Improving K-means by Outlier Removal". In: *Proceedings of the 14th Scandinavian Conference on Image Analysis*, SCIA'05, pp. 978–987, Berlin, Heidelberg. Springer-Verlag. ISBN: 3-540-26320-9, 978-3-540-26320-3. doi: 10.1007/11499145_99. Disponível em: <http://dx.doi.org/10.1007/11499145_99>.

HAYES, D. N., MONTI, S., PARMIGIANI, G., et al., 2006, "Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in

multiple independent patient cohorts", *Journal of Clinical Oncology*, v. 24, n. 31, pp. 5079–5090.

HE, Q., CHANG, K., LIM, E.-P., 2007, "Analyzing feature trajectories for event detection". In: *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07*, p. 207, Amsterdam, The Netherlands. ACM Press. ISBN: 9781595935977. doi: 10.1145/1277741.1277779. Disponível em: <https://dl.acm.org/citation.cfm?doid=1277741.1277779>.

HE, Z., XU, X., DENG, S., 2008, "k-ANMI: A Mutual Information Based Clustering Algorithm for Categorical Data", *Inf. Fusion*, v. 9, n. 2 (abr.), pp. 223–233. ISSN: 1566-2535. doi: 10.1016/j.inffus.2006.05.006. Disponível em: <http://dx.doi.org/10.1016/j.inffus.2006.05.006>.

HECHT, B., HONG, L., SUH, B., et al., 2011, "Tweets from Justin Bieber's Heart: The Dynamics of the Location Field in User Profiles". In: *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, p. 237, Vancouver, BC, Canada. ACM. ISBN: 9781450302289. doi: 10.1145/1978942.1978976. Disponível em: <http://dl.acm.org/citation.cfm?id=1978942.1978976>.

HEIDEMANN, J., KLIER, M., PROBST, F., 2012, "Online Social Networks: A Survey of a Global Phenomenon", *Comput. Netw.*, v. 56, n. 18 (dez.), pp. 3866–3878. ISSN: 1389-1286. doi: 10.1016/j.comnet.2012.08.009. Disponível em: <http://dx.doi.org/10.1016/j.comnet.2012.08.009>.

HEIN, M., SETZER, S., 2011, "Beyond Spectral Clustering - Tight Relaxations of Balanced Graph Cuts". In: *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain.*, pp. 2366–2374. Disponível em: <http://papers.nips.cc/paper/4261-beyond-spectral-clustering-tight-relaxations-of-balanced-graph-cuts>.

HU, Y., NIE, Y., YANG, H., et al., 2010, "Measuring the significance of community structure in complex networks", *Phys. Rev. E*, v. 82 (Dec), pp. 066106. doi: 10.1103/PhysRevE.82.066106. Disponível em: <http://link.aps.org/doi/10.1103/PhysRevE.82.066106>.

IACOBUCCI, D., 1996, *Networks in marketing*. Sage Publications. ISBN: 9780761901396. Disponível em: <https://books.google.com.br/books?id=5DEPAQAAMAAJ>.

ILYAS, I. F., BESKALES, G., SOLIMAN, M. A., 2008, "A Survey of Top-k Query Processing Techniques in Relational Database Systems", *ACM Comput. Surv.*, v. 40, n. 4 (out.), pp. 11:1–11:58. ISSN: 0360-0300. doi: 10.1145/1391729.1391730. Disponível em: <http://doi.acm.org/10.1145/1391729.1391730>.

ISHIZUKA, M., SATTER, A., 2003, *PRICAI 2002: Trends in Artificial Intelligence: 7th Pacific Rim International Conference on Artificial Intelligence, Tokyo, Japan, August 18-22, 2002. Proceedings*. Lecture Notes in Computer Science. Springer Berlin Heidelberg. ISBN: 9783540456834. Disponível em: <https://books.google.com.br/books?id=gqlqCQAAQBAJ>.

IZAKIAN, H., PEDRYCZ, W., 2014, "Agreement-based fuzzy C-means for clustering data with blocks of features." *Neurocomputing*, v. 127, pp. 266–280. Disponível em: <http://dblp.uni-trier.de/db/journals/ijon/ijon127.html#IzakianP14>.

JACCARD, P., 1901, "Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines", *Bulletin de la Société Vaudoise des Sciences Naturelles*, v. 37, pp. 241–272.

JACCARD, P., 1912, "The Distribution of the Flora in the Alpine Zone", *New Phytologist*, v. 11, n. 2 (fev.), pp. 37–50. Disponível em: <http://www.jstor.org/stable/2427226?seq=3>.

JAIN, A. K., MURTY, M. N., FLYNN, P. J., 1999a, "Data Clustering: A Review", *ACM Comput. Surv.*, v. 31, n. 3 (set.), pp. 264–323. ISSN: 0360-0300. doi: 10.1145/331499.331504. Disponível em: <http://doi.acm.org/10.1145/331499.331504>.

JAIN, A. K., MURTY, M. N., FLYNN, P. J., 1999b, "Data clustering: a review", *ACM Comput. Surv.*, v. 31, n. 3 (set.), pp. 264–323. ISSN: 0360-0300. doi: 10.1145/331499.331504. Disponível em: <http://doi.acm.org/10.1145/331499.331504>.

JAIN, A. K., 2010, "Data Clustering: 50 Years Beyond K-means", *Pattern Recogn. Lett.*, v. 31, n. 8 (jun.), pp. 651–666. ISSN: 0167-8655. doi: 10.1016/

j.patrec.2009.09.011. Disponível em: <http://dx.doi.org/10.1016/j.patrec.2009.09.011>.

JOHNSON, N., CARRAN, S., BOTNER, J., et al., 2011, "Pattern in escalations in insurgent and terrorist activity", *Science (New York, N.Y.)*, v. 333, n. 6038, pp. 81–84. ISSN: 0036-8075. doi: 10.1126/science.1205068. Disponível em: <http://www.sciencemag.org/content/333/6038/81.full>.

KARRER, B., LEVINA, E., NEWMAN, M. E. J., 2008, "Robustness of community structure in networks", *Phys. Rev. E*, v. 77 (Apr), pp. 046119. doi: 10.1103/PhysRevE.77.046119. Disponível em: <http://link.aps.org/doi/10.1103/PhysRevE.77.046119>.

KEMPE, D., KLEINBERG, J., TARDOS, E., 2003a, "Maximizing the Spread of Influence Through a Social Network". In: *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pp. 137–146, New York, NY, USA, a. ACM. ISBN: 1-58113-737-0. doi: 10.1145/956750.956769. Disponível em: <http://doi.acm.org/10.1145/956750.956769>.

KEMPE, D., KLEINBERG, J., TARDOS, E., 2003b, "Maximizing the spread of influence through a social network". In: *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 137–146. ACM Press, b. ISBN: 1581137370. doi: 10.1145/956750.956769. Disponível em: <http://dx.doi.org/10.1145/956750.956769>.

KIRKPATRICK, S., GELATT, C. D., VECCHI, M. P., 1983, "Optimization by simulated annealing", *SCIENCE*, v. 220, n. 4598, pp. 671–680.

KOCH, R., 1999, *The 80/20 Principle: The Secret of Achieving More with Less*. A Currency book. Doubleday. ISBN: 9780385491747. Disponível em: <http://books.google.com.br/books?id=Ih359ILyvo8C>.

KUNCHEVA, L. I., 2004, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience. ISBN: 0471210781.

KUNCHEVA, L. I., VETROV, D. P., 2006, "Evaluation of Stability of k-Means Cluster Ensembles with Respect to Random Initialization", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 28, n. 11, pp. 1798–1808. doi: http://doi.ieeecomputersociety.org/10.1109/

TPAMI.2006.226. Disponível em: <http://dx.doi.org/http://doi.ieeecomputersociety.org/10.1109/TPAMI.2006.226>.

KURSUN, O., 2010, "Spectral Clustering with Reverse Soft K-Nearest Neighbor Density Estimation". In: *International Joint Conference on Neural Networks, IJCNN 2010, Barcelona, Spain, 18-23 July, 2010*, pp. 1–8. doi: 10.1109/IJCNN.2010.5596620. Disponível em: <http://dx.doi.org/10.1109/IJCNN.2010.5596620>.

KWAK, H., EOM, Y.-H., CHOI, Y., et al., 2009. "Consistent Community Identification in Complex Networks". arXiv:0910.1508v2.

KWAK, H., LEE, C., PARK, H., et al., 2010, "What is Twitter, a Social Network or a News Media?" In: *WWW '10 Proceedings of the 19th international conference on World wide web*, pp. 591–600, Raleigh, North Carolina, USA. ACM. ISBN: 9781605587998. doi: 10.1145/1772690.1772751. Disponível em: <https://dl.acm.org/citation.cfm?id=1772751>.

LAPPAS, T., VIEIRA, M. R., GUNOPULOS, D., et al., 2012, "On the Spatiotemporal Burstiness of Terms", *Proceedings of the VLDB Endowment*, v. 5, pp. 836–847. ISSN: 2150-8097. doi: 10.14778/2311906.2311911. Disponível em: <http://dl.acm.org/citation.cfm?id=2311906.2311911>.

LAURENT, A., STRAUSS, O., BOUCHON-MEUNIER, B., et al., 2014, *Information Processing and Management of Uncertainty: 15th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2014, Montpellier, France, July 15-19, 2014. Proceedings.* N. pt. 1, Communications in Computer and Information Science. Springer International Publishing. ISBN: 9783319087955. Disponível em: <https://books.google.com.br/books?id=IWIlBAAAQBAJ>.

LAZO, A. C. G. V., RATHIE, P. N., 1978, "On the entropy of continuous probability distributions (Corresp.)." *IEEE Transactions on Information Theory*, v. 24, n. 1, pp. 120–122. Disponível em: <http://dblp.uni-trier.de/db/journals/tit/tit24.html>.

LEE, C. H., YANG, H. C., CHIEN, T. F., et al., 2011, "A novel approach for event detection by mining spatio-temporal information on microblogs". In: *Proceedings - 2011 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2011*, ASONAM '11, pp.

254–259, Kaohsiung, jul. IEEE. ISBN: 9780769543758. doi: 10.1109/ASONAM.2011.74. Disponível em: <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5992610>.

LI, G., GÜNNEMANN, S., ZAKI, M. J., 2013, "Stochastic subspace search for top-k multi-view clustering". In: *Proceedings of the 4th MultiClust Workshop on Multiple Clusterings, Multi-view Data, and Multi-source Knowledge-driven Clustering, in conjunction with KDD 2013, Chicago, IL, USA, August 11, 2013*, p. 3. doi: 10.1145/2501006.2501010. Disponível em: <http://doi.acm.org/10.1145/2501006.2501010>.

LI, T., OGIHARA, M., MA, S., 2010, "On combining multiple clusterings: an overview and a new perspective", *Applied Intelligence*, v. 33, n. 2, pp. 207–219. ISSN: 0924-669X. doi: 10.1007/s10489-009-0160-4. Disponível em: <http://dx.doi.org/10.1007/s10489-009-0160-4>.

LOVÁSZ, L., 1996, "Random Walks on Graphs: A Survey". In: Miklós, D., Sós, V. T., Szőnyi, T. (Eds.), *Combinatorics, Paul Erdős is Eighty*, v. 2, János Bolyai Mathematical Society, pp. 353–398, Budapest.

LUSSEAU, D., NEWMAN, M. E. J., 2004. "Identifying the role that individual animals play in their social network". mar. Disponível em: <http://arxiv.org/abs/q-bio/0403029>.

MACQUEEN, J. B., 1967, "Some Methods for Classification and Analysis of MultiVariate Observations". In: Cam, L. M. L., Neyman, J. (Eds.), *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, v. 1, pp. 281–297. University of California Press.

MANDALA, S., KUMARA, S., YAO, T., 2012, "Detecting alternative graph clusterings." *Phys Rev E Stat Nonlin Soft Matter Phys*, v. 86, n. 1-2, pp. 016111. ISSN: 1550-2376. Disponível em: <http://www.biomedsearch.com/nih/Detecting-alternative-graph-clusterings/23005495.html>.

MANNING, C. D., SCHUTZE, H., 1999, *Foundations of Statistical Natural Language Processing*. MA, MIT Press.

MANNING, C. D., RAGHAVAN, P., SCHÜTZE, H., 2008, *Introduction to Information Retrieval*. New York, NY, USA, Cambridge University Press. ISBN: 0521865719, 9780521865715.

MARKOWITZ, H., 1952, "Portfolio Selection", *The Journal of Finance*, v. 7, n. 1, pp. 77–91. Disponível em: <http://www.jstor.org/stable/2975974>.

MARTÍ, R., GALLEGO, M., DUARTE, A., et al., 2013, "Heuristics and metaheuristics for the maximum diversity problem." *J. Heuristics*, v. 19, n. 4, pp. 591–615. Disponível em: <http://dblp.uni-trier.de/db/journals/heuristics/heuristics19.html#MartiGDP13>.

MASON, W., CONREY, F., SMITH, E., 2007, "Situating Social Influence Processes: Dynamic, Multidirectional Flows of Influence Within Social Networks", *Personality and Social Psychology Review*, v. 11, n. 3, pp. 279.

MCPHERSON, M., SMITH-LOVIN, L., COOK, J. M., 2001, "Birds of a Feather: Homophily in Social Networks", *Annual Review of Sociology*, v. 27, n. 1, pp. 415–444. doi: 10.1146/annurev.soc.27.1.415. Disponível em: <http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.soc.27.1.415>.

MEILĂ, M., 2007, "Comparing Clusterings—an Information Based Distance", *J. Multivar. Anal.*, v. 98, n. 5 (maio), pp. 873–895. ISSN: 0047-259X. doi: 10.1016/j.jmva.2006.11.013. Disponível em: <http://dx.doi.org/10.1016/j.jmva.2006.11.013>.

MEYERHENKE, H., SANDERS, P., SCHULZ, C., 2014, "Partitioning Complex Networks via Size-constrained Clustering", *CoRR*, v. abs/1402.3281. Disponível em: <http://arxiv.org/abs/1402.3281>.

MICHAEL STEINBACH, G. K., KUMAR, V., 2000, "A comparison of document clustering techniques". In: *KDD Workshop on Text Mining*.

MIRSHAHVALAD, A., LINDHOLM, J., DERLEN, M., et al., 2012, "Significant communities in large sparse networks", *PloS one*, v. 7, n. 3, pp. e33721.

MONTI, S., TAMAYO, P., MESIROV, J., et al., 2003, "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data", *Machine learning*, v. 52, n. 1-2, pp. 91–118.

MOOI, E., SARSTEDT, M., 2011, *A Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics*. Springer. ISBN: 9783642125416. Disponível em: <https://books.google.com.br/books?id=vQ316cBxTM8C>.

MOSER, F., COLAK, R., RAFIEY, A., et al., 2009, "Mining Cohesive Patterns from Graphs with Feature Vectors." In: *SDM*, v. 9, pp. 593–604. SIAM.

MÜLLER, E., GÜNNEMANN, S., FÄRBER, I., et al., 2010, "Discovering Multiple Clustering Solutions: Grouping Objects in Different Views of the Data."

In: Webb, G. I., 0001, B. L., Zhang, C., et al. (Eds.), *ICDM*, p. 1220. IEEE Computer Society.

MUNAGA, H., SREE, M. D. R. M., MURTHY, J. V. R., 2012, "Article: DenTrac: A Density based Trajectory Clustering Tool", *International Journal of Computer Applications*, v. 41, n. 10 (March), pp. 17–21. Full text available.

MÜLLER, E., GÜNNEMANN, S., FÄRBER, I., et al., 2012, "Discovering Multiple Clustering Solutions: Grouping Objects in Different Views of the Data". In: *Tutorial at IEEE 28th International Conference on Data Engineering (ICDE 2012), Washington, DC, USA*.

NETTLETON, D. F., 2013, "Data mining of social networks represented as graphs." *Computer Science Review*, v. 7, pp. 1–34. Disponível em: <`http://dblp.uni-trier.de/db/journals/csr/csr7.html`>.

NEUBAUER, N., OBERMAYER, K., 2009, "Towards community detection in k-partite k-uniform hypergraphs". In: *Proceedings of the NIPS 2009 Workshop on Analyzing Networks and Learning with Graphs*, pp. 1–9.

NEWMAN, M. E. J., 2004, "Detecting community structure in networks", *The European Physical Journal B - Condensed Matter and Complex Systems*, v. 38, n. 2 (mar.), pp. 321–330. ISSN: 1434-6028. doi: 10.1140/epjb/e2004-00124-y. Disponível em: <`http://dx.doi.org/10.1140/epjb/e2004-00124-y`>.

NEWMAN, M. E. J., GIRVAN, M., 2003, "Mixing patterns and community structure in networks". In: *in Statistical Mechanics of Complex Networks*, pp. 66–87. Springer,Berlin(2003.

NEWMAN, M. E. J., GIRVAN, M., 2004, "Finding and evaluating community structure in networks", *Phys. Rev. E*, v. 69, n. 2 (fev.), pp. 026113. doi: 10.1103/PhysRevE.69.026113. Disponível em: <`http://link.aps.org/doi/10.1103/PhysRevE.69.026113`>.

NEWMAN, M., 2010, *Networks: An Introduction*. New York, NY, USA, Oxford University Press, Inc. ISBN: 0199206651, 9780199206650.

NGUYEN, N., CARUANA, R., 2007, "Consensus Clusterings". In: *ICDM'07*, pp. 607–612.

NIU, D., DY, J. G., JORDAN, M. I., 2010, "Multiple Non-Redundant Spectral Clustering Views." In: Fürnkranz, J., Joachims, T. (Eds.), *ICML*, pp.

831–838. Omnipress. ISBN: 978-1-60558-907-7. Disponível em: <http://dblp.uni-trier.de/db/conf/icml/icml2010.html>.

NIU, D., DY, J., JORDAN, M., 2014, "Iterative Discovery of Multiple AlternativeClustering Views", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 36, n. 7 (July), pp. 1340–1353. ISSN: 0162-8828. doi: 10.1109/TPAMI.2013.180.

NOH, J. D., RIEGER, H., 2004, "Random Walks on Complex Networks", *Phys. Rev. Lett.*, v. 92 (Mar), pp. 118701. doi: 10.1103/PhysRevLett.92.118701. Disponível em: <http://link.aps.org/doi/10.1103/PhysRevLett.92.118701>.

OVELGONNE, M., GEYER-SCHULZ, A., 2012, "An Ensemble Learning Strategy for Graph Clustering". In: *10th DIMACS Implementation Challenge Graph Partitioning and Graph Clustering*.

P. E. HART, N. J. N., RAPHAEL, B., 1968, "A formal basis for the heuristic determination of minimum cost paths", *IEEE Transactions on Systems, Science, and Cybernetics*, v. SSC-4, n. 2, pp. 100–107.

PALUBECKIS, G., 2007, "Iterated tabu search for the maximum diversity problem", *Applied Mathematics and Computation*, v. 189, n. 1, pp. 371–383. doi: 10.1016/j.amc.2006.11.090. Disponível em: <http://dx.doi.org/10.1016/j.amc.2006.11.090>.

PENNEBAKER, J., 2013, *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury USA. ISBN: 9781608194964. Disponível em: <http://books.google.com.br/books?id=mJ4tLwEACAAJ>.

PERLIGER, A., PEDAHZUR, A., 2011, "Social Network Analysis in the Study of Terrorism and Political Violence", *PS: Political Science & Politics*, v. 44 (1), pp. 45–50. ISSN: 1537-5935. doi: 10.1017/S1049096510001848. Disponível em: <http://journals.cambridge.org/article_S1049096510001848>.

PFITZNER, D., LEIBBRANDT, R., POWERS, D., 2009a, "Characterization and Evaluation of Similarity Measures for Pairs of Clusterings", *Knowl. Inf. Syst.*, v. 19, n. 3 (maio), pp. 361–394. ISSN: 0219-1377. doi: 10.1007/s10115-008-0150-6. Disponível em: <http://dx.doi.org/10.1007/s10115-008-0150-6>.

PFITZNER, D., LEIBBRANDT, R., POWERS, D., 2009b, "Characterization and evaluation of similarity measures for pairs of clusterings", *Knowledge and Information Systems*, v. 19, n. 3, pp. 361–394. ISSN: 0219-1377. doi: 10.1007/s10115-008-0150-6. Disponível em: <http://dx.doi.org/10.1007/s10115-008-0150-6>.

PIEDMONT, R., 1998, *The Revised NEO Personality Inventory: Clinical and Research Applications*. The Springer Series in Social Clinical Psychology. Springer US. ISBN: 9780306459436. Disponível em: <https://books.google.co.in/books?id=Mho9UYWCE_cC>.

RAJARAMAN, A., ULLMAN, J. D., 2011, *Mining of Massive Datasets*. New York, NY, USA, Cambridge University Press. ISBN: 1107015359, 9781107015357.

RENTFROW, P. J., GOSLING, S. D., 2003, "The Do Re Mi's of Everyday Life: The Structure and Personality Correlates of Music Preferences", *Journal of Pers. Soc. Psychology*, v. 84, n. 6.

RIEDY, E. J., MEYERHENKE, H., EDIGER, D., et al., 2012, "Parallel Community Detection for Massive Graphs". In: *Proceedings of the 9th International Conference on Parallel Processing and Applied Mathematics - Volume Part I*, PPAM'11, pp. 286–296, Berlin, Heidelberg. Springer-Verlag. ISBN: 978-3-642-31463-6. doi: 10.1007/978-3-642-31464-3_29. Disponível em: <http://dx.doi.org/10.1007/978-3-642-31464-3_29>.

ROCKLIN, M., PINAR, A., 2011, "On Clustering on Graphs with Multiple Edge Types", *CoRR*, v. abs/1109.1605. Disponível em: <http://dblp.uni-trier.de/db/journals/corr/corr1109.html#abs-1109-1605>.

ROSI, A., MAMEI, M., ZAMBONELLI, F., et al., 2011, "Social sensors and pervasive services: Approaches and perspectives". In: *2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pp. 525–530, Seattle, WA, USA, mar. IEEE. ISBN: 978-1-61284-938-6. doi: 10.1109/PERCOMW.2011.5766946. Disponível em: <http://dx.doi.org/10.1109/PERCOMW.2011.5766946>.

RUAN, Y., FUHRY, D., PARTHASARATHY, S., 2013, "Efficient Community Detection in Large Networks Using Content and Links". In: *Proceedings of the 22Nd International Conference on World Wide Web*, WWW

'13, pp. 1089–1098, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee. ISBN: 978-1-4503-2035-1. Disponível em: <http://dl.acm.org/citation.cfm?id=2488388.2488483>.

SAJJA, P. S., AKERKAR, R., 2012, *Intelligent Technologies for Web Applications*. Chapman & Hall/CRC. ISBN: 1439871620, 9781439871621.

SAKAKI, T., OKAZAKI, M., MATSUO, Y., 2010, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors". In: *Proceedings of the 19th international conference on World wide web - WWW '10*, WWW '10, pp. 851–860, New York, NY, USA. ACM. ISBN: 978-1-60558-799-8. doi: 10.1145/1772690.1772777. Disponível em: <http://doi.acm.org/10.1145/1772690.1772777>.

SATHANUR, A. V., JANDHYALA, V., XING, C., 2013, "PHYSENSE: Scalable sociological interaction models for influence estimation on online social networks." In: Glass, K., Colbaugh, R., Sanfillippo, A., et al. (Eds.), *ISI*, pp. 358–363. IEEE. ISBN: 978-1-4673-6214-6. Disponível em: <http://dblp.uni-trier.de/db/conf/isi/isi2013.html#SathanurJX13a>.

SCHAEFFER, S. E., 2007, "Survey: Graph Clustering", *Comput. Sci. Rev.*, v. 1, n. 1 (ago.), pp. 27–64. ISSN: 1574-0137. doi: 10.1016/j.cosrev.2007.05.001. Disponível em: <http://dx.doi.org/10.1016/j.cosrev.2007.05.001>.

SCHULZ, A., HADJAKOS, A., PAULHEIM, H., et al., 2013, "A Multi-Indicator Approach for Geolocalization of Tweets". In: *Seventh International AAAI Conference on Weblogs and Social Media*, pp. 573–582, Boston, Massachusetts USA. AAAI Press. Disponível em: <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6063>.

SEIFI, M., JUNIER, I., ROUQUIER, J.-B., et al., 2013, "Stable Community Cores in Complex Networks". In: Menezes, R., Evsukoff, A., González, M. C. (Eds.), *Complex Networks*, v. 424, *Studies in Computational Intelligence*, Springer Berlin Heidelberg, pp. 87–98. ISBN: 978-3-642-30286-2. doi: 10.1007/978-3-642-30287-9_10. Disponível em: <http://dx.doi.org/10.1007/978-3-642-30287-9_10>.

SHANNON, C. E., 1948a, "A mathematical theory of communication", *Bell System Technical Journal*, v. 27 (Jul), pp. 379 – 423.

SHANNON, C. E., 1948b, "A mathematical theory of communication", *Bell System Technical Journal*, v. 27 (Jul), pp. 379 – 423.

SHEN, Y., SYU, Y.-S., NGUYEN, D. T., et al., 2012, "Maximizing Circle of Trust in Online Social Networks". In: *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, HT '12, pp. 155–164, New York, NY, USA. ACM. ISBN: 978-1-4503-1335-3. doi: 10.1145/2309996.2310023. Disponível em: <http://doi.acm.org/10.1145/2309996.2310023>.

SHIMBEL, A., 1953, "Structural Parameters of Communication Networks", *Bulletin of Mathematical Biophysics*, v. 15, pp. 501–507.

SKINNER, J., 2011, "Social Media and Revolution: The Arab Spring and the Occupy Movement as Seen through Three Information Studies Paradigms", *Sprouts: Working Papers on Information Systems*, v. 11, pp. 169. Disponível em: <http://sprouts.aisnet.org/11-169>.

SRIVASTAVA, A., SOTO, A. J., MILIOS, E., 2013, "Text Clustering Using One-mode Projection of Document-word Bipartite Graphs". In: *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, pp. 927–932, New York, NY, USA. ACM. ISBN: 978-1-4503-1656-9. doi: 10.1145/2480362.2480539. Disponível em: <http://doi.acm.org/10.1145/2480362.2480539>.

STARBIRD, K., PALEN, L., HUGHES, A. L., et al., 2010, "Chatter on the Red: What Hazards Threat Reveals About the Social Life of Microblogged Information". In: *CSCW '10 Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pp. 241–250, Savannah, Georgia, USA. ACM. ISBN: 9781605587950. doi: 10.1145/1718918.1718965. Disponível em: <https://dl.acm.org/citation.cfm?id=1718965>.

STAUDT, C. L., MEYERHENKE, H., 2013, "Engineering high-performance community detection heuristics for massive graphs". In: *Parallel Processing (ICPP), 2013 42nd International Conference on*, pp. 180–189. IEEE.

STEINLEY, D., BRUSCO, M. J., 2007, "Initializing K-means Batch Clustering: A Critical Evaluation of Several Techniques." *J. Classification*, v. 24, n. 1, pp. 99–121.

STEPANOVA, E., 2011, *The Role of Information Communication Technologies in the "Arab Spring- Implications beyond the region.* Relatório Técnico 159, The George Washington University Elliott School of International Affai,

Washington, US. Disponível em: <http://www.gwu.edu/~ieresgwu/assets/docs/ponars/pepm_159.pdf>.

STONEDAHL, F., R, W., WILENSKY, U., 2010, "Evolving viral marketing strategies". In: *In GECCO '10: Proceedings of the 12th annual conference on genetic and evolutionary computation*.

STREHL, A., GHOSH, J., 2003, "Cluster Ensembles — a Knowledge Reuse Framework for Combining Multiple Partitions", *J. Mach. Learn. Res.*, v. 3 (mar.), pp. 583–617. ISSN: 1532-4435. doi: 10.1162/153244303321897735. Disponível em: <http://dx.doi.org/10.1162/153244303321897735>.

SUN, Y., HAN, J., GAO, J., et al., 2009, "iTopicModel: Information Network-Integrated Topic Modeling". In: *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*, ICDM '09, pp. 493–502, Washington, DC, USA. IEEE Computer Society. ISBN: 978-0-7695-3895-2. doi: 10.1109/ICDM.2009.43. Disponível em: <http://dx.doi.org/10.1109/ICDM.2009.43>.

THEODORIDIS, S., KOUTROUMBAS, K., 2008, *Pattern Recognition, Fourth Edition*. Academic Press. ISBN: 1597492728, 9781597492720.

TOPCHY, A., JAIN, A., PUNCH, W., 2005, "Clustering ensembles: models of consensus and weak partitions", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 27, n. 12 (Dec), pp. 1866–1881. ISSN: 0162-8828. doi: 10.1109/TPAMI.2005.237.

TUMASJAN, A., SPRENGER, T., SANDNER, P., et al., 2010. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment". Disponível em: <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1441>.

VEGA-PONS, S., RUIZ-SHULCLOPER, J., 2011, "A Survey of Clustering Ensemble Algorithms." *IJPRAI*, v. 25, n. 3, pp. 337–372. Disponível em: <http://dblp.uni-trier.de/db/journals/ijprai/ijprai25.html#Vega-PonsR11>.

VERGEER, M., HERMANS, L., SAMS, S., 2013, "Online social networks and micro-blogging in political campaigning The exploration of a new campaign tool and a new campaign style", *Party Politics*, v. 19, n. 3, pp. 477–501.

VON LUXBURG, U., 2007, "A tutorial on spectral clustering", *Statistics and Computing*, v. 17, n. 4, pp. 395–416. ISSN: 0960-3174. doi: 10.1007/s11222-007-9033-z. Disponível em: <http://dx.doi.org/10.1007/s11222-007-9033-z>.

WAGNER, S., WAGNER, D., 2007. "Comparing Clusterings- An Overview". .

WANG, J., 2009, "Mean-Variance Analysis: A New Document Ranking Theory in Information Retrieval." In: Boughanem, M., Berrut, C., Moth, J., et al. (Eds.), *ECIR*, v. 5478, *Lecture Notes in Computer Science*, pp. 4–16. Springer. ISBN: 978-3-642-00957-0. Disponível em: <http://dblp.uni-trier.de/db/conf/ecir/ecir2009.html#Wang09>.

WANG, M.-F., HUANG, C.-S., TSAI, M.-F., et al., 2012, "Generalized Analysis of Message Propagation on Social Network", *International Journal of Future Generation Communication and Networking*, v. 5, n. 2, pp. 11–24.

WANG, Y., 2013, *Metaheuristics for large binary quadratic optimization and its applications*. Theses, Université d'Angers, fev. Disponível em: <https://tel.archives-ouvertes.fr/tel-00936210>.

WASSERMAN, S., FAUST, K., 1994, *Social network analysis: Methods and applications*, v. 8. Cambridge university press.

WATANABE, K., OCHI, M., OKABE, M., et al., 2011, "Jasmine: A Real-time Local-event Detection System Based on Geolocation Information Propagated to Microblogs". In: *International Conference on Information and Knowledge Management, Proceedings*, CIKM'11, pp. 2541–2544, New York, NY, USA. ACM. ISBN: 978-1-4503-0717-8. doi: 10.1145/2063576.2064014. Disponível em: <http://dl.acm.org/citation.cfm?doid=2063576.2064014>.

WATTS, D. J., DODDS, P. S., NEWMAN, M. E. J., 2002, "Identity and search in social networks", *Science*, v. 296, pp. 1302–1305.

WEBSTER, C. M., MORRISON, P. D., 2004, "Network Analysis in Marketing", *Australasian Marketing Journal (AMJ)*, v. 12, n. 2, pp. 8 – 18. ISSN: 1441-3582. doi: http://dx.doi.org/10.1016/S1441-3582(04)70094-4. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1441358204700944>.

WENG, J., LEE, B.-S., 2011, "Event Detection in Twitter". In: *Fifth International AAAI Conference on Weblogs and Social Media*, pp. 401–408, Barcelona,

Spain. AAAI Press. Disponível em: <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2767>.

WHELAN, S., DAVIES, G., 2006, *Profiling Consumers of Own Brands and National Brands Using Human Personality*. Disponível em: <https://books.google.com.br/books?id=bVMhngEACAAJ>.

WILLIAMS, C. B., GIRISH, J., 2012, "Social networks in political campaigns: Facebook and the congressional elections of 2006 and 2008", *New Media & Society*, p. 1461444812457332.

WU, Q., HAO, J.-K., 2013, "A hybrid metaheuristic method for the Maximum Diversity Problem." *European Journal of Operational Research*, v. 231, n. 2, pp. 452–464. Disponível em: <http://dblp.uni-trier.de/db/journals/eor/eor231.html#WuH13>.

XU, R., WUNSCH, I., 2005, "Survey of clustering algorithms", *Neural Networks, IEEE Transactions on*, v. 16, n. 3, pp. 645–678. ISSN: 1045-9227.

XU, S., WANG, Z., LI, X., et al., 2013, "A Novel Cluster Combination Algorithm for Document Clustering". In: Yang, J., Fang, F., Sun, C. (Eds.), *Intelligent Science and Intelligent Data Engineering*, v. 7751, *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 189–195. ISBN: 978-3-642-36668-0. doi: 10.1007/978-3-642-36669-7_24. Disponível em: <http://dx.doi.org/10.1007/978-3-642-36669-7_24>.

ZHANG, Y., LI, T., 2011, "Consensus Clustering + Meta Clustering = Multiple Consensus Clustering." In: Murray, R. C., McCarthy, P. M. (Eds.), *FLAIRS Conference*. AAAI Press. Disponível em: <http://dblp.uni-trier.de/db/conf/flairs/flairs2011.html#ZhangL11>.

ZHOU, Y., CHENG, H., YU, J. X., 2009, "Graph Clustering Based on Structural/Attribute Similarities", *Proc. VLDB Endow.*, v. 2, n. 1 (ago.), pp. 718–729. ISSN: 2150-8097. doi: 10.14778/1687627.1687709. Disponível em: <http://dx.doi.org/10.14778/1687627.1687709>.

ZHOU, Y., CHENG, H., YU, J. X., 2010, "Clustering Large Attributed Graphs: An Efficient Incremental Approach". In: *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM '10, pp. 689–698, Washington, DC, USA. IEEE Computer Society. ISBN: 978-0-7695-4256-0. doi: 10.1109/ICDM.2010.41. Disponível em: <http://dx.doi.org/10.1109/ICDM.2010.41>.